Biochemistry 324

Bioinformatics

Multiple Sequence Alignment (MSA)

Big- Oh notation

Greek *omicron* symbol "O"
The **"Big-Oh" notation** indicates the **complexity** of an **algorithm** in terms of **execution speed** and **storage** needs

1. Algorithm to calculate a^b

```
exp1(a,b):
ans=1
    while(b>0):
    ans *= a
    b -= 1
return ans
```

O(b) linear

2. Algorithm to calculate n*m

```
exp2(n,m):
x=0
    for i in range (n):
        for j in range (m):
        x += 1
return x
```

$O(n*m) \cong O(n^2)$ quadratic

n=1000, nanosecond per n

log	10 nanoseconds
linear	1 microsecond
quadratic	1 millisecond
exponential	10^{284} years!

Approaches to multiple alignments

- The **Needleman-Wunsch** or **Smith-Waterman** pairwise sequence alignments based on dynamic programming is **exact** and guarantees an **optimal** alignment
- The "needle" or "water" algorithms are **O(mn)** and **$O(m^2n)$**, where m and n are the lengths of the 2 sequences (**quadratic**)
- A **multiple alignment** approach based on the "needle" or "water" algorithm will take **$O(2^N L^N)$**, where N is the number of sequences, L is the average sequence length (**exponential**)
- Thus, **dynamic programming** approaches to **multiple alignments** are **not** computationally **feasible**

- Five algorithmic approaches to multiple alignments:

  - **Exact**: "needle", "water"
  - **Progressive**: clustalW $O(N^2)$
  - **Iterative:** Praline, MUSCLE $O(N^2L + NL^2)$
  - **Consistency-based**: MAFFT $O(N \log N)$, T-coffee $O(N^3L)$
  - **Structure based:** Expresso $O(N^3L)$

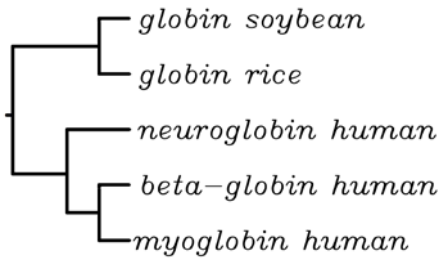
All can be installed on Linux, Mac OS/X or Windows platforms with more options/less restrictions

ClustalW (old)

- Based on **pairwise alignment** of **all combinations**, constructing a **guide tree**, and then **assembling the multiple alignment** based on **best to worst alignment scores**
- http://www.genome.jp/tools-bin/clustalw

1. beta-globin_human
2. myoglobin_human
3. neuroglobin_human
4. globin_soybean
5. globin_rice

```
Sequences (1:2) Aligned. Score: 23.1293
Sequences (1:3) Aligned. Score: 16.3265
Sequences (1:4) Aligned. Score: 11.4504
Sequences (1:5) Aligned. Score: 12.6761
Sequences (2:3) Aligned. Score: 15.894
Sequences (2:4) Aligned. Score: 11.4504
Sequences (2:5) Aligned. Score: 11.2676
Sequences (3:4) Aligned. Score: 12.9771
Sequences (3:5) Aligned. Score: 14.7887
Sequences (4:5) Aligned. Score: 38.1679
```



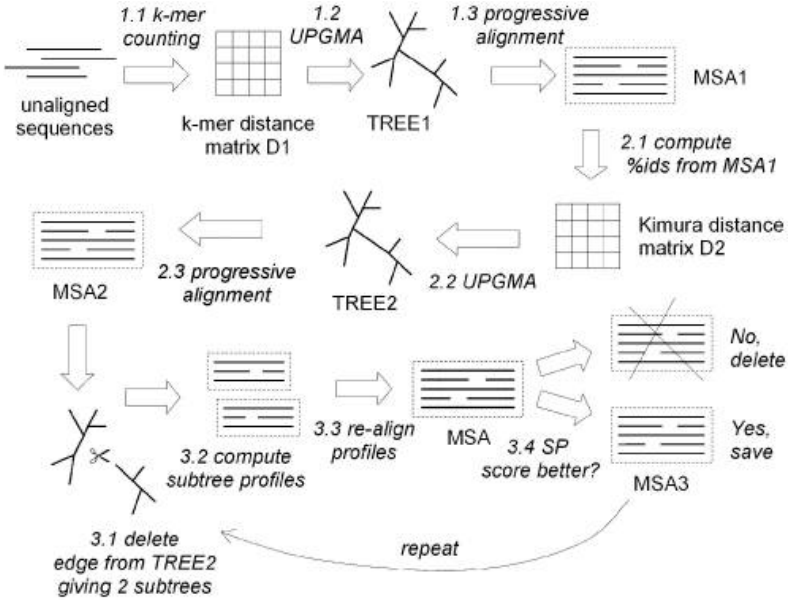
((beta-globin_human:0.38151,myoglobin_human:0.38720):0.04595,neuroglobin_human:0.40859,(globin_soybean:0.31392,globin_rice:0.30440):0.14342);

```
globin_soybean          ------------MTTSDVTTSMFERIGGST--TIDALVDRFYDRMDTLPEAQMIRAMHAD
globin_rice             MKWLKKMMAKPSAERDPQQSNAYDRIGGEE--VIRALAKQFYHQMQTNPDTQALLAMHRS
neuroglobin_human       -------MERPEPELIRQSWRAVSRSPLEH--GTVLFARLFALEPDLLPLFQYNCRQFSS
beta-globin_human       -------MVHLTPEEKSAVTALWGKVNVDE--VGGEALGRLLVVYPWTQRFFESFGDLST
myoglobin_human         --------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKS
                                                 :   .          :

globin_soybean          D---------LGLIRDVLKRYLTEWTGGPKLYTPEKGHPRLRQRHIGFAIGDAERDAWLL
globin_rice             P---------IPESEQKLFEFLSGWLGGPQLFHQRHGHPALRARHMPFSIDETMRDQWLL
neuroglobin_human       PEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGE
beta-globin_human       PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCD--KLHVDPENFRLLGN
myoglobin_human         EDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEF--ISE
                                         :          .        .        . :

globin_soybean          CMRGAMEETVT---DSAARQDLDRAISGLADWMRNRS-------
globin_rice             CMQRALAIEIK---EPQHREAIYQAISTLADHMRNQ--------
neuroglobin_human       SLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWDGE----
beta-globin_human       VLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
myoglobin_human         CIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
                         :   :         .  :    :    . . :
```

Multiple sequence comparison by log-expectation (MUSCLE)

- The "**distance**" based on number of **common k-tuples shared** between sequences are calculated
- A **binary tree** is constructed
- Profiles calculated for child **alignments at each node**, working from outside to root, giving MSA1 at root
- MSA1 is estimate based of k-tuple similarities
- **Kimura distance** is calculated from MSA1 and a **new binary tree** constructed
- **Changed branches** are **re-aligned** to produce MSA2
- Starting from the most distant nodes, **working towards root**, **profiles** are **aligned**
- If new profile **score** is **improved**, it is **retained**
- Continue until **convergence** or reaching set **limit**

http://www.ebi.ac.uk/Tools/msa/muscle/



Kimura model:

$d_{AB} = -\ln(1-f_{AB}-0.2\times f_{AB}^2)$

where f_{AB} = dissimilarity (fraction of observed differences) between sequences A and B,
d_{AB} = estimated evolutionary distance (fraction of expected substitutions) between sequences A and B

Tree-based Consistency Objective Function for alignment Evaluation (T-Coffee)

1. Library construction

```
Seq A: GARFIELD THE LAST FAT CAT
Seq B: GARFIELD THE FAST CAT
Seq C: GARFIELD THE VERY FAST CAT
Seq D: THE FAT CAT
```

- Two **libraries** of all **global** (clustalW) and **local** (Lalign) pairwise alignments are constructed for all possible **sequence pairs** (A-B, A-C, A-D, B-C, etc)
- Individual **symbol pairing** in each alignment is given a **weight** according to the **percentage similarity** of the aligned sequences
- The two **libraries** are **merged** by **adding weights** of duplicate entries

```
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

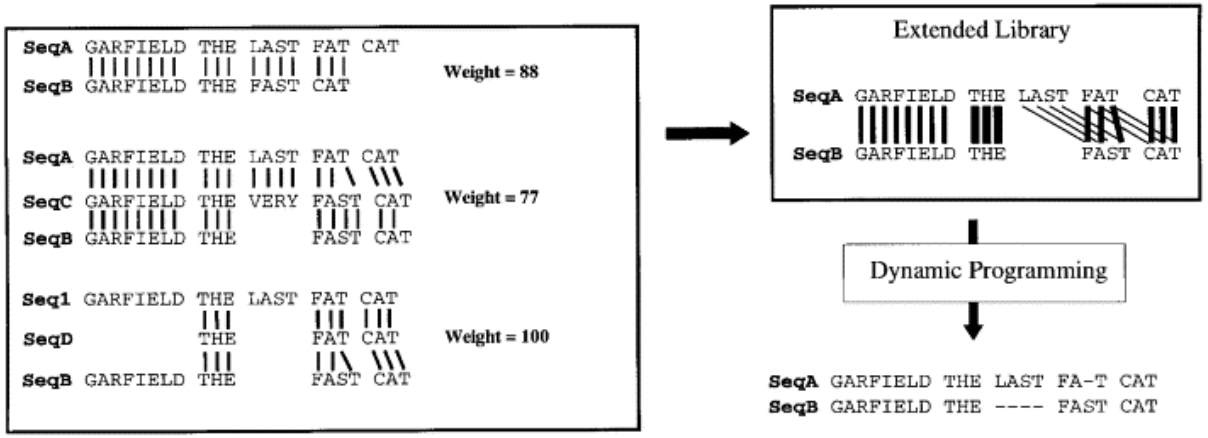
```
SeqB GARFIELD THE ---- FAST CAT
SeqC GARFIELD THE VERY FAST CAT
```
Prim Weight = 100

```
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

T-Coffee

2. Library extension

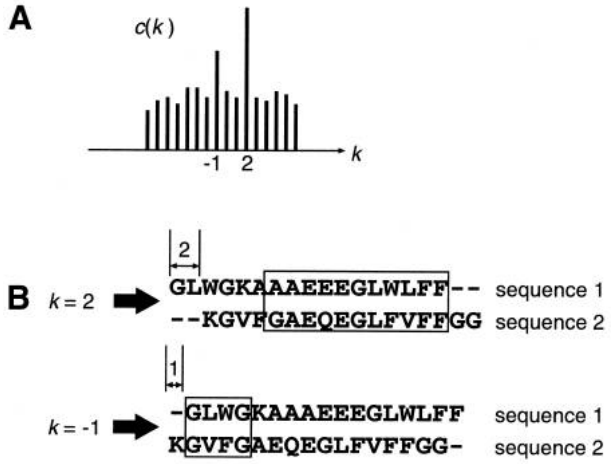


- Aligned triple sequences are considered (A-B, A-C, B-C)
- The **weight** of individual **symbol pairings** that are present in all alignments are **summed**
- We now perform a **dynamic programming** alignment of all possible sequence pairs using the **extended library** as a **scoring matrix**
- A binary tree if calculated based on the scores of the alignments
- Using the **tree as guide**, **alignments** are calculated **from** the **most similar pairs** down to the root of the tree
- **No gap penalties or extension introduced** during MA since these are already accommodated in weight library

http://www.ebi.ac.uk/Tools/msa/tcoffee/

Multiple alignment by Fast Fourier Transform (MAFFT)

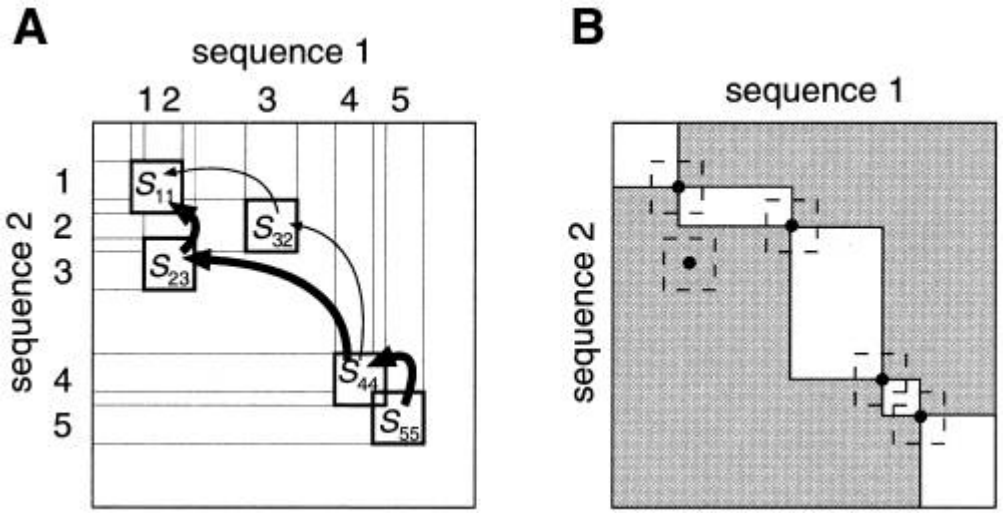
- The volume and charge properties of each amino acid is represented in a vector profile
- The correlation between each position of the profile is calculated
- A Fast Fourier analysis is performed on the correlation to determine the offset (how many residues sequence 1 has been slid past sequence 2) between homologous regions
- A Fourier analysis identifies the dominant frequencies present in a signal composed of the combination of many frequencies
- This analysis is O(N log N) as opposed to $O(N^2)$
- A homology search is performed between the two sequences in a sliding window at the determined offset



Katoh et al. 2002

MAFFT

- The positions of homology defines a constrained path through a homology matrix
- The best alignment path to connect the identified homologous regions is then calculated in this series of smaller, adjacent windows
- MAFFT is very useful to do MA of large numbers (>10,000) of sequences



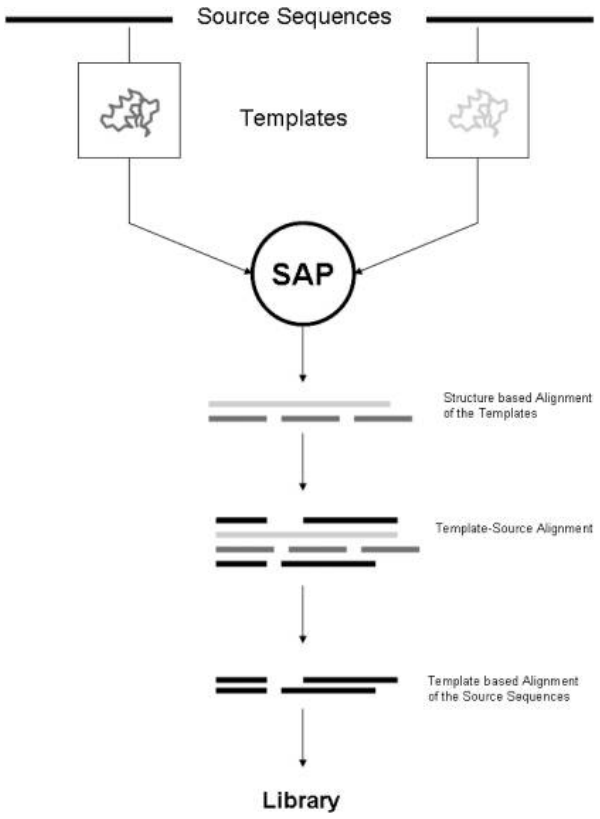
Katoh et al. 2002

http://www.ebi.ac.uk/Tools/msa/mafft/

Expresso

"The term Expresso also conveys the notion of aroma extraction and concentration, a notion that resonates with the way structures are 'expressed' within the MSA" -- developers

- A **BLAST search** of the **PDB** protein structure database with query sequence is performed
- **A hit** with >60% sequence identity and >70% coverage is selected
- The coordinates of the **structures are aligned** with SAP, **without** a need to **superimpose** them
- SAP **identifies structurally equivalent α-carbons** in sequences A and B based on the **similarities of the distance** between the **α-carbon** in structure A and all other α-carbons in A, compared to the distance between an α-carbon in structure B and all other α-carbons in B
- SAP produces a **structural alignment**
- The **sequence A** and **B** are then **aligned** to the paired **structural alignments**, and the **alignment** added to the **library**
- The **library** is then used to produce the MSA using a **progressive alignment** as **implemented** in the **T-coffee** algorithm



http://tcoffee.crg.cat/apps/tcoffee/do:expresso