

Algoritmi genetici

Algoritmi genetici

□ **Corso di laurea in Informatica**

(anno accademico 2024/2025)

- Insegnamento: Apprendimento ed evoluzione in sistemi artificiali
- Docente: Marco Villani

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



Dipartimento di
Scienze Fisiche,
Informatiche
e Matematiche

E' vietata la copia e la riproduzione dei contenuti e immagini in qualsiasi forma. E' inoltre vietata la redistribuzione e la pubblicazione dei contenuti e immagini non autorizzata espressamente dall'autore o dall'Università di Modena e Reggio Emilia

Ispirazione biologica

- In natura vi sono molti esempi di sistemi capaci di apprendere (modificare se stessi in risposta ad opportuni stimoli)
 - Sistema immunitario
 - Colonie di insetti
 - Cervello
 - ...
- Su scala temporale differente (e vedremo cosa questo significhi), altrettanto fanno
 - Le singole specie, all'interno della storia dell'evoluzione biologica

Evoluzione e apprendimento

- Nel corso dell'evoluzione le specie “apprendono” implicitamente alcune caratteristiche del proprio ambiente
- C'è anche un apprendimento individuale nel corso dell'esistenza
 - Scale temporali ben separate
- Da un punto di vista astratto sono abbastanza simili

L'evoluzione biologica

- Convinzioni diffuse prima del '700
 - la terra è giovane
 - le specie viventi sono apparse simultaneamente e si sono mantenute stabili
- La scoperta dell'evoluzione delle specie
 - la geologia mostra che la terra è antica
 - la sistematica e lo studio dei fossili
 - si diffonde la convinzione che le specie viventi siano soggette a cambiamenti

Lamarck

- Ipotesi: cambiamenti che avvengono nel corso della vita di un individuo vengono trasmessi ai discendenti
 - il collo delle giraffe
 - il motore dell'evoluzione è l'“apprendimento” in vita
 - I due fenomeni sono direttamente interconnessi
- in questo modo si ha un graduale adattamento progressivo della specie all'ambiente

Darwin e Wallace

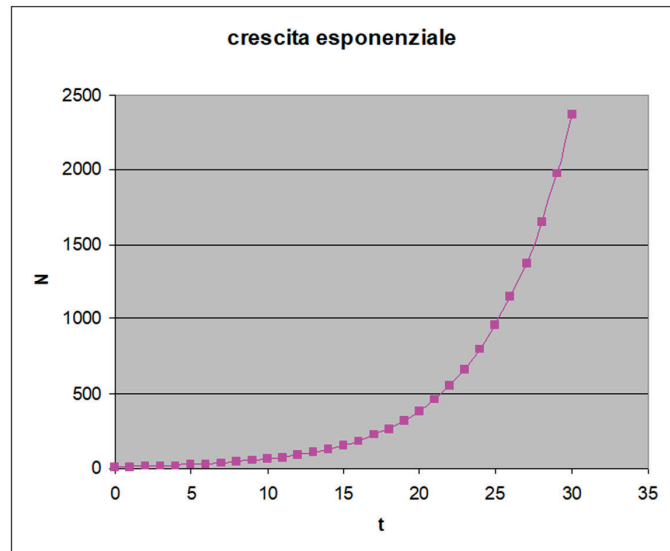
- ❑ I cambiamenti che avvengono durante la vita di un individuo non sono ereditabili
- ❑ le specie cambiano perché vi sono differenze fra gli individui, fin dalla nascita
- ❑ Queste differenze sono fondamentalmente casuali

Darwin e Wallace

- ❑ L'ambiente esercita una pressione selettiva che favorisce gli individui portatori di caratteri vantaggiosi, che invadono la popolazione
- ❑ La pressione selettiva deriva dalla tensione fra la tendenza alla crescita esponenziale delle popolazioni e il vincolo delle risorse finite, che consente solo a una frazione dei nuovi nati di sopravvivere e riprodursi

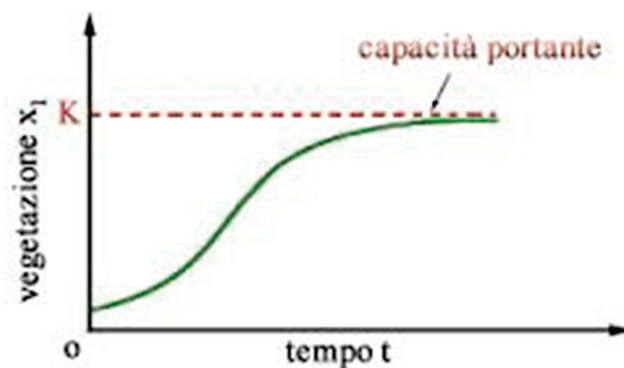
L'ispirazione economica: Malthus

- $N(t+1) = N(t)[1+k]$
- se tutti i discendenti sopravvivevano e si riproducevano, risorse finite verrebbero rapidamente esaurite



L'ispirazione economica: Malthus

- $N(t+1) = N(t)[1+k] - \beta N^2$
- Vi è una capacità portante finita



Sottopopolazioni in competizione

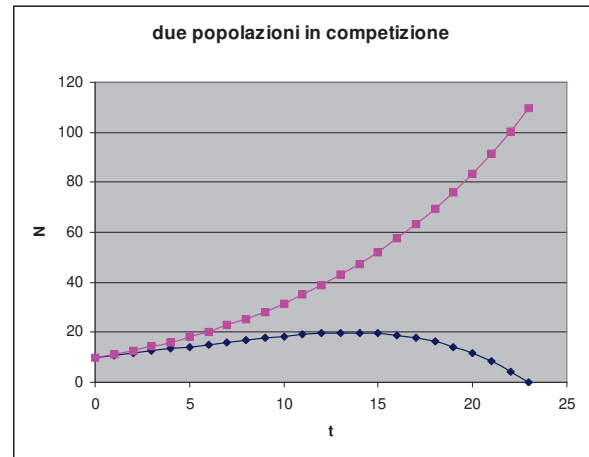
- Due specie che occupano la stessa nicchia competono per le stesse risorse

- $N \equiv N_1 + N_2$

- $N_1(t+1) = N_1(t)[1+k_1] - \beta N(t)^2$

- $N_2(t+1) = N_2(t)[1+k_2] - \beta N(t)^2$

- the survival of the fittest

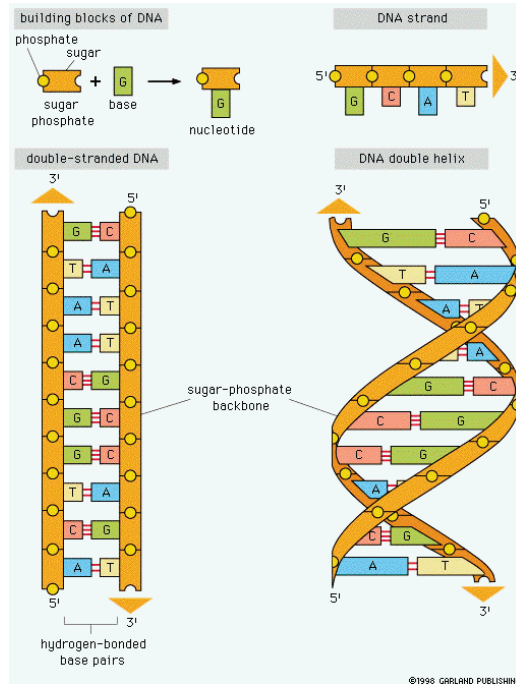


Fase iniziale

Le scoperte della genetica

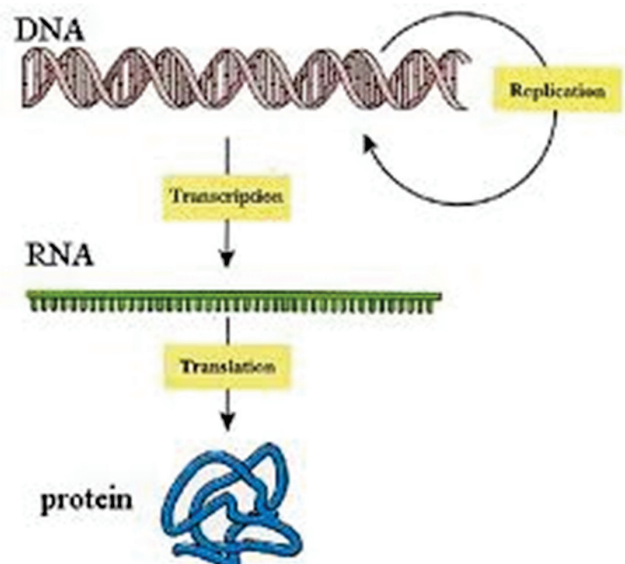
- Darwin aveva intuito le caratteristiche dell'evoluzione, senza conoscerne però il meccanismo
- Mendel scoprì che alcuni caratteri si trasmettono in maniera discreta
 - geni
 - diploidismo: forme dominanti e recessive
- il DNA trasmette l'informazione genetica (Chargaff)
- la doppia elica (Watson e Crick)

1953, Watson e Crick

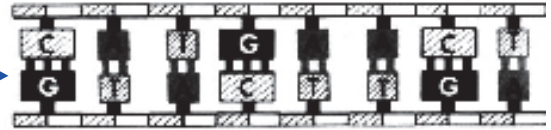


Il dogma centrale

- Il DNA è costituito da una sequenza di nucleotidi, ognuno dei quali contiene una base azotata
 - A, C, G o T
- Il DNA contiene l'informazione necessaria per sintetizzare l'm-RNA e quindi le proteine



Codifica

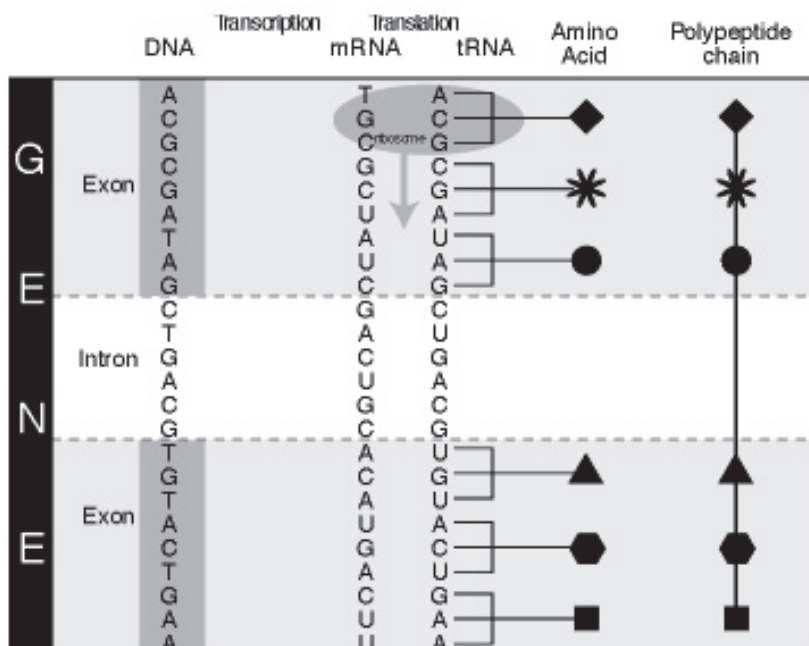


G A T T A C A G G T T A

Prima rappresentazione grafica della doppia elica del DNA. Il disegno, realizzato da Odile Crick, la moglie dello scienziato Francis Crick, è stato pubblicato per la prima volta sulla rivista Nature il 25 Aprile 1953.

□ GATTACA...

Codifica



□ **Codice genetico: traduzione delle triplette in 21 aminoacidi**

□ più codici particolari: parti, fermati, ...

□ **Ogni gene codifica per una proteina**

□ ogni proteina è composta da poche decine a migliaia di aminoacidi

Le proteine

- Svolgono un ruolo strutturale
- Controllano le attività chimiche nella cellula
 - agendo come catalizzatori
- Sono composte da una sequenza di aminoacidi
 - appartenenti a 20 tipi diversi

Le proteine

- Ogni successione di tre nucleotidi corrisponde a un particolare aminoacido
 - codice ridondante
 - stop codons
- Il dogma centrale: un gene, una proteina
 - flusso unidirezionale di informazione
 - limitate eccezioni: retrovirus
- La semplice relazione lineare geni-proteine vale nei batteri
 - negli eucarioti, splicing alternativi, regolazione post-trascrizionale...

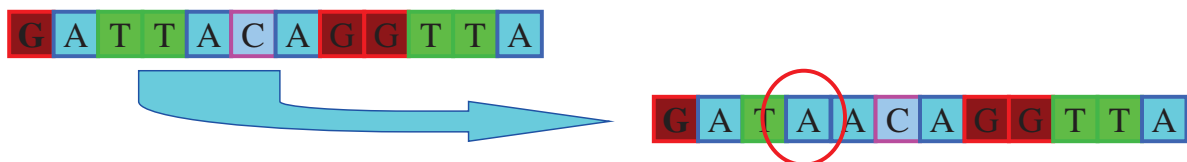
La sintesi neodarwiniana classica

- Meccanismi casuali determinano la modifica dei geni
 - mutazioni puntuali, inserzioni, delezioni

Fonti naturali di variabilità

□ **Mutazione puntiforme**

- **modificazione casuale e puntuale nella sequenza nucleotidica di un gene**



□ **Duplicazione dei geni**

- **e successiva deriva genica (divergenza delle copie)**



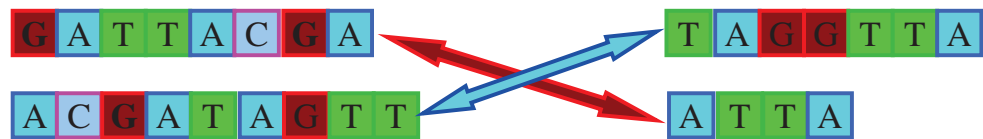
Fonti naturali di variabilità

□ Ricombinazione cromosomica (crossing-over)

1



2



3



La sintesi neodarwiniana classica

- Meccanismi casuali determinano la modifica dei geni
 - mutazioni puntuali, inserzioni, delezioni
 - molti cambiamenti sono dannosi, alcuni sono utili e vengono selezionati dall'ambiente
- Prevale l'ipotesi del gradualismo
 - cambiamenti gradualisti
 - l'assenza di alcuni anelli intermedi è dovuta alla distruzione dei fossili

Alcune ulteriori acquisizioni

- Equilibri punteggiati
 - L'assenza dei fossili è attribuita alla velocità delle transizioni
- Evoluzione neutrale
 - molte mutazioni osservate sono irrilevanti
- DNA non codificante
 - una sofisticata struttura di controllo ?
- Fluidità del genoma
 - trasposoni

Evoluzione e apprendimento

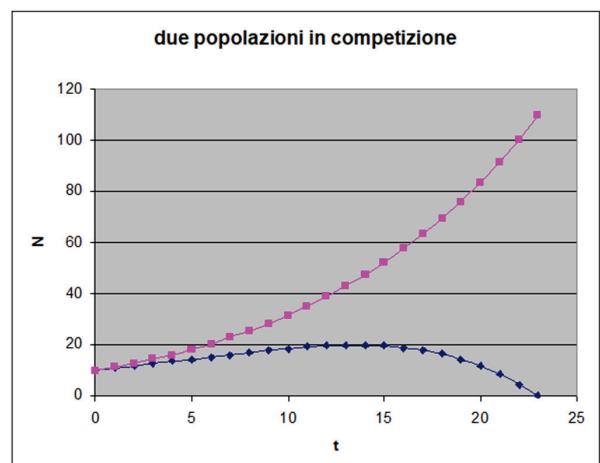
- Evoluzione e apprendimento sono molto simili
- Nel corso dell'evoluzione una specie apprende implicitamente le caratteristiche stabili dell'ambiente in cui vive (ambiente fisico e biologico)
- L'apprendimento può essere visto anche come un processo di generazione di diversi comportamenti alternativi e di selezione di quelli che si rivelano maggiormente adatti

Modelli dell'evoluzione

- Finora, considerazioni qualitative
- Modelli quantitativi possono aiutarci a capire se e come i meccanismi di mutazione/selezione possono funzionare
 - e sotto quali condizioni
- I modelli richiedono l'introduzione di notevoli semplificazioni
 - e propongono una visione astratta dell'evoluzione

La fitness

- La fitness biologica di un individuo è definita come il numero di discendenti fertili
- Fra gruppi di individui che popolano la stessa nicchia, un vantaggio in termini di fitness causa la colonizzazione della popolazione
 - farfalle bianche e grigie



Una visione astratta dell'evoluzione di una specie

In biologia esistono:

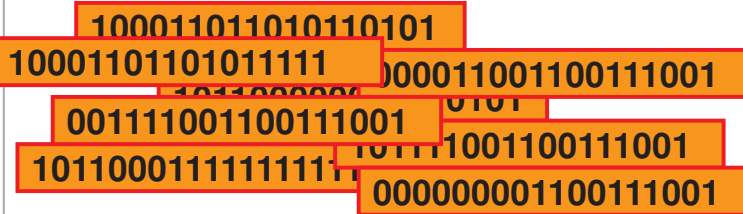
- ❑ Una popolazione composta da individui differenti
- ❑ Un sistema di valutazione della fitness di ogni individuo
 - ❑ è l'ambiente stesso, composto da fattori naturali e umani, altre specie, membri della stessa specie
- ❑ Un meccanismo per generare nuovi individui a partire da quelli a fitness più elevata
 - ❑ in maniera stocastica
- ❑ Un meccanismo per introdurre novità
 - ❑ figli diversi dai genitori
 - ❑ ma mediamente più simili ai genitori che ad individui scelti a caso

Algoritmi genetici

Nell'evoluzione artificiale utilizziamo:

- ❑ Una popolazione composta da «individui» differenti
- ❑ Un sistema di valutazione della fitness di ogni individuo
 - ❑ che dipende dall'obiettivo che ci siamo prefissati
- ❑ Un meccanismo per generare nuovi individui a partire da quelli a fitness più elevata
 - ❑ in maniera stocastica
- ❑ Un meccanismo per introdurre novità
 - ❑ figli diversi dai genitori
 - ❑ Ma mediamente più simili ai genitori che ad individui scelti a caso

Algoritmi genetici



100011011010110101
10001101101011111
000011001100111001
001111001100111001
10110001111111111
000000001100111001

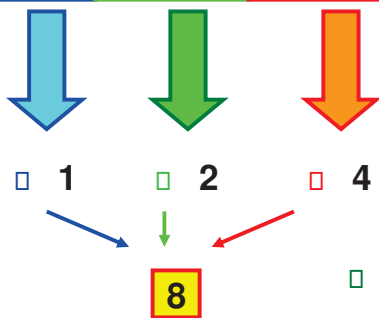
□ Un insieme di soluzioni è detta **popolazione**

□ m stringhe binarie a ciascuna delle quali è associato un valore di fitness



100000001100111001

□ Una particolare sequenza di 0 ed 1 è detta **cromosoma**



□ Ogni cromosoma codifica per un particolare insieme di valori

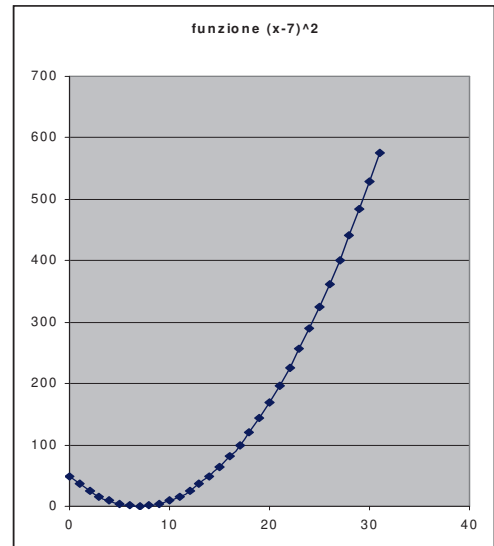
□ Ad ogni insieme di valori è associato un numero, detto **fitness**

L'intuizione di Holland

- Possiamo usare metodi genetici per fare evolvere popolazioni di soluzioni sempre migliori ad un dato problema
- Per far questo è necessario
 - Codificare le soluzioni come "individui"
 - Utilizzare come "fitness" di un individuo una misura della qualità della soluzione che esso rappresenta

Esempio “giocattolo”

- ❑ Il sistema deve cercare il valore minimo di una funzione incognita
- ❑ per x interi compresi nell'intervallo $x \in [0,31]$
- ❑ Noi sappiamo che la funzione è
- ❑ $g(x) = (x-7)^2$
 - ❑ ma il sistema lo ignora!
 - ❑ il problema è banale, ma serve a mostrare come un algoritmo genetico può affrontare un problema di ottimizzazione
- ❑ in questo caso “individuo” coincide con “ipotesi sulla posizione del punto di minimo”



Esempio “giocattolo”

- ❑ $g(x) = (x-7)^2$
- ❑ ogni individuo (soluzione) corrisponde a un particolare valore di x , e può essere rappresentato in forma di numero binario, come un “cromosoma”
 - ❑ usiamo cromosomi binari a 5 bit
 - ❑ la funzione assume valori compresi fra 0 e 576 nell'intervallo considerato
 - ❑ 00111 corrisponde alla soluzione ottimale (numero 7)
- ❑ la popolazione iniziale è generata a caso
- ❑ 10001, 10110, 00101, 11110, 00001, 00011
 - ❑ in cifre arabe
 - ❑ 17 22 5 30 1 3

Esempio “giocattolo”

	■	10001,	10110,	00101,	11110,	00001,	00011
	■	17	22	5	30	1	3
g(x)		100	225	4	529	36	16

□ la fitness $f(x)$ è pari p.es. a $577-g(x)$

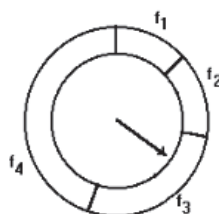
■	477	252	572	48	541	561
	TOTALE 2552					

Scelta dei genitori

- La scelta dei genitori proporzionale alla fitness è quella maggiormente vicina alla metafora biologica
 - ha anche alcuni svantaggi che discuteremo in breve

- E' il metodo di selezione più comune.
- Ad ogni individuo è assegnata una probabilità di essere selezionato

$$p_i = \frac{f_i}{\sum f_i}$$



- Ogni posizione della freccia corrisponde ad un certo numero.
- Si estrae un numero casuale e si seleziona l'individuo puntato dalla freccia.

Esempio “giocattolo”

	■	10001,	10110,	00101,	11110,	00001,	00011
	■	17	22	5	30	1	3
g(x)		100	225	4	529	36	16

- la fitness $f(x)$ è pari p.es. a $577-g(x)$

■	477	252	572	48	541	561
	TOTALE 2552					

- la probabilità di scelta di un individuo come “genitore” è proporzionale alla fitness

- $p(i) = f(i) / \sum_k f(k)$

■	0.19	0.14	0.22	0.02	0.21	0.22
---	------	------	------	------	------	------

La seconda intuizione di Holland: crossover

- i membri della nuova popolazione vengono costruiti a partire da due “genitori”

- due genitori, due figli



- si prendono due genitori diversi e si incrociano

- attorno a un punto scelto a caso



- simula meccanismi biologici di ricombinazione del materiale genico



- la ricombinazione è il motivo per cui la riproduzione sessuata viene preferita in molte specie

La seconda intuizione di Holland: crossover

- supponiamo che la prima coppia scelta sia **00001** (0.21) e 00011 (0.22); incrociando li riotteniamo tali e quali
- supponiamo di avere poi estratto **00101** (0.22) e 00011 (0.22); incrociando dopo la terza posizione otteniamo 000**01** e **001**11
 - il primo è uguale a un individuo già esistente, il secondo è nuovo
- scegliamo infine due genitori **11110** (0.02) e 00101 (0.22), incrociamo dopo la quarta e otteniamo **11111** e 0010**0**
 - entrambi nuovi
- la nuova popolazione è composta da
 - 00001 00011 00001 00111 11111 00100
- vi è un esemplare corrispondente al numero 7, per cui la fitness assume il valore massimo 577

Risolvere i problemi imitando la selezione naturale

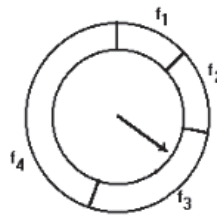


Scelta dei genitori

- La scelta dei genitori proporzionale alla fitness è quella maggiormente vicina alla metafora biologica
 - ha anche alcuni svantaggi che discuteremo in breve

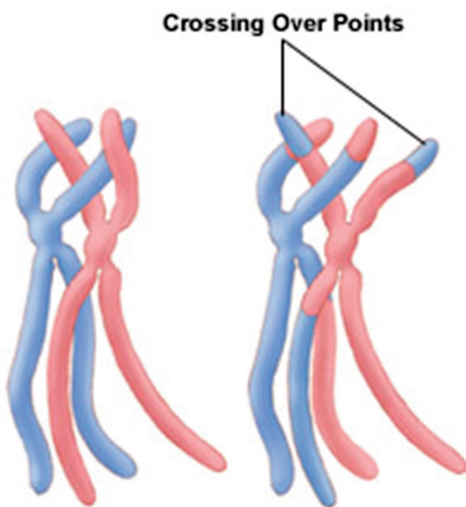
- E' il metodo di selezione più comune.
- Ad ogni individuo è assegnata una probabilità di essere selezionato

$$p_i = \frac{f_i}{\sum f_i}$$



- Ogni posizione della freccia corrisponde ad un certo numero.
- Si estrae un numero casuale e si seleziona l'individuo puntato dalla freccia.

Una buona soluzione...



Crossing Over Points

1010110010101	011100100010
1011001010011	010100101010
1010110010101	010100101010
1011001010011	011100100010
1010110010101	010100101010
1011001010011	011100100010

Una buona soluzione nasce dalla combinazione di parti di soluzione (tale procedimento non si basa sulla bontà delle singole parti)

Il ruolo del crossover

- Il crossover agisce combinando porzioni di soluzioni buone (building blocks)
- per ricombinazione di due individui, si possono generare individui molto migliori di entrambi, individui con fitness simile, o relativamente scadenti
- Esempio particolare: due genitori **11111**(pari a 31) e 00001 (pari a 1), incrociamo dopo la seconda e otteniamo **11**001 (pari a 25) e 00**111** (pari a 7)

La ricombinazione di materiale genetico si rivela molto utile in natura

Il ruolo della mutazione

- Se manca il bit giusto nella popolazione iniziale (o in quella all'istante t), è impossibile crearlo solo col crossover
- ad esempio, una popolazione iniziale del tipo
 - 11000 01000 00101 10100 01101 00001
- non potrà mai generare 00111 per crossover, perché tutti i possibili genitori (e quindi tutti i possibili figli) hanno "0" nella quarta posizione

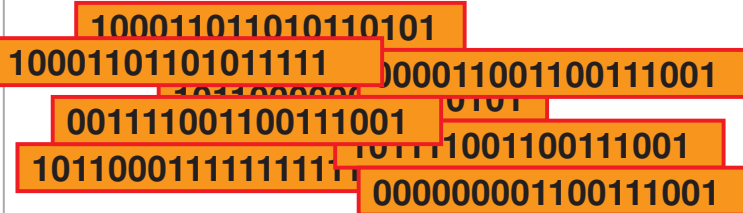
Il ruolo della mutazione

- Per questo si introduce la mutazione puntuale
- Due «metodi»
- con una certa (piccola) probabilità, ogni nuovo individuo generato dal crossover viene mutato in una posizione casuale
- Oppure si genera un figlio per mutazione di un solo genitore della popolazione iniziale

I due operatori genetici principali

- Il crossover consente di esplorare regioni distanti nello spazio dei genomi
 - i figli possono avere una distanza di Hamming elevata dai genitori, se questi sono considerevolmente diversi fra loro
- La mutazione esplora regioni vicine
 - **warning**: esplora regioni in cui la codifica è simile – l'effettiva vicinanza semantica dipende dalla codifica

Riassunto: un GA semplice



100011011010110101
10001101101011111
000011001100111001
001111001100111001
10110001111111111
000000001100111001

□ Un insieme di soluzioni è detta popolazione

□ m stringhe binarie a ciascuna delle quali è associato un valore di fitness



100000001100111001

□ Una particolare sequenza di 0 ed 1 è detta cromosoma

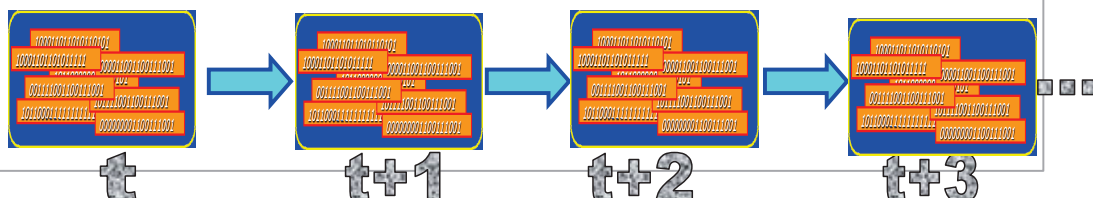
□ Ogni cromosoma codifica per un particolare insieme di valori

□ Ad ogni insieme di valori è associato un numero, detto fitness

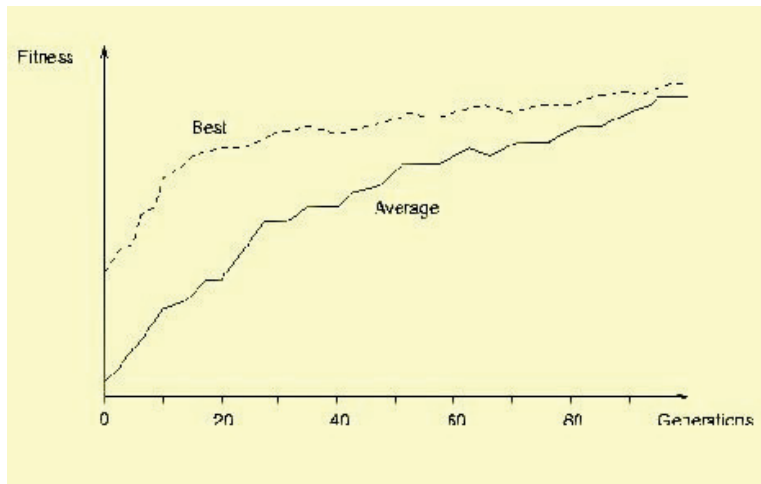
Riassunto: un GA semplice

- 1. $t=0$
- 2. genera una popolazione iniziale P composta da N stringhe binarie di lunghezza L
- 3. finché $t < t_{\max}$
 - valuta la fitness di ogni individuo in P
 - scegli le coppie di genitori in maniera proporzionale alla fitness
 - da ogni coppia di genitori genera una coppia di figli (fino a completare la nuova generazione P') - vedi seguito
 - poni $P' \rightarrow P$
 - incrementa t
 - torna a 3

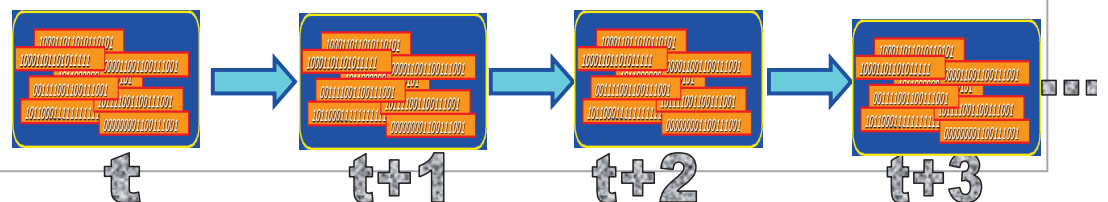
□ end



Riassunto: un GA semplice



- La fitness media cresce con il tempo e tende alla fitness dell'individuo migliore



Generazione della coppia di figli

- Siano G_1 e G_2 i genitori individuati dall'operatore di selezione (membri della popolazione P)
 - e F_1 e F_2 i figli (membri di P')
- per ogni coppia di genitori G_1 e G_2
- applica l'operatore di crossover ottenendo F_1' e F_2'
- con una probabilità p_m applica ad ogni F_k' l'operazione di mutazione puntuale, e ottieni F_1 e F_2
 - se non si applica né crossover né mutazione, i figli sono identici ai genitori

Algoritmo genetico

Una possibile interpretazione del suo funzionamento

Formalizzazione dei building blocks: schemi

- Uno schema è un insieme di individui che hanno una parte in comune
- Può essere definito da una stringa di lunghezza L (pari a quella degli individui) con elementi appartenenti all'alfabeto $\{0,1,\#\}$
 - $\#$ = don't care
- $01\#\#1 = \{01001, 01011, 01101, 01111\}$
- Secondo Holland l'efficacia dei GA è legata al fatto che, ad ogni generazione, la valutazione della fitness di N individui comporta una stima della fitness media associata ad un numero elevato di schemi

Quanti schemi?

- Ad una stringa binaria di lunghezza L corrispondono 3^L schemi
 - in ogni posizione può esserci 1, 0 o #
- ogni individuo appartiene a 2^L schemi diversi
 - quelli che si ottengono assegnando ad ogni posizione il valore che assume in quell'individuo oppure il don't care
 - 11111 appartiene a tutti gli schemi in cui in ogni posizione c'è 1 oppure #
- Esaminando una popolazione di N individui si valutano implicitamente molti schemi (parallelismo implicito)
 - non più di $N \cdot 2^L$
 - molti meno, perché individui diversi possono appartenere al medesimo schema
 - 10000 e 10111 appartengono entrambi a 10####

Come si riproducono gli schemi?

Riproduzione di un singolo individuo

- Sola selezione (proporzionale) senza crossover né mutazioni
 - All'istante t vi è una popolazione P ; all'istante successivo P'
- ogni elemento di P' è generato dall'individuo i di P con probabilità pari a $f_i / \sum_k f_k = f_i / N \langle f \rangle$
- quindi il numero atteso di discendenti di un individuo i è pari a $N f_i / N \langle f \rangle = f_i / \langle f \rangle$

Come si riproducono gli schemi?

- Consideriamo lo schema H composto da $m = m(H, P)$ individui in P
 - Calcoliamo la forza media dello schema: $f(H) = (\sum_k f_k) / m, k \in H$
- Rinumeriamo per comodità gli individui in modo che $H = \{1, 2, \dots, m\}$ (tutti gli individui di H si trovano nella parte iniziale delle popolazione)
- il numero atteso di individui in H alla nuova generazione è
 $\langle m' \rangle = \langle m(H, P') \rangle = (f_1 + f_2 + \dots + f_h) / \langle f \rangle = m f(H) / \langle f \rangle$
Il rapporto rispetto al numero precedente è: $\langle m' \rangle / m = f(H) / \langle f \rangle$
- La proporzione di schemi buoni (schemi con fitness superiore alla media) cresce quindi con la fitness dello schema (in maniera approssimativa esponenziale) !

L'azione del crossover

- Il crossover (classico) può distruggere uno schema buono
- Definiamo la **lunghezza $d(H)$ di uno schema H** come la distanza massima fra due posizioni fissate dello schema
 - $\#1\#1\#\#$ ha lunghezza 2
 - $\#1\#\#\#1$ ha lunghezza 4
 - $0 \leq d(H) \leq L-1$
- La probabilità che il crossover rompa uno schema presente in uno dei due genitori è pari a $d / (L-1)$ - se l'altro genitore non appartiene allo stesso schema

Azione del crossover sugli schemi

- A) 1****
- B) **11*
- C) 1001*
- D) *100*
- E) 1*00*
- F) 1*100
- G) 10101

□ Uno schema è probabile si conservi, se:

- è generale
- è compatto

□ Selezione

- Muta la frequenza degli schemi già presenti
- È efficace nel «rompere» schemi con molti punti fissi

□ Crossover

- Distrugge gli schemi con efficacia dipendente dalla loro lunghezza

Nome	Stringa	F	C
A	1****	4	-
B	**11*	3	1
C	*100*	2	2
D	1*00*	1	3
E	1001*	1	3
F	1*100	-	4
G	10101	-	4

Si favoriscono gli schemi corti a elevata fitness

- Indichiamo con $P(H,t)=m(H,t)/N$ la frazione di individui che appartengono ad H, e con $\phi(H,t)$ il fitness ratio $f(H,t)/\langle f \rangle_t$
- se $d(H)=0$ allora $P(H,t+1) \geq \phi(H,t) P(H,t)$
 - crescita esponenziale degli schemi per $\phi(H,t) \geq 1$
 - finché H non si diffonde al punto di modificare la fitness di una porzione significativa della popolazione, e quindi il proprio fitness ratio
- se $d(H)=L-1$, allora $P(H,t+1) \geq [\phi(H,t) P(H,t)]^2 = \phi(H,t)^2 P(H,t) * P(H,t)$
 - (vedi lucido seguente)
 - condizione sufficiente per avere $P(H,t+1) \geq P(H,t)$ è quindi

$$\phi(H,t)^2 P(H,t) \geq 1$$

- uno schema con fitness doppia della media cresce se è presente in almeno 1/4 della popolazione

Dimostrazione

- Indichiamo con $P(H,t)=m(H,t)/N$ la frazione di individui che appartengono ad H , e con $\phi(H,t)$ il fitness ratio $f(H,t)/\langle f \rangle_t$
- se $d(H)=0$ allora ogni volta che si applica il crossover si ha un figlio in H ; la prob. di selezionare un genitore in H è $\phi(H,t) P(H,t)$ quindi

$$P(H,t+1) = \phi(H,t) P(H,t) + \text{nuovi individui in } H \geq \phi(H,t) P(H,t)$$

- se $d(H)=L-1$, allora solo i figli dell'incrocio di due individui che sono entrambi in H saranno in H

$$P(H,t+1) = [\phi(H,t) P(H,t)]^2 + \text{nuovi individui in } H \geq [\phi(H,t) P(H,t)]^2$$

- La condizione $P(H,t+1) \geq P(H,t)$, escludendo i nuovi nati in H , è quindi $\phi(H,t)^2 P(H,t) \geq 1$

Un possibile problema: convergenza prematura

- Se non vi fosse crossover gli schemi migliori si replicherebbero, ma non ci sarebbe la necessaria creazione di novità
- In generale, si osserva che la popolazione tende a convergere verso una situazione uniforme
 - gli individui si assomigliano tutti
- Questo può portare a fenomeni di “convergenza prematura” in cui una soluzione sub-ottimale (migliore di quelle scoperte fino a un certo punto) colonizza l'intera popolazione
- A questo punto non c'è ulteriore progresso!
 - per controllare il fenomeno è opportuno poter valutare e controllare i tempi di convergenza



Tecniche per rallentare la convergenza



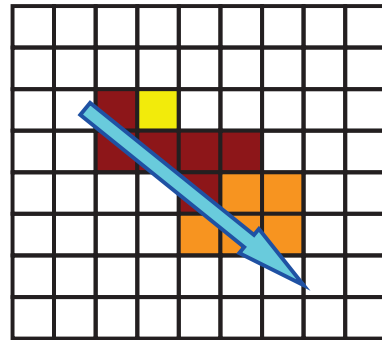
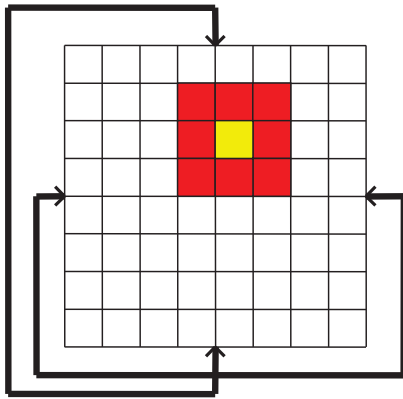
Nicchie

- Una strategia è quella di immaginare che il sistema sia composto da tanti individui collocati in punti diversi nello spazio (celle), e che l'azione abbia luogo all'interno del vicinato di ogni cella
- I figli con fitness elevata possono sostituire individui del vicinato più deboli di loro
- Prima che le «bolle» con individui forti si estendano a tutta la popolazione, le aree distanti hanno il tempo di evolvere in maniera indipendente, rafforzandosi
- In questo modo si consente l'esplorazione locale di alternative che possono essere momentaneamente inferiori rispetto a quelle scoperte altrove - ma che possono fornire il materiale per costruirne di migliori

Nicchie

□ In natura succede che

- I compagni per la riproduzione sono scelti fra quelli appartenenti all'ambiente locale, e proporzionalmente al loro fitness.
- I figli rimangono geograficamente vicino ai propri genitori.
- ambienti favorevoli favoriscono la crescita di individui forti.
- I conflitti sono risolti dipendentemente alla forza relativa dei contendenti.



Nicchie

□ Un genoma con alta fitness, prima di confrontarsi con genomi distanti, deve invadere la zona locale con propri cloni superando molti conflitti

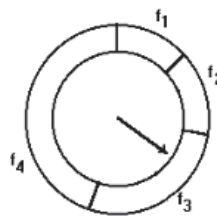
- gli individui migliori necessitano di tempo per «invadere» la popolazione
- questo permette a zone distanti di evolvere in modo (parzialmente) indipendente
- si formano naturalmente delle nicchie ecologiche
- la diversità è preservata a lungo



Scelta dei genitori

- La scelta dei genitori proporzionale alla fitness è quella maggiormente vicina alla metafora biologica, ma ha alcuni svantaggi
 - Possibile scomparsa di individui molto performanti
 - Possibile convergenza prematura se un individuo è molto superiore alla media
- E' il metodo di selezione più comune.
- Ad ogni individuo è assegnata una probabilità di essere selezionato

$$p_i = \frac{f_i}{\sum f_i}$$



- Ogni posizione della freccia corrisponde ad un certo numero.
- Si estrae un numero casuale e si seleziona l'individuo puntato dalla freccia.

Varianti: scelta dei genitori

- La scelta dei genitori proporzionale alla fitness è quella maggiormente vicina alla metafora biologica, ma ha alcuni svantaggi
 - Possibile scomparsa di individui molto performanti
- **Elitismo**: si garantisce la sopravvivenza degli individui con fitness più alta (operatore "copia")
 - serve ad evitare di perdere porzioni di soluzione valide
- **NOTA BENE** L'elitismo non serve per rallentare la convergenza, ma serve a preservare la conoscenza acquisita

Varianti: scelta dei genitori

- La scelta dei genitori proporzionale alla fitness è quella maggiormente vicina alla metafora biologica, ma ha alcuni svantaggi
 - Possibile convergenza prematura se un individuo è molto superiore alla media
- Varianti per contrastare la convergenza prematura
 - **Scaling**
 - **Ranking**
 - **Selezione per torneo**

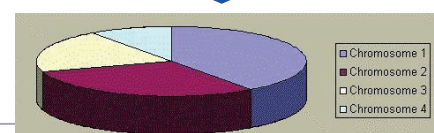
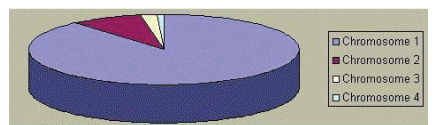
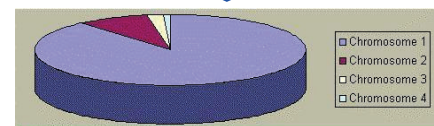
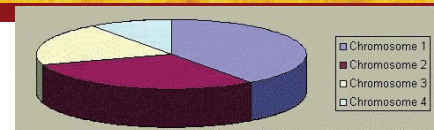
Varianti: scelta dei genitori

■ Scaling

- selezione dipendente in maniera **non lineare dalla fitness**: aumenta o diminuisce la “pressione selettiva” in favore dei più adatti

- la probabilità di essere selezionati non è proporzionale alla fitness, ma ad una funzione di essa

$$p_i = \frac{f_i}{\sum_j f_j} \quad \rightarrow \quad p_i = \frac{F(f_i)}{\sum_j F(f_j)}$$
$$F(x) = x^2 \quad \dots$$



Selezione per rango

Selezione per rango (*rank selection*)

- Si ordinano gli individui per fitness (in modo decrescente).
- Si fissa una distribuzione di punteggi con probabilità decrescente con la posizione occupata, indipendente dai valori di fitness

Vantaggi

- Non si ha convergenza prematura: nessun individuo ha probabilità molto maggiore degli altri di essere selezionato
- Non c'è stagnazione: la distribuzione probabilità non varia.

Svantaggi

Computazionalmente pesante
(convergenza lenta)

Nota: non è biologicamente plausibile.

□ Esempio

□ ranking: probabilità di essere scelti "inversa" rispetto all'ordine di «arrivo»

□ Esempio: punteggio di arrivo nelle gare di formula 1

Scelta per torneo (a s contendenti)

- Nella popolazione di N individui,
 - se ne scelgono a caso s (con **probabilità uniforme**)
 - lo stesso individuo può quindi essere scelto più volte
 - L'**individuo prescelto per la riproduzione** è quello con fitness più alta fra i contendenti
 - Si itera fino a generare N nuovi individui
- Supponiamo vi sia un solo individuo i con fitness superiore a tutti gli altri; alla generazione successiva il numero atteso di copie di i è s
 - Infatti i viene scelto per partecipare in media a s degli N tornei (come tutti gli altri), e li vince tutti
 - al passo successivo, se $s \ll N$, ce ne saranno circa s^2 , etc.
 - ai passi successivi aumenta però la probabilità che due individui superiori partecipino allo stesso torneo

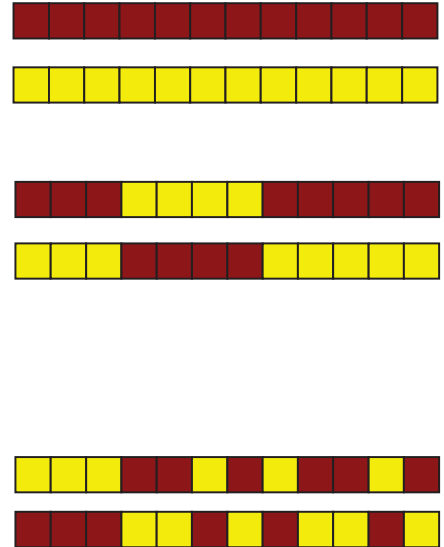
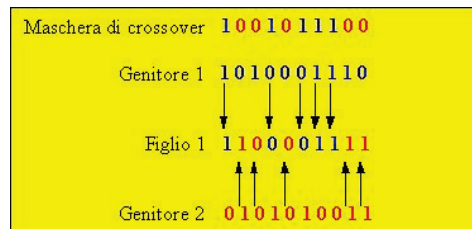
Scelta per torneo (a s contendenti)

- E' possibile studiare il comportamento della selezione per torneo e confrontarlo con quello della selezione proporzionale (dettagli omessi)
 - La selezione proporzionale potrebbe favorire una convergenza rapida qualora uno o due individui fossero molto superiori alla media, mentre con la selezione per torneo il numero atteso di figli è sempre s
- Quindi la selezione per torneo aiuta a prevenire patologie legate alla scoperta accidentale di individui straordinari
 - consente un migliore controllo del progresso della popolazione
 - spesso viene usata al posto di quella proporzionale

Algoritmo genetico
Varianti

Varianti: operatori genetici

- Il crossover tende a distruggere combinazioni di building blocks lontani
- Alternativa di ispirazione biologica: crossover a due punti
- Alternativa artificiale: crossover uniforme



Varianti: codifica

- I GA originali (e molti di quelli attuali) sono codificati in forma binaria
- E' possibile utilizzare anche una **codifica discreta**
 - ogni posizione può essere occupata da un simbolo tratto da un alfabeto a $k > 2$ simboli
- Il crossover resta lo stesso
- L'operatore di **mutazione deve essere modificato**
 - Si sceglie a caso una posizione e si sostituisce il simbolo con uno dei restanti $k-1$ scelto a caso (distribuzione uniforme)

Varianti: codifica

- I GA originali (e molti di quelli attuali) sono codificati in forma binaria
- E' possibile utilizzare anche una **codifica con numeri reali**
 - ogni posizione può essere occupata da un numero reale
- Il crossover può essere realizzato tramite
 - Il classico crossover con punto di taglio (nel caso di taglio separante più geni)
 - Nel caso in cui il taglio coinvolga un unico gene, potrebbe essere utilizzata la media aritmetica dei valori dei due genitori, o la media geometrica
- L'operatore di **mutazione deve essere modificato**
 - Si sceglie a caso una posizione e si aggiunge al valore ivi contenuto un numero casuale (con distribuzione gaussiana a media nulla)

Altre varianti

- Sono state proposte diverse varianti ulteriori ai GA, quali ad esempio
 - selezione dei genitori basata sulla somiglianza
 - rappresentazioni con “regioni codificanti” intervallate da introni
 - rappresentazioni non posizionali
 - ...



Algoritmo genetico

Ottimizzazione



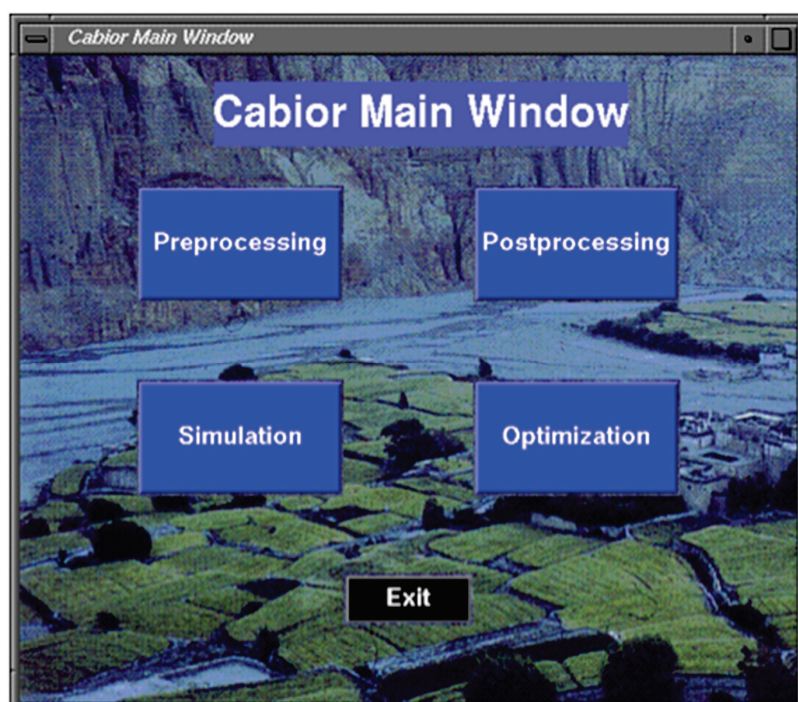
Algoritmi genetici per problemi di ottimizzazione

- I GA possono essere usati per cercare il valore massimo di una data funzione
 - come nell'esempio già visto
- Il **sistema è stocastico**, quindi non è vincolato a cadere in un estremo locale
 - come accadrebbe invece alla discesa secondo gradiente
- **Non vi sono richieste di proprietà particolari** (p.es. continuità, differenziabilità) da imporre a priori alla funzione da ottimizzare

Algoritmi genetici per problemi di ottimizzazione

- ❑ I GA possono essere utilizzati anche quando **non è nota la funzione da ottimizzare**
 - ❑ è sufficiente disporre di un metodo per valutare la fitness di ogni soluzione proposta
- ❑ Esempio: ricerca dei **valori ottimali dei parametri di un modello**
 - ❑ per confronto fra l'andamento del modello e un insieme di dati sperimentali
 - ❑ il cromosoma contiene una codifica dei valori dei diversi parametri
 - ❑ la fitness è legata alla somiglianza fra l'andamento del modello e i dati sperimentali

Algoritmi genetici per il biorisanamento



Algoritmi genetici per problemi di ottimizzazione

- Selezione, crossover, mutazione sono operatori genetici che consentono di affrontare molti problemi difficili di ottimizzazione
 - ottimizzazione non lineare, con diversi massimi locali
 - soluzione ottimale vs. soluzione soddisfacente
- Il metodo è particolarmente efficace quando
 - il crossover non genera troppe soluzioni illegali
 - la struttura dello spazio delle soluzioni è tale per cui le posizioni dei massimi sono correlate (e il crossover può combinare efficacemente porzioni di soluzioni)
 - è sufficiente una buona approssimazione di una buona soluzione

Problemi insolubili: un ago nel pagliaio

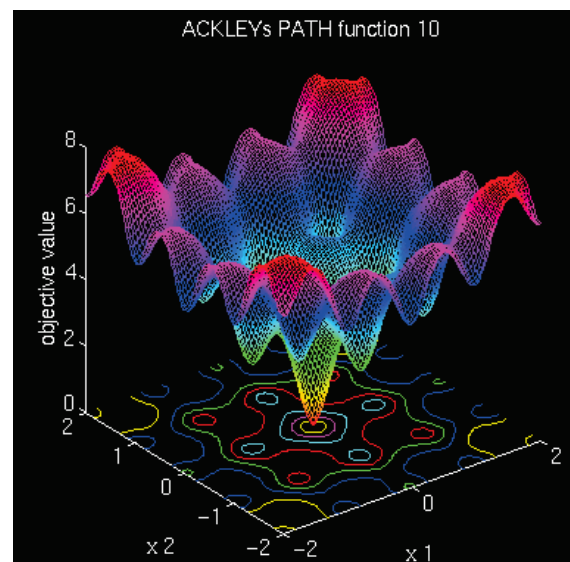
- P.es. trovare il massimo, per x, y interi compresi fra 0 e 1000, di
$$f(x, y) = 1 \text{ solo se } x = 33 \text{ e } y = 701$$
$$f(x, y) = 0 \text{ altrimenti}$$
- La valutazione di diversi punti della funzione non fornisce alcuna informazione sulla localizzazione del punto di massimo
- L'unica tecnica di esplorazione efficace è la ricerca esaustiva
 - che diventa impossibile al crescere della dimensione dello spazio di soluzioni

Problemi solubili

- Il caso più semplice è quello di funzioni monotone, con un solo massimo (relativo ed assoluto)
- Diversi metodi consentono di localizzare il massimo
 - vi sono problemi se vi sono estese regioni con pendenza quasi nulla
 - i GA possono localizzare regioni con fitness elevate, ma la convergenza verso un valore molto preciso non è particolarmente veloce
 - sono utili quando è sufficiente una individuazione approssimata dei valori
 - oppure possono essere usati in combinazione con una tecnica efficace per la ricerca di un valore estremo in una porzione convessa della funzione da ottimizzare

Problemi solubili

- I GA sono particolarmente efficaci quando vi sono diversi massimi locali e quando la posizione dei massimi locali contiene informazioni utilizzabili dall'algoritmo per cercare soluzioni ancora migliori
 - p.es. i massicci nelle alpi





Algoritmo genetico

Conoscenza sul dominio



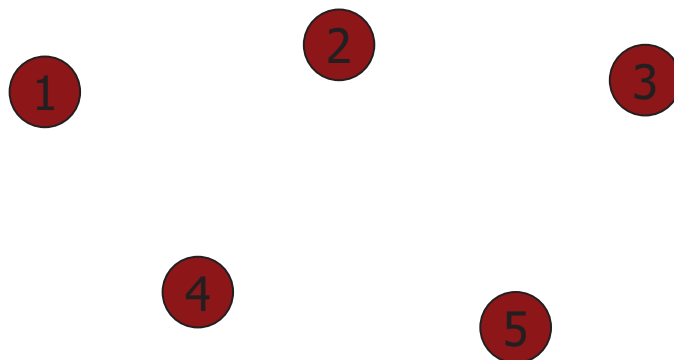
Conoscenza sul dominio

- I GA consentono anche di fare uso delle conoscenze sullo specifico problema di cui si cerca la soluzione ottimale
- Sia nella codifica degli individui
- Che nella scelta degli operatori genetici

Operatori genetici ad hoc

- **Operatori genetici ad hoc**
 - incorporano conoscenze sul dominio
 - consentono di evitare la proliferazione di soluzioni illegali
- Esempio classico è quello del problema del commesso viaggiatore
 - si tratta di visitare L città in modo da minimizzare la lunghezza complessiva del percorso
 - ogni coppia di città è connessa direttamente e la distanza è nota
 - ogni città deve essere visitata una e una sola volta
 - problema di ottimizzazione combinatoria NP hard

Problema del commesso viaggiatore



cromosoma = [prima città visitata, seconda città visitata ..]

esempi [1 2 3 4 5], [3 4 1 2 5] ...

Problema del commesso viaggiatore

- $d(i,j)$ = distanza fra la città i e la città j
- $x = [x_1 \ x_2 \ .. \ x_L]$ $d = d(x_1, x_2) + d(x_2, x_3) + \dots + d(x_{L-1}, x_L)$
- La funzione di **fitness è una funzione decrescente** della lunghezza complessiva del cammino
 - $f = 1/d$, oppure $f = \text{cost} - d$
- Nota: i cromosomi $[12345]$ e $[34512]$ descrivono lo stesso percorso
 - per **evitare valutazioni multiple** si può p.es. convenire che tutti i cammini abbiano la città "1" in prima posizione

Crossover

- Può generare tour illegali
1 2 3 4 5 1 5 3 4 2
- diventano
1 2 3 4 2 1 5 3 4 5
- In uno dei figli la stessa città (2 e 5) è presente più volte, nell'altro manca
- è necessario modificare il crossover per evitare che l'algoritmo sia oberato dal carico di lavoro consistente nell'eliminazione delle soluzioni illegali

Esempio: crossover a mappa parziale

1 2 5 6 4 3

1 3 6 4 5 2

- la parte a destra del punto di taglio è la sezione di matching: 4<->5, 2<->3 definisce la mappa di scambio
- ogni elemento del primo genitore viene verificato e sostituito se appartiene alla mappa di scambio

1 3 4 6 5 2

- idem per il secondo

1 2 6 5 4 3

- genera sempre tour legali
- si può utilizzare anche col crossover a due punti

Algoritmo genetico Applicazioni

Applicazioni

- ❑ Usando diversi tipi di GA e di operatori genetici sono state realizzate applicazioni in diversi settori
- ❑ Problemi classici di ottimizzazione combinatoria
 - ❑ travelling salesman
- ❑ Ottimizzazione di parametri di modelli dinamici
 - ❑ biorisanamento di terreni contaminati
- ❑ Ricerca di topologie ottimali di reti neurali
- ❑ Logistica
 - ❑ istradamento di veicoli entro finestre temporali
- ❑ Scheduling
 - ❑ orari scolastici

Applicazioni

- ❑ **Ottimizzazione di funzioni numeriche**
 - ❑ i GA si sono rivelati essere in grado di superare tecniche convenzionali di ottimizzazioni su funzioni complicate, discontinue e disturbate
- ❑ **Image Processing**
 - ❑ allineamento di immagini
 - ❑ creazione di immagini di sospetti criminali; il testimone agisce come la "funzione fitness" nel GA e controlla la convergenza verso l'immagine corretta
- ❑ **Bin packing**
 - ❑ cioè determinare come disporre un numero di oggetti su uno spazio limitato (circuiti integrati VLSI, od il job shop scheduling dove il problema è allocare un insieme di risorse per portare a termine un insieme di compiti; l'ottima allocazione è quella che permette di finire il lavoro nel minor tempo possibile, o nel minimo tempo di inattività per ogni risorsa.
- ❑ **Machine Learning**
 - ❑ sistemi di apprendimento dove il modello usuale è quello del sistema classificatore
- ❑ **CAD**
 - ❑ Ottimizzazione del sistema di pompe per la lubrificazione di un aereo, progettazione di lame per turbine

Applicazioni finanziarie

- Previsione di serie temporali
 - ottimizzazione di parametri di modelli
- Ottimizzazione di portafoglio
- Trading di divise estere
 - fondamentale
 - tecnico (ottimizzazione di finestre per metodi a media mobile e di soglie)
- Previsione di fallimenti
- Credit scoring
- Induzione di “regole” per il comportamento dei mercati