Tool di allineamento multiplo a confronto

Bioinformatica a.a. 2007/08

Andrea Renieri Matteo Tanca

Università di Pisa, Dipartimento di Informatica

11 Dicembre 2007

Schema della presentazione

Allineamento multiplo di sequenze

Definizione del problema

- **Generalizzazione dell'allineamento fra due sequenze**
  Dato un insieme di sequenze $\{s_1, ..., s_k\}$, definite sul medesimo alfabeto, un allineamento $s'_1, ..., s'_k$ è ottenuto inserendo degli spazi nelle sequenze, in modo che:
  1) $|s'_1| = |s'_2| = ... = |s'_k| = n$
  2) Nessuna colonna sia costituita da soli spazi

Allineamento multiplo di sequenze

Definizione del problema

- **Generalizzazione dell'allineamento fra due sequenze**
  Dato un insieme di sequenze $\{s_1, ..., s_k\}$, definite sul medesimo alfabeto, un allineamento $s'_1, ..., s'_k$ è ottenuto inserendo degli spazi nelle sequenze, in modo che:
  1) $|s'_1| = |s'_2| = ... = |s'_k| = n$
  2) Nessuna colonna sia costituita da soli spazi

Esempio

```
-ACTTGT-
C-CT-GT-
ACACTGGT
```

Campi d'applicazione

- Costruzione di alberi filogenetici
- Generazione di profili
- Caratterizzazione di proteine con funzione sconosciuta (ed identificazione di domini funzionali)

Complessità del problema

Complessità computazionale

- Trovare l'allineamento ottimo fra più di due sequenze è un problema computazionalmente difficile (più esattamente *NP*-completo)
- Il tempo richiesto dalla risoluzione del problema cresce esponenzialmente rispetto alla dimensione dei dati da allineare
- Esistono algoritmi esatti (*Smith-Waterman, Needleman-Wunsch*), ma sono computazionalmente impraticabili: non è possibile calcolare una soluzione esatta in tempo ragionevole!

Complessità del problema

Soluzioni approssimate

- Non potendo calcolare una soluzione esatta, ci si accontenta di un'approssimazione sufficientemente "buona"
- Approssimazioni delle soluzioni possono essere calcolate mediante:
  - euristiche
  - metodi statistici e probabilistici
  - metodi alternativi (FFT, algoritmi genetici)

Complessità del problema

Funzione di score

- Assegnare un punteggio ad un allineamento multiplo è più complicato rispetto al caso a 2 sequenze
- **Desiderata**
  - Indipendenza dall'ordine delle sequenze
  - Malus per spazi e segmenti scorrelati e bonus per segmenti correlati
  - Sensibilità (pochi falsi negativi)
  - Specificità (pochi falsi positivi)
- **SP** (**S**um of **P**airs): "classica" funzione di valutazione, data dalla somma dei punteggi di allineamento coppia a coppia sui simboli di una data colonna

ALLINEAMENTI BASATI SU STRUTTURA

- Gli **allineamenti basati su struttura** limitano il numero di allineamenti considerati, imponendo un ordine di valutazione, attraverso l'utilizzo di un'opportuna struttura
  - ***Stella***
  - ***Albero***
  - ***Grafo***

Allineamenti basati su struttura

Star alignment

- **Idea**: selezionare una sequenza fra quelle in esame ed utilizzarla come “cardine”

1. La sequenza di indice i è selezionata come centro della stella
2. Per ogni coppia $(i, j), j \neq i$ si calcola la distanza di edit fra le due sequenze
3. Si minimizza la funzione di score d

Allineamenti basati su struttura

Star alignment

- **Idea**: selezionare una sequenza fra quelle in esame ed utilizzarla come “cardine”

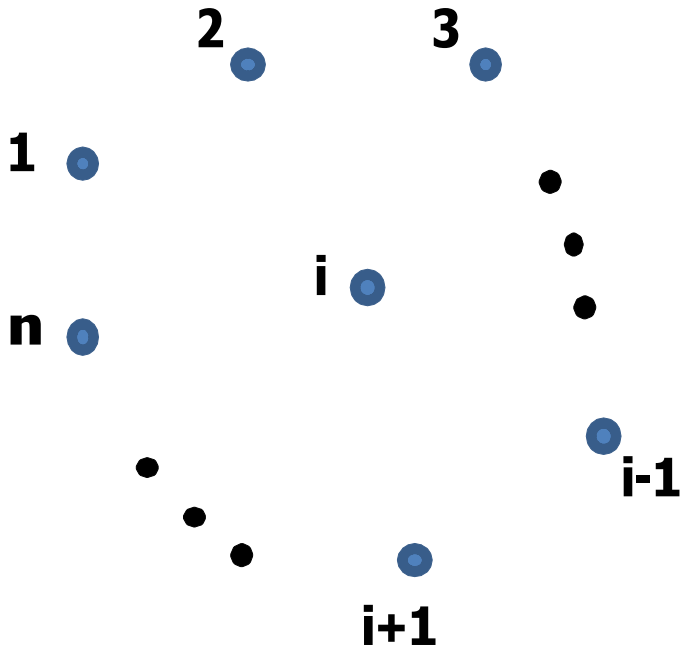
1. La sequenza di indice i è selezionata come centro della stella
2. Per ogni coppia $(i, j), j \neq i$ si calcola la distanza di edit fra le due sequenze
3. Si minimizza la funzione di score d

Allineamenti basati su struttura

Star alignment

- **Idea**: selezionare una sequenza fra quelle in esame ed utilizzarla come “cardine”

1. La sequenza di indice i è selezionata come centro della stella
2. Per ogni coppia $(i, j), j \neq i$ si calcola la distanza di edit fra le due sequenze
3. Si minimizza la funzione di score d

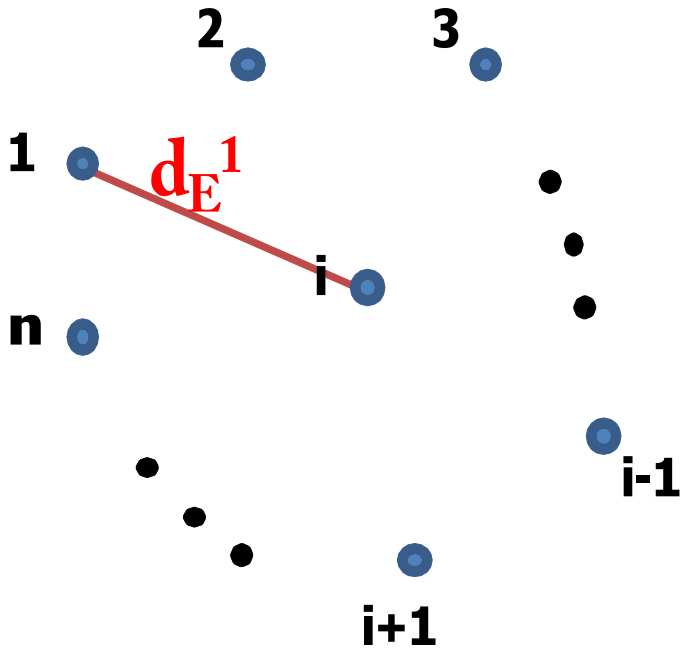


Allineamenti basati su struttura

Star alignment

- **Idea**: selezionare una sequenza fra quelle in esame ed utilizzarla come "cardine"

1. La sequenza di indice i è selezionata come centro della stella
2. Per ogni coppia $(i, j), j \neq i$ si calcola la distanza di edit fra le due sequenze
3. Si minimizza la funzione di score d

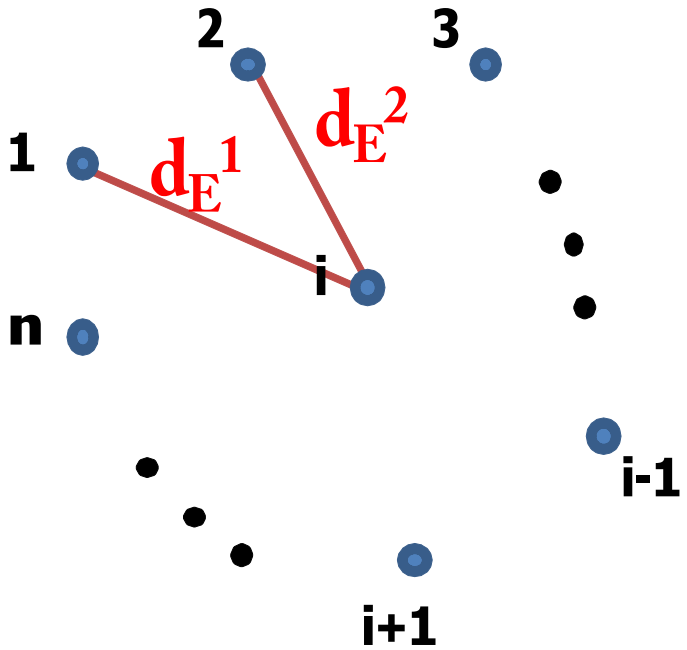


Allineamenti basati su struttura

Star alignment

- **Idea**: selezionare una sequenza fra quelle in esame ed utilizzarla come "cardine"

1. La sequenza di indice i è selezionata come centro della stella
2. Per ogni coppia $(i, j), j \neq i$ si calcola la distanza di edit fra le due sequenze
3. Si minimizza la funzione di score d

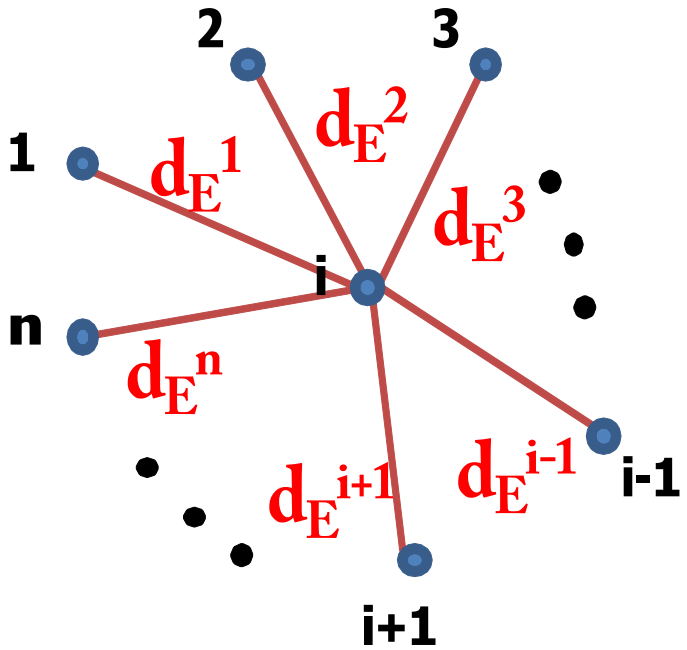


Allineamenti basati su struttura

Star alignment

- **Idea**: selezionare una sequenza fra quelle in esame ed utilizzarla come “cardine”

1. La sequenza di indice i è selezionata come centro della stella
2. Per ogni coppia $(i, j), j \neq i$ si calcola la distanza di edit fra le due sequenze
3. Si minimizza la funzione di score d

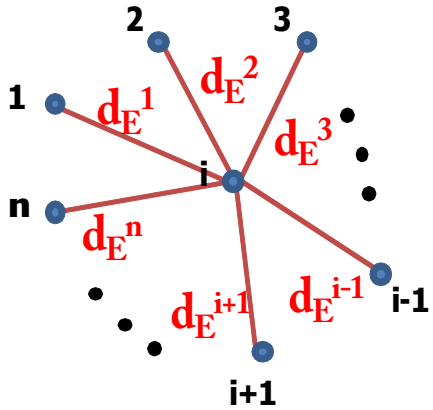


Allineamenti basati su struttura

Star alignment

- **Idea**: selezionare una sequenza fra quelle in esame ed utilizzarla come "cardine"

1. La sequenza di indice i è selezionata come centro della stella
2. Per ogni coppia $(i, j), j \neq i$ si calcola la distanza di edit fra le due sequenze
3. Si minimizza la funzione di score d



$$\min \sum_{k=1}^{n} d_E(k, i)$$

Allineamenti basati su struttura

Allineamenti basati su grafi

- *Should multiple sequence alignment be linear?* [*Lee et al.* - 2002]
- Appiattire troppo la struttura del problema può determinare la perdita di alcune informazioni (specie nel caso dell'allineamento di proteine)
- La struttura di valutazione a grafo è stata proposta nel tentativo di rispettare la struttura tridimensionale delle proteine

ALLINEAMENTI BASATI SU STRUTTURA

I TOOL

- **Allineamento a stella**: *SAlign*, *Modeller*
- **Allineamento ad albero**: *CLUSTAL*, *Multi-Lagan*, *T-Coffee*, **...**
- **Allineamenti su grafi**: *POA (Partial Order Alignment)*, *ABA (A-Brujn Alignment)*

APPROCCI PROBABILISTICI

- Il problema è formulato in termini di un modello matematico
- **HMM** (**H**idden **M**arkov **M**odel): modello probabilistico, in cui si assume che il sistema modellato sia un processo di Markov a parametri sconosciuti
- L'informazione probabilistica può (eventualmente) essere sfruttata per costruire strutture di valutazione (alberi, grafi...)

Approcci probabilistici

I tool

- *ProbCons, ProbAlign*
- *SAM*
- *AMAP*

Approcci alternativi

1/2

- **MAFFT**: metodo basato sulla codifica dell'allineamento multiplo in termini di trasformate discrete di Fourier (DFT)
- L'algoritmo FFT (***F**ast **F**ourier **T**ransform*) permette di calcolare le trasformate in maniera molto efficiente (complessità in tempo $O(N \cdot logN)$)
- Consente di ottenere molto rapidamente buone approssimazioni della soluzione esatta

Approcci alternativi
2/2

- **SAGA** (**S**equence **A**lignment by **G**enetic **A**lgorithm): metodo di allineamento multiplo basato sull'uso di un algoritmo genetico
- **Algoritmi genetici**: usati per risolvere problemi di ricerca e ottimizzazione, simulano l'evoluzione di una popolazione di possibili soluzioni, in base al principio biologico della selezione naturale
- La popolazione iniziale è generata (pseudo-)casualmente e la qualità dei singoli individui è misurata mediante una *funzione di fitness*

MLAGAN E CLUSTALW

- **MLAGAN** e **CLUSTALW** sono tool che realizzano l'allineamento di più sequenze (multiple alignment)
- Entrambi si basano su:
  - Allineamento progressivo
  - Alberi filogenetici

Allineamento progressivo

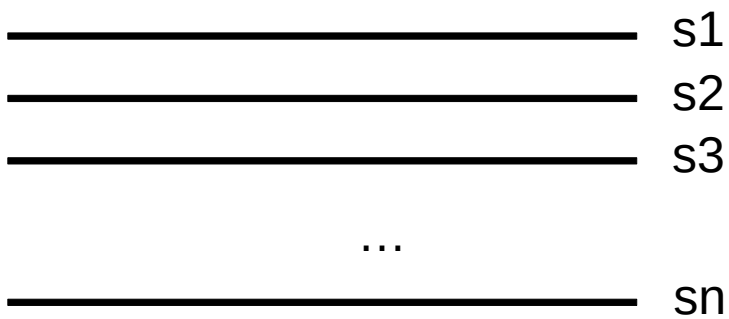
1/3

- Per affrontare il problema dell'allineamento multiplo in tempi accettabili c'è bisogno di metodi che facciano uso di euristiche
- Il metodo più usato è quello dell'**allineamento progressivo**
- L'allineamento progressivo consiste nella costruzione progressiva di allineamenti di coppie di sequenze

Allineamento progressivo

2/3

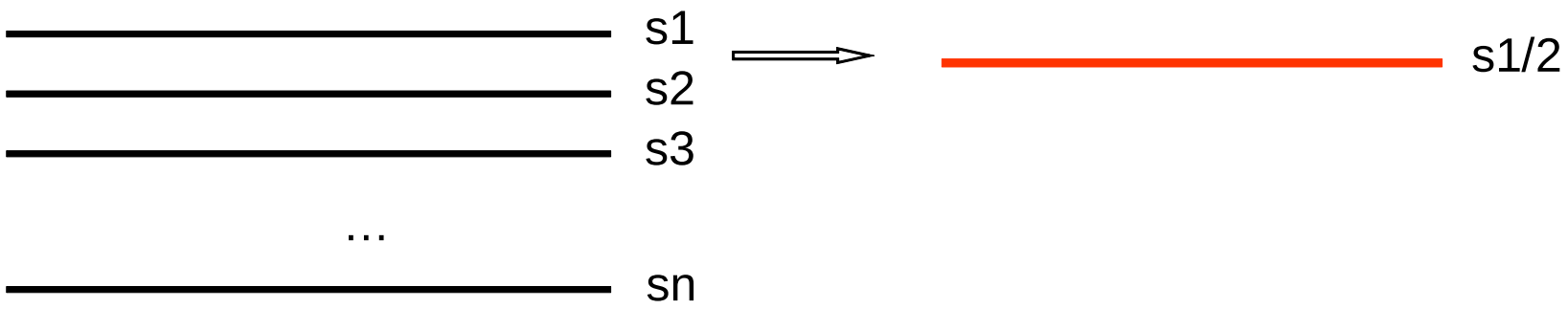
- Dato un insieme di sequenze da allineare, vengono scelte e allineate due sequenze s1 ed s2, da cui si ottiene una nuova sequenza
- Viene scelta poi una terza sequenza s3 da allineare al precedente allineamento, e così via



Allineamento progressivo

2/3

- Dato un insieme di sequenze da allineare, vengono scelte e allineate due sequenze s1 ed s2, da cui si ottiene una nuova sequenza
- Viene scelta poi una terza sequenza s3 da allineare al precedente allineamento, e così via



Allineamento progressivo

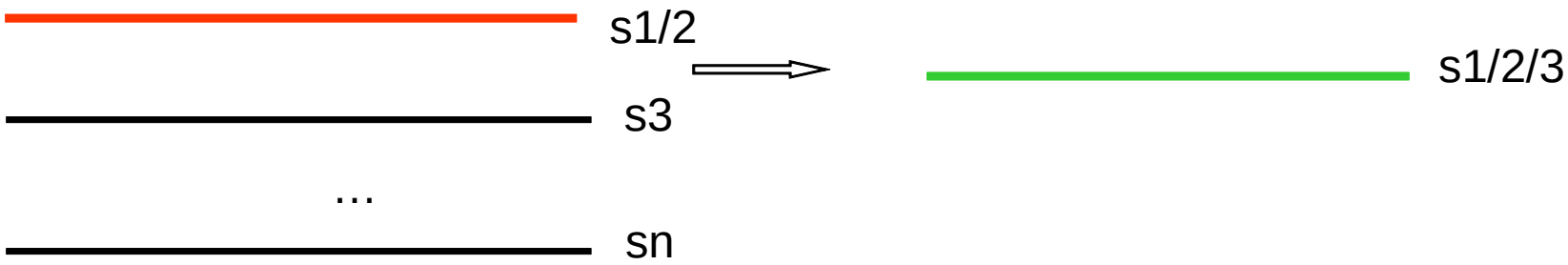
2/3

- Dato un insieme di sequenze da allineare, vengono scelte e allineate due sequenze s1 ed s2, da cui si ottiene una nuova sequenza
- Viene scelta poi una terza sequenza s3 da allineare al precedente allineamento, e così via

Allineamento progressivo

2/3

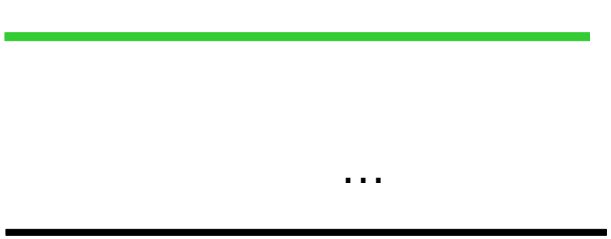
- Dato un insieme di sequenze da allineare, vengono scelte e allineate due sequenze s1 ed s2, da cui si ottiene una nuova sequenza
- Viene scelta poi una terza sequenza s3 da allineare al precedente allineamento, e così via



Allineamento progressivo

2/3

- Dato un insieme di sequenze da allineare, vengono scelte e allineate due sequenze s1 ed s2, da cui si ottiene una nuova sequenza
- Viene scelta poi una terza sequenza s3 da allineare al precedente allineamento, e così via



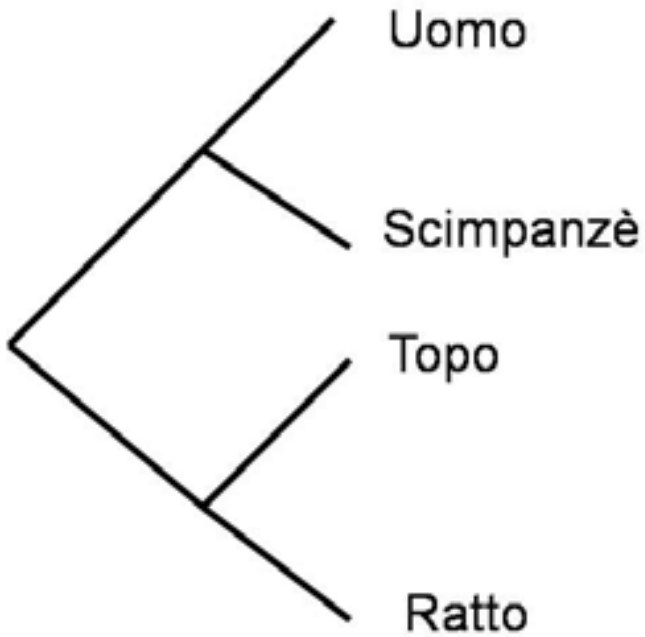
Considerazione: sequenze genetiche simili derivano da organismi “parenti”

L’allineamento progressivo “sfrutta” questa considerazione

ALLINEAMENTO PROGRESSIVO

3/3

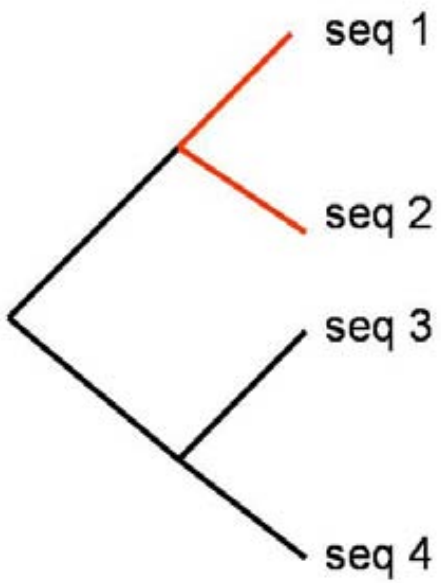
- **Regola**: ad ogni passo allineare le sequenze più simili
- Un albero filogenetico è un albero binario che mette in relazione di “parentela” specie odierne e antenati comuni
  - Foglie: specie attuali
  - Nodi: specie ancestrali



ALLINEAMENTO PROGRESSIVO

3/3

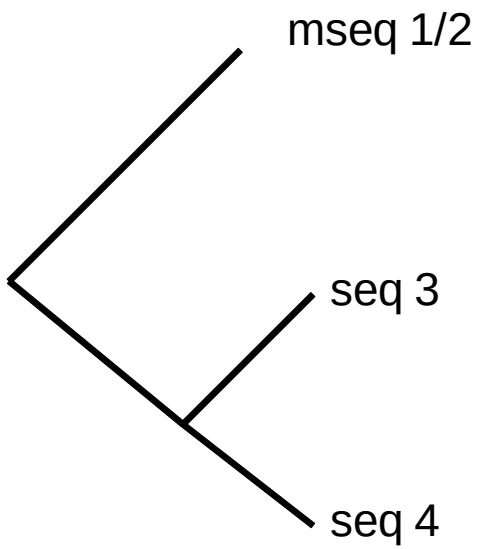
- **Regola**: ad ogni passo allineare le sequenze più simili



Allineamento progressivo

3/3

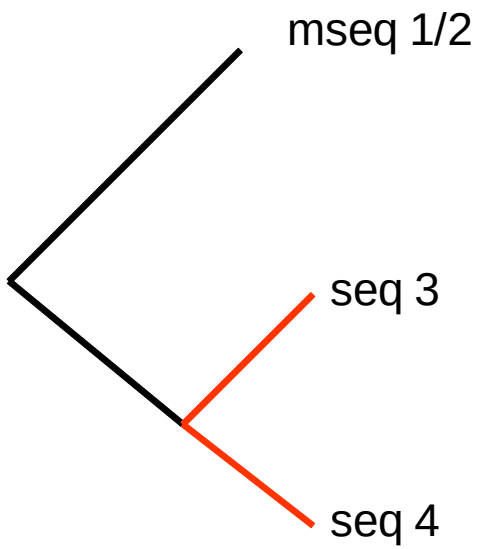
- **Regola**: ad ogni passo allineare le sequenze più simili



Allineamento progressivo

3/3

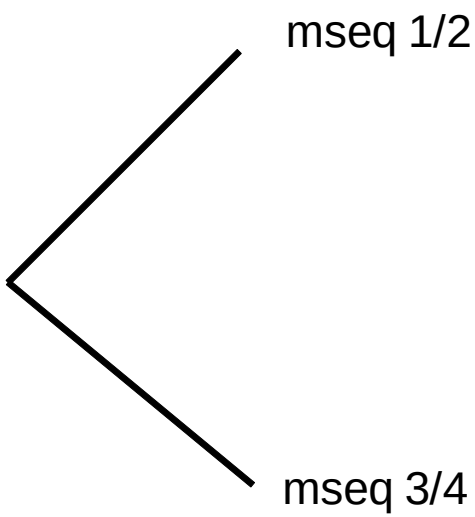
- **Regola**: ad ogni passo allineare le sequenze più simili



Allineamento progressivo

3/3

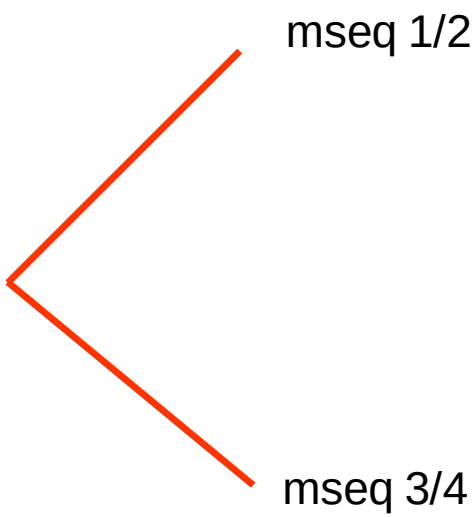
- **Regola**: ad ogni passo allineare le sequenze più simili



Allineamento progressivo

3/3

- **Regola**: ad ogni passo allineare le sequenze più simili



Allineamento progressivo

3/3

- **Regola**: ad ogni passo allineare le sequenze più simili

mseq 1/2/3/4

MultiLagan

Lagan e Mlagan

- **LAGAN**: realizza l'allineamento globale tra coppie di sequenze (*global pairwise alignment*)
- **MLAGAN**: realizza l'allineamento globale di più sequenze (*global multiple alignment*) effettuando progressivamente gli allineamenti pairwise, tramite LAGAN (*progressive alignment*)

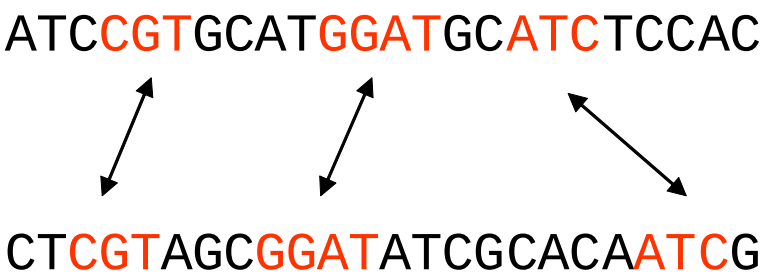
LAGAN 1/2

- LAGAN (Limited Area Global Alignment of Nucleotides) è un tool per l'allineamento di due sequenze
- È utile che le sequenze da analizzare siano ortologhe
  - *LAGAN and MLAGAN assume that one has already identified apparent orthologous regions between two species, and that there are no genomic rearrangements*

LAGAN 2/2

LAGAN allinea una coppia di sequenze in tre passi principali:

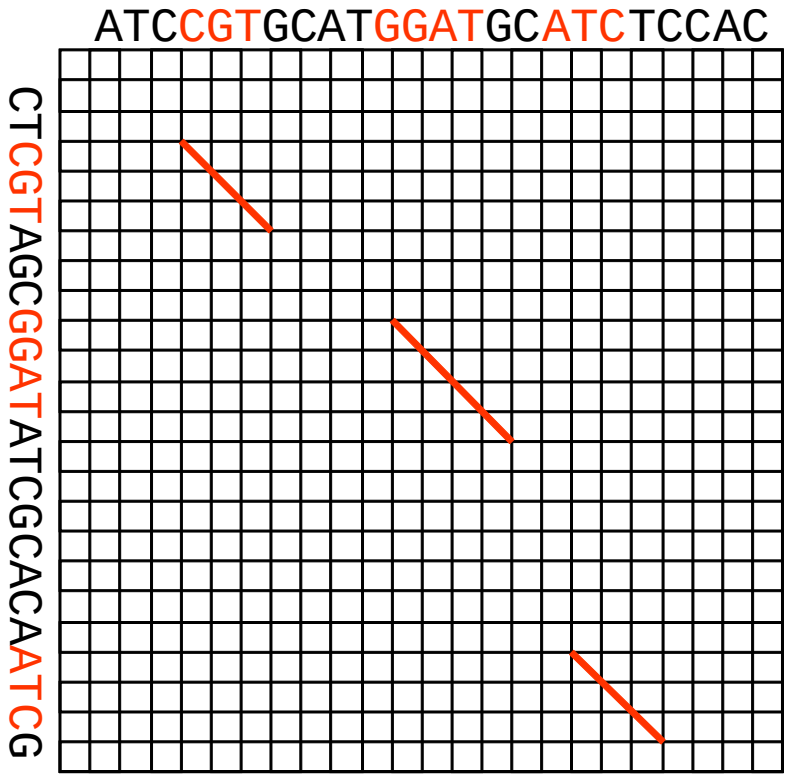
1. Generazione allineamenti locali tra 2 sequenze (CHAOS)
2. Concatenazione sottoinsieme degli allineamenti locali e costruzione di una prima global map
3. Calcolo dell'allineamento globale utilizzando la global map



LAGAN 2/2

LAGAN allinea una coppia di sequenze in tre passi principali:

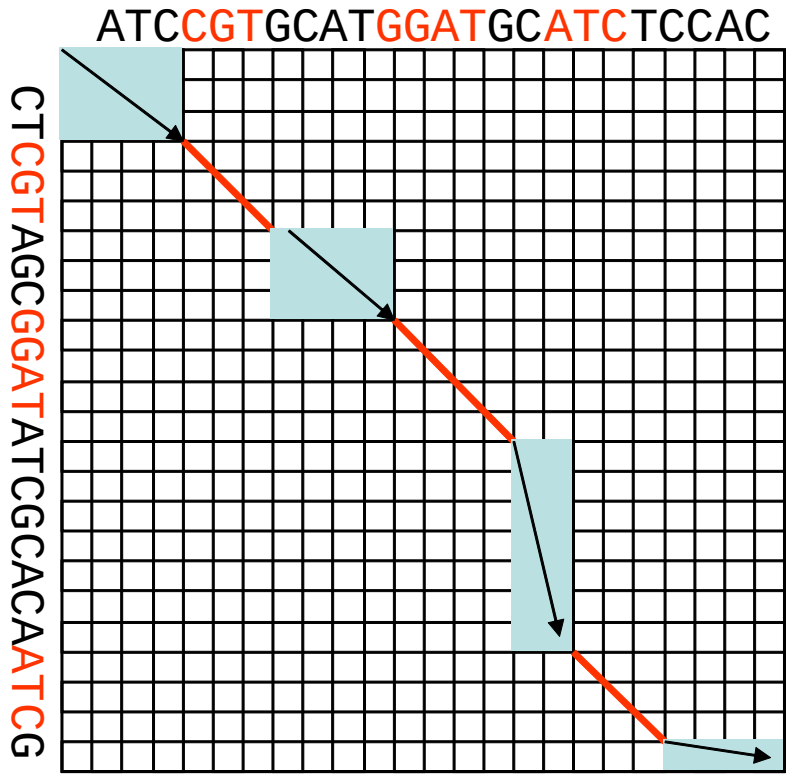
1. Generazione allineamenti locali tra 2 sequenze (CHAOS)
2. Concatenazione sottoinsieme degli allineamenti locali e costruzione di una prima global map
3. Calcolo dell'allineamento globale utilizzando la global map



LAGAN 2/2

LAGAN allinea una coppia di sequenze in tre passi principali:

1. Generazione allineamenti locali tra 2 sequenze (CHAOS)
2. Concatenazione sottoinsieme degli allineamenti locali e costruzione di una prima global map
3. Calcolo dell'allineamento globale utilizzando la global map



Prima fase: allineamenti locali

1/5

- **CHAOS** (CHAins Of Seeds): algoritmo utilizzato di default da LAGAN per gli allineamenti locali
- Trova omologie locali tra due sequenze e le concatena, costruendo allineamenti locali (anchor)
- Per prima cosa sono trovate piccole sotto-sequenze (**seed**) comuni alle due sequenze principali
- Un seed non deve necessariamente presentarsi identico in entrambe le sequenze

PRIMA FASE: ALLINEAMENTI LOCALI

- Dati:
  - una lunghezza k,
  - un numero massimo di differenze c,
- un $(k, c) - seed$ è un seed lungo k che può avere c differenze in entrambe le sequenze

PRIMA FASE: ALLINEAMENTI LOCALI

- Dati:
  - una lunghezza k,
  - un numero massimo di differenze c,
- un $(k, c) - seed$ è un seed lungo k che può avere c differenze in entrambe le sequenze

ESEMPIO

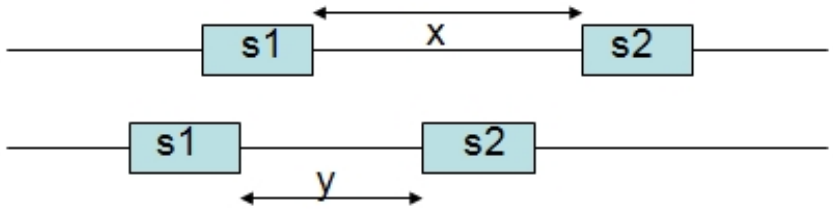
...GGTGCTTGTA...
...CAGATTATCT...
(6,2)-seed = (GCTTGT, GATTAT)

PRIMA FASE: ALLINEAMENTI LOCALI

3/5

- CHAOS dopo aver trovato i seed, cerca di concatenarli creando allineamenti locali (anchor). Dati:
  - d: massima distanza
  - s: massimo shift
- Due seed distanti x e y, rispettivamente nella prima e seconda sequenza, possono essere concatenati se:

- $x \leq d$
- $y \leq d$
- $|x - y| \leq s$



PRIMA FASE: ALLINEAMENTI LOCALI

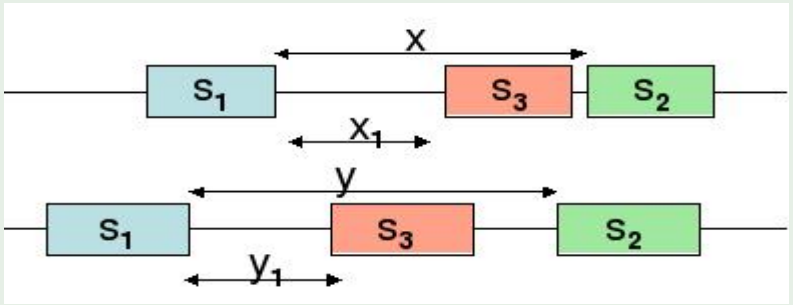
- È possibile che un seed s_1 possa soddisfare le relazioni appena viste con più di un singolo altro seed

ESEMPIO

$\mathbf{x \leq d}$ $\quad$ $\mathbf{x_1 \leq d}$

$\mathbf{y \leq d}$ $\quad$ $\mathbf{y_1 \leq d}$

$\mathbf{|x - y| \leq s}$ $\quad$ $\mathbf{|x_1 - y_1| \leq s}$



PRIMA FASE: ALLINEAMENTI LOCALI

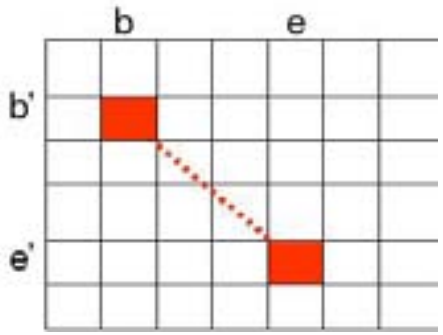
5/5

- In tal caso s_1 è concatenato ad un seed con cui crei una catena di score massimo
- **Score di una catena**: viene assegnata in base ai soliti princìpi generali di punteggio
  - Bonus per match fra caratteri dei seed
  - Malus per mismatch fra caratteri dei seed
  - Malus per gap

SECONDA FASE: GLOBAL MAP

1/3

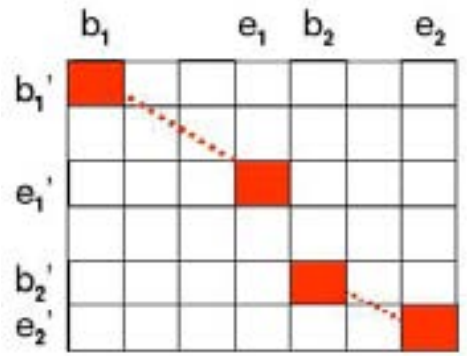
- LAGAN ordina gli allineamenti locali prodotti da CHAOS in una global map
- Un allineamento locale è un vettore (b, e, b', e', s) che rappresenta
  1. la posizione iniziale e finale (**b**egin, **e**nd) dell'allineamento nelle due sequenze
  2. lo score dell'allineamento



SECONDA FASE: GLOBAL MAP

2/3

- Dati due allineamenti locali:
  - $A_1 = (b_1, e_1, b'_1, e'_1, s_1)$
  - $A_2 = (b_2, e_2, b'_2, e'_2, s_2)$



- $A_1 < A_2$ sse:
  - $e_1 < b_2$
  - $e'_1 < b'_2$
- Una catena di allineamenti locali $A_1 < A_2 < ... < A_k$ ha score $s_1 + s_2 + ... + s_k$

SECONDA FASE: GLOBAL MAP

3/3

- La global map ottima è quella con lo score più alto
- Può essere calcolata usando Sparse Dynamic Programming
- Il calcolo ha complessità in tempo $\mathbf{O}(\mathbf{n} \cdot \mathbf{logn})$, dove n è il numero di allineamenti locali considerati

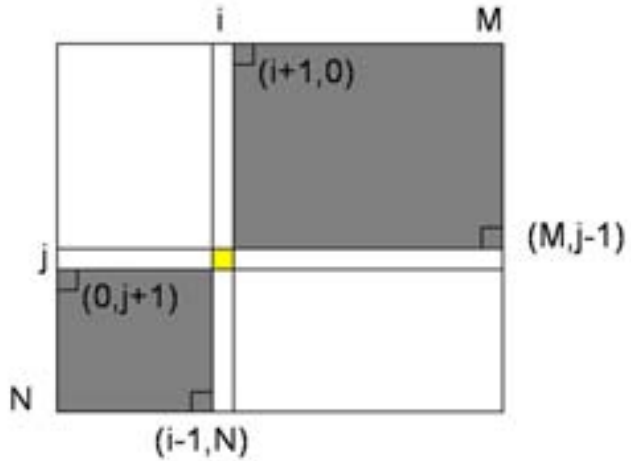
COSTRUZIONE DELL'ALLINEMENTO GLOBALE
1/2

- L'algoritmo di Needleman-Wunsch utilizza una matrice $(M+1) \times (N+1)$
- M e N sono le lunghezze delle sequenze da allineare
- L'algoritmo calcola il valore di ogni cella della matrice e calcola il path con lo score più alto dalla cella $[0,0]$ fino alla cella $[N,M]$
- L'algoritmo ha complessità in tempo $\mathbf{O(N \cdot M)}$

COSTRUZIONE DELL'ALLINEMENTO GLOBALE 2/2

Lagan, per velocizzare il calcolo, adotta la seguente strategia:

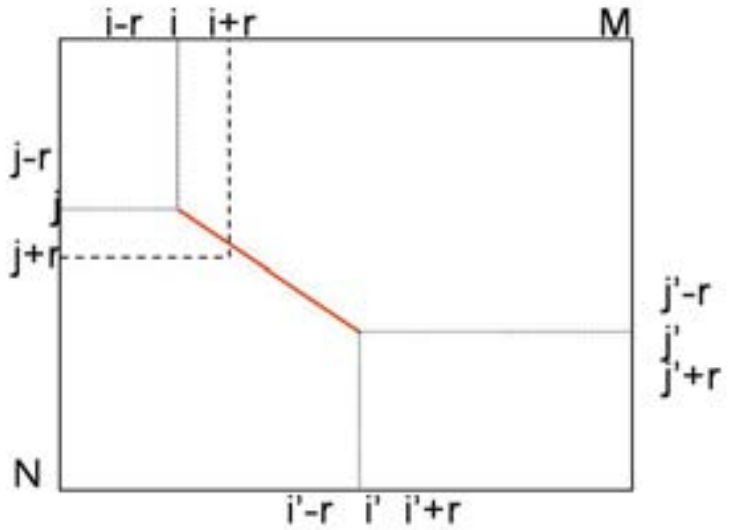
- qualora sia noto che un certo allineamento passa per la casella $[i, j]$ si evita di calcolare il valore delle celle all'interno dei rettangoli delimitati dalle celle $[i+1, 0], [M, j-1]$ e da $[0, j+1], [i-1, N]$



TERZA FASE: ALLINEMENTO GLOBALE
1/2

- L'algoritmo riceve in input un parametro **r** (che delimita l'area di ricerca) e per ogni allineamento da $[i, j]$ a $[i', j']$ valuta lo score delle celle incluse nell'unione formata:

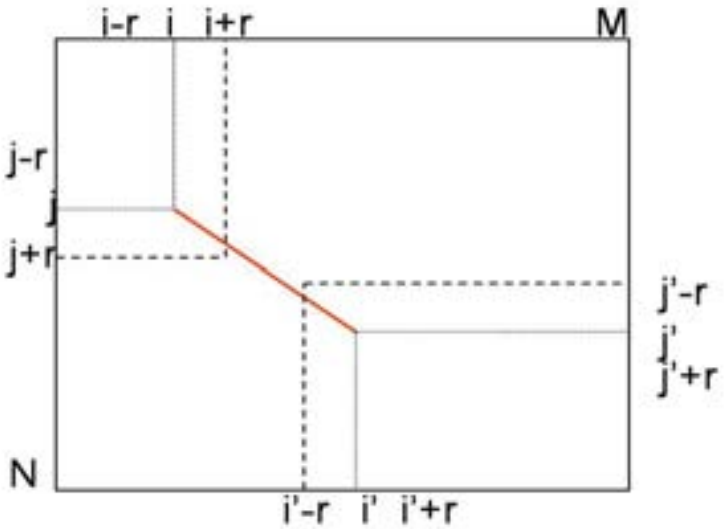
- dal rettangolo da $[0, 0]$ a $[i + r, j + r]$
- dal rettangolo da $[i' - r, j' - r]$ a $[M, N]$
- dalle due diagonali da $[i - r, j + r]$ a $[i' - r, j' + r]$ e da $[i + r, j - r]$ a $[i' + r, j' - r]$



TERZA FASE: ALLINEMENTO GLOBALE
1/2

- L'algoritmo riceve in input un parametro **r** (che delimita l'area di ricerca) e per ogni allineamento da $[i, j]$ a $[i', j']$ valuta lo score delle celle incluse nell'unione formata:

- dal rettangolo da $[0, 0]$ a $[i + r, j + r]$
- dal rettangolo da $[i' - r, j' - r]$ a $[M, N]$
- dalle due diagonali da $[i - r, j + r]$ a $[i' - r, j' + r]$ e da $[i + r, j - r]$ a $[i' + r, j' - r]$

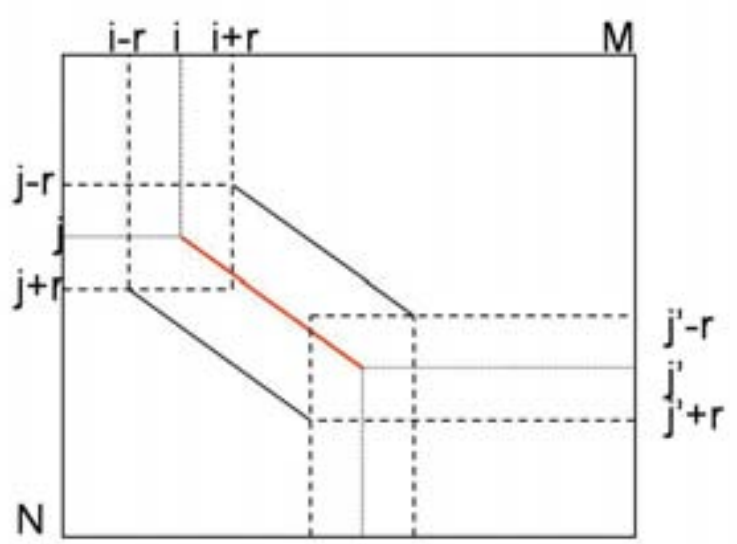


TERZA FASE: ALLINEMENTO GLOBALE

1/2

- L'algoritmo riceve in input un parametro **r** (che delimita l'area di ricerca) e per ogni allineamento da $[i, j]$ a $[i', j']$ valuta lo score delle celle incluse nell'unione formata:

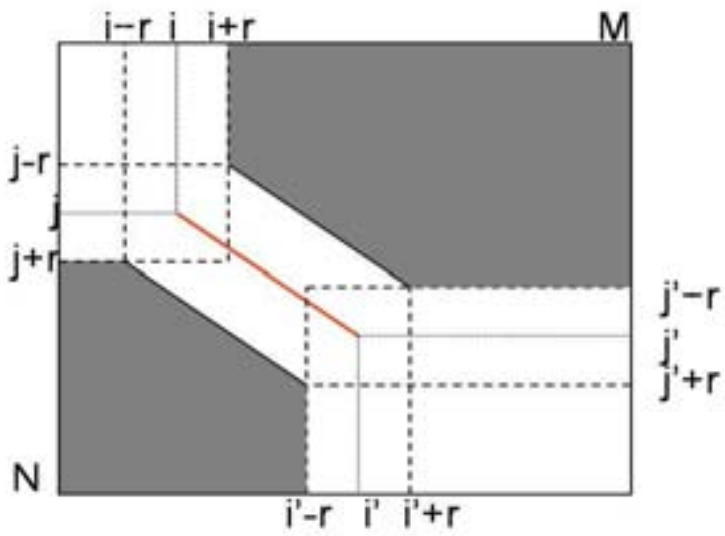
- dal rettangolo da $[0, 0]$ a $[i + r, j + r]$
- dal rettangolo da $[i' - r, j' - r]$ a $[M, N]$
- dalle due diagonali da $[i - r, j + r]$ a $[i' - r, j' + r]$ e da $[i + r, j - r]$ a $[i' + r, j' - r]$



TERZA FASE: ALLINEMENTO GLOBALE
1/2

- L'algoritmo riceve in input un parametro **r** (che delimita l'area di ricerca) e per ogni allineamento da $[i, j]$ a $[i', j']$ valuta lo score delle celle incluse nell'unione formata:

- dal rettangolo da $[0, 0]$ a $[i + r, j + r]$
- dal rettangolo da $[i' - r, j' - r]$ a $[M, N]$
- dalle due diagonali da $[i - r, j + r]$ a $[i' - r, j' + r]$ e da $[i + r, j - r]$ a $[i' + r, j' - r]$



TERZA FASE: ALLINEMENTO GLOBALE
2/2

- Gli anchor forniscono aree della global map dalle quali deve passare l'allineamento
- La complessità in tempo dell'algoritmo dipende dal numero di celle comprese fra anchor consecutivi

MULTILAGAN

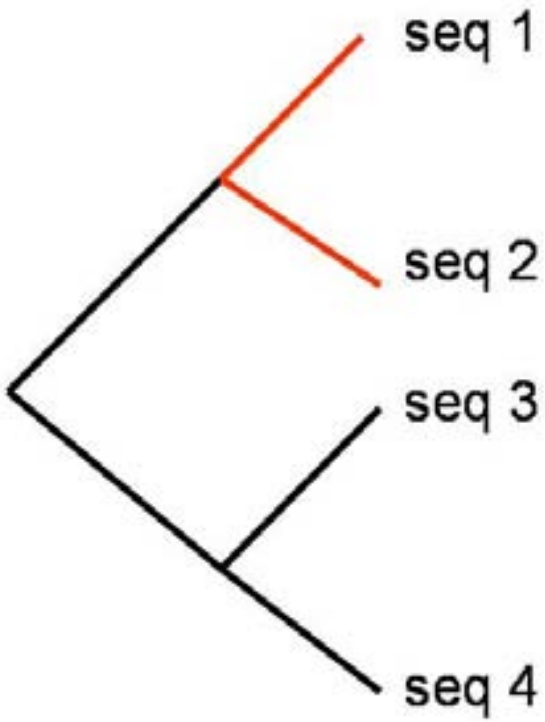
- **MLAGAN** (MultiLagan) è un tool per l'allineamento globale di più sequenze
- Basato su allineamenti progressivi effettuati mediante LAGAN
- Presuppone un albero filogenetico in ingresso, su cui basare gli allineamenti progressivi
- La richiesta ha un senso qualora si cerchino allineamenti fra sequenze ortologhe appartenenti a specie di cui sia noto l'albero filogenetico

MLAGAN

Algoritmo

- Dati *N* sequenze ed un albero filogenetico

1. **Allineamento globale pairwise di sequenze o multi-sequenze**
2. Iterazione punto 1
3. Miglioramento (opzionale)

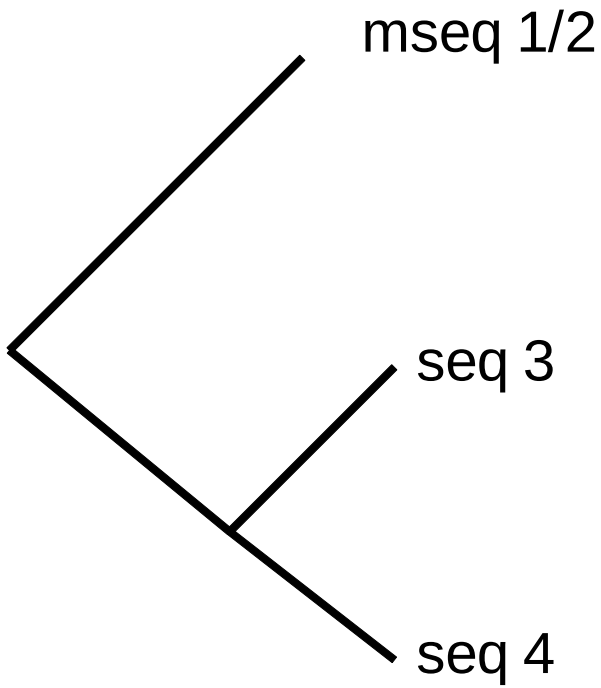


MLAGAN

Algoritmo

- Dati *N* sequenze ed un albero filogenetico

1. **Allineamento globale pairwise di sequenze o multi-sequenze**
2. Iterazione punto 1
3. Miglioramento (opzionale)

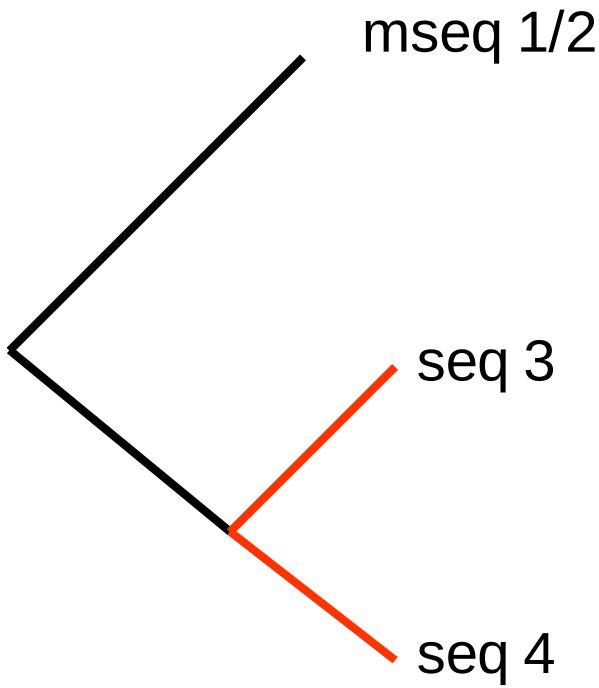


MLAGAN

Algoritmo

- Dati *N* sequenze ed un albero filogenetico

1. **Allineamento globale pairwise di sequenze o multi-sequenze**
2. **Iterazione punto 1**
3. **Miglioramento (opzionale)**

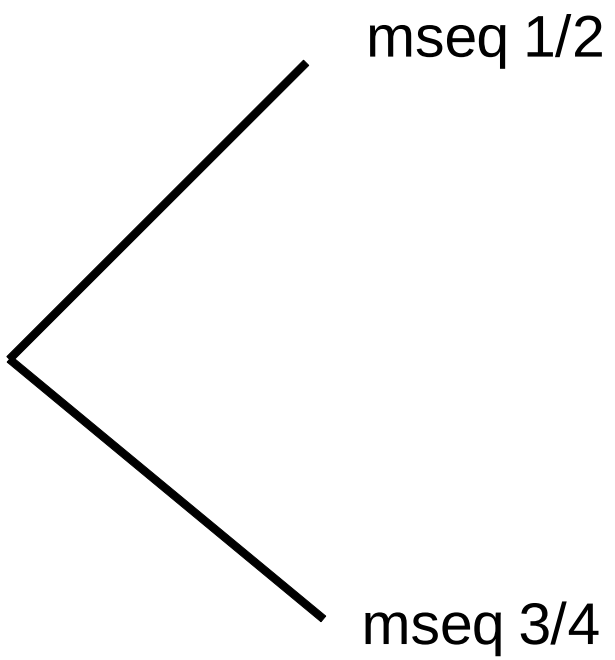


MLAGAN

Algoritmo

- Dati *N* sequenze ed un albero filogenetico

1. **Allineamento globale pairwise di sequenze o multi-sequenze**
2. **Iterazione punto 1**
3. **Miglioramento (opzionale)**

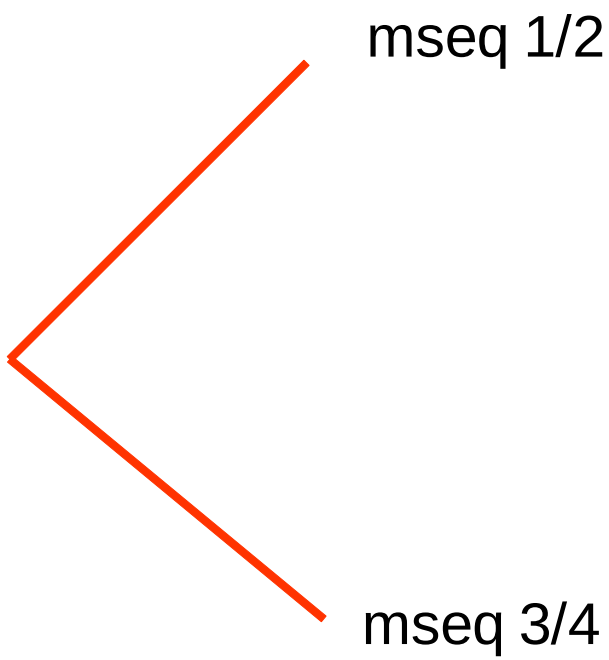


MLAGAN

Algoritmo

- Dati *N* sequenze ed un albero filogenetico

1. **Allineamento globale pairwise di sequenze o multi-sequenze**
2. **Iterazione punto 1**
3. **Miglioramento (opzionale)**



MLAGAN

Algoritmo

- Dati *N* sequenze ed un albero filogenetico

1. **Allineamento globale pairwise di sequenze o multi-sequenze**
2. **Iterazione punto 1**
3. **Miglioramento (opzionale)**

mseq 1/2/3/4

MLAGAN

Algoritmo

- Dati *N* sequenze ed un albero filogenetico

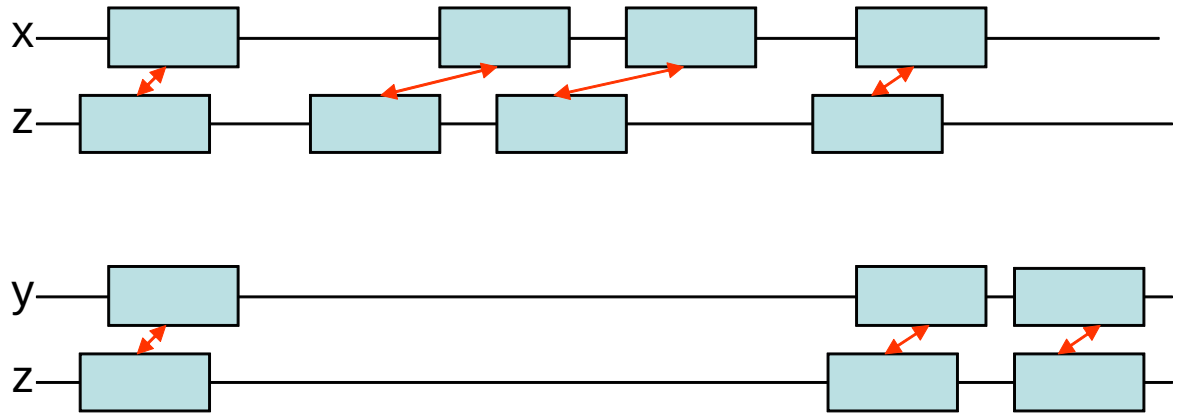
1. **Allineamento globale pairwise di sequenze o multi-sequenze**
2. **Iterazione punto 1**
3. **Miglioramento (opzionale)**

mseq 1/2/3/4

Allineamento globale pairwise (passi 1 e 2)

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi

- **Passo 1**
  - **A**: gli anchor tra X e Z che non si sovrappongono con gli anchor tra Y e Z, divengono anchor tra X/Y e Z, mantenendo il proprio score

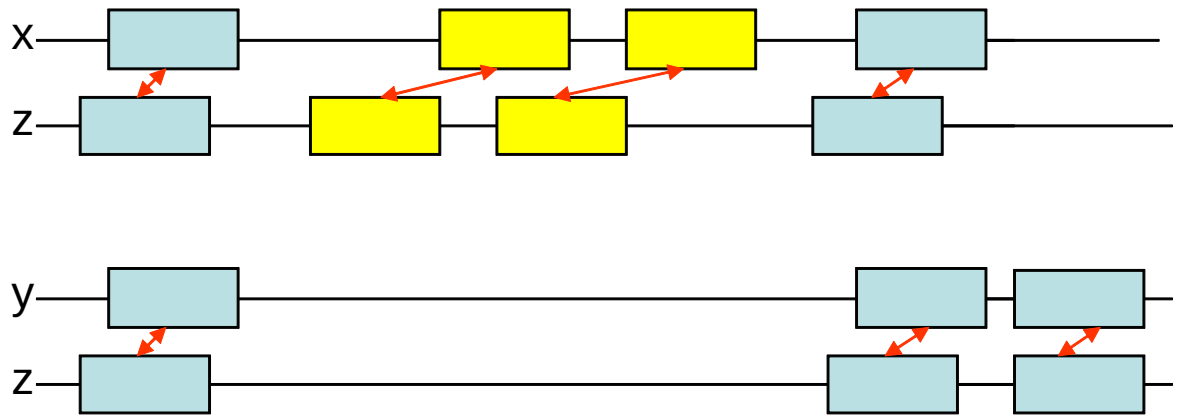


Allineamento globale pairwise (passi 1 e 2)

1/3

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi

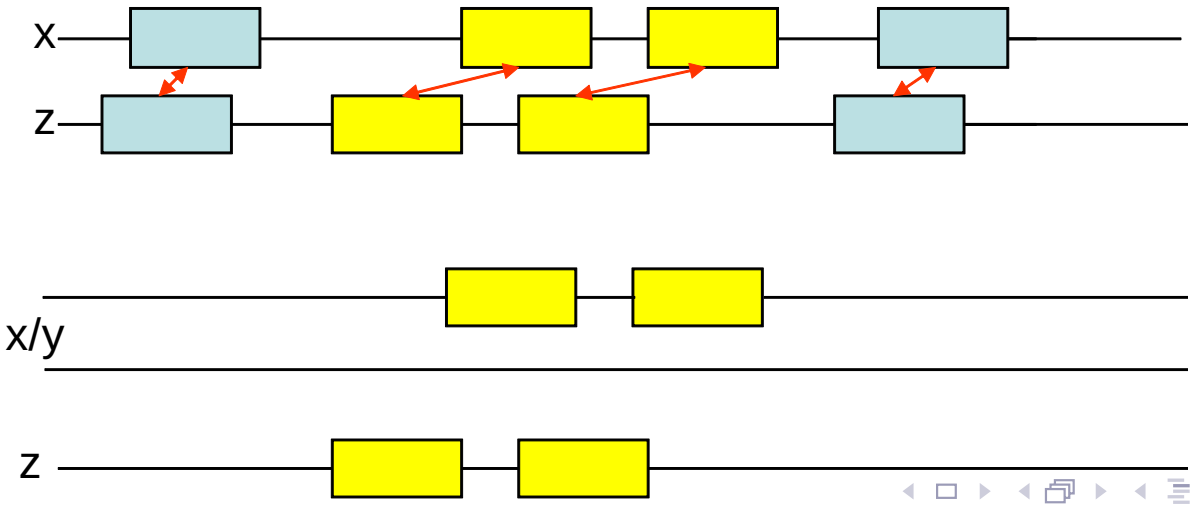
- **Passo 1**
  - **A**: gli anchor tra X e Z che non si sovrappongono con gli anchor tra Y e Z, divengono anchor tra X/Y e Z, mantenendo il proprio score



Allineamento globale pairwise (passi 1 e 2)

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi

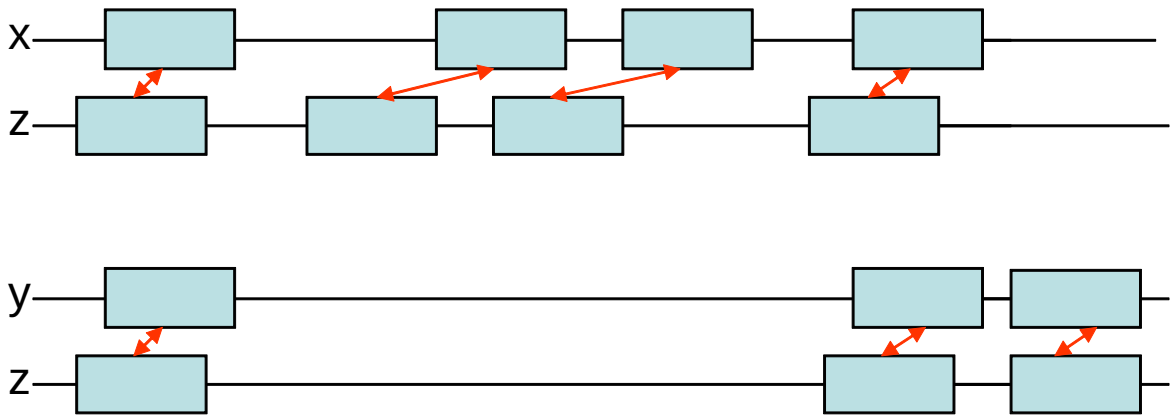
- **Passo 1**
  - **A**: gli anchor tra X e Z che non si sovrappongono con gli anchor tra Y e Z, divengono anchor tra X/Y e Z, mantenendo il proprio score



ALLINEAMENTO GLOBALE PAIRWISE (PASSI 1 E 2)

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi

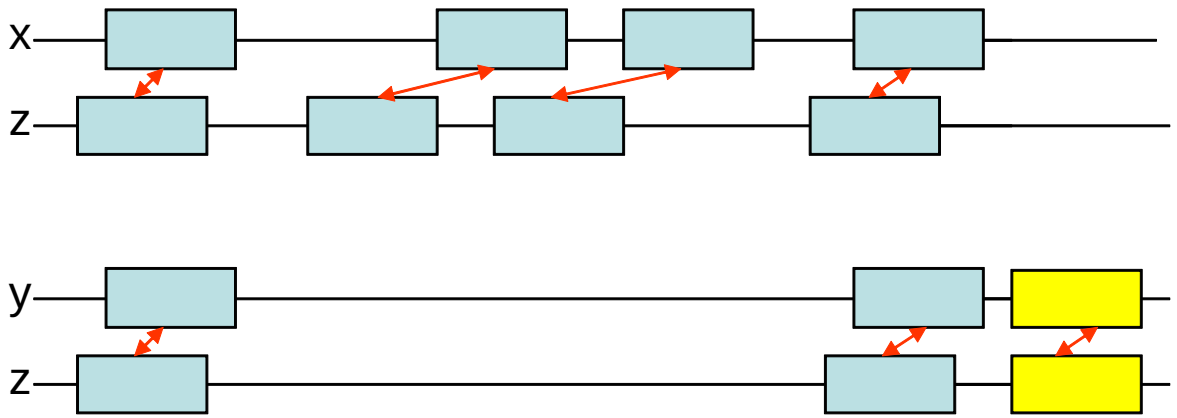
- **Passo 1**
  - **B**: gli anchor tra Y e Z che non si sovrappongono con gli anchor tra X e Z, divengono anchor tra X/Y e Z



Allineamento globale pairwise (passi 1 e 2)

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi

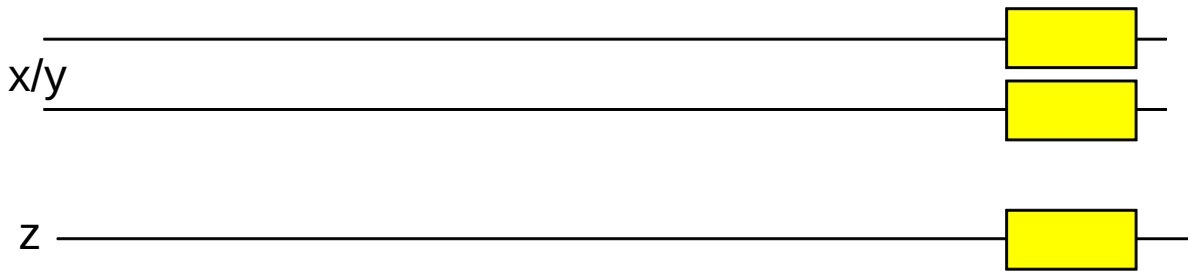
- **Passo 1**
  - **B**: gli anchor tra Y e Z che non si sovrappongono con gli anchor tra X e Z, divengono anchor tra X/Y e Z



ALLINEAMENTO GLOBALE PAIRWISE (PASSI 1 E 2)

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi

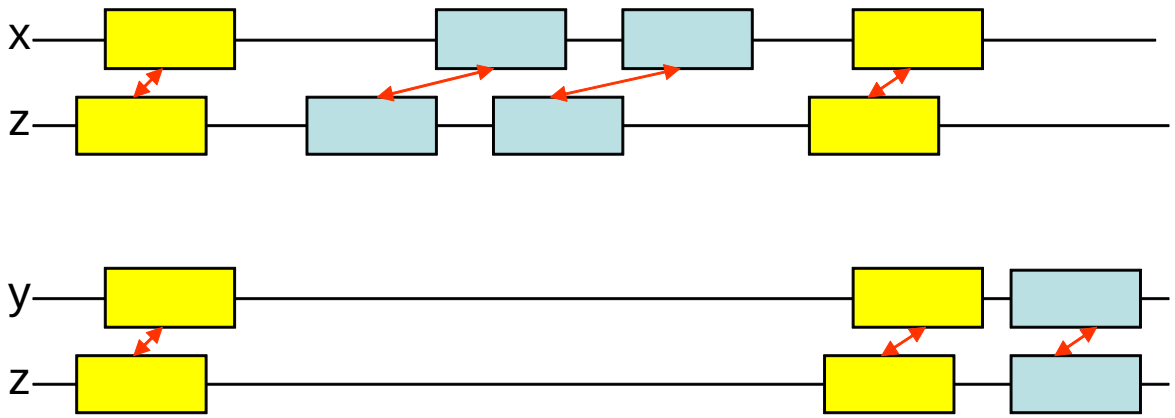
- **Passo 1**
  - **B**: gli anchor tra Y e Z che non si sovrappongono con gli anchor tra X e Z, divengono anchor tra X/Y e Z



Allineamento globale pairwise (passi 1 e 2)

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi

- **Passo 2**: di ogni anchor tra X e Z che si sovrappone anche in parte con un anchor tra Y e Z, viene ricalcolato lo score

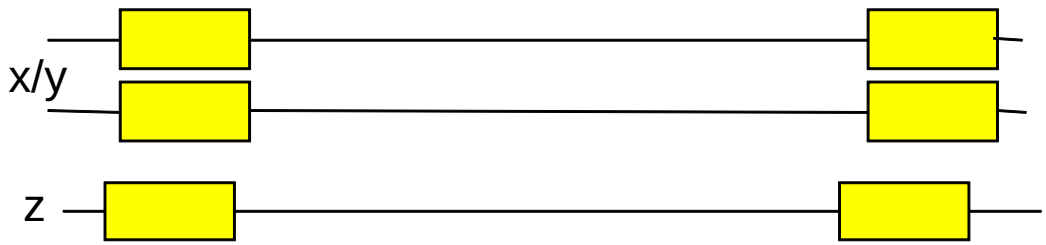


Allineamento globale pairwise (passi 1 e 2)

1/3

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi

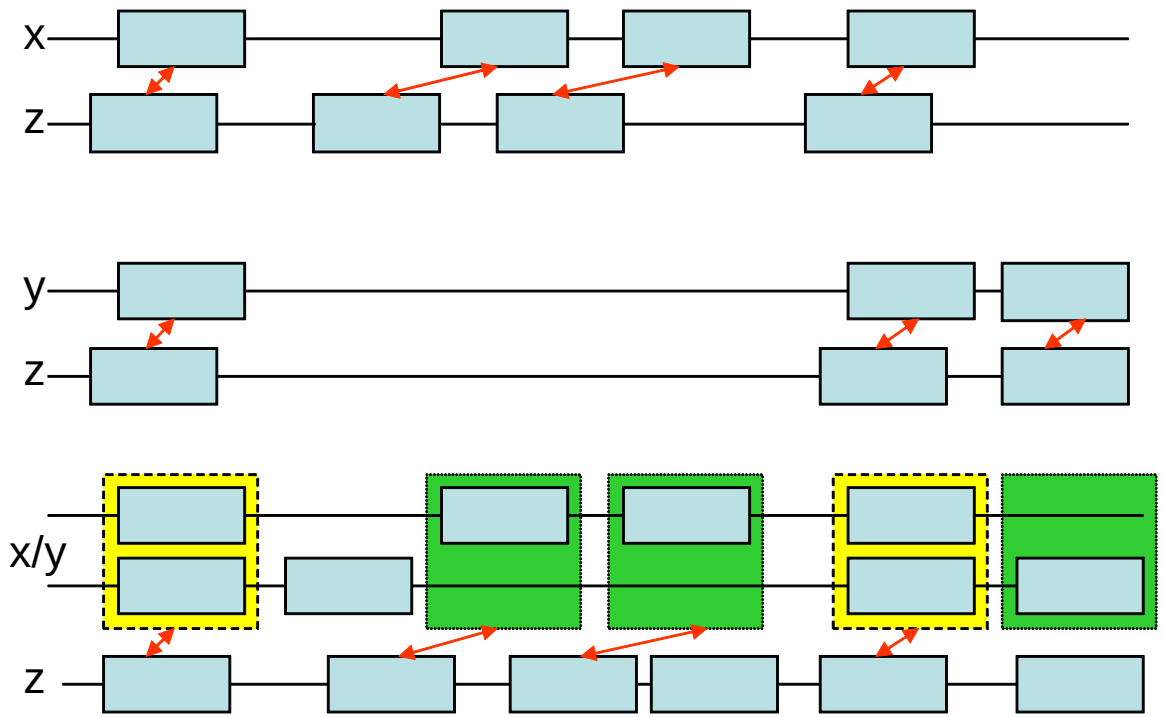
- **Passo 2**: di ogni anchor tra X e Z che si sovrappone anche in parte con un anchor tra Y e Z, viene ricalcolato lo score



Allineamento globale pairwise (passi 1 e 2)

1/3

Per allineare una multiseq. X/Y con una (multi)seq. Z è necessario costruire una global map, che è generata in 2 passi



Miglioramento [opzionale] (passo 3)

1/2

- Un difetto dell'allineamento progressivo è che gli allineamenti pairwise iniziali sono immodificabili, anche se, andando avanti negli allineamenti, vengono trovate alcune inesattezze negli allineamenti precedenti
- Soluzione: raffinamento iterativo
- MLAGAN introduce una versione del raffinamento circoscritta ad aree ridotte (*limited area version of iterative refinement*)

Miglioramento [opzionale] (passo 3)

2/2

- Iterativamente ogni sequenza viene rimossa dall'allineamento globale
- Ogni regione della sequenza rimossa che migliora significativamente lo score dell'allineamento diventa un anchor
- La sequenza è riallineata all'allineamento multiplo (utilizzando LAGAN)

SCORE DELL'ALLINEAMENTO MULTIPLO

- MLAGAN definisce una propria funzione di score
- Il punteggio di un allineamento multiplo è dato da una combinazione lineare di *consensus* e *Sum of Pairs*

CLUSTALW

1/2

- CLUSTAL è un tool per l'allineamento globale di più sequenze (global multiple alignment)
- Utilizza la tecnica dell'allineamento progressivo
- CLUSTALW è spesso usato per l'allineamento di sequenze proteiche divergenti
  - W sta per Weights

CLUSTALW

2/2

- Clustal W esegue l'allineamento progressivo in tre passi principali:
  1. Costruzione di una matrice delle distanze, valutata su tutte le coppie delle sequenze
  2. Costruzione di un albero filogenetico guida delle sequenze mediante il metodo *neighbour joining*
  3. Allineamento pairwise progressivo delle sequenze rispetto alla similarità data dall'albero

Passo 1: costruzione della matrice delle distanze

1/3

- Date n sequenze da allineare, Clustal W allinea tutte le coppie di sequenze separatamente e costruisce una matrice delle distanze tra ogni coppia di sequenze
- Lo score dell'allineamento pairwise viene convertito in "distanza" ($\in [0, 1]$) tra due sequenze

	Seq. 1	**Seq. 2**	**Seq. 3**	**Seq. 4**
Seq. 1	0.00			
Seq. 2	0.11	0.00		
Seq. 3	0.32	0.43	0.00	
Seq. 4	0.17	0.18	0.57	0.00

Passo 1: costruzione della matrice delle distanze

2/3

- I programmi della serie CLUSTAL calcolano la matrice delle distanze con un metodo approssimato ma veloce
- CLUSTALW dà anche la possibilità di calcolare la matrice con un metodo più accurato, ma computazionalmente lento (decine di minuti)
  - basato su allineamenti effettuati con programmazione dinamica
  - matrici per attribuire il peso ai match/mismatch (matrici *PAM* o *BLOSUM*)

Passo 1: costruzione della matrice delle distanze

3/3

- **PAM** (**P**ercent **A**ccepted **M**utations): l'entry (i, j) contiene lo score assegnato alla coppia di aminoacidi (A_i, A_j)
- Lo score è proporzionale alla frequenza con cui ci si aspetta che A_i sostituisca A_j
- Alcune sostituzioni di aminoacidi occorrono più facilmente di altre
- Proteine omologhe non devono necessariamente avere gli stessi aminoacidi in ogni posizione

Passo 2: costruzione albero guida

1/3

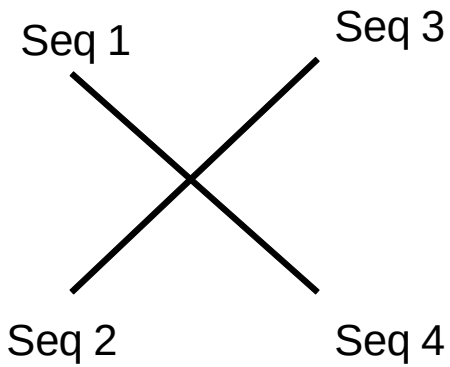
- L'euristica in CLUSTALW consiste nell'allineare per prime le sequenze più vicine
- Viene costruito un albero filogenetico (a partire dalla matrice delle distanze)
  - Metodo neighbour joining
- Importanza della "precisione" della matrice delle distanze

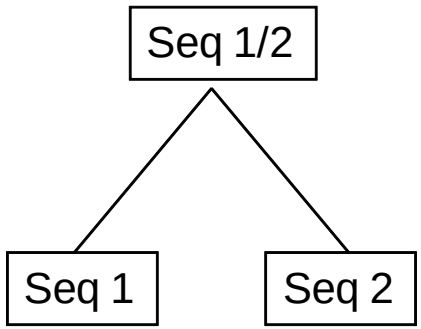
Passo 2: costruzione albero guida

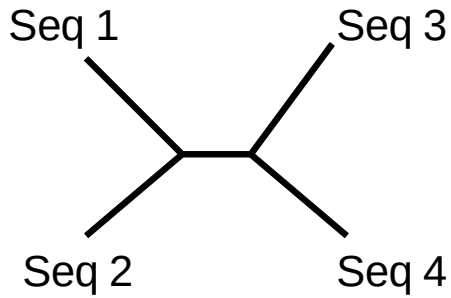
2/3

- Dalla matrice viene scelta la coppia che diverge meno: essa formerà il primo sottoalbero

	Seq. 1	Seq. 2	Seq. 3	Seq. 4
Seq. 1	0.00			
Seq. 2	0.11	0.00		
Seq. 3	0.32	0.43	0.00	
Seq. 4	0.17	0.18	0.57	0.00







Passo 2: costruzione albero guida

2/3

- Nella matrice viene sostituita la entry Seq 1/2 alle singole entry Seq 1 e Seq 2
- Vengono calcolate le distanze di Seq 1/2 dalle sequenze rimanenti (media aritmetica)

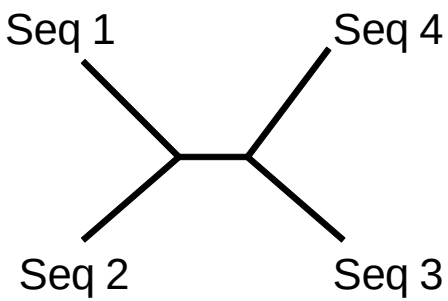
	Seq. 1/2	**Seq. 3**	**Seq. 4**
Seq. 1/2	0.00		
Seq. 3	**0.375**	0.00	
Seq. 4	**0.175**	0.57	0.00

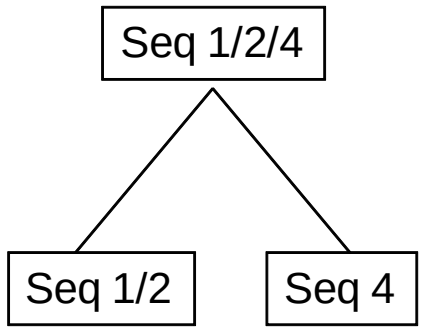
PASSO 2: COSTRUZIONE ALBERO GUIDA

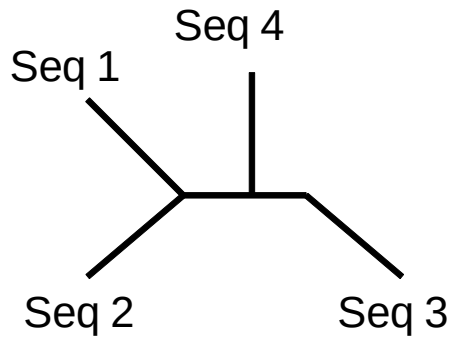
2/3

- Dalla matrice viene scelta la coppia che diverge meno: essa formerà il secondo sottoalbero

	Seq. 1/2	**Seq. 3**	**Seq. 4**
Seq. 1/2	0.00		
Seq. 3	0.375	0.00	
Seq. 4	0.175	0.57	0.00



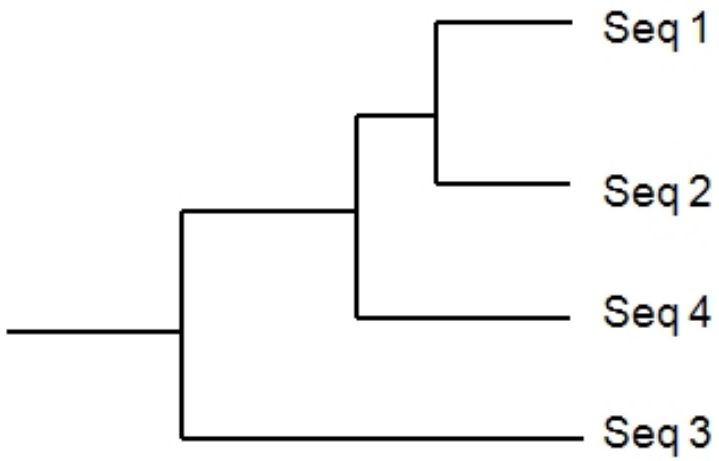




Passo 2: costruzione albero guida

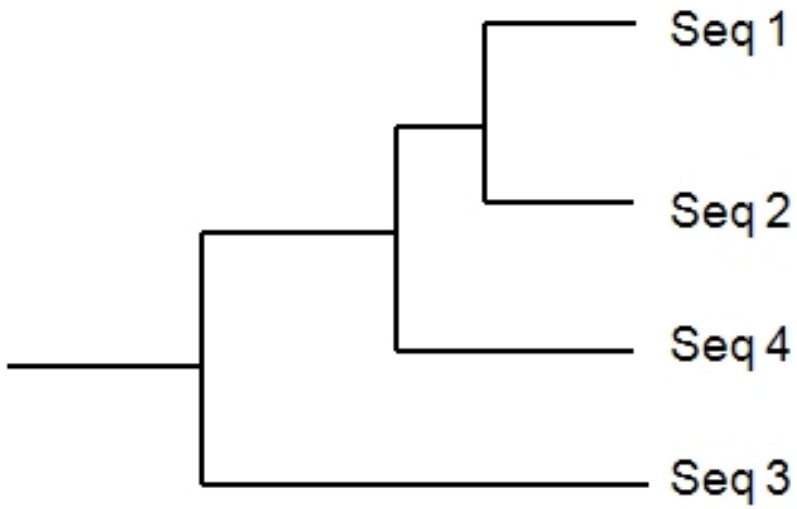
3/3

- Si ottiene un albero senza radice
- La radice viene posta tramite il metodo “mid-point”
- La lunghezza dei rami è proporzionale alla divergenza delle sequenze che essi rappresentano



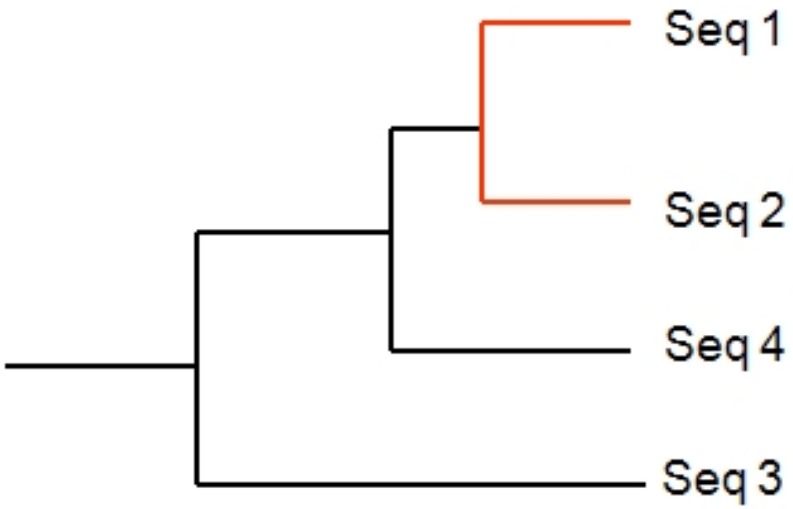
Passo 3: allineamento multiplo

- CLUSTAL W realizza l'allineamento multiplo tramite allineamenti pairwise progressivi
  - Le coppie vengono allineate seguendo l'ordine dato dall'albero a partire dalle sequenze più simili



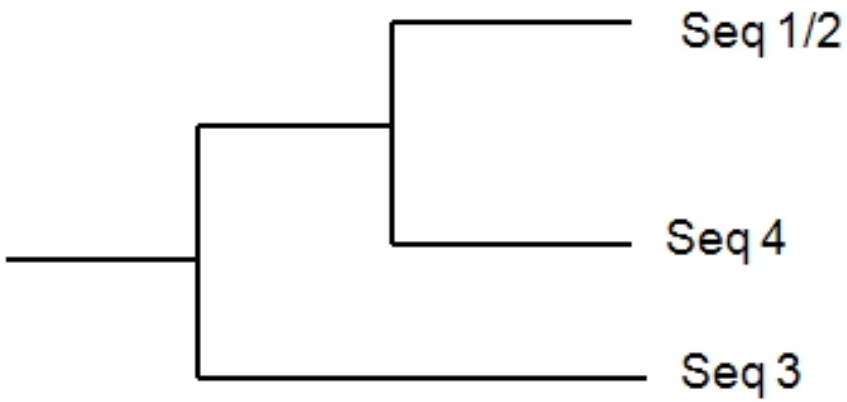
PASSO 3: ALLINEAMENTO MULTIPLO

- CLUSTAL W realizza l'allineamento multiplo tramite allineamenti pairwise progressivi
  - Le coppie vengono allineate seguendo l'ordine dato dall'albero a partire dalle sequenze più simili



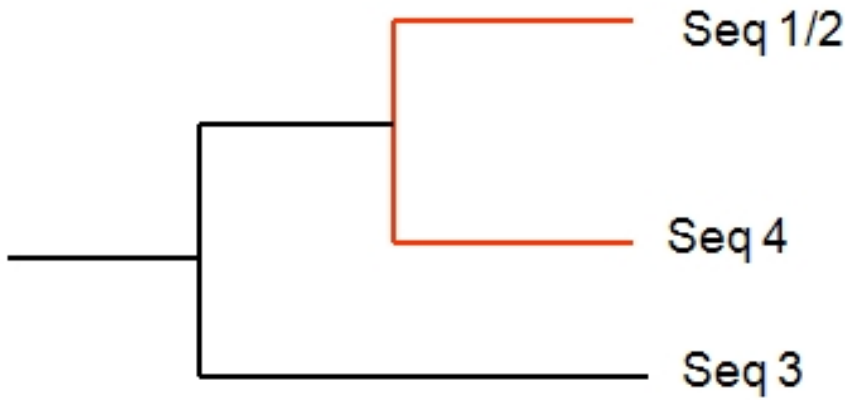
PASSO 3: ALLINEAMENTO MULTIPLO

- CLUSTAL W realizza l'allineamento multiplo tramite allineamenti pairwise progressivi
  - Le coppie vengono allineate seguendo l'ordine dato dall'albero a partire dalle sequenze più simili



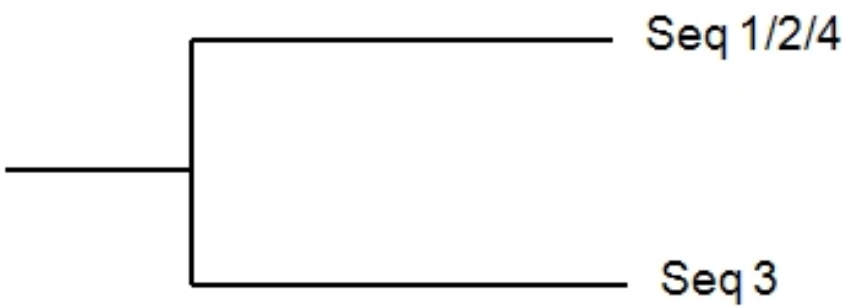
Passo 3: allineamento multiplo

- CLUSTAL W realizza l'allineamento multiplo tramite allineamenti pairwise progressivi
  - Le coppie vengono allineate seguendo l'ordine dato dall'albero a partire dalle sequenze più simili



PASSO 3: ALLINEAMENTO MULTIPLO

- CLUSTAL W realizza l'allineamento multiplo tramite allineamenti pairwise progressivi
  - Le coppie vengono allineate seguendo l'ordine dato dall'albero a partire dalle sequenze più simili



PASSO 3: ALLINEAMENTO MULTIPLO

- CLUSTAL W realizza l'allineamento multiplo tramite allineamenti pairwise progressivi
  - Le coppie vengono allineate seguendo l'ordine dato dall'albero a partire dalle sequenze più simili

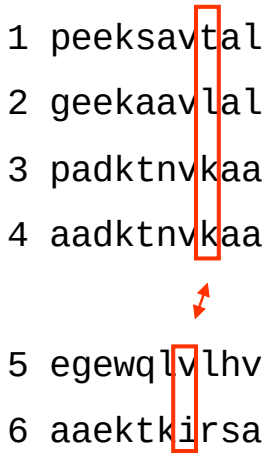
PASSO 3: ALLINEAMENTO MULTIPLO

- CLUSTAL W realizza l'allineamento multiplo tramite allineamenti pairwise progressivi
  - Le coppie vengono allineate seguendo l'ordine dato dall'albero a partire dalle sequenze più simili

FUNZIONE DI SCORE

1/3

- Alignment weighting
  - Per dare uno score agli allineamenti vengono utilizzate matrici di score e il punteggio dato alle sequenze nell'albero



Score =
M(t,v) * w1 * w5 +
M(t,i) * w1 * w6 +
M(l,v) * w2 * w5 +
M(l,i) * w2 * w6 +
M(k,v) * w3 * w5 +
M(k,i) * w3 * w6 +
M(k,v) * w4 * w5 +
M(k,i) * w4 * w6

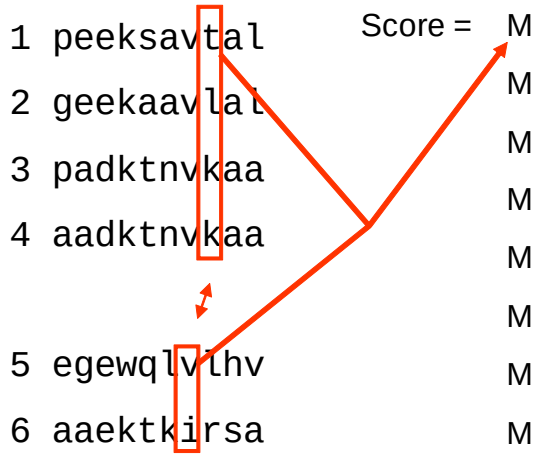
M(x,y) è la entry nella matrice di score per l'aminoacido X vs Y

w n è lo score dato alla sequenza n

FUNZIONE DI SCORE

1/3

- Alignment weighting
  - Per dare uno score agli allineamenti vengono utilizzate matrici di score e il punteggio dato alle sequenze nell'albero



Score =
M(t,v) * w1 * w5 +
M(t,i) * w1 * w6 +
M(l,v) * w2 * w5 +
M(l,i) * w2 * w6 +
M(k,v) * w3 * w5 +
M(k,i) * w3 * w6 +
M(k,v) * w4 * w5 +
M(k,i) * w4 * w6

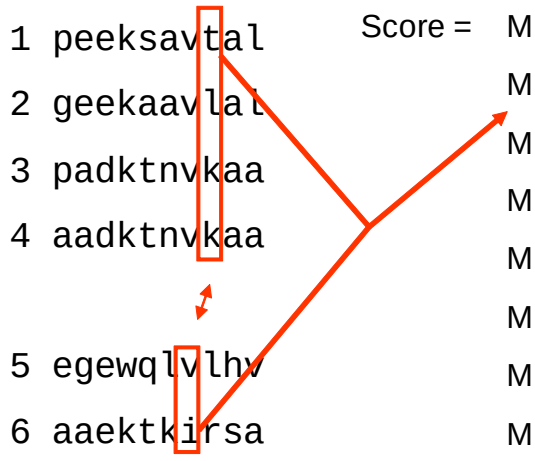
M(x,y) è la entry nella matrice di score per l'aminoacido X vs Y

w n è lo score dato alla sequenza n

FUNZIONE DI SCORE

1/3

- Alignment weighting
  - Per dare uno score agli allineamenti vengono utilizzate matrici di score e il punteggio dato alle sequenze nell'albero



Score =
M(t,v) * w1 * w5 +
M(t,i) * w1 * w6 +
M(l,v) * w2 * w5 +
M(l,i) * w2 * w6 +
M(k,v) * w3 * w5 +
M(k,i) * w3 * w6 +
M(k,v) * w4 * w5 +
M(k,i) * w4 * w6

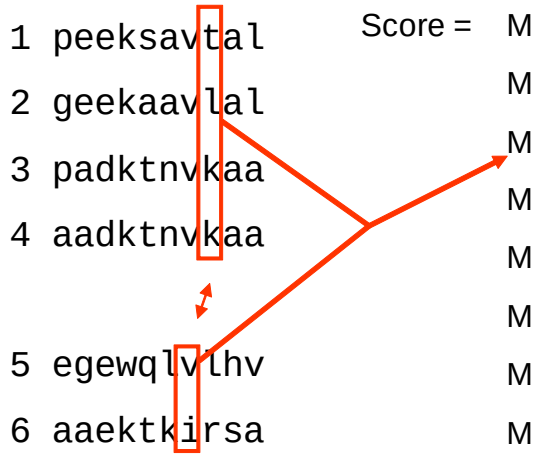
M(x,y) è la entry nella matrice di score per l'aminoacido X vs Y

w n è lo score dato alla sequenza n

FUNZIONE DI SCORE

1/3

- Alignment weighting
  - Per dare uno score agli allineamenti vengono utilizzate matrici di score e il punteggio dato alle sequenze nell'albero



Score =
M(t,v) * w1 * w5 +
M(t,i) * w1 * w6 +
M(l,v) * w2 * w5 +
M(l,i) * w2 * w6 +
M(k,v) * w3 * w5 +
M(k,i) * w3 * w6 +
M(k,v) * w4 * w5 +
M(k,i) * w4 * w6

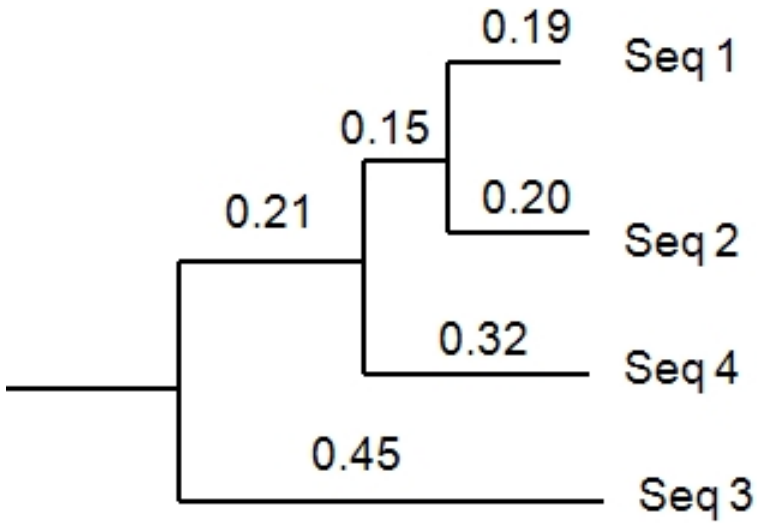
M(x,y) è la entry nella matrice di score per l'aminoacido X vs Y

w n è lo score dato alla sequenza n

FUNZIONE DI SCORE

2/3

- Ad ogni diramazione dell'albero viene dato un valore che dipende:
  - dalla distanza della diramazione dalla radice
  - dalla divergenza delle sequenze che la diramazione rappresenta

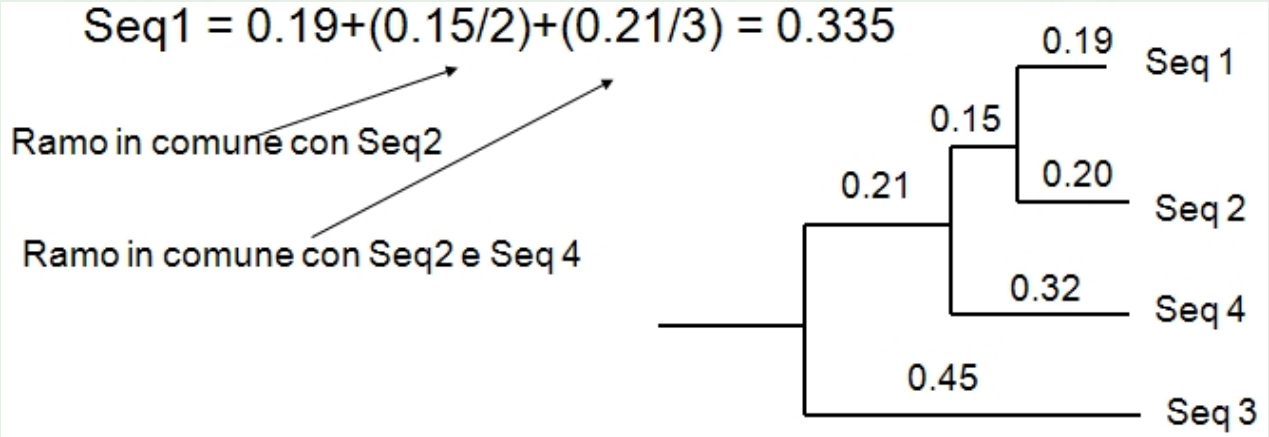


Funzione di score

2/3

- Tramite questi valori viene dato uno score ad ogni sequenza
- Lo score è calcolato a partire dalla distanza della sequenza dalla radice tenendo conto dei rami in comune con altre sequenze

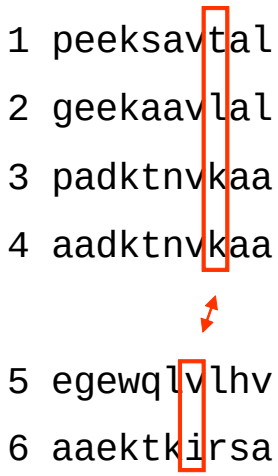
Esempio



FUNZIONE DI SCORE

3/3

- Per ogni allineamento pairwise viene usato un algoritmo di programmazione dinamica. Vengono usate:
  - Matrici di score PAM o BLOSUM
  - Penalità per gap
- Alignment weighting
  - Per dare uno score agli allineamenti vengono utilizzate matrici di score e il punteggio dato alle sequenze nell'albero



Score =
M(t,v) * **w1** * **w5** +
M(t,i) * w1 * w6 +
M(l,v) * w2 * w5 +
M(l,i) * w2 * w6 +
M(k,v) * w3 * w5 +
M(k,i) * w3 * w6 +
M(k,v) * w4 * w5 +
M(k,i) * w4 * w6

M(x,y) è la entry nella matrice di score per l'aminoacido X vs Y

w n è lo score dato alla sequenza n

Riepilogo

Osservazioni

- L'allineamento progressivo è un approccio euristico, quindi non è detto che il metodo restituisca l'allineamento migliore
- Se le sequenze sono molto simili, allora l'allineamento progressivo è più che ragionevole
- Se le sequenze sono molto diverse l'allineamento progressivo diviene meno attendibile
- Problema del ***local minimum***: un errore introdotto nei primi allineamenti pairwise non può essere corretto e può corrompere l'allineamento multiplo

RIEPILOGO

MLAGAN E CLUSTALW

- MLAGAN
  - Usa algoritmi di programmazione dinamica a partire dai seed
  - Necessita di un albero filogenetico passato come parametro
  - Buona funzione di score
  - Cerca di ovviare al *local minimum*
- CLUSTAL W
  - Usa algoritmi di programmazione dinamica
  - Per allineamento di sequenze aminoacidiche utilizza matrici di score per aminoacidi
  - Si basa su un albero filogenetico creato autonomamente
  - Buona funzione di score
  - Non risolve il problema del *local minimum*

AMAP

Protein multiple alignment by sequence annealing

- La più comune metrica prestazionale per tool di allineamento multiplo è data dalla **sensibilità** (*recall*)
- **Sensibilità**: dato un insieme di sequenze di benchmark (di cui si conoscono gli allineamenti ottimi), la sensibilità è la percentuale di posizioni omologhe correttamente individuate dal tool
- La **specificità** (capacità di evitare falsi positivi) dello strumento è però (quasi) altrettanto importante

AMAP

Caratteristiche generali

- **AMAP** adotta un approccio algoritmico del tipo *tempra simulata*, ottenendo buoni risultati tanto per sensibilità, quanto per specificità
- L'allineamento non è progressivo : viene costruito un match alla volta!
- Il comportamento dell'algoritmo è probabilistico e modellato da **PHMM** (**P**air **H**idden **M**arkov **M**odel)

ALLINEAMENTO MULTIPLO GLOBALE PARZIALE

- Un **allineamento multiplo globale parziale** (PGMA) di sequenze $\sigma^1, \sigma^2, \ldots, \sigma^k$ è:
  - un insieme parzialmente ordinato (*poset*) $P = \{c_1, c_2, \ldots, c_m\}$ e
  - una funzione surgettiva $\varphi : S_{\sigma^1,\sigma^2,\ldots,\sigma^k} \rightarrow P$ tale che: $i \leq j \Rightarrow \varphi(\sigma_i^a) \leq \varphi(\sigma_j^a)$
- $\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_m}$: colonne dell'allineamento multiplo
- $\sigma_{\mathbf{i}}^{\mathbf{n}}$: *i-esima* posizione della *n-esima* sequenza
- Nota: un insieme parzialmente ordinato può essere rappresentato in maniera equivalente con un grafo diretto aciclico (DAG)

ALLINEAMENTO MULTIPLO GLOBALE

- Un **allineamento multiplo globale** può essere espresso in termini di allineamenti multipli parziali
- Un'**estensione lineare** di un insieme parzialmente ordinato $P = \{c_1, c_2, \ldots, c_m\}$ è una permutazione degli elementi tale che $c_i < c_j \Rightarrow i < j$
- Un allineamento globale è un PGMA più un'estensione lineare dell'insieme P ad esso associato

ALLINEAMENTO MULTIPLO GLOBALE

ESEMPIO

N G Y E

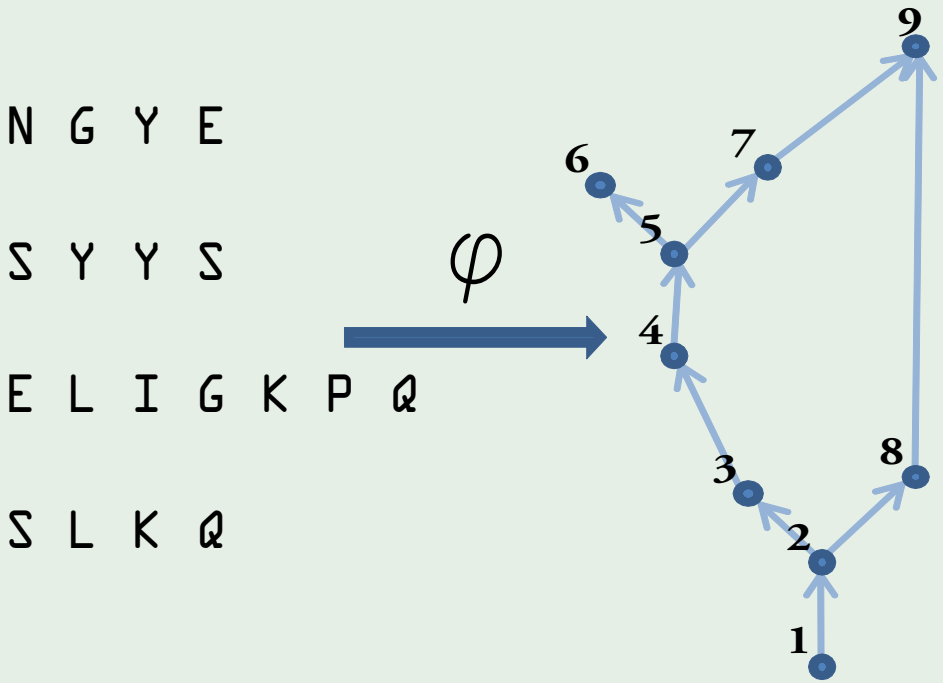
S Y Y S

E L I G K P Q

S L K Q

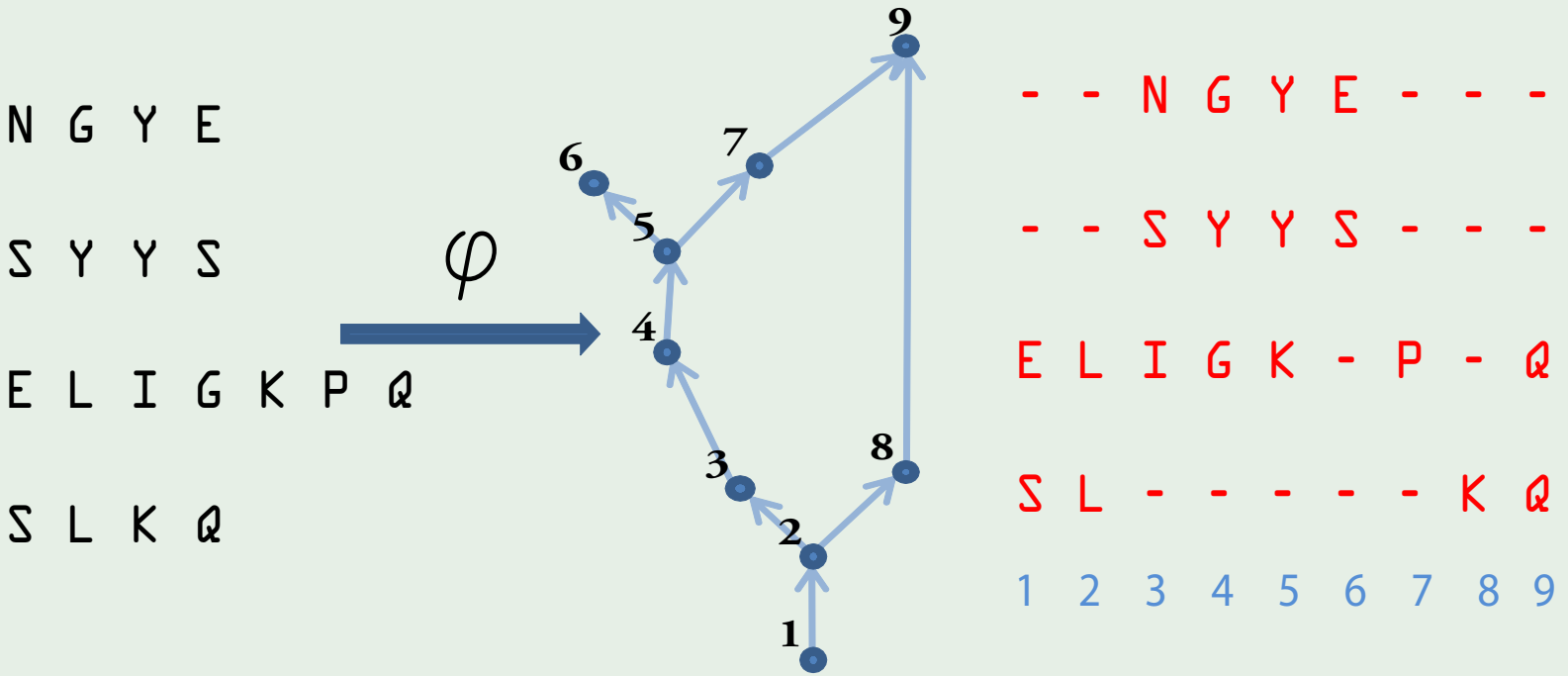
ALLINEAMENTO MULTIPLO GLOBALE

ESEMPIO



ALLINEAMENTO MULTIPLO GLOBALE

ESEMPIO



Algoritmo

Intuizione

- Ad ogni passo il numero di colonne dell'allineamento parziale attuale viene ridotto di 1, operando sulle posizioni che hanno maggiore probabilità di essere omologhe
- Si utilizzano probabilità a posteriori, che esprimono la probabilità di un'ipotesi **in seguito** all'osservazione di un certo evento
- La funzione di score è una combinazione della *Development score* f_D (equivalente alla SP) con una metrica (AMA) che preserva maggiormente la specificità
- Due colonne sono fuse solo se il PGMA generato ha una valutazione non peggiore del precedente

Algoritmo

Pseudocodice

$M_L \leftarrow M_{Null}$
$i \leftarrow L$
while $\exists\, c_k^{M_i}, c_l^{M_i} : \; merge(c_k^{M_i}, c_l^{M_i}) = M'$ **and** $f(M') \geq f(M_i)$ **do**
$M_{i-1} \leftarrow M'$
$i \leftarrow i - 1$
end while

Algoritmo

Pseudocodice

$M_L \leftarrow M_{Null}$
$i \leftarrow L$
while $\exists\, c_k^{M_i}, c_l^{M_i} : \; merge(c_k^{M_i}, c_l^{M_i}) = M'$ **and** $f(M') \geq f(M_i)$ **do**
$M_{i-1} \leftarrow M'$
$i \leftarrow i - 1$
end while

Algoritmo

Pseudocodice

$M_L \leftarrow M_{Null}$
$i \leftarrow L$
while $\exists\, c_k^{M_i}, c_l^{M_i} : \; merge(c_k^{M_i}, c_l^{M_i}) = M'$ **and** $f(M') \geq f(M_i)$ **do**
$M_{i-1} \leftarrow M'$
$i \leftarrow i - 1$
end while

Algoritmo

Pseudocodice

$M_L \leftarrow M_{Null}$
$i \leftarrow L$
while $\exists\, c_k^{M_i}, c_l^{M_i} : \; merge(c_k^{M_i}, c_l^{M_i}) = M'$ **and** $f(M') \geq f(M_i)$ **do**
$M_{i-1} \leftarrow M'$
$i \leftarrow i - 1$
end while

SEQUENCE ANNEALING

- **Sequence annealing**: dato un insieme di sequenze S ed una funzione di score f, un SA è una catena di PGMA

$$\mathbf{M_L} \supset \mathbf{M_{L-1}} \supset \mathbf{M_{L-2}} \supset \cdots \supset \mathbf{M_r} \text{ tale che}$$

- M_i è associato all'insieme P_i e $|P_i| = i$ (il numero di colonne del PGMA di indice i è pari ad i)
- $f(M_{i+1}) \leq f(M_i)$
- M_i è ottenuto da M_{i+1} tramite fusione di due colonne c_j^{i+1} e c_k^{i+1} in una c_h^i

MERGE

APPROCCIO TEORICO

1. Controlla se due colonne possono essere fuse
2. In caso affermativo, dopo aver effettuato la fusione, aggiorna l'insieme parzialmente ordinato

- Il problema di trovare un'estensione lineare può essere risolto in maniera efficiente risolvendo l'*online topological ordering problem* sul grafo equivalente
- **Online topological ordering problem**: problema dell'ordinamento topologico in cui i vertici del grafo compaiono uno per volta

Merge

Implementazione pratica

- Ad ogni coppia di colonne è assegnato un peso
- Le coppie sono poste in uno heap
- Ad ogni iterazione la coppia di peso più elevato viene estratta dallo heap (in tempo costante)
- I pesi variano dinamicamente: logicamente decrescono ad ogni fusione, ma per ragioni di efficienza sono ricalcolati solo al momento dell'estrazione

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTWKPELPKCVR----------------------------------------
SEQ2 ---------------CQANNMWGPTRLPTCVS-----------------------------
SEQ3 ---------------------------------IWSGKPPICEKV-------------------
SEQ4 ----------------------------------------CLISGSSVQWSDPLPECREH
```

Sequence annealing

Esempio

```
SEQ1 SPGNTWKPELPKCVR----------------------------------------
SEQ2 ---------------CQANNMWGPTRLPTCVS-----------------------------
SEQ3 ---------------------------------IWSGKPPICEKV-------------------
SEQ4 ----------------------------------------CLISGSSVQWSDPLPECREH
```

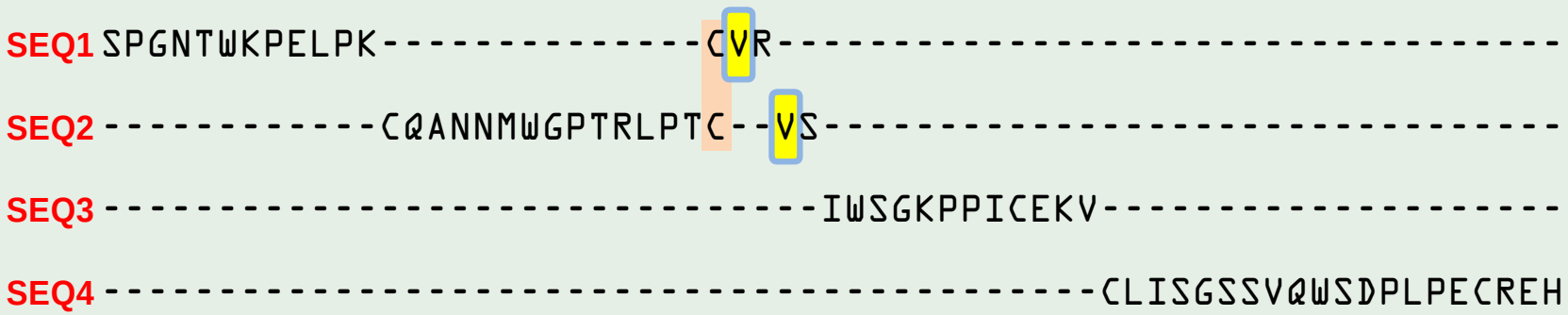
Sequence annealing

Esempio

```
SEQ1 SPGNTWKPELPK-------------CVR--------------------------------
SEQ2 ------------CQANNMWGPTRLPTC--VS--------------------------------
SEQ3 ------------------------------IWSGKPPICEKV--------------------
SEQ4 ----------------------------------------CLISGSSVQWSDPLPECREH
```

Sequence annealing

Esempio



SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTWKPELPK-------------CVR--------------------------------
SEQ2 ------------CQANNMWGPTRLPTCV-S--------------------------------
SEQ3 -------------------------------IWSGKPPICEKV--------------------
SEQ4 ----------------------------------------CLISGSSVQWSDPLPECREH
```

Sequence annealing

Esempio

```
SEQ1 SPGNTWKPELPK-------------CVR-------------------------------
SEQ2 ------------CQANNMWGPTRLPTCVS-------------------------------
SEQ3 -----------------------------IWSGKPPICEKV--------------------
SEQ4 ------------------------------------------CLISGSSVQWSDPLPECREH
```

Sequence annealing

Esempio

```
SEQ1 SPGNTWKPELPK--------------CVR-----------------------------
SEQ2 ------------CQANNMWGPTRLPTCVS-----------------------------
SEQ3 -------------------------------IWSGKPPI----------------CEKV---
SEQ4 ------------------------------------CLISGSSVQWSDPLPEC---REH
```

Sequence annealing

Esempio

```
SEQ1 SPGNTWKPELPK-------------CVR----------------------------
SEQ2 ------------CQANNMWGPTRLPTCVS----------------------------
SEQ3 -----------------------------IWSGKPPI----------------CEK--V
SEQ4 ------------------------------------CLISGSSVQWSDPLPEC--REH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTWKPELPK-------------CVR----------------------------
SEQ2 ------------CQANNMWGPTRLPTCVS----------------------------
SEQ3 -------------------------------IWSGKPPI----------------CE-KV
SEQ4 -----------------------------------CLISGSSVQWSDPLPEC-REH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTWKPELPK-------------CVR--------------------------
SEQ2 ------------CQANNMWGPTRLPTCVS--------------------------
SEQ3 -----------------------------IWSGKPPI----------------CEKV
SEQ4 ------------------------------------CLISGSSVQWSDPLPECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTWKPE----------------------------------LPKCVR------
SEQ2 -----------CQANNMWGPTRLP---------------------TCVS------
SEQ3 ----------------------------IWSGKPPI---------------CEKV
SEQ4 ---------CL-------------ISGS--------SVQWSDPL-----PECREH
```

Sequence annealing

Esempio

```
SEQ1 SPGNTWKPE----------------------------------LPKCVR-----
SEQ2 -----------CQANNMWGPTRLP---------------------TCVS-----
SEQ3 ----------------------------IWSGKPPI--------------CEKV
SEQ4 ---------CL-------------ISGS--------SVQWSDPLP----ECREH
```

Sequence annealing

Esempio

```
SEQ1 SPGNTWKPE-----------------------------------LPKCVR-----
SEQ2 -----------CQANNMWGPTRLP--------------------TCVS-----
SEQ3 ----------------------------IWSGKPP-------------ICEKV
SEQ4 ---------CL-------------ISGS-------SVQWSDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTWKPE----------------------------------LPKCVR-----
SEQ2 -----------CQANNMWGPTRLP--------------------TCVS-----
SEQ3 ----------------------------IWSGKP--------P----ICEKV
SEQ4 ---------CL-------------ISGS------SVQWSDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTWKP----------------------------ELPKCVR-----
SEQ2 ----------CQANNMWGPTRLP-------------------TCVS-----
SEQ3 --------------------------IWSGKP--------P----ICEKV
SEQ4 --------CL-------------ISGS------SVQWSDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTWK----------------------------PELPKCVR-----
SEQ2 ---------CQANNMWGPTRLP-------------------TCVS-----
SEQ3 -------------------------IWSGKP--------P----ICEKV
SEQ4 -------CL-------------ISGS------SVQWSDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNTW----------------------------KPELPKCVR-----
SEQ2 --------CQANNMWGPTRLP-------------------TCVS-----
SEQ3 ------------------------IWSGKP--------P----ICEKV
SEQ4 ------CL-------------ISGS------SVQWSDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNT--------------------------WKPELPKCVR-----
SEQ2 -------CQANNMWGPTRLP-------------------TCVS-----
SEQ3 ------------------------IWSGKP--------P----ICEKV
SEQ4 -----CL-------------ISGS------SVQWSDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNT----------------------W----KPELPKCVR-----
SEQ2 -------CQANNMWGPTRLP------------------TCVS-----
SEQ3 ----------------------I---WSGKP----P----ICEKV
SEQ4 -----CL-------------ISGS-SVQW----SDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNT----------------------W--KPE-LPKCVR-----
SEQ2 -------CQANNMWGPTRLP-----------------TCVS-----
SEQ3 ----------------------I---WSG--KP-P----ICEKV
SEQ4 -----CL-------------ISGS-SVQW--SDP-LP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNT----------------------W--KPELPKCVR-----
SEQ2 -------CQANNMWGPTRLP----------------TCVS-----
SEQ3 ----------------------I---WSG--KPP----ICEKV
SEQ4 -----CL-------------ISGS-SVQW--SDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNT----------------------WK-PELPKCVR-----
SEQ2 -------CQANNMWGPTRLP---------------TCVS-----
SEQ3 ------------------------I---WSG-KPP----ICEKV
SEQ4 -----CL-------------ISGS-SVQWS-DPLP----ECREH
```

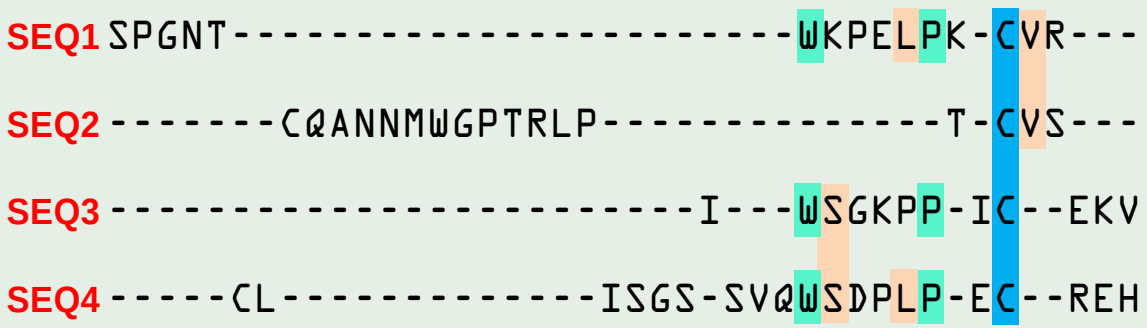
SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPGNT----------------------WKPELPKCVR-----
SEQ2 -------CQANNMWGPTRLP--------------TCVS-----
SEQ3 ------------------------I---WSGKPP----ICEKV
SEQ4 -----CL-------------ISGS-SVQWSDPLP----ECREH
```

SEQUENCE ANNEALING

ESEMPIO



SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPG------NT----------------WKPELPK-CVR---
SEQ2 ---CQ--ANN-MWGPTRLP-------------T-CVS---
SEQ3 ----------------------I---WSGKPP-IC--EKV
SEQ4 -----CL------------ISGS-SVQWSDPLP-EC--REH
```

SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPG------NT----------------WKPELPKCVR---
SEQ2 ---CQ--ANN-MWGPTRLP-------------TCVS---
SEQ3 ----------------------I---WSGKPPIC--EKV
SEQ4 -----CL------------ISGS-SVQWSDPLPEC--REH
```

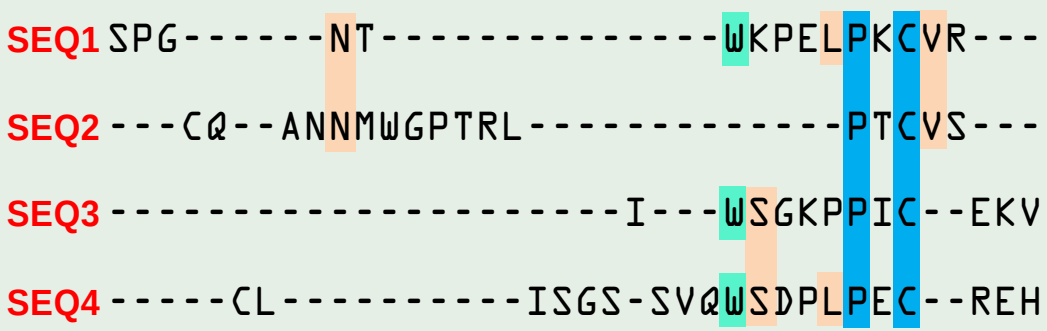
SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 SPG------NT---------------WKPELPKCVR---
SEQ2 ---CQ--ANN-MWGPTRL-------------PTCVS---
SEQ3 ----------------------I---WSGKPPIC--EKV
SEQ4 -----CL-----------ISGS-SVQWSDPLPEC--REH
```

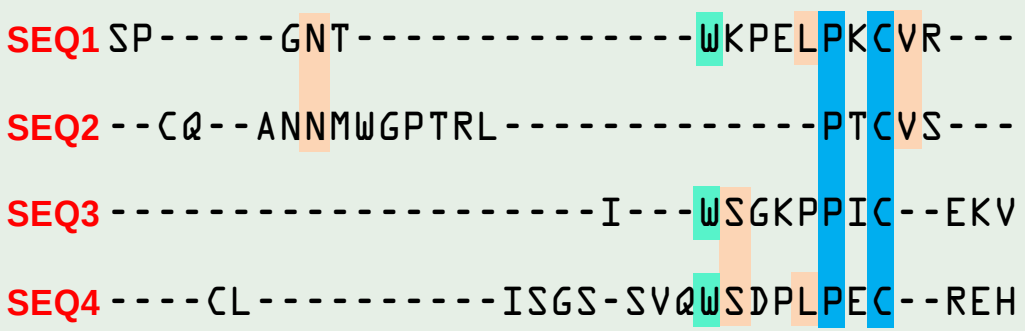
SEQUENCE ANNEALING

ESEMPIO



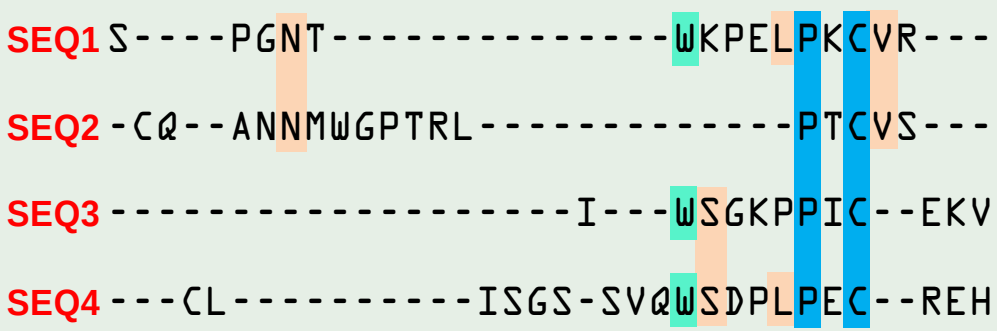
SEQUENCE ANNEALING

ESEMPIO



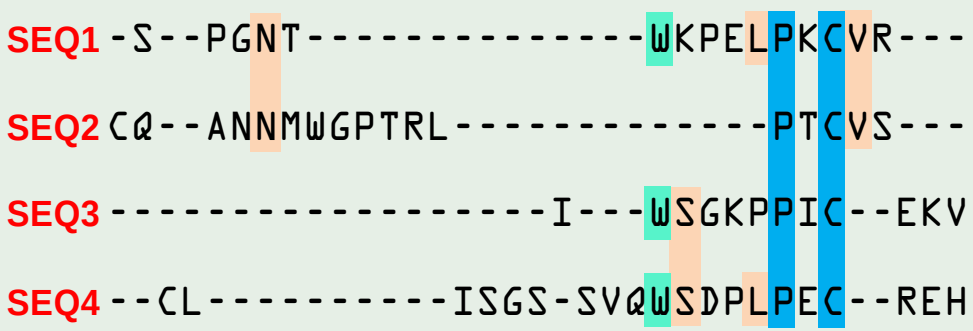
SEQUENCE ANNEALING

ESEMPIO



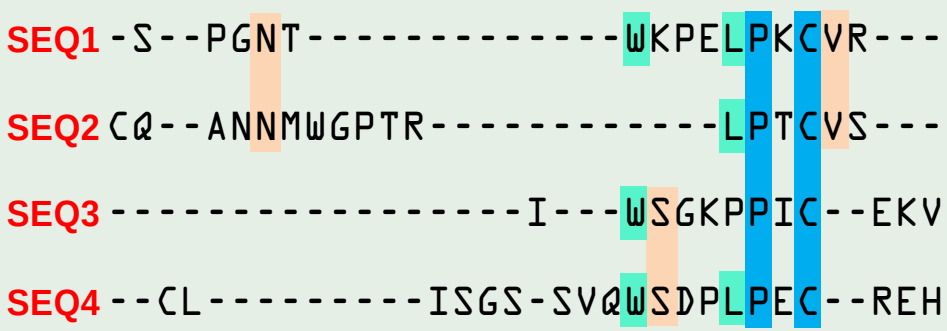
SEQUENCE ANNEALING

ESEMPIO



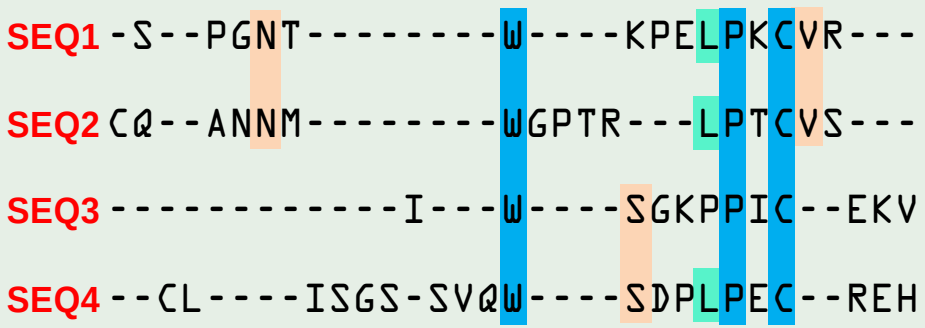
SEQUENCE ANNEALING

ESEMPIO



SEQUENCE ANNEALING

ESEMPIO



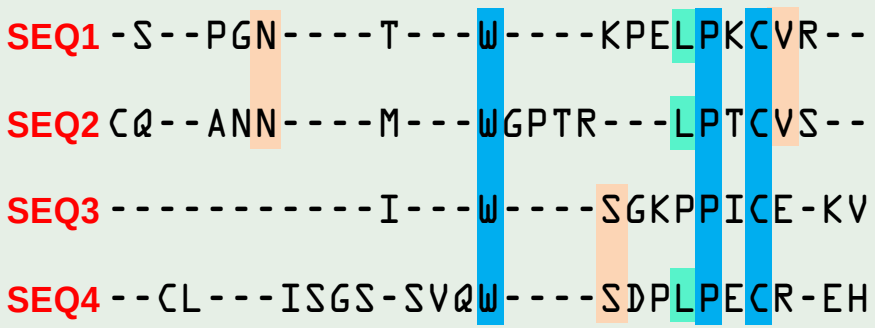
SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 -S--PGN----T---W----KPELPKCVR---
SEQ2 CQ--ANN----M---WGPTR---LPTCVS---
SEQ3 ----------I---W----SGKPPIC--EKV
SEQ4 --CL---ISGS-SVQW----SDPLPEC--REH
```

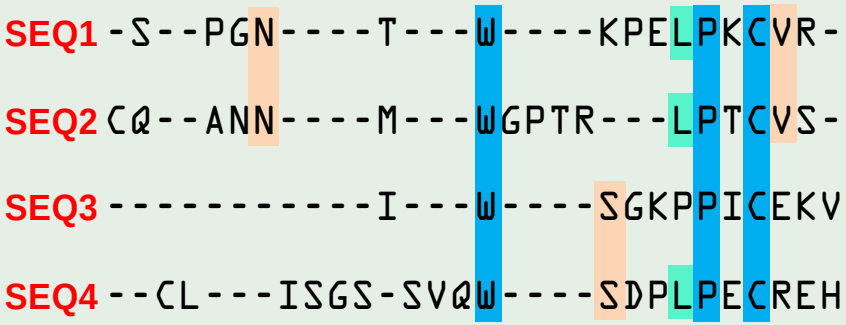
SEQUENCE ANNEALING

ESEMPIO



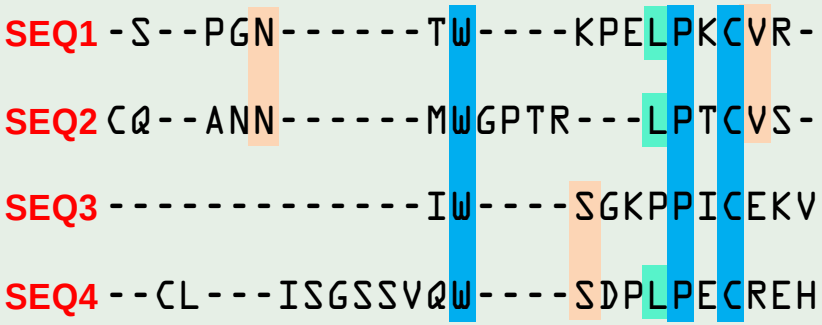
SEQUENCE ANNEALING

ESEMPIO



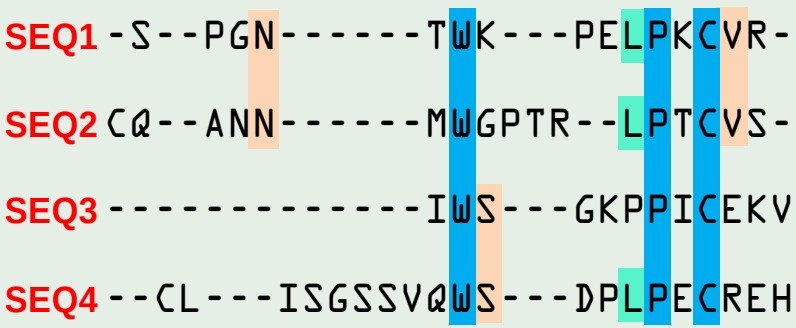
SEQUENCE ANNEALING

ESEMPIO



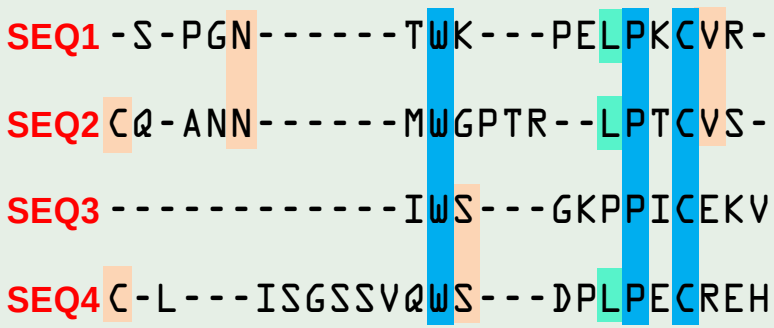
SEQUENCE ANNEALING

ESEMPIO



SEQUENCE ANNEALING

ESEMPIO



SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 -S-PGN------TWKP--ELPKCVR-
SEQ2 CQ-ANN------MWGPTR-LPTCVS-
SEQ3 ------------IWSG--KPPICEKV
SEQ4 C-L---ISGSSVQWSD--PLPECREH
```

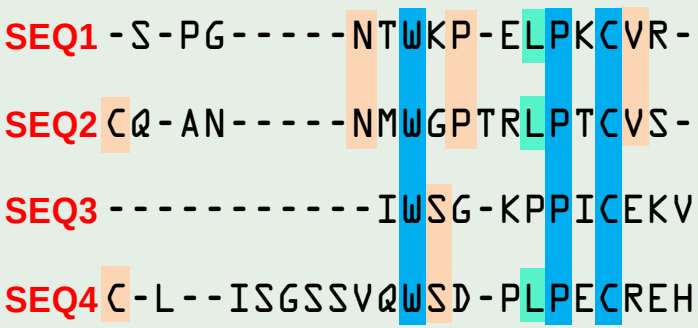
SEQUENCE ANNEALING

ESEMPIO

```
SEQ1 -S-PGN------TWKP-ELPKCVR-
SEQ2 CQ-ANN------MWGPTRLPTCVS-
SEQ3 ------------IWSG-KPPICEKV
SEQ4 C-L---ISGSSVQWSD-PLPECREH
```

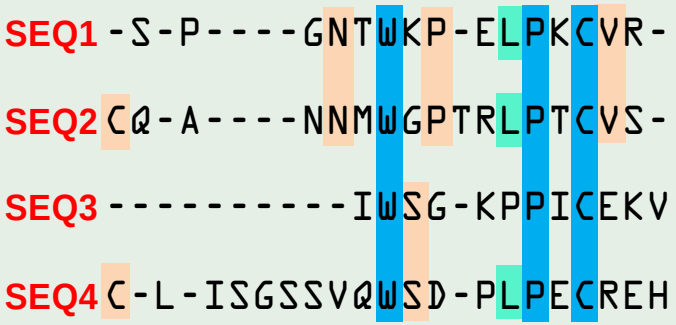
SEQUENCE ANNEALING

ESEMPIO



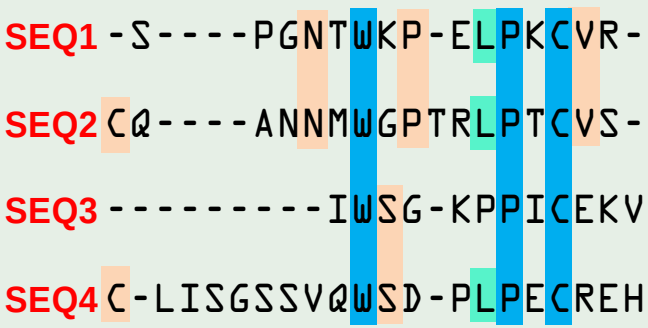
SEQUENCE ANNEALING

ESEMPIO



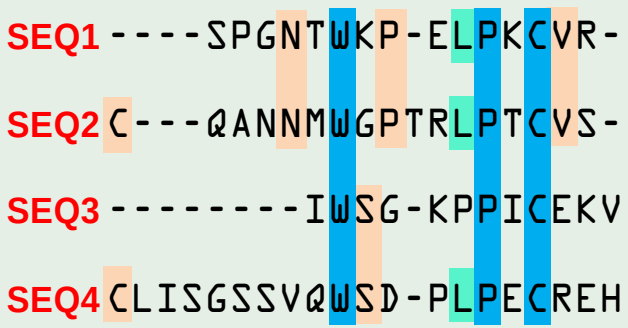
SEQUENCE ANNEALING

ESEMPIO



SEQUENCE ANNEALING

ESEMPIO



FUNZIONI DI SCORE

- **AMA** (**A**lignment **M**etric **A**ccuracy): somma, su tutte le coppie di sequenze, delle frazioni di residui correttamente allineati
- Un valore elevato dell'AMA indica un'alta specificità dell'allineamento prodotto
- **Development score f_D**: somma dei punteggi di allineamento coppia a coppia sui simboli di una data colonna
- Un valore elevato della funzione f_D indica un'alta sensibilità del risultato

Funzioni di score

- **AMA** (**A**lignment **M**etric **A**ccuracy): somma, su tutte le coppie di sequenze, delle frazioni di residui correttamente allineati
- Un valore elevato dell'AMA indica un'alta specificità dell'allineamento prodotto
- **Development score f_D**: somma dei punteggi di allineamento coppia a coppia sui simboli di una data colonna
- Un valore elevato della funzione f_D indica un'alta sensibilità del risultato

FUNZIONI DI SCORE

- **AMA** (**A**lignment **M**etric **A**ccuracy): somma, su tutte le coppie di sequenze, delle frazioni di residui correttamente allineati
- Un valore elevato dell'AMA indica un'alta specificità dell'allineamento prodotto
- **Development score f_D**: somma dei punteggi di allineamento coppia a coppia sui simboli di una data colonna
- Un valore elevato della funzione f_D indica un'alta sensibilità del risultato

Funzioni di score

Valore atteso

- Sarebbe auspicabile massimizzare entrambe le funzioni di score ad ogni passo
- L'obiettivo è difficilmente realizzabile
- Ci si accontenta di massimizzare le medie delle due funzioni
- A seconda della fase dell'algoritmo, sarà privilegiata l'una o l'altra
- Tipicamente gli allineamenti migliori sono ottenuti partendo con valori di sensibilità ridotti (a favore di un'elevata specificità), per poi massimizzarla nelle fasi finali

FUNZIONE OBIETTIVO

- Famiglia di funzioni, parametriche rispetto al fattore di gap $\mathbf{G_f}$, calcolate mediante probabilità a posteriori
- Il valore di G_f influenza la qualità dell'allineamento prodotto:
  - $G_f = 0 \Rightarrow$ sensibilità massima (massimizza lo score f_D)
  - $G_f = 0.5 \Rightarrow$ massimizza la media AMA
  - $G_f > 0.5 \Rightarrow$ sensibilità fortemente ridotta a favore di una elevata specificità
- Diverse istanze della famiglia sono usate nelle varie fasi dell'algoritmo

PROBABILITÀ A POSTERIORI

- Ricavate dal modello *Pair Hidden Markov Model*
- $\mathbf{P_{Match}}(\sigma_{\mathbf{i}}^{\mathbf{q}}, \sigma_{\mathbf{j}}^{\mathbf{t}})$: probabilità che l'i-esimo carattere della q-esima sequenza sia allineato col j-esimo carattere della t-esima sequenza
- $\mathbf{P_{Match}}(\sigma_{\mathbf{i}}^{\mathbf{q}}, -^{\mathbf{t}})$: probabilità che l'i-esimo carattere della q-esima sequenza sia allineato con uno spazio della t-esima sequenza
- I pesi associati alle coppie di colonne sono calcolati dinamicamente in base a:
  - probabilità a posteriori
  - fattore di gap (ridotto progressivamente durante l'esecuzione)

SAGA

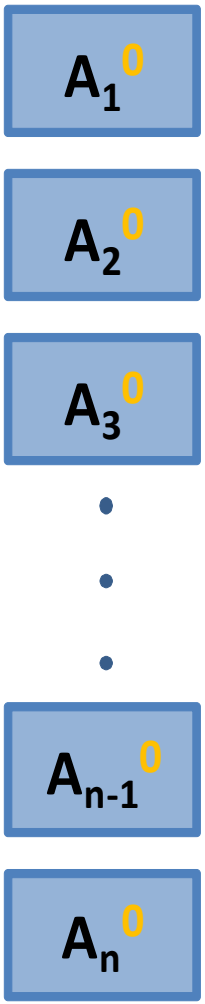
SAGA (Sequence Alignment by Genetic Algorithm)

- Inizialmente si crea una popolazione di allineamenti ($\mathbf{G_0}$) in maniera casuale
- La dimensione della popolazione sarà mantenuta costante per l'intera evoluzione
- Ad ogni iterazione la generazione attuale si evolve nella successiva: i nuovi individui derivano da un singolo "genitore" (mutazione) o da una coppia (ricombinazione)
- Un individuo ha un numero di figli proporzionale alla bontà della sua funzione obiettivo
- Se dopo un certo numero di iterazioni la funzione obiettivo non ha subito variazioni significative (popolazione stabile) l'algoritmo termina

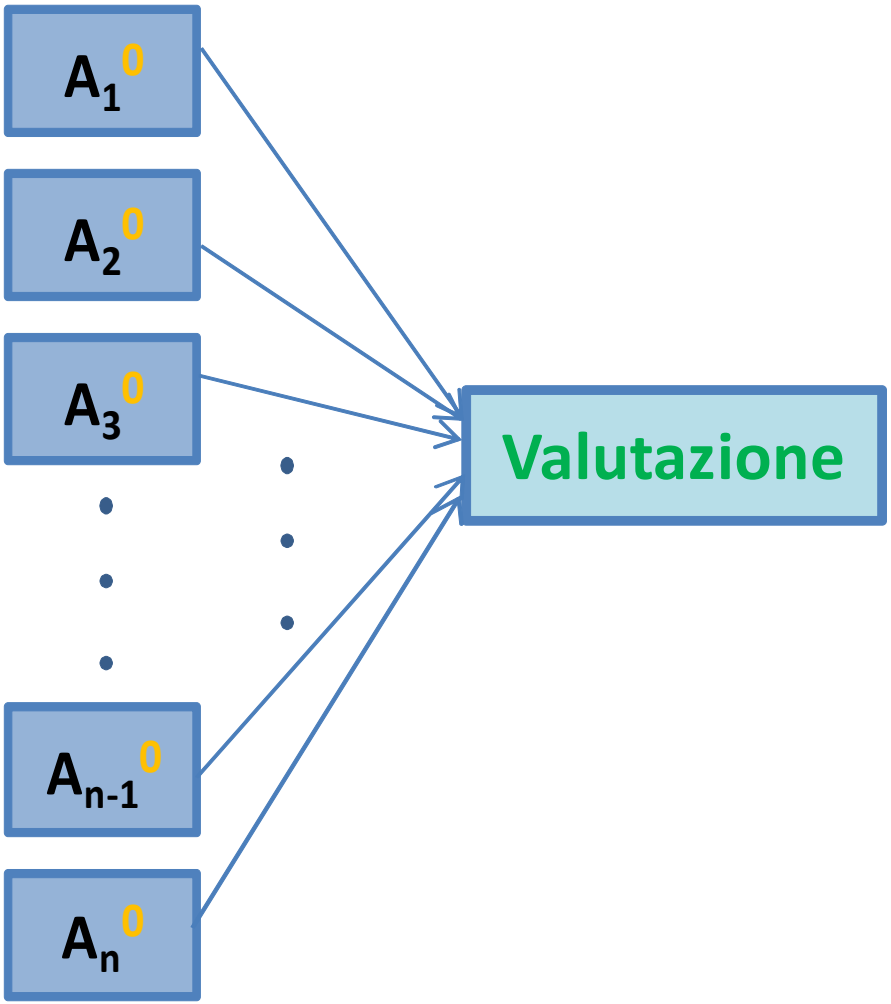
Funzione obiettivo

- La scelta della funzione obiettivo è cruciale: l'algoritmo procede fino a raggiungere una popolazione che contenga allineamenti al di sotto di un certo costo
- Gli autori propongono l'utilizzo di due funzioni:
  1. **WSP** (**W**eighted **S**um of **P**airs): funzione SP "pesata" rispetto alla similarità fra coppie di sequenze
  2. **Affine gap penalties**: determina il costo dei gap
- Il costo dell'allineamento è dato da una combinazione lineare delle due

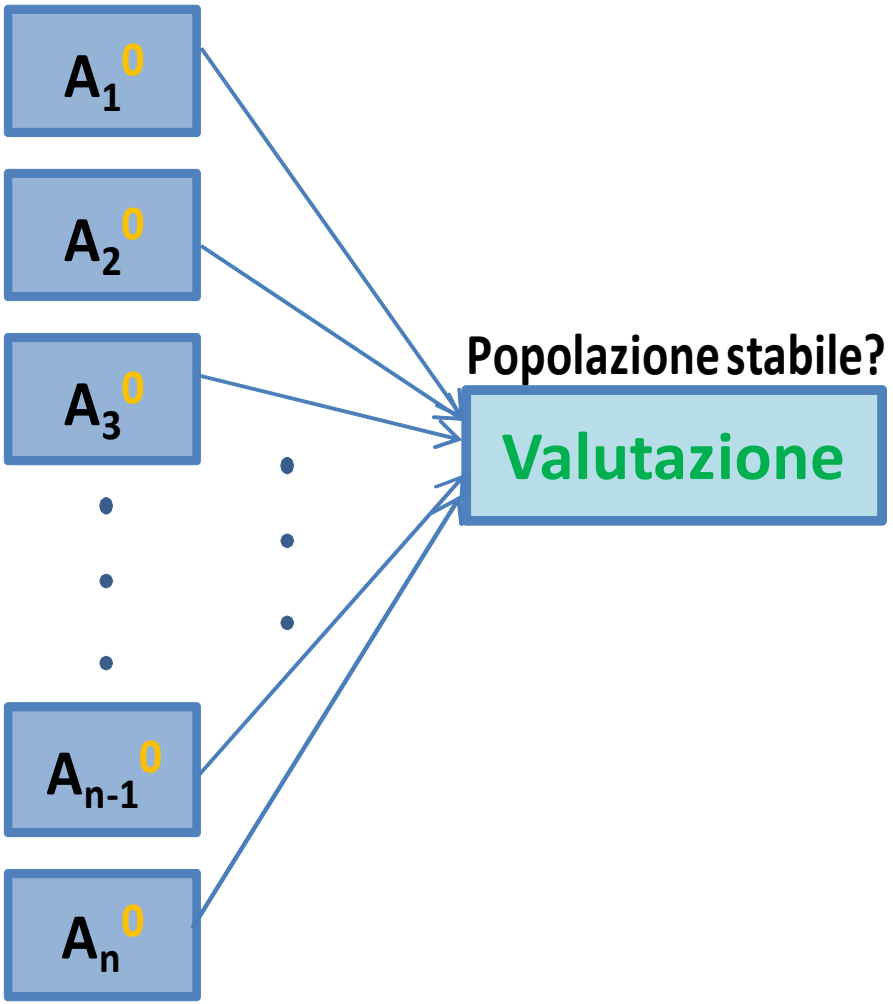
SCHEMA GENERALE



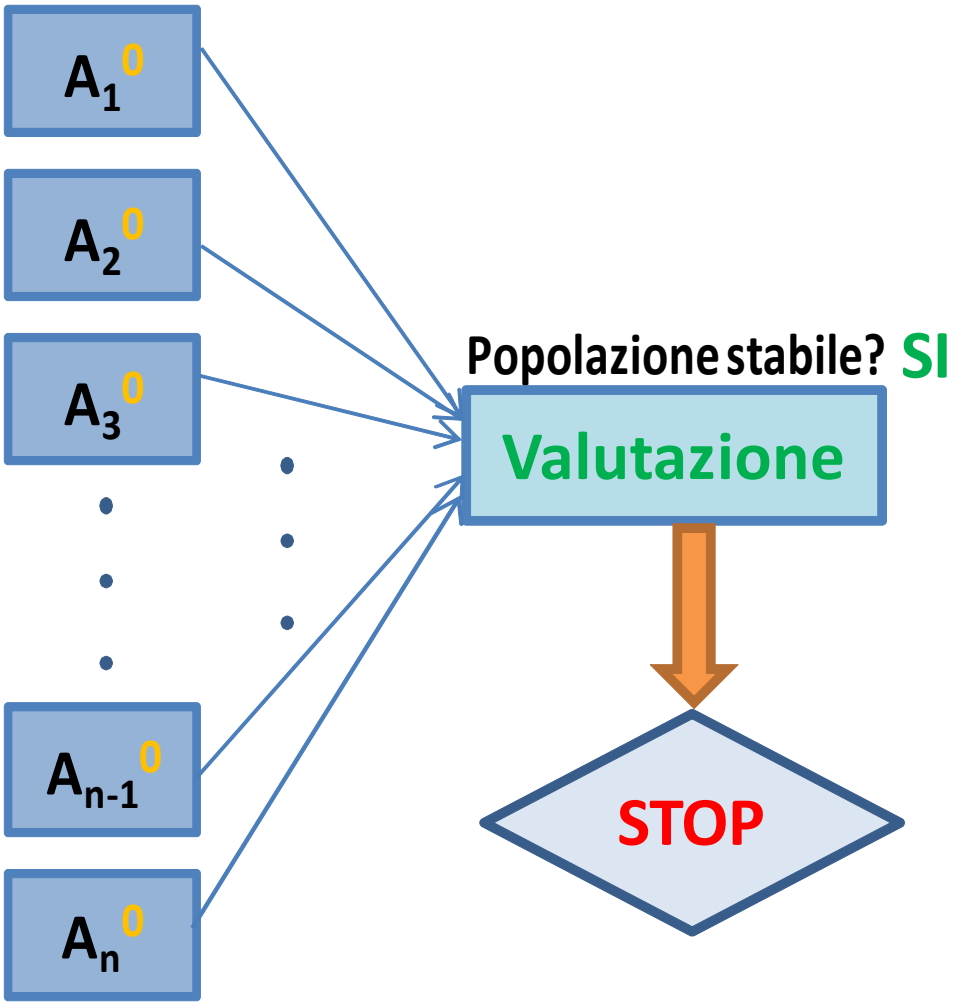
SCHEMA GENERALE



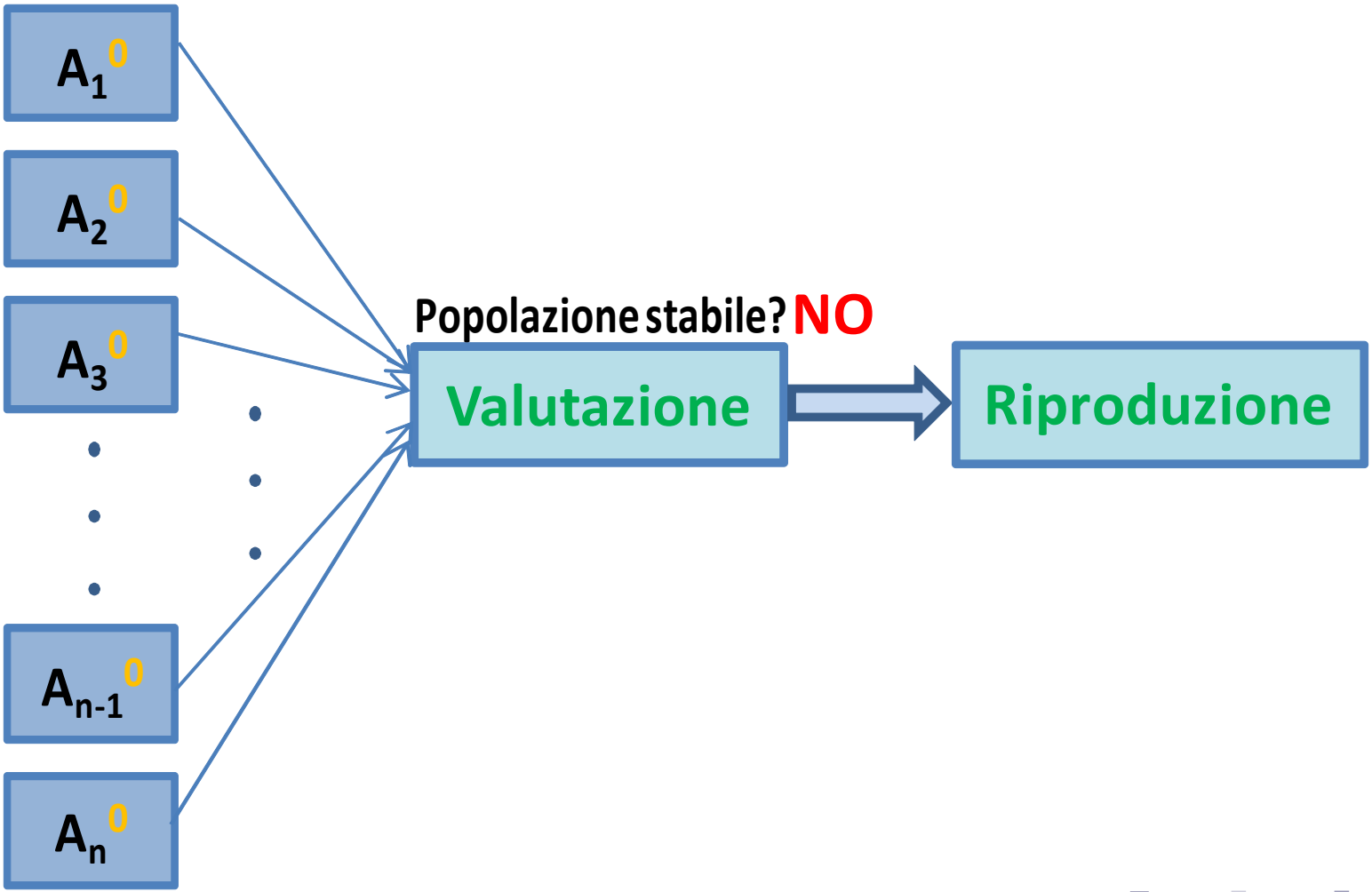
SCHEMA GENERALE



SCHEMA GENERALE



SCHEMA GENERALE



RIPRODUZIONE

- Overlapping generation: ad ogni passo solo una percentuale limitata della popolazione viene rimpiazzata (50%)
- La scelta degli individui da rimpiazzare avviene gratuitamente durante la valutazione dello score: gli $N/2$ allineamenti di costo più elevato sono destinati alla sostituzione
- Expected Offspring (EO): tasso di riproduzione di un individuo, utilizzato per una selezione probabilistica dei "genitori": se un individuo viene selezionato il suo EO è diminuito per l'intero turno di riproduzione

Operatori

In SAGA gli operatori *naturali* di **mutazione** e **ricombinazione** sono implementati tramite:

- Crossover: one-point, uniform
- Gap insertion
- Block shuffling
- Block searching
- Local optimal rearrangements

Operatori

In SAGA gli operatori *naturali* di **mutazione** e **ricombinazione** sono implementati tramite:

- Crossover: one-point, uniform
- Gap insertion
- Block shuffling
- Block searching
- Local optimal rearrangements

Operatori

In SAGA gli operatori *naturali* di **mutazione** e **ricombinazione** sono implementati tramite:

- Crossover: one-point, uniform
- Gap insertion
- Block shuffling
- Block searching
- Local optimal rearrangements

Crossover (ricombinazione)

- One point crossover: ricombina due allineamenti mediante un singolo scambio. Vengono prodotti due nuovi allineamenti (quello con lo score peggiore viene eliminato)
- Uniform crossover: meno distruttivo del precedente, promuove scambi fra zone di omologia. I blocchi da scambiare sono scelti fra posizioni consistenti
- **Consistenza**: dati due allineamenti, due posizioni sono consistenti se e solo se contengono in ogni riga il medesimo residuo

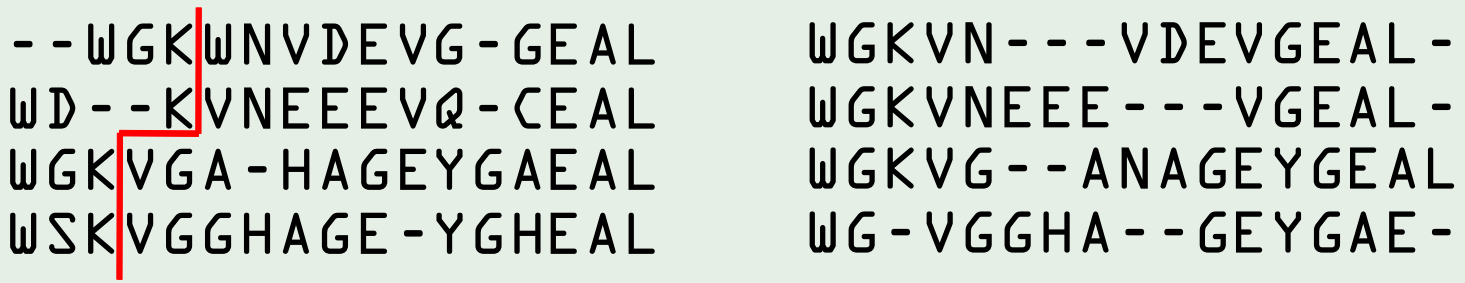
Crossover

Esempio

```
--WGKWNVDEVG-GEAL      WGKVN---VDEVGEAL-
WD--KVNEEEVQ-CEAL      WGKVNEEE---VGEAL-
WGKVGA-HAGEYGAEAL      WGKVG--ANAGEYGEAL
WSKVGGHAGE-YGHEAL      WG-VGGHA--GEYGAE-
```

Crossover

Esempio

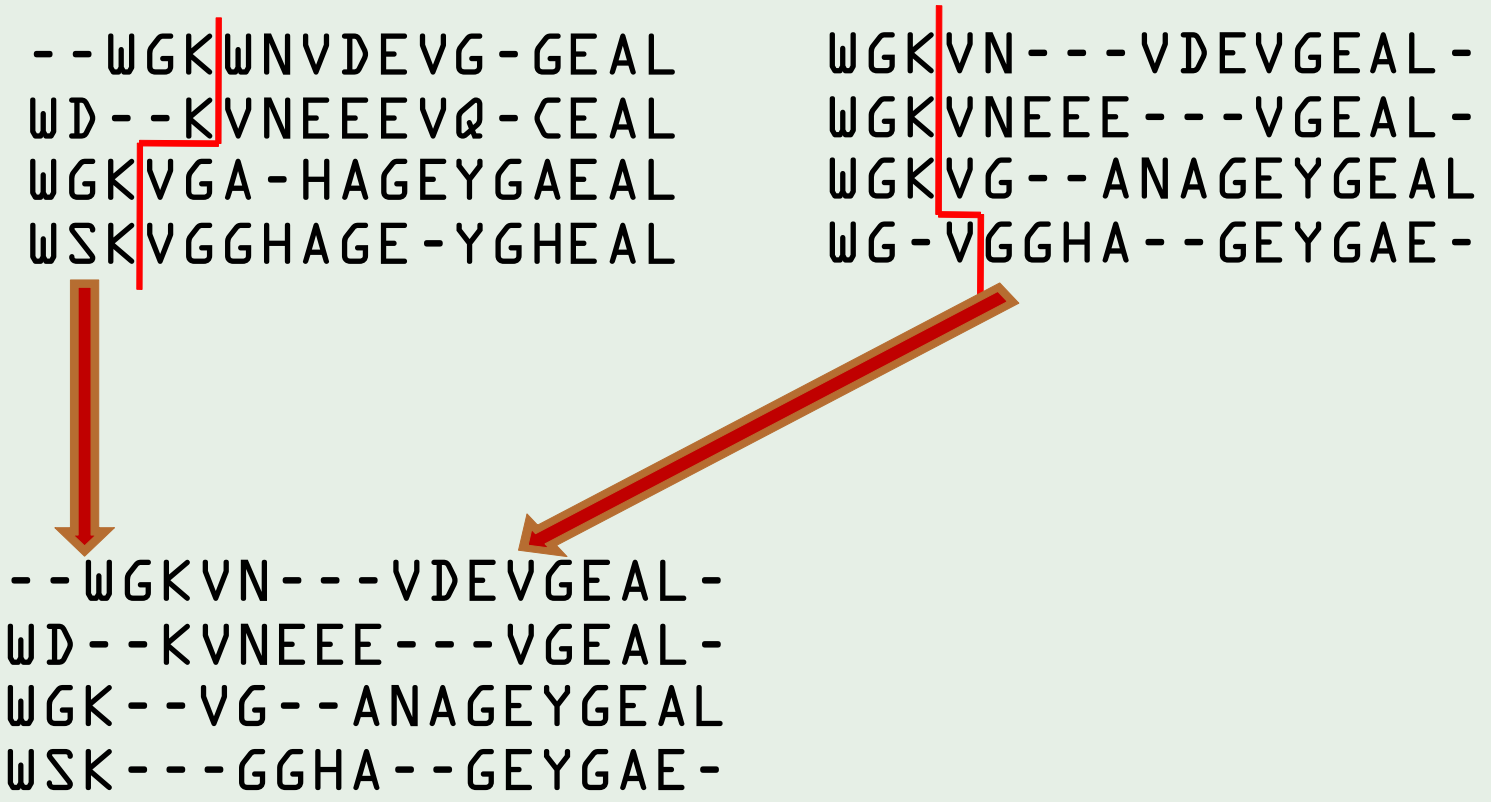


Crossover

Esempio

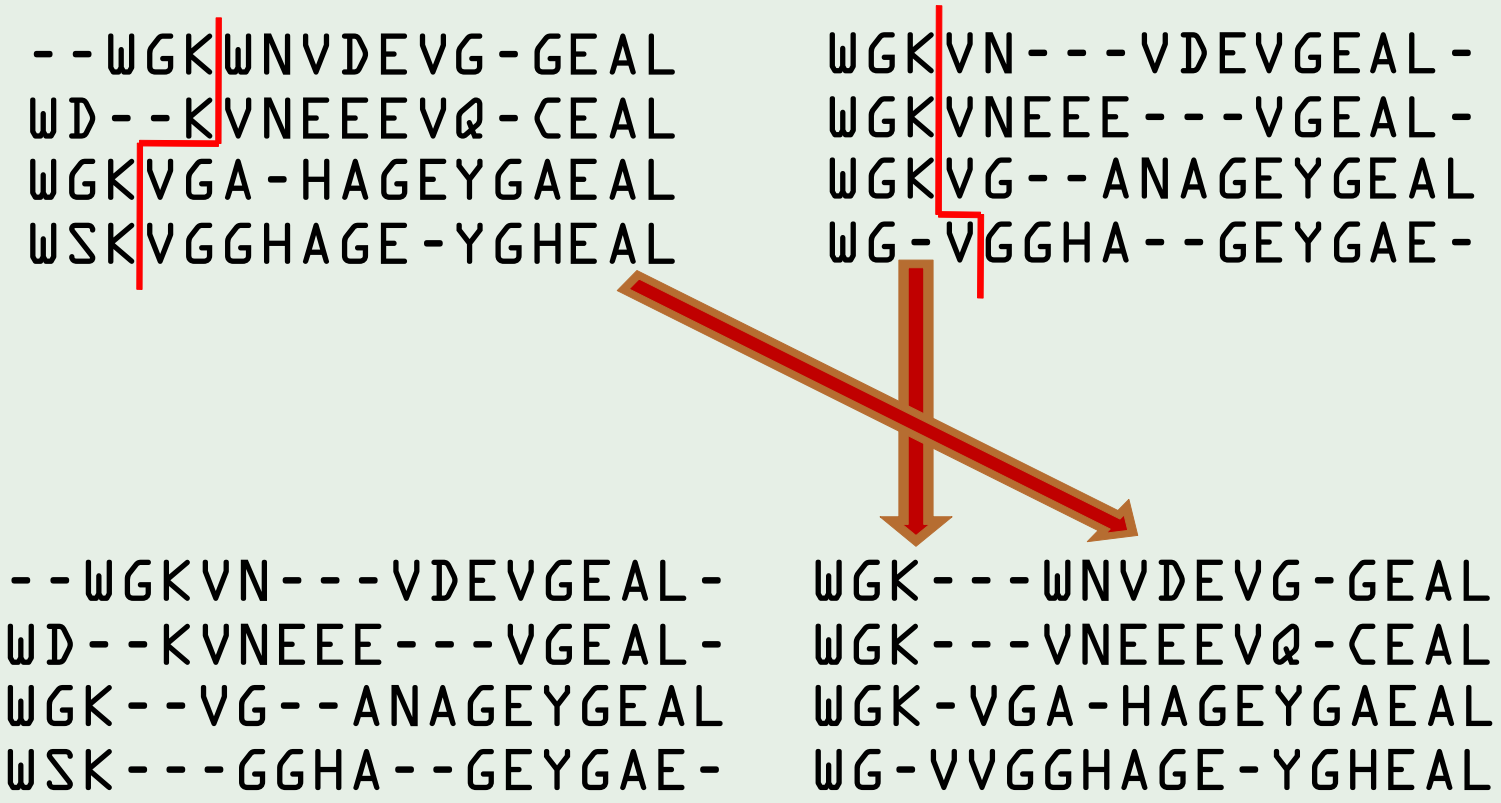
Crossover

Esempio



Crossover

Esempio



Crossover

Esempio

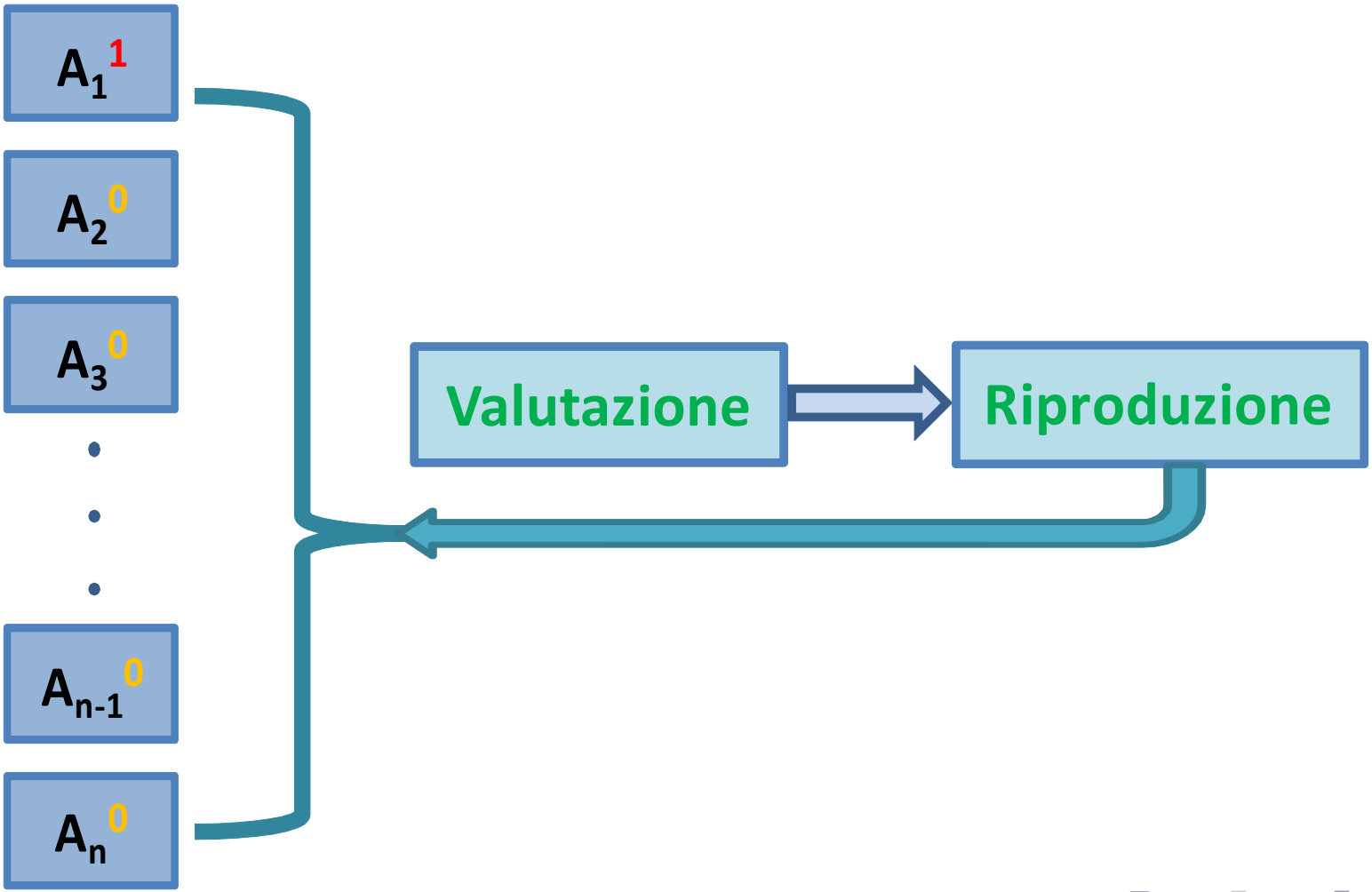
```
--WGKWNVDEVG-GEAL      WGKVN---VDEVGEAL-
WD--KVNEEEVQ-CEAL      WGKVNEEE---VGEAL-
WGKVGA-HAGEYGAEAL      WGKVG--ANAGEYGEAL
WSKVGGHAGE-YGHEAL      WG-VGGHA--GEYGAE-
```

```
WGK---WNVDEVG-GEAL
WGK---VNEEEVQ-CEAL
WGK-VGA-HAGEYGAEAL
WG-VVGGHAGE-YGHEAL
```

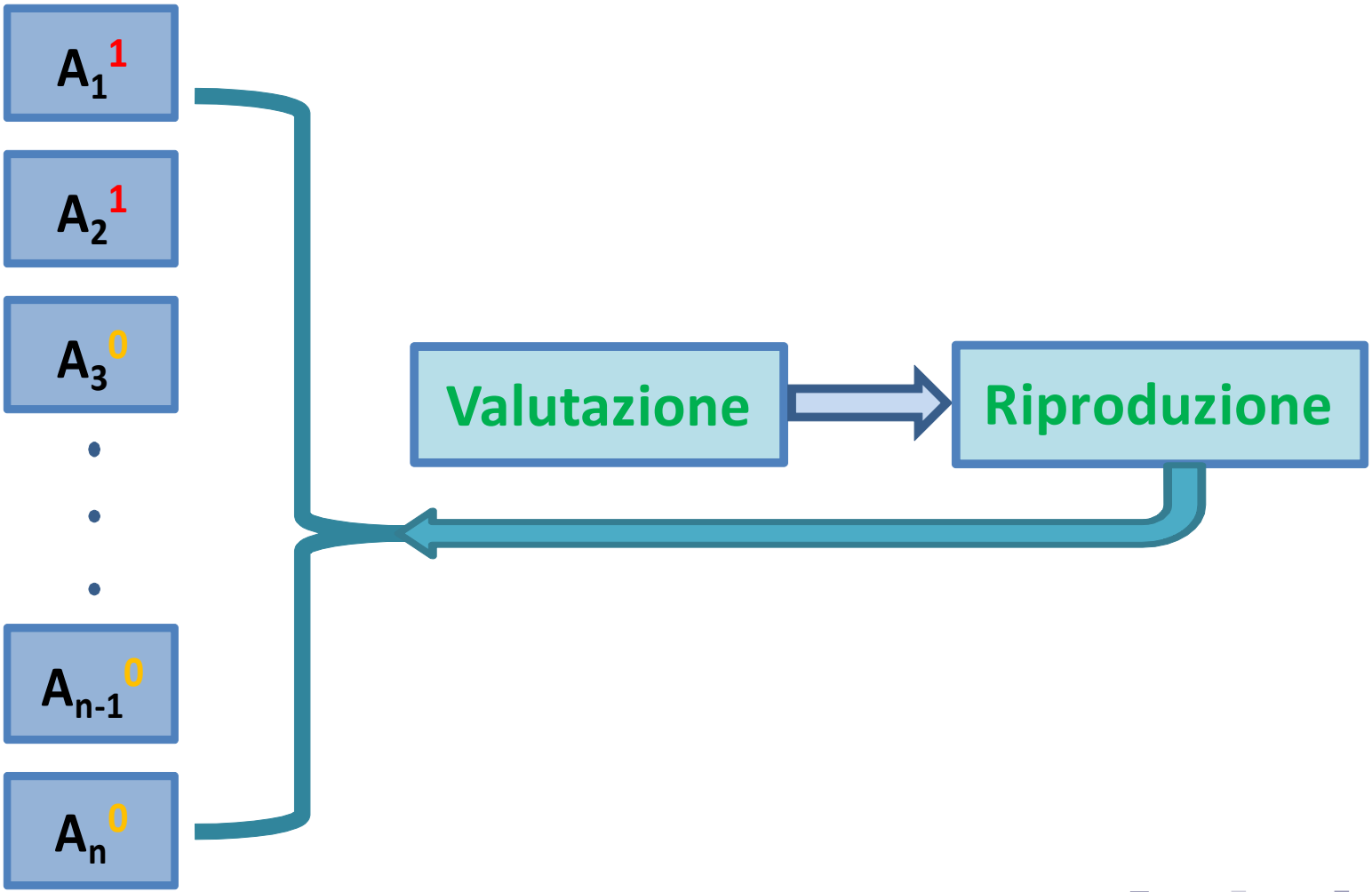
SCHEDULAZIONE DINAMICA DEGLI OPERATORI

- Ad ogni iterazione l'algoritmo sceglie l'operatore da utilizzare in maniera probabilistica
- Inizialmente gli operatori sono equiprobabili
- Le probabilità iniziali (non necessariamente ottimali) sono modificate dinamicamente durante l'esecuzione
- La probabilità associata all'operatore *op* è **proporzionale all'efficienza** nelle ultime 10 generazioni (miglioramento della qualità degli allineamenti prodotti tramite *op*)
- La probabilità associata ad un operatore resta in ogni caso strettamente maggiore di zero (per evitarne la scomparsa)

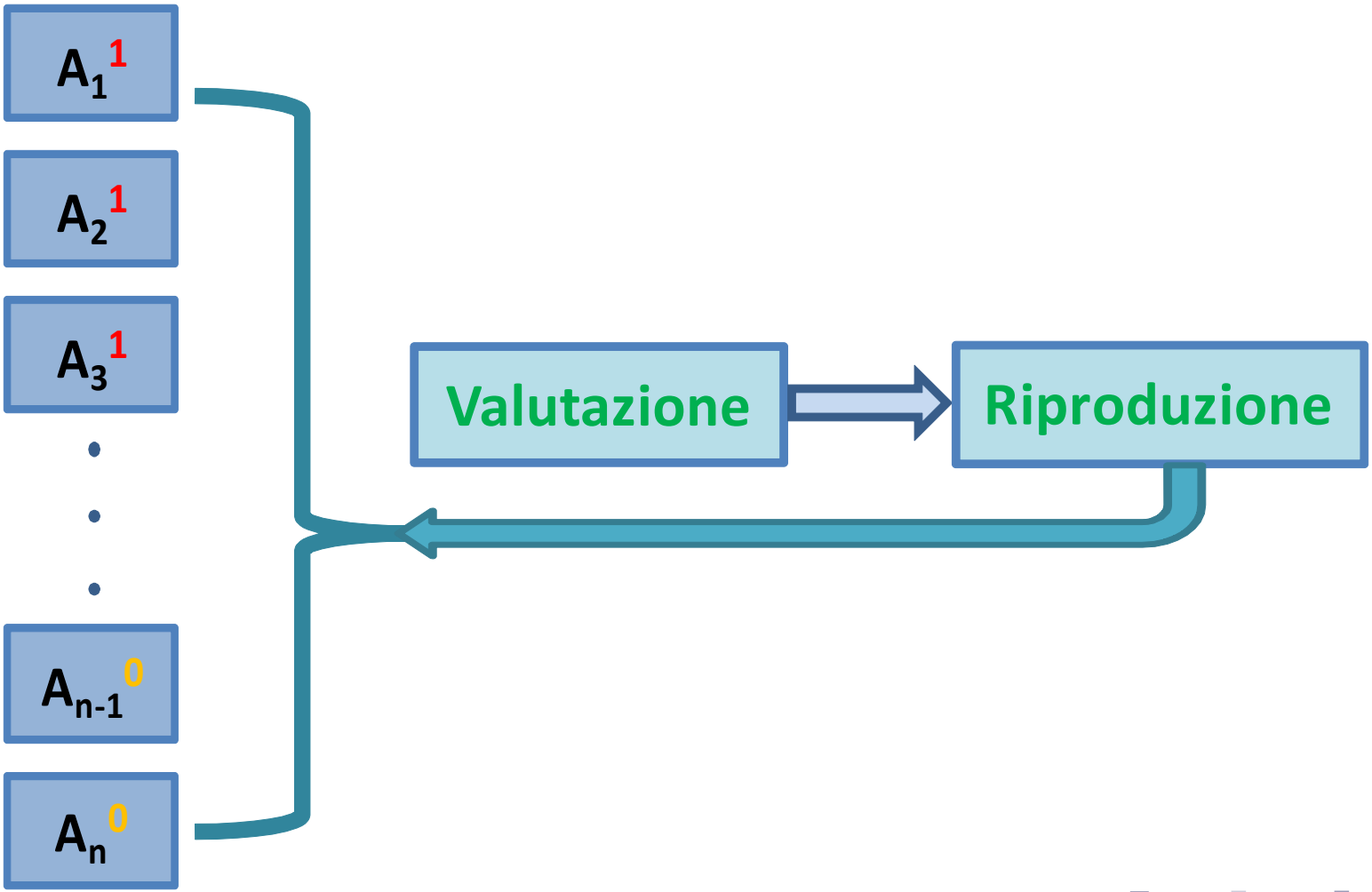
SCHEMA GENERALE



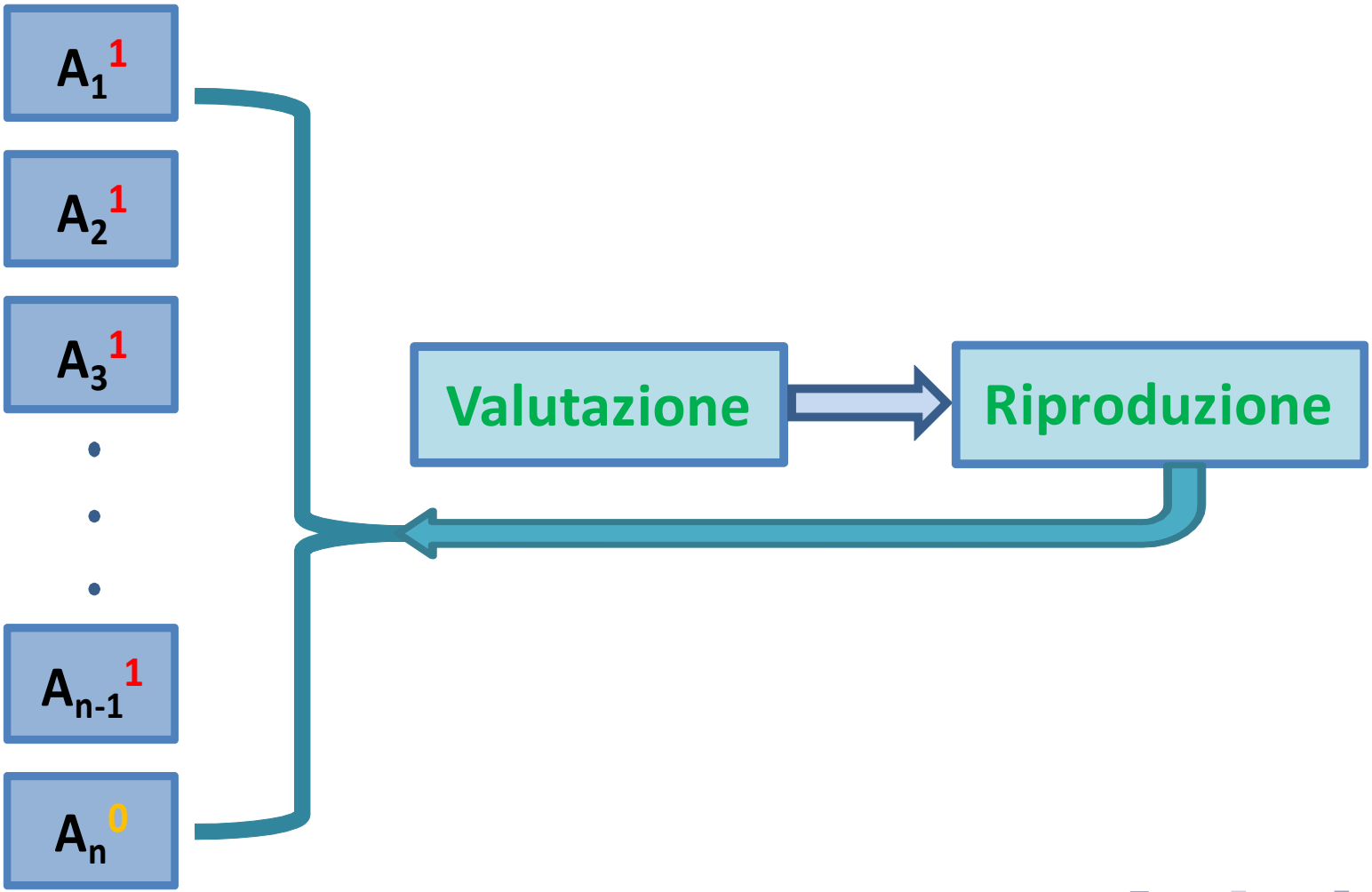
SCHEMA GENERALE



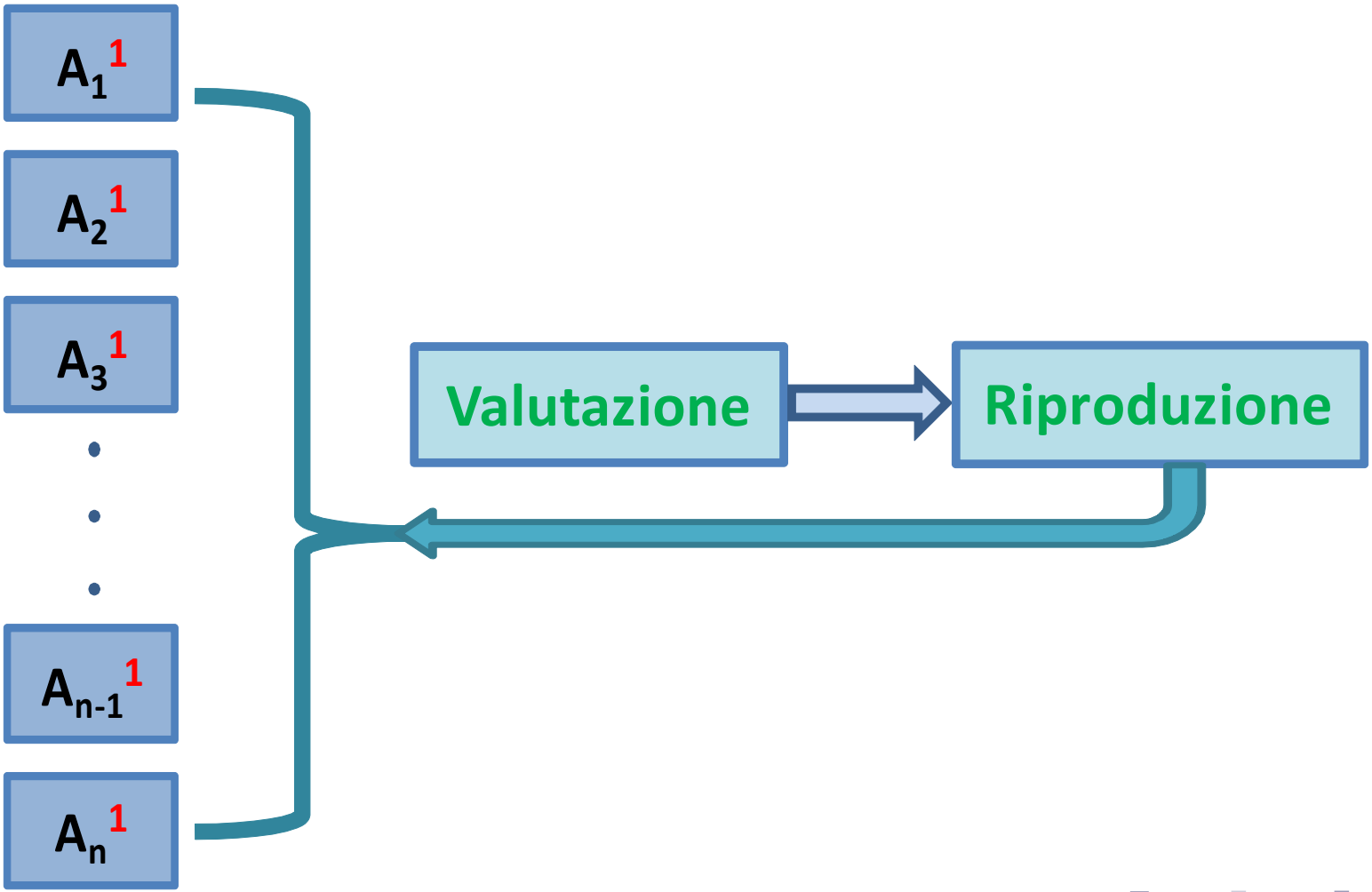
SCHEMA GENERALE



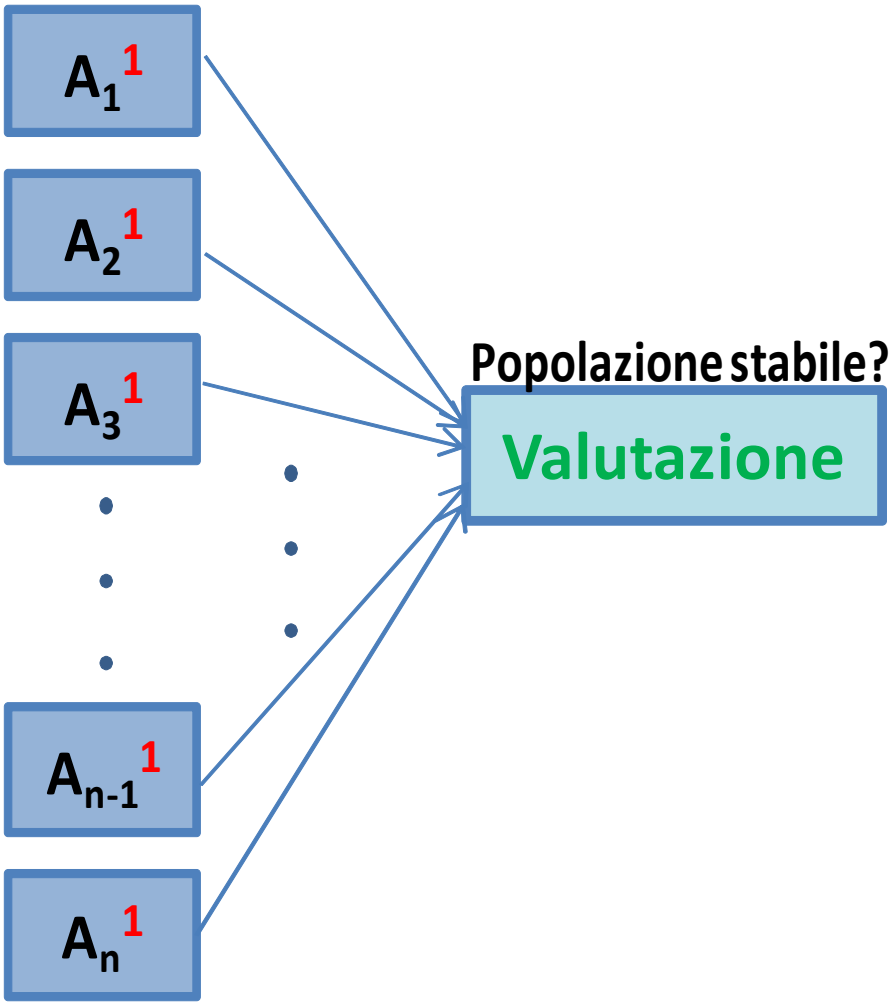
SCHEMA GENERALE



SCHEMA GENERALE



SCHEMA GENERALE



RISULTATI
AMAP

	Twilight-FP(209)		Superfamilies-FP(425)		
Tool	f_D	***AMA***	f_D	***AMA***	**Tempo**
CLUSTALW	20.4	35.5	50.9	37.0	1.7 sec
DIALIGN	17.0	74.1	46.7	71.5	5.7 sec
ProbCons	26.8	55.6	56.0	55.0	28.5 sec
T-Coffee	13.0	56.5	42.5	56.6	61.2 sec
AMAP$_{sens}$	27.3	68.3	56.1	63.8	13.5 sec
AMAP	19.2	84.4	46.4	84.2	11.2 sec

RISULTATI

SAGA vs CLUSTALW

		CLUSTAL W			
Proteina	**# seq**	**Lunghezza**	**Score**	**% ALLIN.**	**Tempo**
Igb	32	144	31.812.824	55,86	60 sec
Ac-Protease2	10	186	10.514.101	41,02	16 sec
S-Protease2	12	281	16.354.800	64,37	21 sec
Globin2	12	171	5.249.682	94,90	18 sec

		SAGA			
Proteina	**# seq**	**Lunghezza**	**Score**	**% ALLIN.**	**Tempo**
Igb	32	144	31.417.736	55,97	41.135 sec
Ac-Protease2	10	186	10.393.145	43,50	12.236 sec
S-Protease2	12	281	16.282.179	66,18	20.537 sec
Globin2	12	171	5.233.058	94,01	2.538 sec

Bibliografia I

- Christopher Lee et al.: **Multiple sequence alignment using partial order graphs**
  Bioinformatics, 2002
- Chuong B. Do et al.: **ProbCons: Probabilistic consistency-based multiple sequence alignment**
  Genome Research, 2005
- Michael Brudno et al.: **Lagan and Multi-Lagan: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA**
  Genome Research, 2003

Bibliografia II

- Michael Brudno et al.: **The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences**
  Nucleic Acids Research, 2004
- Julie D.Thompson et al.: **CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**
  Nucleic Acids Research, 1994
- Julie D. Thompson et al.: **The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**
  Nucleic Acids Research, 1997

Bibliografia III

- Ramu Chenna et al.: **Multiple sequence alignment with the Clustal series of programs**
  Nucleic Acids Research, 2003
- Ariel S. Schwartz, Lior Pachter: **Multiple alignment by sequence annealing**
  Bioinformatics, 2006
- Cédric Notredame, Desmond G. Higgins: **SAGA: sequence alignment by genetic algorithm**
  Nucleic acid Research, 1996