



UNIVERSITA' DEGLI STUDI DE L'AQUILA

Facoltà Di Scienze mm.ff.nn

Dipartimento di Informatica

Corso di Laurea Specialistica in Informatica

Alberi Filogenetici

Studenti:

Giuseppe Carota Mat. 158787

Graziella Iezzi Mat. 140442

Marco Mezzaluna Mat. 159341

Professore:

Prof. Pasquale Caianiello

Anno Accademico 2004/2005

Alberi filogenetici

La differenza tra le specie è sempre stato un interessante argomento di studio per molti scienziati. Secondo i sostenitori della teoria darwiniana, tutti gli organismi si sono evoluti a partire da microrganismi, pertanto hanno tutti antenati comuni. E da essi si sono evoluti i vari gruppi di organismi. Allora si è introdotto il concetto di albero filogenetico, per indicare uno schema secondo il quale sarebbe avvenuta l'evoluzione.

Il primo albero filogenetico fu disegnato da Ernst Haeckel (1834-1919) nel 1866, mostrata nella Figura 1.

Negli ultimi anni con la scoperta del DNA i biologi hanno avuto una solida base di studio. Confrontando i DNA delle diverse specie presenti nell'ambiente, infatti, si può costruire un albero filogenetico in cui gli apici dei rami rappresentano le specie e i nodi interni ipotetici predecessori incogniti delle specie iniziali.

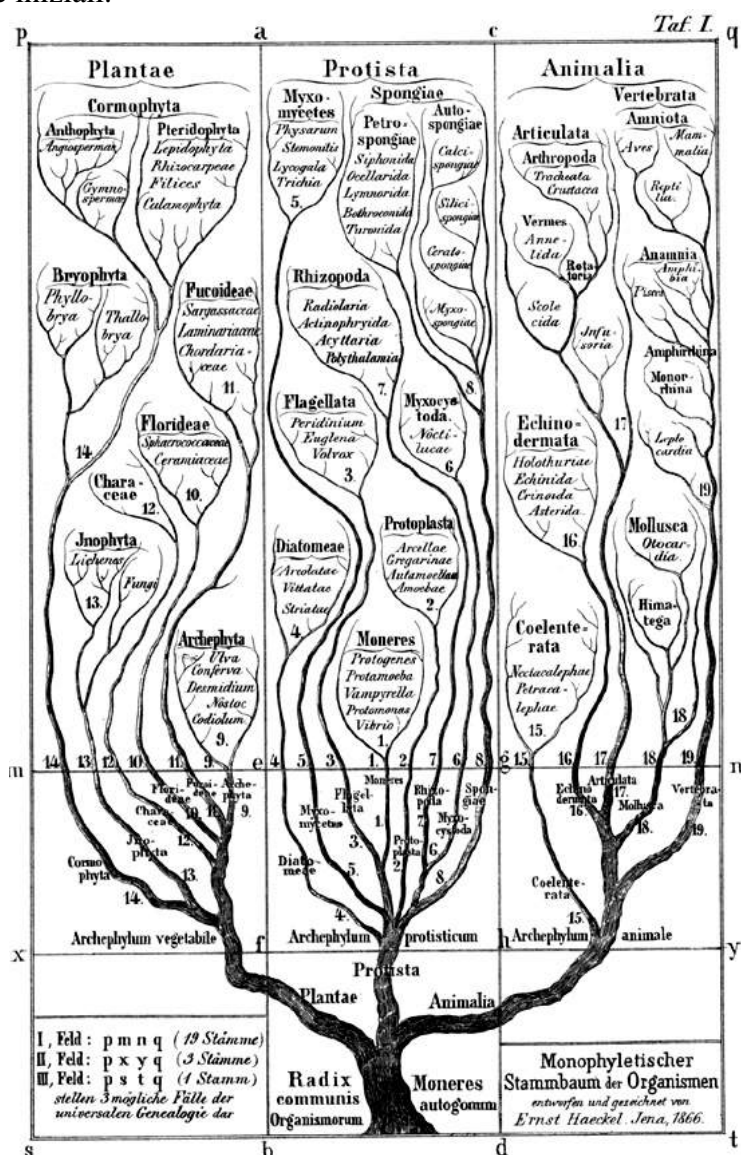


Figura 1

Possiamo utilmente confrontarlo con il moderno albero della vita, presentato nella Figura 2, sia dal punto di vista della struttura quanto da quello del contenuto (l'avanzamento delle conoscenze in poco più di un secolo di biologia è stupefacente!).

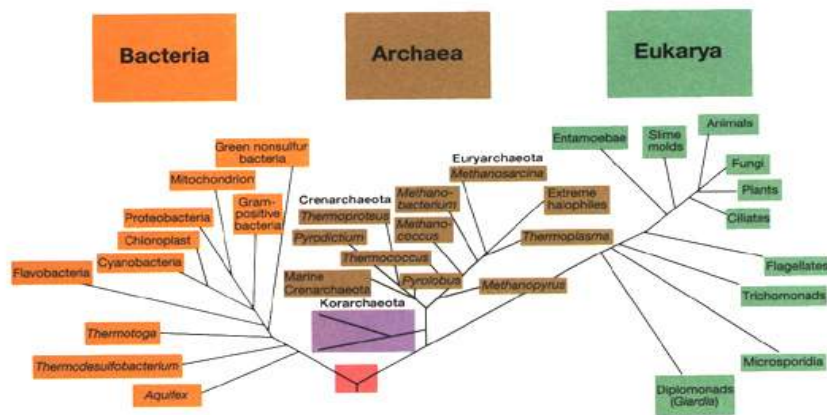


Figura 2

Il *codice genetico* è un sistema per cui le informazioni genetiche codificate nel DNA arrivano a operare la sintesi di tutti i tipi di proteine necessarie alla vita degli organismi. E' una sorta di linguaggio molecolare, basato sull'ordine in cui si susseguono nella molecola di DNA le quattro diverse basi azotate $G = \{U, C, A, G\}$, G rappresenta il nostro alfabeto. Il DNA, quindi, dispone di un alfabeto di quattro lettere per specificare i circa 20 *amminoacidi* (Figura 3) da cui possono essere costituite, secondo un preciso ordine di successione, le *proteine*. Gli amminoacidi si combinano in parole di lunghezza 3, dette triplette .

		2nd base in codon					
		U	C	A	G		
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G	3rd base in codon
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G	
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G	
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G	

Figura 3

Per codificare le proteine la sequenza di DNA viene letta una tripletta alla volta, essa viene tradotta in amminoacido (nella sintesi proteica) finché si arriva a particolari triplette STOP, che cioè non corrispondono ad alcun amminoacido e che pongono termine alla sintesi proteica, così viene formata una proteina.

Ogni individuo viene rappresentato univocamente dal suo DNA. Queste sequenze vengono memorizzate in matrici P di dimensione $n \times m$, dove n rappresenta gli individui e m la cardinalità dei simboli del DNA. Diamo un esempio qui di seguito

1. AGGATGAATGGGCGAACAG...

2. TGCTCGCGGGTAGAAGAAC...
3. TAGATGAATGGTAGAACAA...
4. TGCAGCGTGATAGAACAAC...
5. TGGAGAAATGATAGAACA...
6. TGCACGCGGCATAGAACGA...
7. TGGATAGATGATACCACAA...
8.

Questa matrice viene rappresentata come *albero filogenetico*, per capire come ogni individuo “dista” da un altro. Ogni individuo rappresenta una foglia dell’albero e partendo da esse si aggregano quelle tra di loro più “vicine” formando un albero che mette in relazione ogni individuo con gli altri.

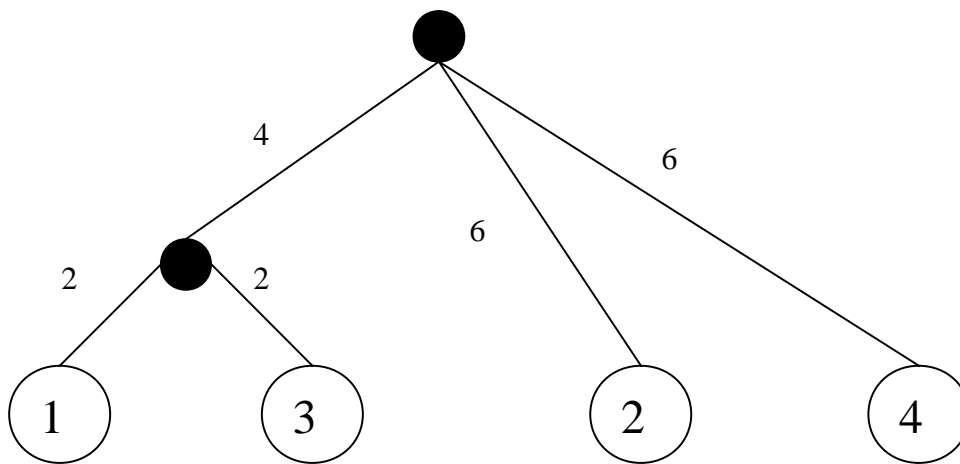
Esempio:

Prendiamo un sottoinsieme della precedente matrice

1. AGGATGAA
2. TGCTCGCG
3. TAGATGAA
4. TGCAGCGT

La distanza tra i vari individui è definita dal numero di caratteri diversi tra individuo e individuo:

- 1 dista 6 da 2, 2 dista 6 da 3, 3 dista 6 da 4.
 2 da 3, 6 da 4.
 6 da 4.



Dall’albero possiamo costruire una grammatica generica che genera una stringa che descrive univocamente ogni albero:

albero -> (seq)
 seq -> [albero , c],seq | [IND , c],seq | [albero , c] | [IND , c]

dove IND è il valore che indica l’individuo,
 c è la distanza dell’individuo o del sotto albero dal resto dell’albero.

Quindi riprendendo l’esempio precedente avremo una stringa di questo genere:

$((([1,2],[3,2]),4),[2,6],[4,6])$

In generale:

$[i,j]$ dove

i nodo,

j arco.

$[(..),j]$ sottoalbero collegato ad un nodo intermedio

$(..)$ albero o sotto albero completo