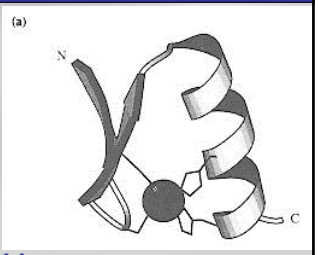
1

Multiple Sequence Alignment

Marco Botta
Dipartimento di Informatica
Università di Torino
botta@di.unito.it
www.di.unito.it/~botta/didattica/

2

Protein Families



- Consider Zinc Fingers:
- All have the same function:
  - Bind to DNA
- All have similar structure
- They constitute a Protein Family
- In a protein family some parts of the sequence (the functional parts) are more conserved than others.

Multiple Alignment

- Proteins can be classified into families:
  - Common structure.
  - Common function.
  - Common evolutionary origin.
- For a set of sequences belonging to some family
  - Each pair has some differences
  - But, there are some common motifs in almost all sequences of the family
- A multiple alignment carries more information than pairwise alignment

Multiple Sequence Alignment – MSA

- Motivations
  - Molecular Phylogenesis

    Build phylogenetic trees that show distances and evolutionary relation among molecules, based on sequence comparisons.
  - Study of genome evolution
  - Characterization of genes and proteins with unknown function

    Through the identification of recurrent motifs and functionally relevant sites.
  - Identification of regulatory elements

    Through the identification of common pattern among different organisms.

Definition

5

A **multiple alignment** of strings $S_1, S_2, \ldots, S_k$ is a series of strings with blanks $S'_1, S'_2, \ldots, S'_k$ such that:

- $|S'_1| = |S'_2| = \ldots = |S'_k|$
- S'_j is an extension of S_j obtained by insertion of blanks.

MSA: An Example

```
1pamA       TDVIYQIFTD RFSDGNPANN P---TGAAFD GSC-TNLRLY CGGDWQGIIN
cdgt_bacli  TDVIYQVFTD RFLDGNPSNN P---TGAAFD GTC-SNLKLY CGGDWQGLVN
amy_thetu   TDVIYQIVTD RFVDGNTSNN P---TGDLYD PTH-TSLKKY FGGDWQGIIN
cdg2_bacma  TDTVYQIVTD RFVDGNSANN P---TGAAFS SDH-SNLKLY FGGDWQGITN
cdg1_bacma  TDVIYQIVTD RFADGDRTNN P---AGDAFS GDR-SNLKLY FGGDWQGIID
cdgt_bacst  SDVVYQIVVD RFVDGNTSNN P---SGALFS SGC-TNLRKY CGGDWQGIIN
cdgt_bacs2  KDVIYQIVTD RFSDGNPGNN P---SGAIFS QNC-IDLHKY CGGDWQGIID
amym_bacst  GDVIYQIIID RFYDGDTTNN NPAKSYGLYD PTK-SKWKMY WGGDLEGVRQ
cdgt_klepn  KETIYFLFLD RFSDGDPSNN A---GFNSAT YDP-NNLKKY TGGDLRGLIN
amyb_bacpo  KQSIYFIMTD RFSNGDPSND N---YGG-FN SN-NSDQRKW HGGDFQGIIN
amy1_schpo  RRSIYQIITD RFSLEEGATE ---------R IPCDPVRFMY CGGTWNGIRN
2aaa        TQSIYFLLTD RFGRTDNSTT ---------- ATCNTGNEIY CGGSWQGIID
amya_aspor  SQSIYFLLTD RFARTDGSTT ---------A TC-NTADQKY CGGTWQGIID
amy1_schoc  DQSIYQIVTD RFARSDGSTT ---------- ADCLVSDRKY CGGSYKGIID
amy1_sacfi  SQSIYQIVTD RFARTDGDTS ---------A SC-NTEDRLY CGGSFQGIIK
ydd2_schpo  KQVIYQVLTD RFALDEDN-- ---------- FYAKASGNLY LGGTWKGITR
amy_bacci   TDVIYQIVTD RFVDGNTANN P---AGSAYD ATCSTNLKLY CGGDWQGIMN
1jdc        GD---EIILQ GFHWNVVREA P--------- ---------- --NDWYNILR
```

Sum of Pairs

- The sum of pairwise distances between all pairs of sequences for some scoring matrix

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

- Not only assumes that alignment of each column is independent, but also each pair of sequences.
  - Each sequence is scored as if descended from *k-1* sequences instead of one common ancestor.

Sum-Of-Pairs Score: an Example

```
A A C T G - T - - A G
A A C - G - T A T A C
A A C T - A T A - - G
```

$$\sigma(m) = \sum_{k<l} S(m_k, m_l)$$

- If we choose an indel model that assigns 1 to matches and 0 to mismatches we have:

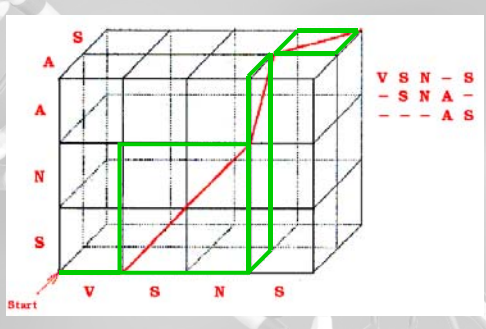
$$\sigma(m) = S(m_1, m_2) + S(m_1, m_3) + S(m_2, m_3) = 6 + 6 + 5 = 17$$

MSA Algorithms

- Exact solution: Dynamic Programming
- MSA Euristic Approaches
  - Center Star Method
  - Profiles
  - Iterative Alignment
  - Progressive Alignment: Feng-Doolittle
  - ClustalW
  - Consistency-based Methods
  - T-Coffee
  - MSA by HMM: Probcons
- Scoring Functions and Alignment Evaluation

Dynamic Programming: Hypercube

- Given the sequences S_1=VSNS, S_2=SNA and S_3=AS a 3-dimensional hypercube is obtained:



- The optimal alignment can be calculated exactly using r-dimensional dynamic programming.
  - Space complexity $O(n^k)$
  - Time complexity $O(2^k n^k)$

Center Star Method

- The Center-Star method is an approximate algorithm based on the Sum-Of-Pairs Score (SP).
- Given as input a set of sequences S = {S_1, S_2, … S_k}, we want to find the multiple alignment that minimizes the SP distance (or maximizes the SP score).

Center Star Method: Definitions

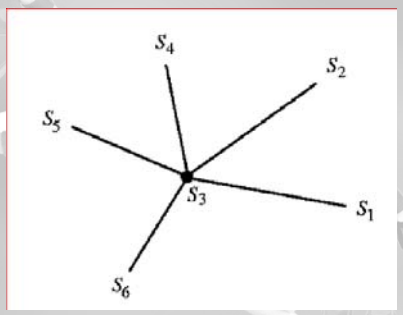
- Given a set S of k sequences, the central sequenzce $S_c \in S$, is the sequence that minimizes the function:

$$\sum_{S_j \in S} D(S_c, S_j)$$

- i.e., the sum of the distances of every sequence from S_c is minimal.

Center Star Method: Definitions

- Center-Star is a tree with k nodes, in which S_c is the central node and the remaining k-1 nodi are labelled with different sequence in S \ {S_c}



- The MSA M_c of the set of sequences S is the multiple alignment consistent with such tree.

Center Star Method: Algorithm

- Find the sequence $S_t \in S$ that minimizes $\sum_{i \neq t} D(S_i, S_t)$ and $M = \{S_t\}$
- Add the sequences in S\{S_t} to M one by one, according to their distance to S_t (closest first), aligning every new sequence to S_t and possibly adding new gaps.
- Complexity: $O(k^2n^2)$, where k is the number of sequences and n is the maximum length.
- The distance SP of the resulting alignment is guaranteed to be less than twice the optimal distance.

Iterative Alignment

- This approach makes use of pairwise scores in order to add sequences to a multiple alignment.
- It starts by aligning the most similar sequences, according to a given distance metric.
- Then, at each step, it chooses the sequence that os the closest to all sequences already aligned, and add it to the multiple alignment.
- Possibly, new spaces "-" are added to the aligned sequences.

Progressive Alignment

- The basic idea is that the most reliable biological information obtainable from a set of sequences, derives from the alignment of the closest pair of sequences.
- Therefore, every gap "-" that occurs in this alignment must be preserved in the multiple alignment construction process (differently from what happens in iteratve alignment methods).
- Several tools for MSA are based on this approach, among them ClustalW and T-Coffee.

Progressive Alignment: Feng-Doolittle Algorithm

- Compute $\binom{k}{2}$ pairwise alignments and convert their scores in distances.
- Build a phylogenetic tree.
- Align the sequences in the order given by the tree, starting from the closest sequences.

ClustalW

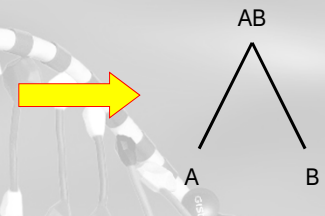
- ClustalW is the most popular tool for multiple alignment of biosequences
- It implements Feng-Doolittle progressive approach.
- Given a set *S* of *n* sequences, ClustalW performs all pairwise alignments between pairs of sequences in *S* and build a distance matrix.

	Seq. A	Seq. B	Seq. C	Seq. D
Seq. A	0.00			
Seq. B	0.11	0.00		
Seq. C	0.32	0.43	0.00	
Seq. D	0.17	0.18	0.57	0.00

ClustalW

- Then, it build a guide tree using the *neighbour-joining* method.
- It chooses the closest pair and these will form the first subtee:

	Seq. A	Seq. B	Seq. C	Seq. D
Seq. A	0.00			
Seq. B	0.11	0.00		
Seq. C	0.32	0.43	0.00	
Seq. D	0.17	0.18	0.57	0.00



ClustalW

- The entries A and B will substituted by a single entry AB and distances from the remaining sequences are computed as a simple average:

	Seq. AB	Seq. C	Seq. D
Seq. AB	0.00		
Seq. C	0.375	0.00	
Seq. D	0.175	0.57	0.00

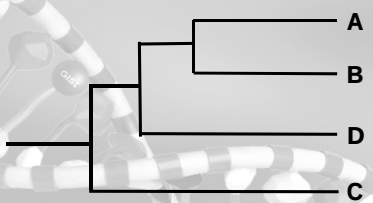
$$D(AB,D) = \frac{D(A,D)+D(B,D)}{2} = \frac{0.17+0.18}{2} = 0.175$$

$$D(AB,C) = \frac{D(A,C)+D(B,C)}{2} = \frac{0.32+0.43}{2} = 0.375$$

- Repeating this process, the full tree is built.

ClustalW: Phylogenetic Tree

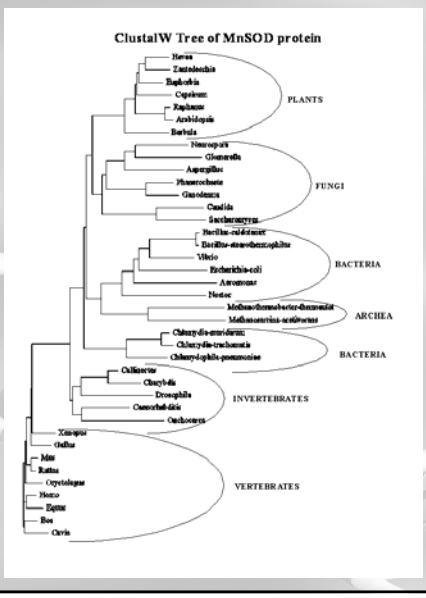
- Tree branches have length proportional to the distance between sequences:



- This tree will be used to guide the progressive alignment.
- In our example, sequences A and B will aligned first, then D will be added and finally C will be added.

Phylogenetic Tree: An Example

- The phylogenetic tree depicted is build by ClustalW from sequences of the protein mnSOD from different organisms: the obtained clustering reflect quite reliably the known phylogenesis (i.e., based on geopaleotological data).



ClustalW

- This is a sample output of ClustalW.

- The * symbol at the bottom of a column means 100% match
- The : symbol means high similarity (75%).
- The . symbol means medium similarity (50%-75%).

Consistency-based Methods

- The first algorithm for MSA that was *consistency-based* has been presented by Kececioglu in 1993.
- Given a set of sequences S, the "optimal" alignment must be the most consistent with the optimal pairwise alignment of the sequences in S.
- Computing such alignment is an NP-Complete problem and it can only be exactly solved for a small number of sequences.

Advantages of Consistency-based Method

•Objective functions do not depend on specific scoring matrix, rather on pairwise alignment methids.

•Consistency-based methods depend upon the positions of the amminoacids in the pairwise alignment; this means that the score depends in their positions rather than on their chemical-physical properties.

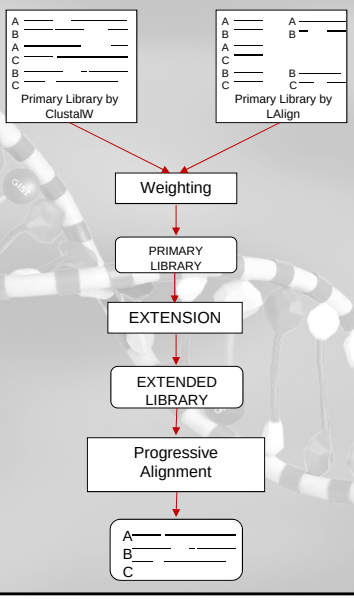
SAGA

- One of the first euristic tools consistency-based was SAGA (1996).
- SAGA uses COFFEE (Consistency-based Objective Function For alignmEnt Evaluation) objective function, that reflect the consistency level between a multiple alignment and a library of pairwise alignment of the same set of sequences.
- The COFFEE-Score is optimized by means of a genetic algorithm.
- Even though SAGA is able to produce interesting results, the genetic algorithm approach is quite slow.

T-Coffee

- T-Coffee (Tree-based COFFEE) is a heuristic approach based on COFFEE objective function.
- The multiple alignment is computed from a collection of pairwise local and global alignments by using a progressive method based on guide tree (as in ClustalW)
- Due to this procedure, T-Coffee obtains a quite impressive precision when multiple aligning low similarity sequences.

T-Coffee Algorithm



Scoring Functions and Alignment Evaluation

- There are a number of scoring functions besides the Sum-Of-Pairs, we consider two of them:
  - Entropy
  - Circular-Sum
- The *right* choice of the objective function is very important in the design of a good alignment algorithm
- Unluckily, there are no universal functions that fully capture the biological meaning of the alignment

Entropy

- Entropy $E(A) = \sum_{C \in A} E(C)$

where C are the columns of the alignment $E(C) = -\left(\sum_{X \in \Sigma} p_X \log p_X \right)$
and p_X is the frequency of symbol X in column C.

```
A A C T G – T - - A G
A A C – G – T A T A C
A A C T – A T A - - T
```

$$E(1) = -\left(\sum_{X \in (A,C,G,T,-)} p_X \log p_X \right) = -(p_A \log p_A + p_C \log p_C + p_G \log p_G + p_T \log p_T + p_- \log p_-) = -\left(\frac{3}{3} \cdot \log \frac{3}{3} + 0 + 0 + 0 + 0 \right) = 0$$

$$E(11) = -\left(0 + \frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3} + 0 \right) = -(0 - 0{,}15 - 0{,}15 - 0{,}15 + 0) = 0{,}45$$

$$E(A) = 0 + 0 + 0 + 0{,}11 + 0{,}11 + 0{,}16 + 0 + 0{,}11 + 0{,}16 + 0{,}11 + 0{,}45 = 1{,}21$$

- A highly conserved column has low variability and high information content. The *better* the alignment, the lower the entropy.

Circular Sum

- Circular-Sum: $CS(A) = \frac{1}{2}\sum_{i=1}^{n} MPA(a_{C_i}, a_{C_{i+1}})$
- Where $MPA(a_i, a_j) = \sum_{m=1}^{k} S(a_i[m], a_j[m])$ and $C_{n+1} = c_1$ is the pairwise-alignment score induced by the MSA.

```
A A C T G – T - - A G
A A C – G – T A T A C
A A C T – A T A - - T
```

$$MPA(a_1, a_2) = \sum_{m=1}^{11} S(a_1[m] + a_2[m]) = 1+1+1+0+1+0+1+0+0+1+0 = 6$$

$$MPA(a_2, a_3) = \sum_{m=1}^{11} S(a_2[m] + a_3[m]) = 1+1+1+0+0+0+1+1+0+0+0 = 5 \qquad CS(A) = 6+5+5 = 16$$

$$MPA(a_3, a_1) = \sum_{m=1}^{11} S(a_3[m] + a_1[m]) = 1+1+1+1+0+0+1+0+0+0+0 = 5$$

What's next ?

- A multiple alignment is only the beginning of the analysis !
- Extract a representation of the multiple alignment
  - Consensus sequence: the most frequent character in each position (e.g. DxNDNxPxF)
  - WebLogo: graphical representation

Profiles

- Profiles are structures used to summarize common properties of groups of sequences, and are the basis of many multiple alignment methods.

- Let M be a multiple sequence alignment of length *l*.
- The profile of M is a matrix $l \times |\Sigma \cup \{-\}|$ where Σ is the alphabet of sequences in M, whose columns report the frequency of every symbol in the corresponding column of the alignment

Profiles: an Example

A C A - - G – T C A
A C - - T G C T – A
- C A A T G C T G A

	A	C	G	T	-
1	2/3	0	0	0	1/3
2	0	3/3	0	0	0
3	2/3	0	0	0	1/3
4	1/3	0	0	0	2/3
5	0	0	0	2/3	1/3
6	0	0	3/3	0	0
7	0	2/3	0	0	1/3
8	0	0	0	3/3	0
9	0	1/3	1/3	0	1/3
10	3/3	0	0	0	0

Alignment of a Sequence to a Profile

- To align a sequence to a profile the dynamic programming global alignment algorithm (Needleman-Wunsch) is used, with the following scoring function:
- Let *p(i,j)* be a profile, i=1…l and j=1…|Σ|+1 and let S = {S_1, S_2, …, S_n}.
- We can define the Scoring Function:

$$\sigma_{sp} : (\Sigma \cup \{-\}) \times \{1,2,...,l\} \to \Re$$

$$\sigma_{sp}(b,i) = \sum_{a \in \Sigma} p_{i,a} \sigma(a,b)$$

Alignment of two profiles

- Let $P_1 = (p'_{ij})$ and $P_2 = (p''_{ij})$ i=1…l and j=1…|Σ|+1 two profiles.
- The scoring function is the following:

$$\sigma_{pp} : \{1,2,...,l\} \times \{1,2,...,l\} \to \Re$$

$$\sigma_{pp}(i,j) = \sum_{k=1}^{|\Sigma|+1} f\left(p'_{i,k}, p''_{j,k}\right)$$

- where *f* is a function that assigns a score to pairs of columns according to the frequency of symbols in the alphabet.

PSSM: Position Specific Scoring Matrix

- Each element of a PSSM represents a probability of a specific character in a given position of the multiple alignment

	0	1	2	3	4	5	6	7	8	9	10
-	0.8	0.8	0.8	0.5	0.2	0.0	0.0	0.0	0.6	0.7	0.8
B	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0
D	0.2	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0
E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.3	0.0
F	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
G	0.0	0.0	0.0	0.3	0.0	0.0	0.2	0.0	0.0	0.0	0.0
I	0.0	0.2	0.0	0.0	0.0	1.0	0.0	0.2	0.0	0.0	0.0
L	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
N	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0
O	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.2
S	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
U	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0

45

Creating a PSSM

- After aligning the sequences we see that there are some conserved regions.
- We use the multiple alignment of Blast results to create a Position Specific Scoring Matrix.
- This matrix represents information from a whole family, it is more strict in highly conserved regions.

46

PSI- BLAST (Position Specific Iterated)

- BLAST provides a new automatic "profile like" search.
- Iterative procedure:
  - Perform BLAST on database.
  - Use Significant alignments to construct a "position specific" score matrix.
  - This matrix replaces the query sequence in the next round of database searching.
- The program may be iterated until no new significant alignments are found.
- Most commonly used search method today.

PHMM: Profile Hidden Markov Model

- A Profile HMM is a linear state machine consisting of a series of nodes, each of which corresponds roughly to a position (column) in the string it represents