

Uliniowienia
wielu
sekwencji

Bartek
Wilczyński

Reminder on
alignments

Multiple
alignments

Uliniowienia wielu sekwencji

Bartek Wilczyński

27. marca 2018

There are no “official” notes for the lecture, but each lecture is (loosely) based on a chapter in one of the three textbooks.

- Lecture 1: Chapter **5** of “Computational Molecular Biology” by *P. Pevzner*
- Lecture 2: Chapters **13-14** of “Inferring Phylogenies” by *J. Felsenstein*
- Lecture 3: Chapter **2** of “Biological sequence analysis” by *Durbin, Eddy, Krogh and Mitchison*
- Lecture 4: Chapter **11** of “Inferring Phylogenies” by *J. Felsenstein*
- Lecture 5: Chapter **6** of “Biological sequence analysis” by *Durbin, Eddy, Krogh and Mitchison*

These books should be available in the library. If you have problems getting them, I can lend the books for a few moments, so that people can copy the relevant pages.

How sequences evolve?

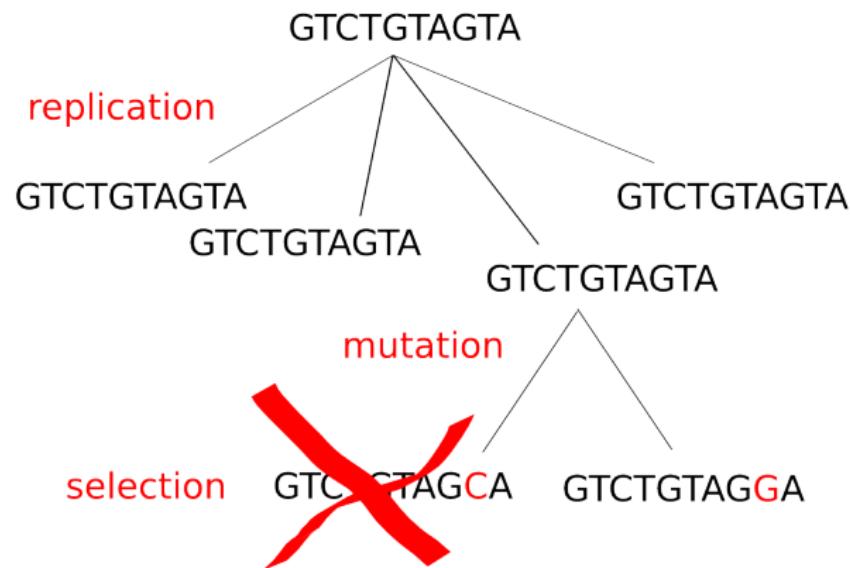


image (c) BW

Reconstructing alignments

Reminder on
alignments

Multiple
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	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗
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	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗
A	-40 -22 -8 ← -16 -16 -9 -12 -15 -7 3 ← -5	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗
E	-48 -30 -16 -3 ← -11 -11 -12 -12 -15 -5 2	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗	↑ ↗
	-56 -38 -24 -11 -6 -12 -14 -15 -12 -9 1									

HEAGAWGHE-E

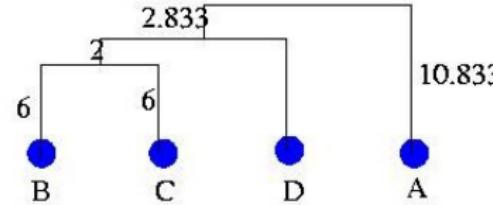
--P-AW-HEAE

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, AACAA, ACAAA

- CAAC- AACAA- ACAAA-
- -AACAA -ACAA -CAAC

Which C in CAAC was in the ancestral sequence?

Reminder on alignments

Multiple alignments

What is the solution to those inconsistencies?

Can we make a generalization of the pairwise alignment idea?

Q5E940_BOVIN	MREDRATRKTSNLYLKLII	LLDQDPKCFVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-PALE
ELAO_HUMAN	MREDRATRKTSNLYLKLII	LLDQDPKCFVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-PALE
ELAO_MOUSE	MREDRATRKTSNLYLKLII	LLDQDPKCFVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-PALE
ELAO_RAT	MREDRATRKTSNLYLKLII	LLDQDPKCFVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-PALE
ELAO_CHICK	MREDRATRKTSNLYLKLII	LLDQDPKCFVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-PALE
ELAO_RAMSY	MREDRATRKTSNLYLKLII	LLDQDPKCFVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-SALI
Q7ZUG3_BRARE	MREDRATRKTSNLYLKLII	LLDQDPKCFVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-PALE
ELAO_ICTP6	MREDRATRKTSNLYLKLII	LWDQDPKCFVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-PALE
ELAO_DROME	MRENNKAKAQAQYIKVY	DLDFEPPCKVYIADWMSK	KMQLQIMSLSLGK	RVLMLGK	MMRAKIEGHLEHN	-POLI
ELAO_DCIDI	MSAG-SKEEKFLIEKATLFTT	DTORMIVAYADWMSK	SDLOXQTSKTSIGI	GAVLMGK	MIRVKYIRDLSK	-FELD
Q54L0P_DCIDI	MSAG-SKEEKFLIEKATLFTT	DTORMIVAYADWMSK	SDLOXQTSKTSIGI	GAVLMGK	MIRVKYIRDLSK	-FELD
ELAO_PLAF8	MAKLSKQKQFQMYLIEKLSS	IQOQSILKLVHYDMS	NMASVYIMSLSLGK	AFLMGK	RTALAKHNL	-RAY
ELAO_SULAC	MIGLAVITTTKLAQWVDE	VALFELKTTKLIQI	GFPADLKHEIILKKRLKG	ADIVLKV	QIFNLNFAL	-YOK
ELAO_SULTO	MIRIMAVITQE	TRIAKWWIEVEKL	DTETRHTTLLI	GFPADLKHD	HKKRGMG	-LUDV
ELAO_SULSO	MKRLLALALKRQVSKAS	VWEELTELK	IKNSNTI	GFLNGEFTADFLKHEIILKKRLKG	ADIVLKV	-LUDV
ELAO_ARPER	MGSVSYLVGOMYKRE	GPIDPEWKTLMRLE	LEFSKVRVLFADLT	TYFVYV	MMWAKKBLI	-LUDV
ELAO_PYRAB	MHLALICRKYRYYETB	QDAPARVYKIS	SEATHLLOK	TYFVFLDHL	RLILWHEYRHLR	-GIV
ELAO_METAC	MAEERHTEHTEIPQW	DELENIKPK	QSTQHKEVY	PGFEC	LLATKHMDE	-ETIP
ELAO_METMA	MAEERHTEHTEIPQW	QKQKDDEI	JENIJKEL	QSTQHKEVY	PGFEC	-ESIP
ELAO_ARFCU	MAAVYRS	-PEP	TYRATVEL	IKRM	SSKPVVAVI	-SFV
ELAO_METKA	MAYKAKG	OPPSCSY	DYKAEWV	WRKREY	DLDE	-DYL
ELAO_METHH	MAHVAE	WV	WV	WV	WV	-WV
ELAO_METTL	MITAESEA	HKAPWPKIEWV	YVWKL	QWV	QWV	-WV
ELAO_METVA	MIDAKSE	HKAPWPKIEWV	YVWKL	QWV	QWV	-WV
ELAO_METJJA	METKVKAYAPKIEEY	YVWKL	QKSPVYI	WV	WV	-WV
ELAO_PYRAB	MAHVAE	WV	WV	WV	WV	-WV
ELAO_PYRHO	MAHVAE	WV	WV	WV	WV	-WV
ELAO_PYRFU	MAHVAE	WV	WV	WV	WV	-WV
ELAO_PYREK	MAHVAE	WV	WV	WV	WV	-WV
ELAO_HALMA	MCAESERKET	IDEWPKQYE	DAVIE	MEISTES	CVNNYACID	-DCEL
ELAO_HALVO	MSESEQRV	EYIWPQKRE	WV	WV	WV	-DCEP
ELAO_HALSA	MSEAEQRTTE	EWEEPKQRE	YAEV	DLV	DSV	-DCLD
ELAO_THEAC	MKEVSPQQKKEVLY	ETRISAKRS	YAVID	ACIE	PRODQGKNGKQK	-EKL
ELAO_THEVO	MRKRCV	WV	YSELAD	DTKS	YVTKGVR	-EKL
ELAO_PICTO	MTEP	POKQW	IDFYKLN	EME	INSRKYAVAS	-KALDSD
rule11.10.20.30.40.50.
					60.
					70.
					80.
					90.

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 parsimonious mutations)?
hint: Consider a column with all characters different.
- How much does it overestimate the number of necessary mutations?

Can we use dynamic programming?

Reminder on
alignments

Multiple
alignments

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

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)

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

$$S(a^{kl}) \geq \beta^{kl}$$

where $\beta^{kl} = \sigma(a) + S(\hat{a}^{kl}) - \sum_{k' < l'} S(\hat{a}^{k'l'})$.

image (c) Durbin et al

Carillo-Lippman lower bound method

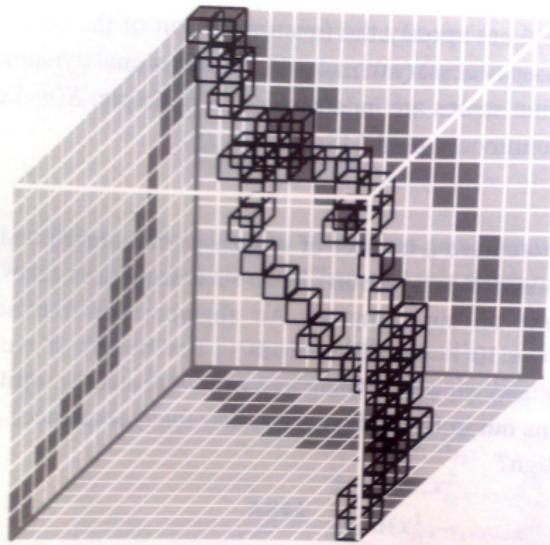


Figure 6.3 Carrillo & Lipman's algorithm allows the search for optimal alignments to be restricted to a subset of the multidimensional programming matrix, shown here as three-dimensional. The sets B^{kl} are shown in dark grey, and the cells in the matrix to which the search can be confined are outlined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

image (c) Durbin et al

Feng-Doolittle greedy approach

- 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. ◇

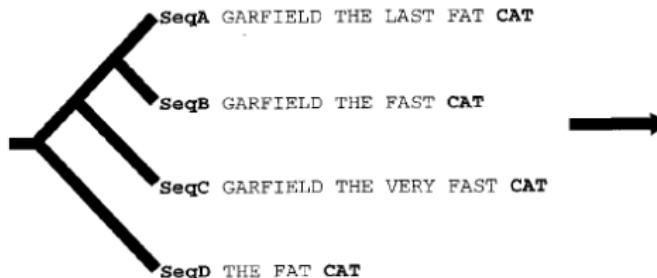
image (c) Durbin et al

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.
 $s(-, x) \neq s(-, y)$

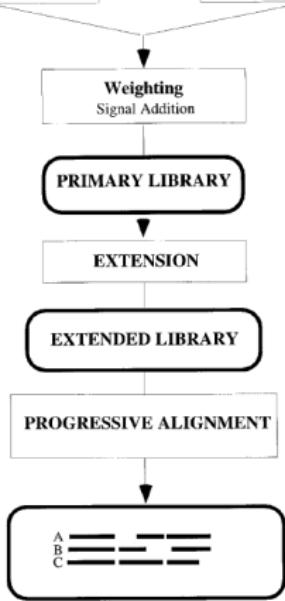
Incremental alignment problems

a) Regular Progressive Alignment Strategy



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

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



b) Primary Library

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

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

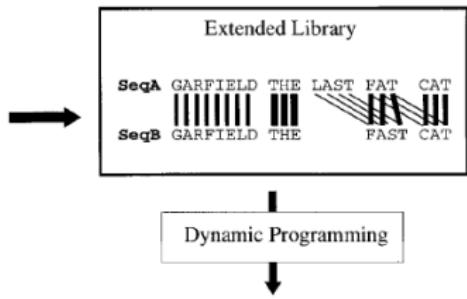
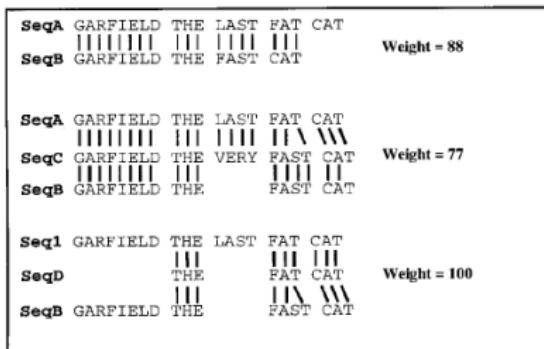
SeqA GARFIELD THE LAST FA-T CAT Prim. Weight = 77
SeqC GARFIELD THE VERY FAST CAT

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

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

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

c) Extended Library for seq1 and seq2



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

Muscle algorithm (Edgar 2004)

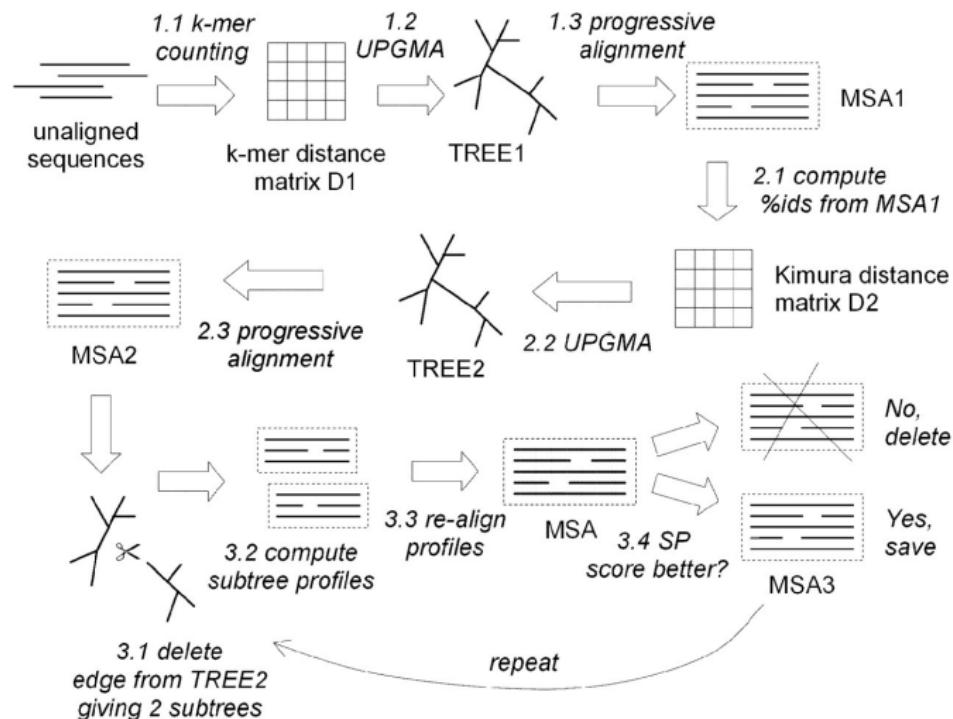


image (c) RC. Edgar, NAR 2004