SAGA: sequence alignment by genetic algorithm

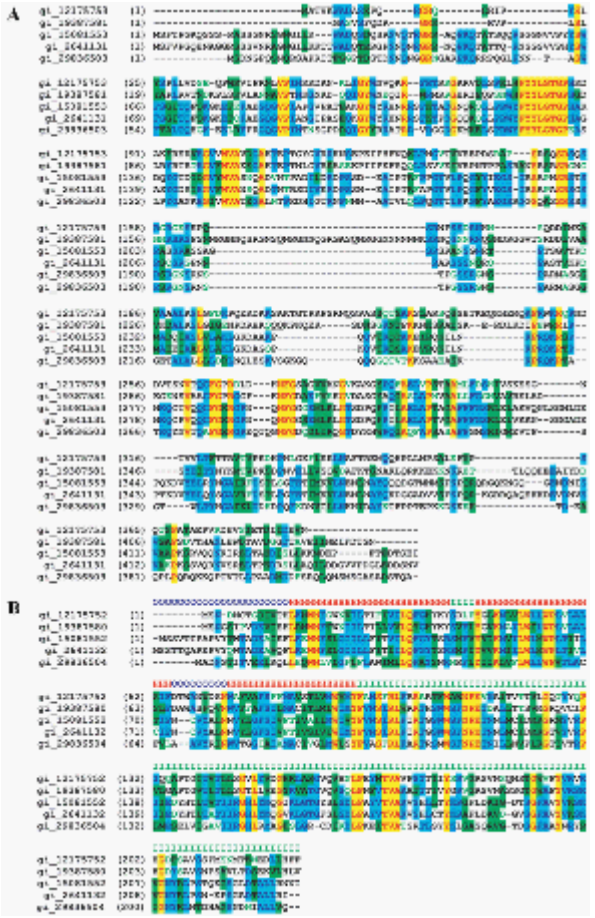
ALESSANDRO PIETRELLI

Soft Computing

Bologna, 25 Maggio 2007

Multi Allineamento di Sequenze (MSAs)

- Cosa sono?
- A cosa servono?
- Come vengono calcolati



Multi Allineamento di Sequenze (MSAs)

Cosa sono?

- Viene effettuato un **confronto tra sequenze** (principalmente omologhe)
- Approccio bioinformatico più rilevante per la caratterizzazione funzionale delle sequenze nucleotidiche e proteiche
- I siti funzionalmente più rilevanti, vengono **conservati** durante l'evoluzione

Multi Allineamento di Sequenze (MSAs)

A cosa servono?

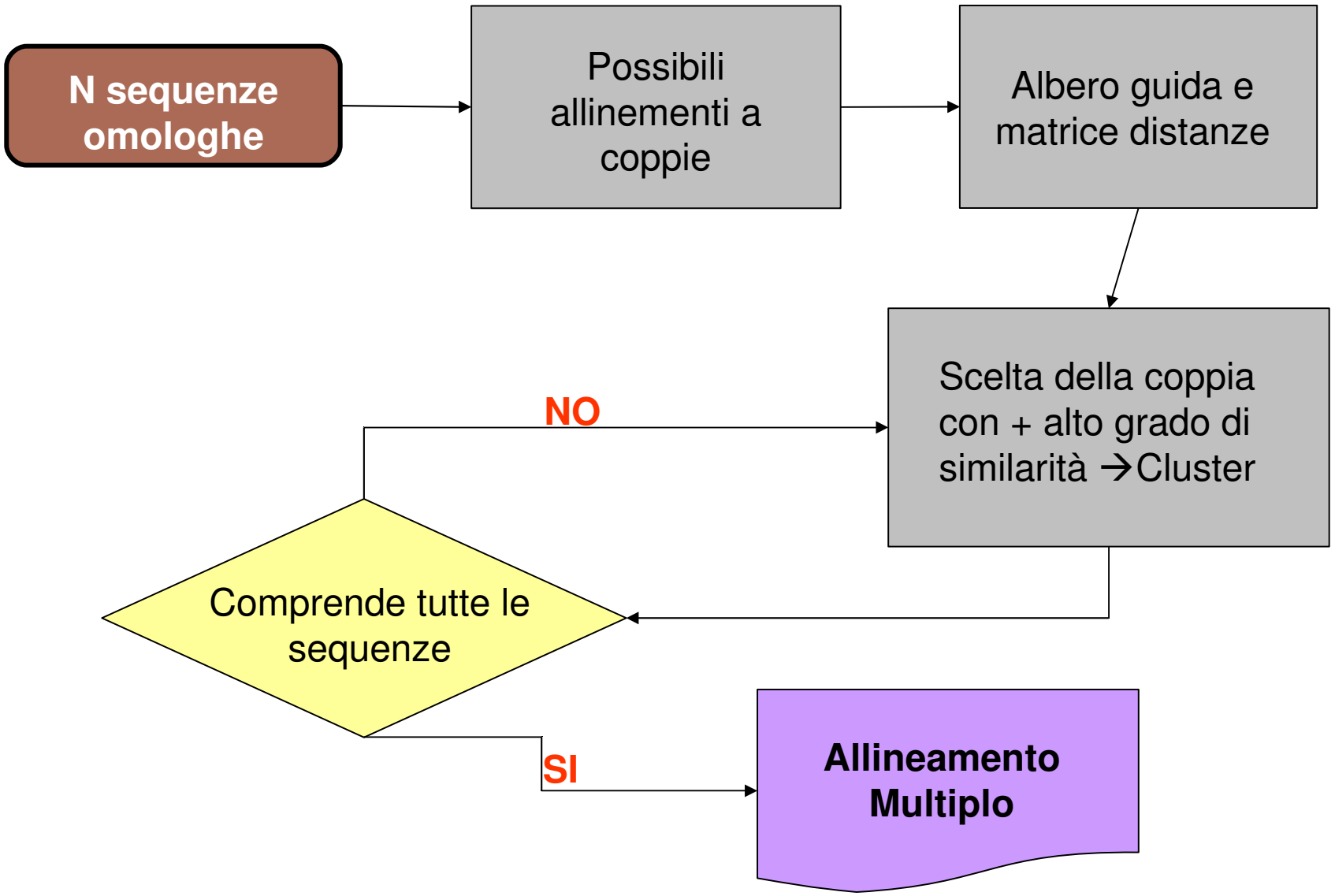
- Caratterizzazione di famiglie proteiche, **identificazione di regioni di omologia**
- Aiuto alla **predizione** di strutture secondarie e terziarie
- Step preliminare per la costruzione di **mappe filogenetiche** a livello molecolare
- Step preliminare per un disegno di esperimento volto a testare **funzioni proteiche**

Multi Allineamento di Sequenze (MSAs)

Calcolo

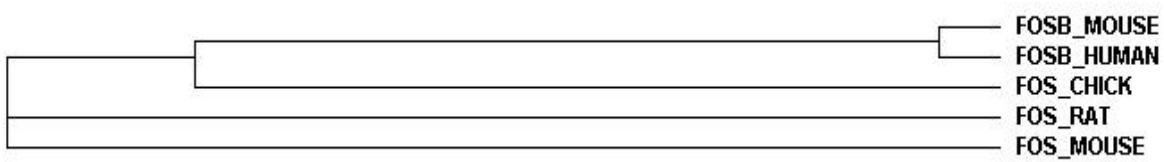
- Allineare solamente 2 sequenze è computazionalmente semplice e diretto
- Allineare 3 o più sequenze diventa un problema difficile da risolvere
- Il problema aumenta la propria **complessità esponenzialmente** a seconda delle sequenze coinvolte nell'allineamento

Allineamento Multiplo Progressivo



Schema

SFM137	0.211	SFM137																				
SFW113	0.213	0.003	SFW113																			
SFW112	0.213	0.003	0.000	SFW112																		
SFW80	0.214	0.002	0.001	0.001	SFW80																	
SFW79	0.212	0.002	0.000	0.000	0.000	SFW79																
SFW78	0.214	0.003	0.002	0.000	0.001	0.001	SFW78															
SFW77	0.216	0.005	0.002	0.002	0.003	0.002	0.002	SFW77														
SMMY45	0.228	0.175	0.178	0.178	0.177	0.176	0.177	0.181	SMMY45													
SMMY46	0.226	0.180	0.182	0.182	0.182	0.181	0.182	0.186	0.003	SMMY46												
SL126	0.238	0.167	0.172	0.172	0.172	0.171	0.172	0.177	0.163	0.161	SL126											
SO125	0.239	0.168	0.166	0.166	0.165	0.165	0.166	0.170	0.156	0.158	0.144	SO125										
SBM131	0.238	0.162	0.168	0.168	0.166	0.166	0.168	0.170	0.160	0.161	0.154	0.112	SBM131									
SBM132	0.236	0.164	0.168	0.168	0.166	0.166	0.168	0.170	0.160	0.161	0.151	0.112	0.008	SBM132								
SBB117	0.231	0.168	0.172	0.172	0.170	0.171	0.172	0.174	0.167	0.168	0.161	0.112	0.017	0.018	SBB117							
SBB118	0.231	0.168	0.172	0.172	0.170	0.171	0.172	0.174	0.167	0.168	0.161	0.112	0.017	0.018	0.000	SBB118						
SMN37	0.228	0.150	0.155	0.155	0.153	0.153	0.154	0.157	0.161	0.163	0.161	0.117	0.066	0.067	0.066	0.066	SMN37					
SMN38	0.230	0.160	0.164	0.164	0.163	0.163	0.164	0.167	0.168	0.167	0.161	0.123	0.068	0.069	0.070	0.070	0.015	SMN38				
SMN39	0.228	0.159	0.163	0.163	0.161	0.162	0.163	0.165	0.167	0.166	0.157	0.122	0.068	0.069	0.070	0.070	0.015	0.002	SMN39			
SMN40	0.230	0.160	0.164	0.164	0.163	0.163	0.164	0.167	0.168	0.167	0.161	0.123	0.068	0.069	0.070	0.070	0.015	0.000	0.002	SMN40		
SMM133	0.220	0.153	0.158	0.158	0.156	0.156	0.157	0.160	0.163	0.161	0.153	0.120	0.070	0.072	0.073	0.073	0.017	0.021	0.021	0.021	SMM133	
SMM134	0.219	0.155	0.159	0.159	0.157	0.158	0.159	0.161	0.161	0.160	0.153	0.120	0.070	0.072	0.073	0.073	0.017	0.021	0.021	0.021	0.002	




```
FOS_RAT          KCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPNDLGF 212
FOS_MOUSE        KCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGF 212
FOS_CHICK        KCRNRRRELTDTLQAETDQLEEEKSALQAEIANLLKEKEKLEFILAAHRPACKMPEELRF 211
FOSB_MOUSE       KCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEG- 229
FOSB_HUMAN       KCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEG- 229
                 ********** *********:**: *::***:* ****:***:*.**:*.**:* :
```

1

Matrice delle distanze

2

Albero guida

3

Allineamento multiplo progressivo

ClustalW

- Uno dei maggiori tool di multiallineamento utilizzati nella rete
- Utilizza un **algoritmo di tipo progressivo**
- Genera un albero (cladogramma) per evidenziare relazioni evoluzionistiche tra le sequenze in questione
- **Non ho uno score** relativo alla bontà dell'allineamento

ClustalW

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	FOSB_MOUSE	338	2	FOSB_HUMAN	338	95
1	FOSB_MOUSE	338	3	FOS_CHICK	367	43
1	FOSB_MOUSE	338	4	FOS_RAT	380	43
1	FOSB_MOUSE	338	5	FOS_MOUSE	380	44
2	FOSB_HUMAN	338	3	FOS_CHICK	367	43
2	FOSB_HUMAN	338	4	FOS_RAT	380	43
2	FOSB_HUMAN	338	5	FOS_MOUSE	380	45
3	FOS_CHICK	367	4	FOS_RAT	380	74
3	FOS_CHICK	367	5	FOS_MOUSE	380	75
4	FOS_RAT	380	5	FOS_MOUSE	380	96

- Punteggio determinato dalle matrici di sostituzione (PAM, Blosum, ecc..)
- Gap Score (ex., term.)

PLEASE NOTE: Some scores may be missing from the above table if the alignment was done using multiple CPU mode. Please check the output.

Sort by | Sequence Number | View Output File

Phylogram

Show as Cladogram Tree | Hide Distances | View DND File

Matrici di Sostituzione

• Forniscono dei **punteggi** appropriati ad ogni coppia di amminoacidi appaiati in un allineamento

PAM (Percent Accepted Mutation):

• Basate sulla percentuale di sostituzione amminoacidica durante l'evoluzione

Blosum:

• Basate sulla banca dati BLOCKS, contenente una collezione di allineamenti senza gap (30%→95% di identità)

• Da ognuno di questi blocchi è possibile ricavare la frequenza relativa di sostituzione amminoacidica

Matrici di Sostituzione

Es. PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Pro / Contro

PRO:

- Algoritmo rapido e semplice
- Supporta molte sequenze da allineare
- Alta sensitività

CONTRO

- Minimo locale
- Similarità aa > 50% ; Similarità DNA > 70%
- Parametro per la bontà dell'allineamento

Alternative

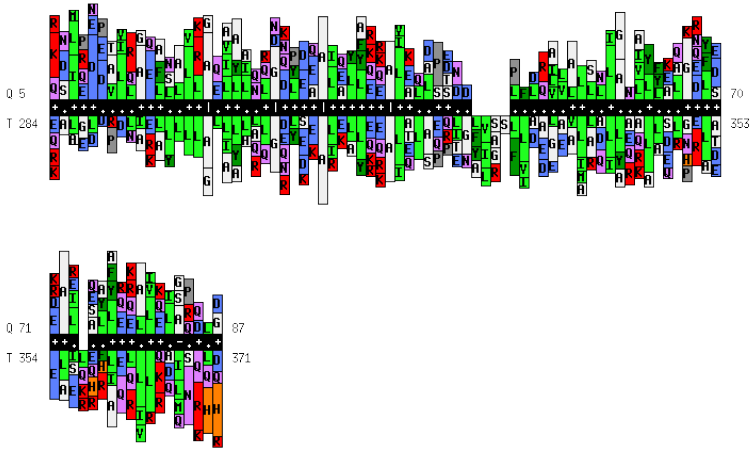
HMM: Metodo statistico nel quale il sistema è *modellato* sulle catene di Markov

Questo approccio ha bisogno di molte sequenze per l'addestramento (più di 100)

Objective Functions: Misurano la qualità del multiallineamento

Trova il migliore allineamento anche a **livello biologico**

Approccio **computazionalmente impossibile**

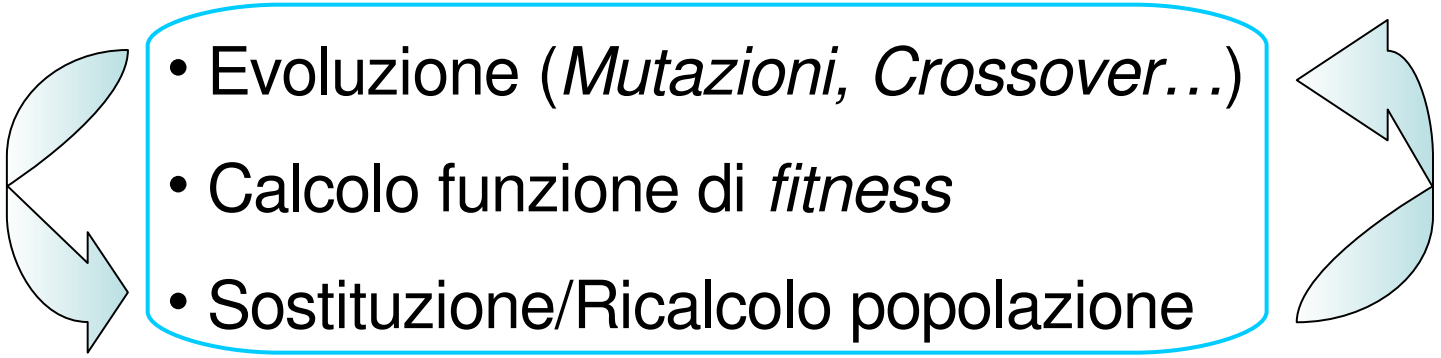


Genetic Algorithm

- È un metodo euristico di ricerca ed ottimizzazione, ispirato al principio della selezione naturale di Charles Darwin che regola l'evoluzione biologica.

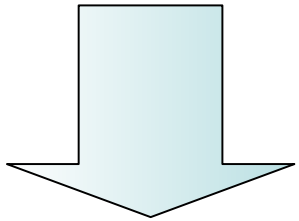
<u>Principi di funzionamento:</u>

- Soluzione iniziale casuale (*Individui*)
- Evoluzione (*Mutazioni, Crossover…*)
- Calcolo funzione di *fitness*
- Sostituzione/Ricalcolo popolazione



OF + GA = **SAGA**

- Scopo: Ricerca di OFs che riescano a descrivere al meglio gli allineamenti generati
- Metodo: Uso di 2 differenti OFs attraverso l'ottimizzazione degli algoritmi genetici



Somma pesata delle coppie di amminoacidi (WSP)

+

Penalità dei gap

… nel dettaglio

$$\text{ALIGNMENT COST(A)} = \sum_{i=2}^{N} \sum_{j=1}^{i-1} W_{i,j} \, \text{COST}(A_i, A_j)$$

Punteggio totale dell'allineamento

Peso relativo dato alle sequenze

Costo delle sequenze Ai,Aj

OF1: WSP + PAM250 + quasi-natural gap penalties (MSA)

OF2: WSP + PAM250 + natural gap penalties (ClustalW)

Algoritmo

1. *Inizializzazione*
2. *Valutazione*
3. *Incrocio/Evoluzione*
4. *Fine*

Generazione casuale della popolazione (G_0)

Dimensione = 100

Consiste in set di allineamenti contenenti solo gap terminali

Algoritmo

1. *Inizializzazione*
2. *Valutazione*
3. *Incrocio/Evoluzione*
4. *Fine*

Viane valutata la *fitness* di ogni individuo attraverso la OF

↓

Solo una parte della popolazione (50%) viene rimpiazzata dalla progenie

Algoritmo

1. *Inizializzazione*
2. *Valutazione*
3. *Incrocio/Evoluzione*
4. *Fine*

La nuova generazione (G_1) viene creata dal 50% “buono” della generazione precedente

+

“Progenie” creata da modificazioni dei genitori selezionati per generare nuovi individui

Algoritmo

1. *Inizializzazione*
2. *Valutazione*
3. *Incrocio/Evoluzione*
4. *Fine*

La creazione dei nuovi individui avviene attraverso la scelta casuale degli operatori implementati

La popolazione non produce duplicati, aumentando così la variabilità

Algoritmo

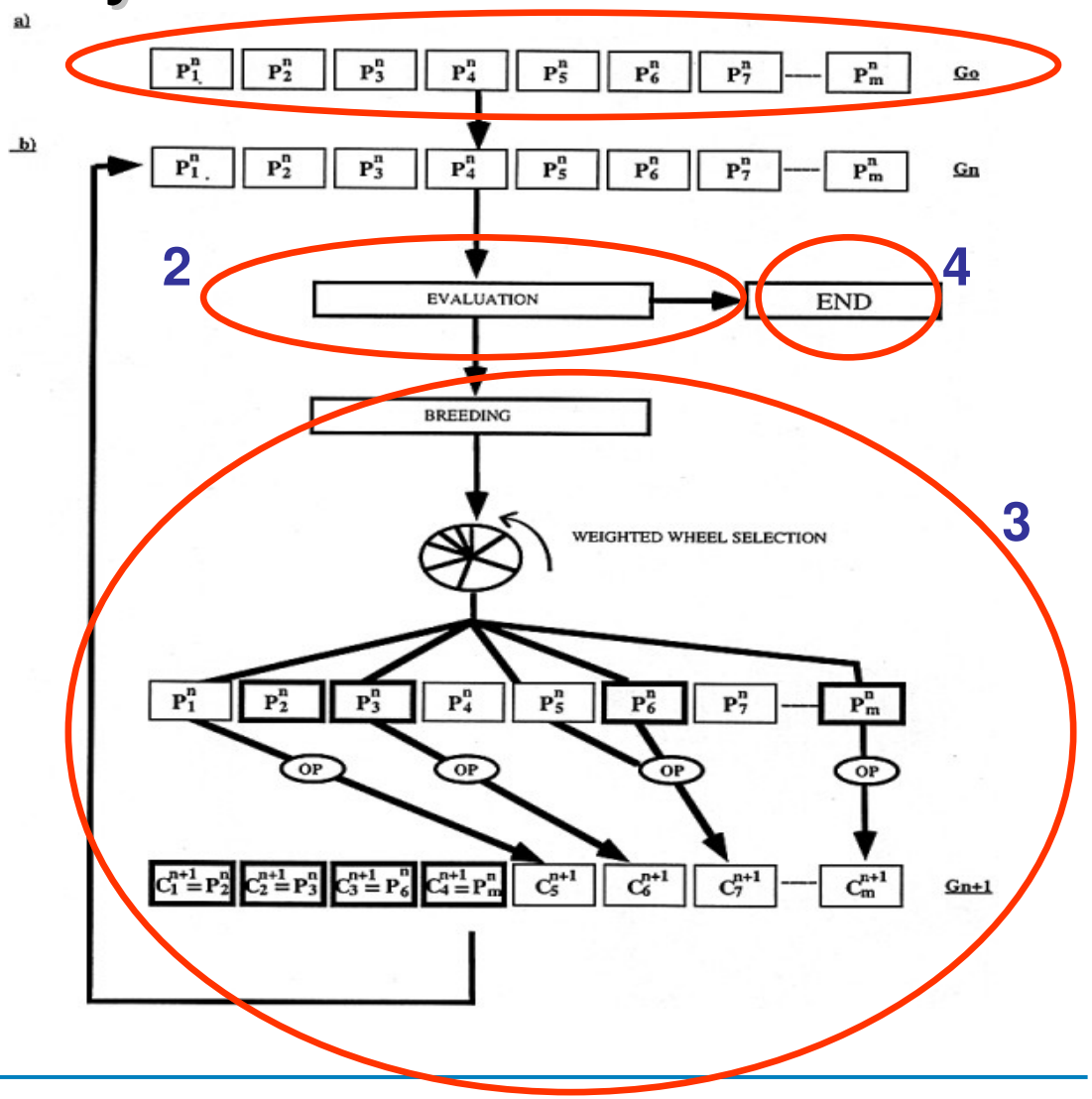
1. *Inizializzazione*
2. *Valutazione*
3. *Incrocio/Evoluzione*
4. *Fine*

Criterio euristico di STOP:

100 generazioni

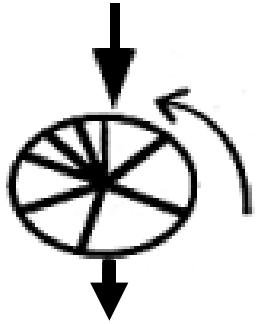
Layout

1. *Inizializzazione*
2. *Valutazione*
3. *Incrocio/Evoluzione*
4. *Fine*



Scelta dei genitori

- EO = Expected Offspring (0..2)
- Valore attribuito ad ogni individuo, deriva dal valore di fitness
- Probabilità di essere scelto come “genitore”



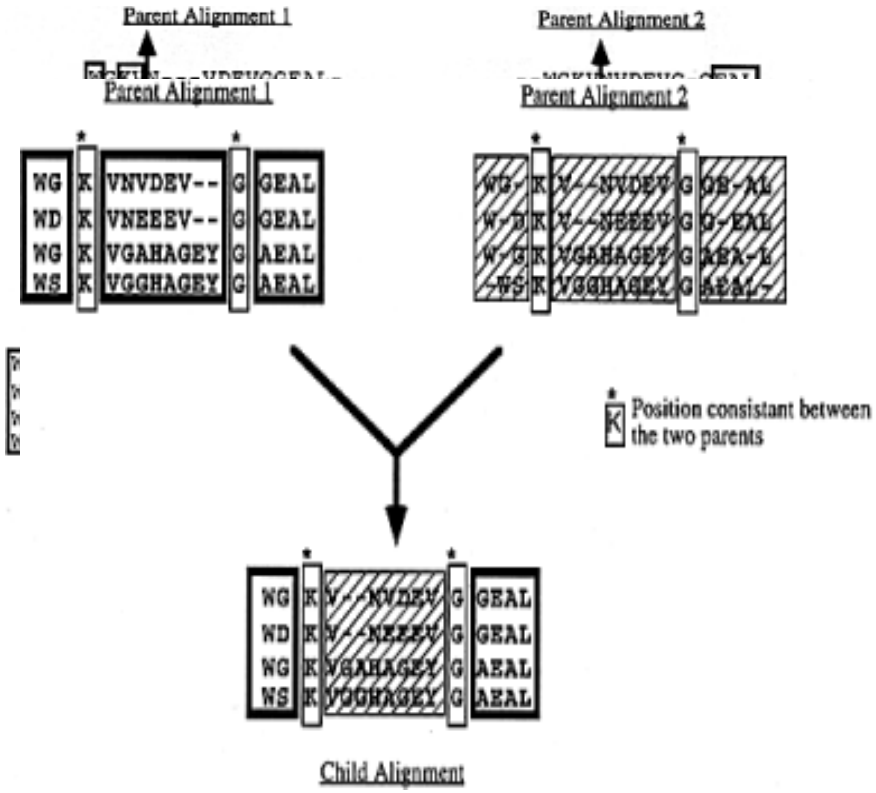
WEIGHTED WHEEL SELECTION

- Scelti i genitori, i loro punteggi di EO vengono decrementati

Gli operatori

Crossover:

1. One-Point: Combina 2 individui in 1 attraverso un singolo taglio
2. Uniforme: Swap di blocchi dei 2 individui genitori



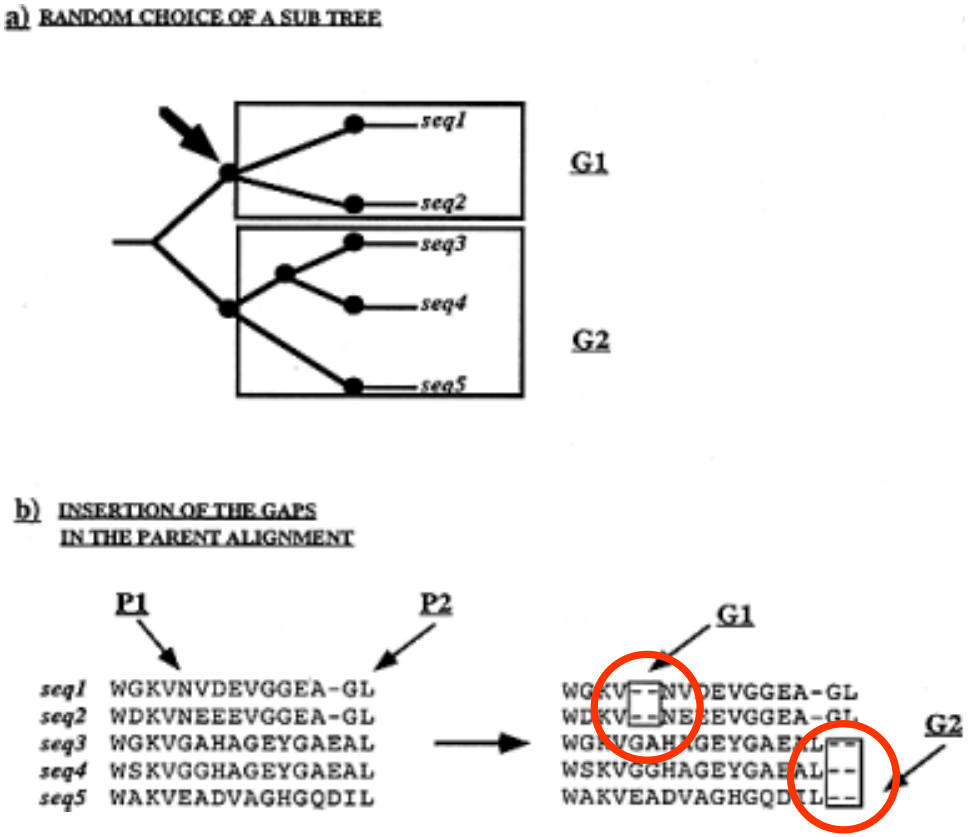
Gli operatori

Gap insertion:

Estende gli allineamenti inserendo gaps

Le sequenze vengono divise in 2 gruppi

In ogni gruppo il gap è nella medesima posizione

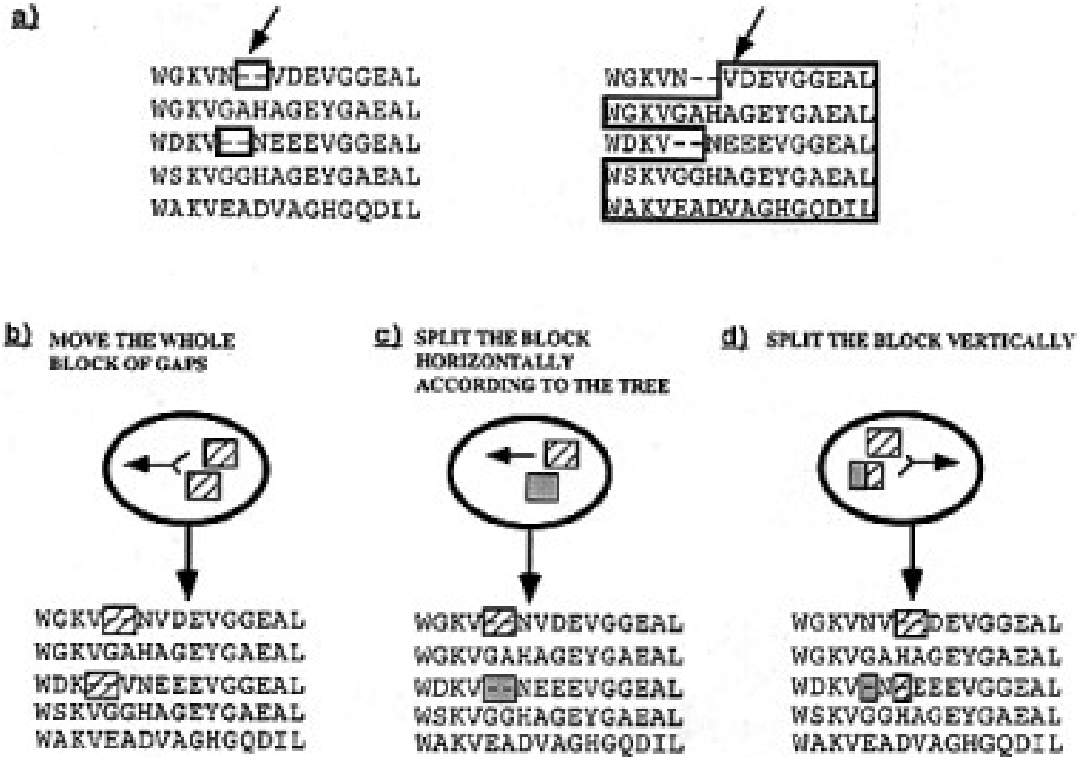


Gli operatori

Block shuffling:Muove blocchi di gap/residui all'interno dell'allineamento

Blocco = residui/gap sovrapposti delimitati da gap/residui o termine sequenza

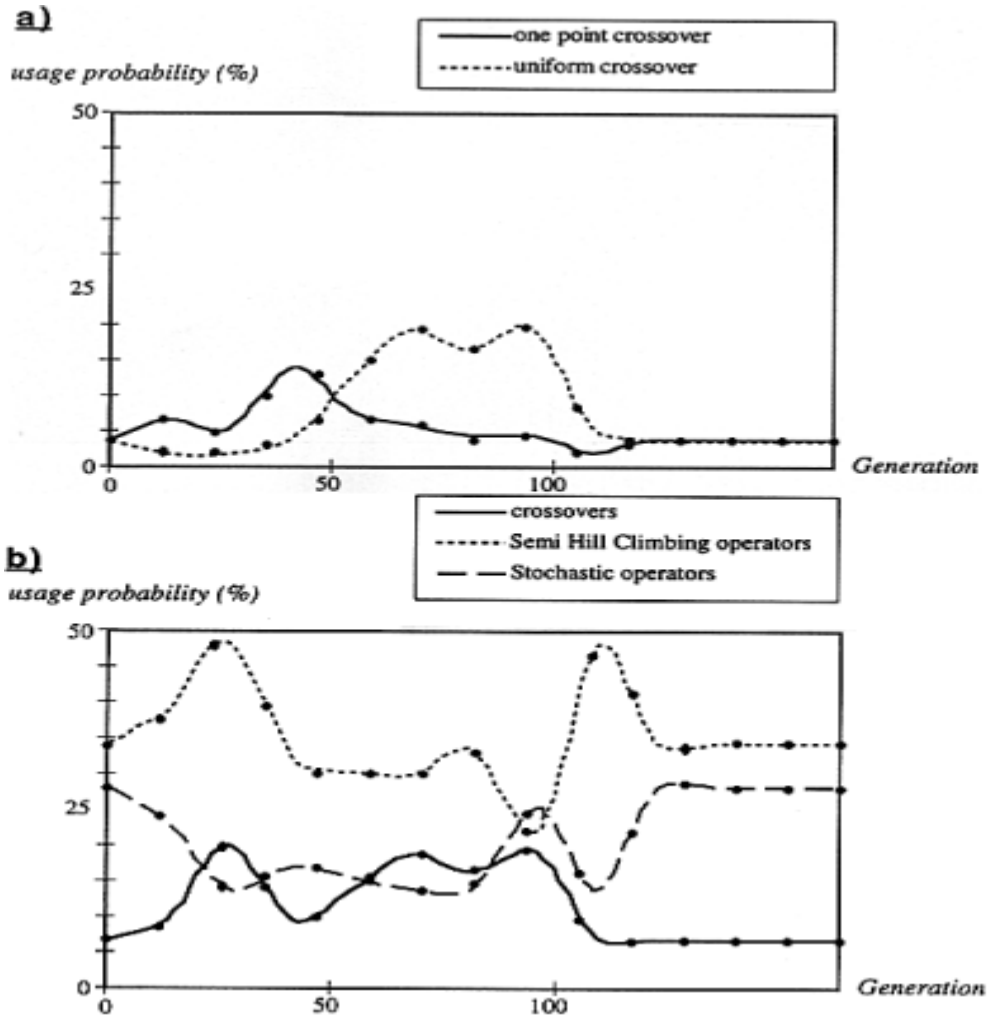
Il blocco deriva dalla scelta casuale della posizione del primo residuo/gap che viene selezionato



Scelta dinamica degli operatori

- 22 diversi operatori implementati
- Ad ogni operatore è data una probabilità di scelta $G_0=1/22$
- Ad ogni generazione viene calcolato un punteggio in base alla *qualità* degli individui generati da quel determinato operatore
- Ogni 10 generazioni→ P_{op} = Score/Child
- Se un operatore produce allineamenti migliori→ Aumenta la sua probabilità di essere usato
- Minimo 1/44

Scelta dinamica degli operatori



Abilità di *Self-Tuning*

La probabilità ritorna ai livelli iniziali quando non porta più nessun miglioramento

Test case – OF1

Table 1. The performance of MSA and SAGA on nine test cases

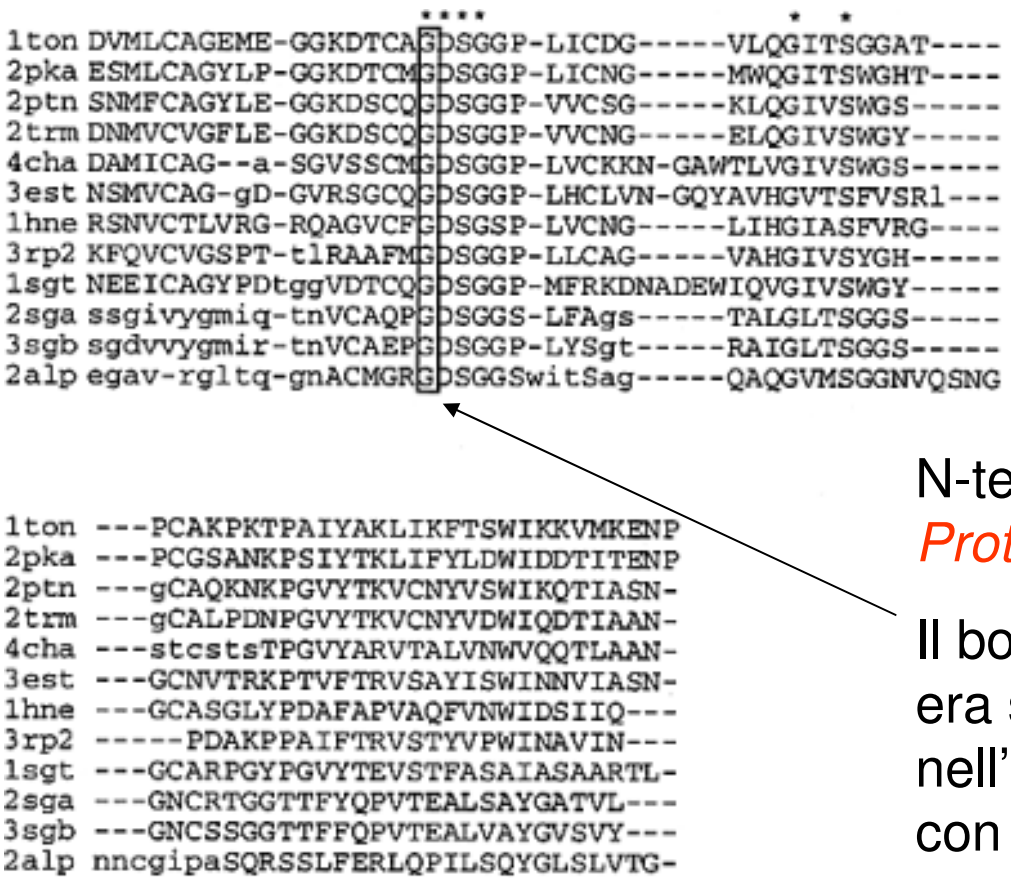
Test case	Nseq	Length	MSA score	MSA versus structure (%)	CPU-time	SAGA score	SAGA versus structure (%)	CPU-time
Cyt c	6	129	1 051 257	74.26	7	1 051 257	74.26	960
Gcr	8	60	371 875	75.05	3	371 650	82.00	75
Ac protease	5	183	379 997	80.10	13	379 997	80.10	331
S protease	6	280	574 884	91.00	184	574 884	91.00	3500
Chtp	6	247	111 924	*	4525	111 579	*	3542
Dfr secstr	4	189	171 979	82.03	5	171 975	82.50	411
Sbt	4	296	271 747	80.10	7	271 747	80.10	210
Globin	7	167	659 036	94.40	7	659 036	94.40	330
Plasto	5	132	236 343	54.03	22	236 195	54.05	510

Test case – OF2

Table 2. The performance of CLUSTAL W and SAGA on four test cases

Test case	Nseq	Length	CLUSTAL W score	CLUSTAL W versus structure (%)	CPU-time	SAGA score	SAGA versus structure (%)	CPU-time
Igb	32	144	31 812 824	55.86	60	31 417 736	55.97	41 135
Ac Protease2	10	186	10 514 101	41.02	16	10 393 145	43.50	12 236
S Protease2	12	281	16 354 800	64.37	21	16 282 179	66.18	20 537
Globin2	12	171	5 249 682	94.90	18	5 233 058	94.01	2538

Test case – OF2



N-terminale della
Proteasi S2

Il box di Glicine marcato non era stato trovato nell'allineamento ottenuto con ClustalW

Conclusioni

✓ Tool molto flessibile

✓ Ottima combinazione per la scelta degli operatori→ Aumento della variabilità

✓ Ricerca consistente dell'allineamento ottimo→ Confronto con le strutture terziarie

✓ La strategia con gli *algoritmi genetici* porta a testare diverse OFs

✓ Nessun problema di minimo locale

✘ Lento rispetto ad un algoritmo Greedy