UNIVERSITÀ DEGLI STUDI DI CATANIA
FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
CORSO DI LAUREA IN FISICA

Roberta Sinatra

ANALISI STATISTICA DEL PROTEOMA E FILOGENESI

TESI DI LAUREA

RELATORE:
CHIAR.MO DOTT. V. LATORA

ANNO ACCADEMICO 2004/2005

Indice

Introduzione

Negli anni ’50 si scopre il DNA: *il codice della vita*. La poeticità dell’espressione trasmette tutta l’importanza di quella che è stata una delle più grandi conquiste scientifiche del XX secolo.

In un primo momento vi era la convinzione secondo la quale si sarebbe conquistata una forte capacità predittiva nell’individuare le caratteristiche biologiche di un organismo, soprattutto dell’uomo, non appena tutti i “messaggi” contenuti nel DNA e nelle proteine fossero stati decifrati. I dati fornitici dalla ricerca smentiscono un’ipotesi di lavoro tanto semplicistica, mentre confermano come caratteristica peculiare della “vita” la convivenza di ordine e casualità. Il DNA, e di conseguenza il sistema proteico che da esso è codificato[1], si presenta infatti come un sistema complesso che richiede, per essere studiato, un’interazione tra discipline diverse quali la biologia, la fisica, la matematica e l’informatica. Nascono in questo modo la biofisica, la bioinformatica e la filogenesi molecolare.

Negli anni ’70 si assiste ad una massiccia attività di laboratorio che procede al sequenziamento di tratti di DNA e alla purificazione di diverse proteine[2]. È per immagazzinare questa immensa mole di dati che tra la fine degli anni ’70 e l’inizio degli anni ’80 nascono i primi database “biologici”[3].

La facile accessibilità alle sequenze nucleotidiche e proteiche ha permesso l’incentivazione e lo studio della filogenesi molecolare. Questa disciplina permette, tramite particolari analisi statistiche e computazionali delle sequenze nucleotidiche e proteiche, di ricostruire l’albero evolutivo delle specie cui le sequenze appartengono. La filogenesi molecolare è anche utiliz-

[1] Una piccola percentuale del DNA, mediante i complessi processi di trascrizione e traduzione, codifica per le proteine. Questi tratti di DNA si chiamano geni. Il processo di traduzione sostituisce ad ogni tripletta di nucleotidi un certo aminoacido (secondo la tabella di codifica riportata in appendice B).

[2] Vedi la voce sequenziamento nel glossario.

[3] Oggi esistono diversi database; la maggior parte di essi cooperano tra di loro mettendo i dati in comune. Il database più frequentato e utilizzato è quello dell’NCBI, disponibile all’indirizzo www.ncbi.nlm.nih.gov.

zata, ad esempio, per ricostruire come si è evoluta una proteina o una famiglia di proteine cui compete una funzione[4] che è rimasta invariata nel corso dell'evoluzione.

In questo lavoro vogliamo presentare i rudimenti della filogenesi molecolare, nonché una particolare analisi statistica che, applicata ai proteomi, permette l'introduzione di un nuovo concetto di distanza per la creazione degli alberi filogenetici.

[4]Negli organismi viventi ogni proteina svolge una particolare funzione biologica. Alcune funzioni sono essenziali per qualunque organismo vivente, per cui ritroviamo anche in organismi molto diversi alcune similarità a livello di sequenze nucleotidiche e aminoacidiche.

Capitolo 1

Filogenesi molecolare

1.1 Il processo evolutivo

Gli “errori” nella tramissione genetica sono alla base dei processi evolutivi che, a partire da una forma primitiva hanno prodotto nel tempo l’enorme diversità delle forme di vita attuali, pur partendo da un unico progenitore comune: la radice dell’albero della vita (Figura 2.1). La trasmissione dell’informazione genetica si ottiene attraverso il processo di replicazione del DNA. Anche se l’apparato di replicazione è molto accurato è possibile che, sebbene con una probabilità molto piccola, si verifichino degli errori, ovvero mutazioni della sequenza di DNA che possono poi essere eventualmente “fissati” in tutta la popolazione degli individui di quella specie o in una larga frazione di essa. Oltre alla sostituzione di un nucleotide con un altro, lungo la sequenza di DNA, possono intervenire altri cambiamenti dovuti all’inserzione o alla delezione di tratti più o meno lunghi di DNA, oppure a riarrangiamenti di vario tipo. Questo spiega perché gli organismi viventi, pur discendendo da un unico progenitore comune, posseggono genomi di dimensioni molto diverse tra loro, da alcuni milioni di nucleotidi nei batteri a circa tre miliardi nell’uomo [1].

1.2 Mutazioni

Le mutazioni subìte dal DNA si riflettono inevitabilmente nei suoi prodotti: nel transcrittoma e nel proteoma. Le sostituzioni di nucleotidi, per motivi chimici, non sono equiprobabili. Per esempio, è più alta la probabilità che si verifichi una transizione, cioé la sostituzione di una purina con una purina e di una pirimidina con un’altra pirimidina, piuttosto che una trasver-

sione, cioé la sostituzione di una purina con una pirimidina e viceversa[1]. Tuttavia non tutte le mutazioni incidono nella stessa maniera sul processo evolutivo. Esistono infatti:

- mutazioni vantaggiose;
- mutazioni svantaggiose;
- mutazioni neutrali.

La selezione naturale favorisce le prime, contrasta le seconde e non ha alcuna influenza sulle ultime. Nel caso infatti delle mutazioni neutrali vi è sì la sostituzione di un nucleotide, ma questa non porta ad un cambiamento in termine di composizione aminoacidica nella codifica della proteina[2]. Per questo motivo in alcuni studi filogenetici si preferisce effettuare dei confronti tra sequenze proteiche piuttosto che genomiche.

1.3 Geni ortologhi e paraloghi

Sulla base di cosa si costruisce un albero filogenetico? Normalmente si operano dei confronti su *geni omologhi*. Due geni o due proteine si dicono omologhi/e se derivano da un progenitore comune. Alla luce di questa definizione è evidente che l'omologia non coincide con la similarità, che si ha quando due sequenze hanno molti siti in comune. Due geni o due proteine possono essere omologhe, ma poco simili. Quasi sempre invece due proteine simili sono anche omologhe. In quei rari casi in cui non lo sono si parla di *convergenza evolutiva*. Ci sono due diversi tipi di omologia. Due sequenze omologhe si definiscono *ortologhe* se appartengono a due specie diverse e il loro processo di divergenza ha avuto origine in seguito al processo di speciazione da cui le due specie suddette hanno avuto origine. In tal caso la sequenza originale da cui le due sequenze derivano era presente nel più recente progenitore delle due specie. Due sequenze si dicono *paraloghe* se il loro processo di divergenza ha avuto origine in seguito ad un processo di duplicazione genica. Solo nel primo caso l'evoluzione dei geni segue l'evoluzione degli organismi e la filogenesi delle sequenze dovrebbe riprodurre quella degli organismi da cui queste derivano. Si assume che i prodotti di geni ortologhi conservino la stessa funzione, mentre quelli di geni paraloghi spesso si specializzano in funzioni differenti.

[1] Sono purine l'adenina (indicata con A) e la guanina (indicata con G). Sono pirimidine la citosina (indicata con C) e la timina (indicata con T).

[2] Ricordiamo che non c'è una corrispondenza biunivoca tra triplette e aminoacidi. Il numero totale di possibili triplette è $4^3 = 64$, mentre gli aminoacidi sono in tutto 20. Ciò implica che un aminoacido può essere codificato da più di una tripletta (vedi appendice B)

1.4 Gli alberi filogenetici

Le relazioni evolutive tra gli organismi, o più in generale tra geni omologhi, possono essere modellizzate mediante *alberi filogenetici*. Un albero filogenetico è un grafico costituito da nodi, rami e foglie. Le foglie (nodi esterni) sono etichettate con le specie o le sequenze note che si vogliono confrontare; i nodi interni rappresentano ipotetici predecessori incogniti degli oggetti iniziali. I rami definiscono le relazioni in termini di discendenza evolutiva. Da ogni nodo si dipartono sempre tre rami: due discenti ed uno ascendente verso il nodo progenitore. Nella maggior parte dei casi non si hanno alberi *politomici*, ovvero alberi che abbiano in un nodo più di due rami discendenti e in tal caso l'albero si dice *completamente risolto*.

Se un albero filogenetico descrive esclusivamente le relazioni filogenetiche tra i vari nodi e la lunghezza dei diversi rami non ha alcun significato, si parla di *cladogramma*. Se invece la lunghezza dei rami è proporzionale alla distanza evolutiva tra i nodi, l'albero è detto *filogramma*. Gli alberi si classificano anche in *rooted* e *unrooted* (con o senza radice). Un albero rooted possiede un nodo particolare, la radice appunto, che rappresenta il comune progenitore di tutti i nodi rappresentati nell'albero (vedi figura 1.2). In questo caso i rami dell'albero sono orientati in funzione del tempo. Un albero unrooted descrive esclusivamente le relazioni evolutive tra le unità tassonomiche senza fornire alcuna informazione circa il processo evolutivo in funzione del tempo (vedi figura 1.2). In altre parole, sappiamo soltanto quanto una specie è lontana da un'altra in termini di evoluzione. Solitamente la forma rooted di un albero filogenetico viene utilizzata solo se si assume la validità dell'ipotesi di orologio molecolare[3]. Quando vi è una diversa velocità di evoluzione tra le specie, viene determinato soltanto l'albero unrooted.

1.5 Determinazione delle distanze genetiche tra sequenze e matrici delle distanze

Una categoria di metodi per la costruzione di alberi si basa sull'osservazione che gli alberi stessi possono essere rappresentati dalle distanze. Tali metodi sono detti *metodi-distanza* e cercano di convertire la distanza tra due sequenze in alberi filogenetici.

Alla base di tali metodi c'è la supposizione che, secondo un qualche criterio biologico, si sia associata ad un insieme di sequenze $(S_1, S_2, ...S_N)$ una distanza $d(S_i, S_j) = d_{ij}$ tale che:

[3]Si parla di orologio molecolare quando si ha proporzionalità diretta tra numero di sostituzioni nucleotidiche o aminoacidiche che si accumulano tra geni o proteine omologhe e tempo intercorso per la loro divergenza.

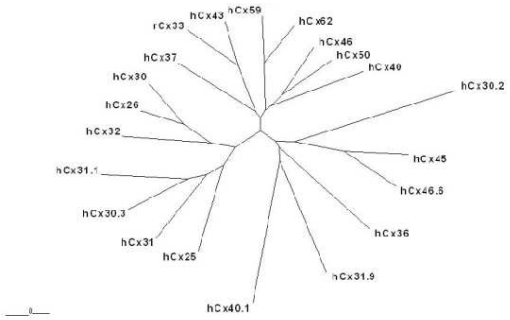

Figura 1.1: Esempio di albero unrooted. L'albero rappresenta le relazioni filogenetiche tra le connessine umane [2].

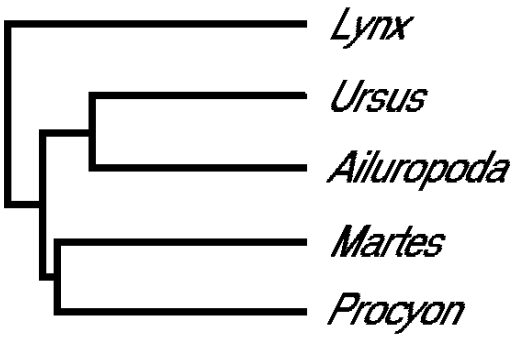

Figura 1.2: Esempio di albero rooted [2].

- $d_{ij} \geq 0 \; \forall i, j$
- $d_{ij} = 0 \iff i = j$
- $d_{ij} = d_{ji}$

La matrice D i cui coefficienti sono le d_{ij} è detta matrice distanza. La matrice è quadrata, simmetrica e di dimensione N (numero di sequenze del set) [4]. A questo punto ci serve introdurre e definire il concetto di distanza genetica tra sequenze. Solitamente la distanza genetica tra sequenze può essere definita solo se si effettua un allineamento delle sequenze. Come abbiamo già visto parlando di mutazioni, le sequenze posso subire inserzioni o delezioni di uno o più siti nucleotidici o aminoacidici. Ciò porta ad una differenza di lunghezza tra le sequenze che comporta dei problemi nella definizione e nell'implementazione del concetto di distanza. Le sequenze vanno allora allineate[4] e mediante l'inserzione di spazi vuoti (normalmente rappresentati con il simbolo "-") riportate ad avere la stessa lunghezza. In questo modo è possibile parlare di distanza genetica tra due sequenze, aminoacidiche o nucletodiche: essa è definita come il numero di sostituzioni necessarie per poter "sovrapporre" una sequenza sull'altra. La distanza viene poi normalizzata rispetto alla lunghezza delle sequenze e pertanto l'unità di misura più naturale da utilizzare è data dal numero di sostituzioni per sito.

La distanza genetica così definita non è tuttavia coincidente con la reale distanza evolutiva. Infatti a causa della possibilità di sostituzioni multiple sullo stesso sito (multiple hits), di sostituzioni convergenti e di retromutazioni, il numero di sostituzioni che viene osservato tra una coppia di sequenze è inferiore rispetto al numero di sostituzioni che effettivamente hanno avuto luogo. Per questo motivo, al fine di ricostruire il giusto processo evolutivo stimando l'effettivo numero di sostituzioni avvenute, si ricorre a metodi di natura stocastica.

I modelli matematici utilizzati mirano a costruire una matrice delle probabilità delle sostituzioni basandosi su alcune assunzioni aprioristiche derivanti da considerazioni di tipo biologico. La maggior parte di questi modelli è stata sviluppata per valutare le sostituzioni nucleotidiche. In questo caso si ha a che fare, infatti, con matrici 4x4[5] più semplici da definire e gestire. I modelli più noti per sostituzioni nucleofile sono quelli di Jukes & Cantor, kimura, Tamura, Lanave e Saccone. Le matrici di sostituzione utilizzate per le sequenze aminoacidiche sono le

[4] Gli algoritmi di allineamento costituiscono uno dei più importanti campi di applicazione della bioinformatica. Quando si hanno più di due sequenze si parla di multiallineamento. Essendo impossibile trovare un allineamento esatto tra più sequenze a causa delle difficoltà computazionali, si è costretti a ricorrere ad algoritmi dinamici ed euristici. Uno dei primi algoritmi per il multiallineamento è stato il *pairwise alingment*.

[5] La matrice è di dimensione 4 perché ogni nucleotide può essere sostituito con uno degli altri tre.

PAM e le BLOSUM, costruite in base ad osservazioni sperimentali di famiglie di proteine la cui evoluzione è nota[6]. Per calcolare la distanza genetica tra sequenze aminoacidiche si può anche usare la formula proposta da Kimura:

$$d_{ij} = -\ln(1 - p - 0.2p^2) \tag{1.1}$$

dove p è la frazione di aminoacidi diversi tra la sequenza i e la sequenza j. È stato empiricamente dimostrato che tale formula fornisce una buona approssiamzione per sequenze non troppo divergenti ($p \leq 0.7$). Un requisito fondamentale per l'applicabilità dei modelli stocastici è la stazionarietà della composizione in basi o in aminoacidi. Tale condizione implica che le sequenze considerate nell'analisi devono avere, entro le normali fluttuazioni statistiche, la stessa composizione in basi o aminoacidi. È necessario effettuare la verifica della stazionarietà prima di applicare qualunque modello per l'analisi evolutiva delle sequenze in quanto l'inclusione nell'analisi di sequenze composizionalmente divergenti porterebbe a risultati non accurati. È possbile introdurre un ulteriore parametro, che può essere utilizzato per modellare il processo di evoluzione molecolare, e che serve a tener conto che i siti considerati non sono ugualmente variabili. Per esempio, nelle sequenze aminoacidiche alcuni residui importanti per l'attività funzionale della proteina sono del tutto invarianti o mostrano una limitatissima variabilità, mentre altri siti funzionalmente meno rilevanti mostrano una variabiltà più o meno marcata.

1.6 Metodi per costruire gli alberi filogenetici

I metodi che utilizzano la distanza genetica per costruire l'albero filogenetico di un set di unità tassonomiche sfruttano particolari algoritmi, detti di *clustering*. Esiste anche un alto tipo di metodi: sono quelli di ottimizzazione, che si basano sulla massimizzazione di una funzione obiettivo di qualità dell'albero. Sono metodi di clustering l'*UPGMA* e il *Neighbor-Joining*; sono metodi di ottimizzazione il metodo della *massima parsimonia* e il metodo della *massima verosimiglianza*.

[6] vedi appendice A per dettagli sulle matrici PAM e BLOSUM.

1.6.1 UPGMA

È uno dei più semplici algoritmi di clustering esistenti. UPGMA sta per Unweighted Pair Group Method with Arithmetic mean. Il metodo utilizza un algoritmo di clusterizzazione iterativo che procede associando via via le sequenze o cluster di sequenze più simili tra loro. Data la matrice delle distanze *pairwise* di un set di sequenze, vogliamo rappresentare in forma di filogramma il suo contenuto di informazione. L'algoritmo è di tipo iterativo e le operazioni da svolgere in ogni ciclo sono le seguenti:

1. identificare la minima distanza tra tutte le possibili coppie di sequenze. Supponiamo di avere quattro seguenze A, B, C e D. Calcoliamo le distanze tra ogni possibile coppia, e sia minima la distanza tra le unità tassonomiche B e C (d_{BC});

	A	**B**	**C**
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

2. questa coppia di unità tassonomiche formerà un cluster;

3. calcolare la nuova matrice delle distanze, defininendo come distanza dal cluster la media delle distanze da ciascuna sequenza del cluster. Proseguendo con il nostro esempio:

	A	**BC**
BC	$d_{(A,BC)}$	
D	d_{AD}	$d_{(BC,D)}$

   dove $d_{(A,BC)} = \frac{(d_{AB}+d_{AC})}{2}$ e $d_{(BC,D)} = \frac{(d_{BD}+d_{CD})}{2}$;

4. si ricomincia l'iterazione cercando nuovamente la minima distanza.

In tutto le iterazioni sono $N-1$, dove N è il numero di sequenze alle quali si sta applicando il metodo.
L'UPGMA presuppone la validità dell'ipotesi dell'orologio molecolare, ovvero la costanza nel tempo di numero di sostituzioni per sito.

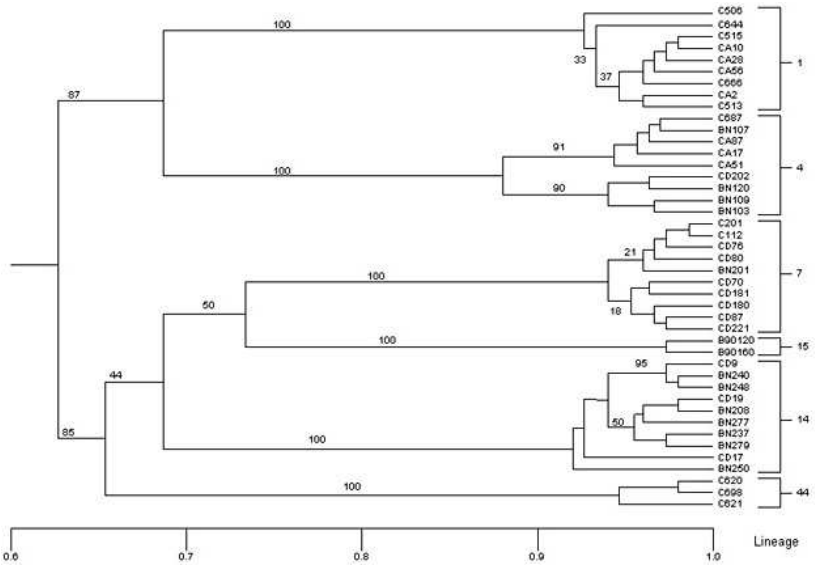

Figura 1.3: Esempio di fenogramma che può essere costruito con il metodo UPGMA. In basso è indicato in che verso scorre il tempo [3].

1.6.2 Neighbor-joining

L'idea di base di questo algoritmo è quella di confrontare tra loro coppie di unità tassonomiche e di costruire dei clusters che però non siano contenuti l'uno nell'altro, ma che anzi siano "separati" dal resto. Per far ciò i clusters costruiti dal Neighbor-Joining sono formati dai cosiddetti vicini (neighbors) che sono definiti come due foglie che sono in relazione tra loro più che con tutte le altre e che per questo sono connesse attraverso un solo nodo all'albero. Lo scopo dell'algoritmo è quello di minimizzare la lunghezza di ogni ramo, basato sul *minimum evolution criterion*[7].

Anche questo algoritmo è iterativo ($N-3$ steps, dove N è, come sempre, il numero di sequenze utilizzate nell'analisi). Ad ogni ciclo le operazioni da svolgere sono:

1. per ogni sequenza i calcolare $r_i = \sum_{k \neq i} d_{ik}$;
2. scegliere la coppia di sequenze (i, j) per cui la quantità $d_{ij} - \frac{r_i + r_j}{N-2}$ è minima;
3. unire i e j ad un cluster (i, j) mediante un nodo dell'albero e calcolare la lunghezza dei

[7] Il criterio vuole che siano più vicine evolutivamente specie che differiscono meno

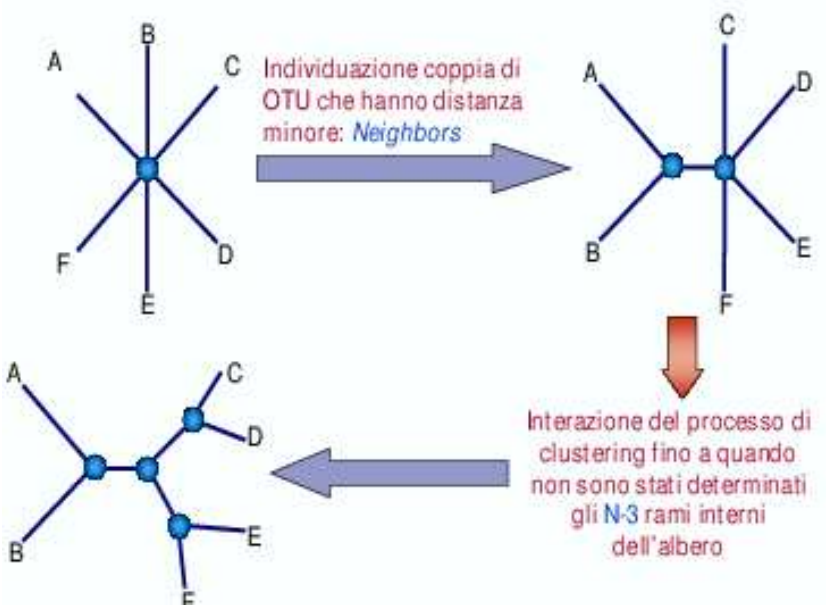

Figura 1.4: Illustrazione dell'algoritmo del Neighbor-Joining. A partire da un albero completamente non risolto (*star phylogeny*) vengono via via determinati, con $N-3$ steps, tutti i rami interni dell'albero in modo che la lunghezza complessiva di tutti i rami sia la più piccola possibile [5].

rami che collegano i e j a questo nuovo nodo (i,j) nel modo seguente:

$$d_{i,(ij)} = \frac{1}{2}d_{ij} + \frac{(r_i - r_j)}{2(N-2)} \quad d_{j,(ij)} = \frac{1}{2}d_{ij} + \frac{(r_j - r_i)}{2(N-2)};$$

4. calcolare così la distanza del nuovo cluster da ogni altro cluster k:

$$d_{(ij),k} = \frac{d_{ik} + d_{jk} - d_{ij}}{2};$$

5. così come fatto per il metodo UPGMA, sostituire i e j con il cluster (ij).

1.6.3 Metodo della massima parsimonia

Una delle maggiori limitazioni dei metodi basati sulla matrice delle distanze è che sintetizzare l'informazione filogenetica presente nelle sequenza multiallineate in distanze tra coppie di sequenze comporta una consistente perdita di informazione. Per evitare di perdere questa

informazione, l'analisi filogenetica può essere operata sulle stesse sequenze piuttosto che sulle distanze genetiche. Il metodo della massima parsimonia utilizza questo tipo di approccio.

Il metodo non assume esplicitamente alcun modello di evoluzione molecolare ed è un metodo essenzialmente qualitativo in quanto consente di determinare la topologia dell'albero che descrive le relazioni filognetiche tra le sequenze in esame. La lunghezza di ciascun ramo dell'albero, pari al numero minimo di sostituzione occorse tra i nodi che esso congiunge, non stima in modo accurato l'effettiva distanza genetica in quanto non tiene conto della possibilità di sostituzioni multiple o convergenti.

Le operazioni da compiere per applicare il metodo sono le seguenti:

1. selezionare i siti informativi. Un sito è informativo se favorisce uno o più alberi tra tutti i possibili e se contiene almeno due differenti caratteri, ciascuno dei quali è presente almeno in due sequenze;

2. una volta selezionati i siti informativi, si calcola il numero minimo di sostituzioni richiesto da ciascun dei possibili alberi senza radice che descrivono le relazioni filogenetiche tra le unità tassonomiche in esame;

3. l'albero (o gli alberi) di massima parsimonia è quello che richiede il numero minimo di sostituzioni totalizzate fra tutti i siti informativi considerati.

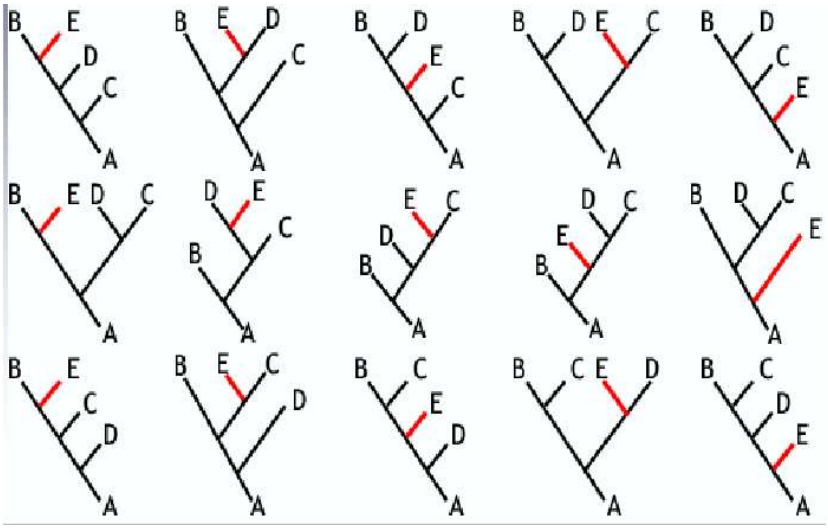

Figura 1.5: Alcuni dei possibili alberi ottenibili con cinque sequenze A, B, C, D, E [5].

Il metodo dovrebbe prendere in considerazione tutte le possibili topologie, ma dato che queste aumentano in modo esponenziale all'aumentare delle unità tassonomiche, vengono di solito considerate solo quelle che sulla base delle distanze genetiche risultano essere più probabili. Questo comporta che l'albero determinato potrebbe non essere il più "parsimonioso" in assoluto.

Il metodo della massima parsimonia ha delle limitazioni legate, oltre che alla difficoltà computazionale, alla mancanza di un modello di evoluzione molecolare. Esso non tiene per esempio conto di sostituzioni parallele, convergenti o multiple che sono tanto più frequenti quanto maggiore è la distanza tra le sequenze. Per di più tutte le sostituzioni vengono considerate equivalenti e questo non corrisponde al reale processo evolutivo. Inoltre il metodo spesso non individua una soluzione univoca: vengono trovati molti alberi ugualmente "parsimoniosi". Non è possibile adottare nessun criterio per discernere la soluzione migliore e pertanto si determina la topologia comune a tutti gli alberi egualmente parsimoniosi, determinando così l'albero *consensus*[8]. Tale topologia presenterà un numero più o meno grande di nodi non risolti (politomie), che corrisponderanno alle differenze topologiche negli alberi filogenetici di partenza.

Tutte queste limitazioni sono comunque maggiori per le sequenze nucleotidiche, per cui il metodo viene usato soprattutto per l'analisi di sequenze aminoacidiche.

1.6.4 Metodo della massima verosimiglianza

Il metodo della massima verosimiglianza, così come quello della massima parsimonia, non utilizza la matrice delle distanze ma direttamente le sequenze. Questo metodo cerca di quantificare quale sia la probabilità che ad una certa ipotesi H, nel nostro caso un albero filogenetico, corrisponda un certo insieme di dati D, nella fattispecie un allineamento multiplo. Indichiamo questa probabilità con la scrittura seguente:

$$Pr(D|H)$$

L'albero che ottiene il massimo valore di probabilità rappresenta la stima di massima verosimiglianza della filogenesi tra le sequenze considerate. L'albero di massima verosimiglianza è quindi quello che meglio giustifica il set di dati in esame, ovvero il multiallineamento.

[8] Uno dei pacchetti più noti per questo tipo di analisi è il PHYLIP.

La probabilità viene calcolata come prodotto delle probabilità che ha ciascun sito del multiallineamento di presentare un certo carattere se ha avuto luogo un particolare processo evolutivo (rappresentato dall'albero filogenetico in questione). Tutte le variabili del modello, come, per esempio, rate di sostituzione, topologia dell'albero, lunghezza dei rami, vengono calibrate per massimizzare il valore di verosimiglianza.

La principale limitazione del metodo sta nell'elevata complessità computazionale. Di fatto è impossibile applicare il metodo se si ha un set di più di 20-30 sequenze.

	Minima evoluzione	**Massima parsimonia**	**Massima verosimiglianza**
vantaggi	Molto veloce; può essere usato con un numero molto elevato di unità tassonomiche.	Consente di utilizzare molte unità tassonomiche sfruttando completamente l'informazione dei singoli caratteri.	Estremamente robusta da un punto di vista statistico. Consente l'impiego di modelli di evoluzione molecolare sofisticati e realistici.
svantaggi	Si perde informazione filogenetica nel passaggio da caratteri a distanze genetiche.	Può produrre più alberi equamente parsimoniosi non facilmente interpretabili; non tiene conto delle sostituzioni multiple.	Estremamente dispendiosa dal punto di vista computazionale. Non impiegabile per un alto numero di sequenze.

Tabella 1.1: Schema sui vantaggi e gli svantaggi dei metodi illustrati.

1.7 Il bootstrap

L'attendibilità di un'analisi filogenetica può essere valutata mediante alcuni test che valutano la significatività statistica dei vari nodi che compongono l'albero in questione. La tecnica più utilizzata a questo scopo è il cosiddetto *bootstrap*.

Il bootstrap è un test utilizzato in diversi ambiti. In questo caso consiste nell'effettuare un certo nuemro di ricampionamenti del multiallineamento in analisi, estraendone a caso i siti. Si considerino N sequenze multiallineate S_i $(i = 1, N)$ di lunghezza L. Il multiallineamento si può

rappresentare con la seguente matrice:

$$A = \begin{pmatrix} a_{11} & a_{12} & .. & a_{1L} \\ a_{21} & .. & .. & a_{2L} \\ .. & .. & .. & .. \\ a_{N1} & .. & .. & a_{NL} \end{pmatrix} \tag{1.2}$$

dove a_{ij} è il j-mo residuo dell'i-ma sequenza. Un sito del multiallineamento corrisponde a una colonna di questa matrice. Se si scelgono a caso L siti, anche con ripetizione, si otterrà un multiallineamento simulato che puèssere analizzato con gli stessi metodi utilizzati per il multiallineamneto reale. Si immagini allora di effettuare un gran numero (ordine del centinaio) di questi allineamenti simulati e di costruire per ciascuno di essi un albero filogenetico, in modo da avere alla fine un numero di alberi pari al numero di multiallineamenti simulati effettuati. A questo punto si cerca di ricostruire l'albero *consenso*: si calcola per ciascun nodo la percentuale di alberi simulati in cui viene riprodotto lo stesso nodo. Due alberi riproducono lo stesso nodo quando esso condivide le stesse unità tassonomiche discendenti. La percentuale così calcolata viene definita il valore di bootstrap del nodo considerato. Un alto valore di bootstrap corrisponde ad un'alta significatività statistica del nodo. Normalmente si applica la *majority rule consensus*, vale a dire la regola per cui un nodo viene considerato significativo se presenta almeno il 50% di bootstrap.

Bisogna comunque essere cauti: la tecnica del bootstrap ci dà informazioni sulla precisione ma non sull'accuratezza dell'analisi. La teoria del bootstrap dà inoltre una stima della topologia dell'albero: mostra quanto le diverse parti di una sequenza convergono verso una stessa topologia filogenetica. Se il processo evolutivo si è manifestato in modo uniforme su tutta l'estensione della sequenza considerata allora i valori di bootstrap tenderanno ad essere più alti. Al contrario, se parti della sequenza hanno subito processi evolutivi differenti i valori di bootstrap tenderanno ad essere molto bassi.

Capitolo 2

Peptidi over-represented: una nuova analisi filogenetica

Analizzando particolari geni o famiglie di geni sono state scoperte importanti relazioni evolutive. Mediante per esempio lo studio della piccola subunità ribosomiale (SSU rRNA), Doolittle ha scoperto la tripartizione fondamentale dell'albero della vita in Batteri, Archea e Eucarioti [8].

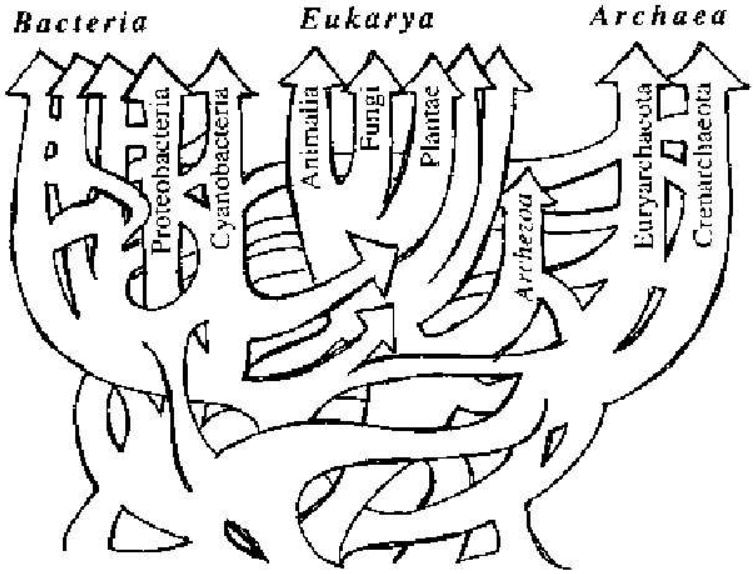

Figura 2.1: L'albero della vita con le tre principali ramificazioni: batteri, archea e eucarioti [8].

Non sempre tuttavia l'albero costruito analizzando i geni corrisponde all'albero che descrive le relazioni tra le specie. L'incongruenza tra "albero dei geni" e "albero delle specie" può essere determinata da varie cause. La più frequente sta nel fatto che alcuni dei geni considerati non sono realmente ortologhi ma paraloghi (vedi 1.3). Altre possibili cause di differenza possono essere eventi di trasferimento genico orizzontale, fenomeni di evoluzione concertata o di pressione mutazionale direzionata che produce eterogeneità composizionale tra le sequenze. I metodi di analisi filogenetica che illustriamo ([6] e [7]) mirano ad aggirare questo problema, cercando di condurre lo studio non su una sola proteina (o su un gene), ma sull'intero proteoma. In questo modo si può guadagnare "potere risolutivo" nella conoscenza e nello studio delle discendenza dai tre rami fondamentali prima citati avendo a disposizione un'analisi globale piuttosto che una locale.

2.1 Primo metodo

I proteomi non sono un insieme random di peptidi. In passato è stata dimostrata l'esistenza di fenomeni lontani dalla "casualità", come il clustering di aminoacidi o le forti correlazioni tra segmenti diversi di proteoma e di genoma. Tenendo conto di ciò, supponiamo che a livello molecolare l'evoluzione biologica si possa manifestare tramite alcuni segmenti "significativi". Normalmente si considerano significativi quei tratti di proteina che deviano sensibilmente dall'essere un "assemblaggio" random di aminoacidi. Il metodo che mostriamo si basa sulla ricerca e sullo studio di peptidi di lunghezza k che in un proteoma reale emergono molte più volte rispetto a quanto farebbero in un proteoma random. Questi peptidi sottostanno a correlazioni statistiche che esprimono le distanze evolutive.

2.1.1 Ricerca dei k-peptidi significativi

Una proteoma P è costituito da un insieme di n_P sequenze proteiche (o proteine). Le sequenze proteiche sono stringhe di diversa lunghezza formate da un alfabeto A di 20 lettere[1]:

$$A = \{\sigma_1, \sigma_2, ..., \sigma_{20}\};$$

[1] 21 lettere se si considera che le zone a bassa complessità (zone a scarso significato biologico, come le poli-A nel caso del DNA) vengono mascherate con il carattere X.

ogni σ rappresenta un aminoacido diverso. Se il proteoma è costituito in tutto da N_P aminoacidi, si può calcolare la frequenza di ogni carattere nel modo seguente:

$$f(\sigma_i) = \frac{n_i}{N_P} \; con \; i = 1, ..., 20 \tag{2.1}$$

dove n_i è il numero di volte che l'*i*-esimo aminoacido compare nel proteoma.

Indichiamo con L_i con $i = 1, ..., n_P$ la lunghezza dell'*i*-esima proteina. Chiamiamo *k*-peptide una sequenza di *k* lettere contigue. Con un alfabeto di 20 simboli, si possono formare in tutto 20^k *k*-peptidi diversi. Nel nostro proteoma selezioniamo i *k*-peptidi in maniera *overlapping*, ma senza che essi si ritrovino a "cavallo" di proteine diverse. Tenendo conto di questo, nel proteoma ci troviamo difronte a un numero di *k*-peptidi pari a:

$$N_P^{(k)} = \sum_{i=1}^{n_P} (L_i - k + 1) = N_P - n_P(k-1) \tag{2.2}$$

L'apice *k* per $N_P^{(k)}$ sta ad indicare che stiamo calcolando il numero totale di *k*-peptidi nel nostro proteoma *P*.

Indicando con $p_j^{(k)} = \{\sigma_{1j}, \sigma_{2j}, ..., \sigma_{kj}\}$ il *j*-esimo *k*-peptide, possiamo contare il numero di volte che compare nel proteoma. Indichiamo questo valore con $N_j^{(o)}$. Diamo inoltre una stima del numero di volte in cui lo stesso *k*-peptide comparirebbe in un proteoma random di stessa lunghezza, con lo stesso numero di k-peptidi e tale da mantenere la stessa frequenza aminoacidica calcolata con la (2.1). Indichiamo questa quantità con $N_j^{(e)}$ e la stimiamo nel modo seguente:

$$N_j^{(e)} = N_P^{(k)} \cdot Pr[p_j^{(k)}] \tag{2.3}$$

dove possiamo stimare la probabilità di occorrenza del *j*-esimo peptide $Pr[p_j^{(k)}]$ con il prodotto delle frequenze delle lettere che lo compongono, ovvero:

$$Pr[p_j^{(k)}] = f(\sigma_{1j}) \cdot f(\sigma_{2j}) \cdot ... \cdot f(\sigma_{kj}) = \prod_{i=1}^{k} f(\sigma_{ij}) \tag{2.4}$$

Per ognuno dei *k*-peptidi la cui occorrenza attesa è $N_j^{(e)}$, possiamo calcolare la probabilità che esso sia osservato *N* volte in un proteoma random usando la distribuzione di Poisson:

$$Pr_{N_j^{(e)}}[N] = \frac{[N_j^{(e)}]^N}{N!} \cdot e^{-N_j^{(e)}} \tag{2.5}$$

Stiamo utilizzando la statistica di Poisson, ovvero la statistica degli eventi rari, perchè si ha che $N_j^{(e)} \ll [N_P^k]$ per i valori di k che vogliamo utilizzare.

Definiamo *statisticamente rilevanti* quei k-peptidi che si trovano "nella coda" (=5% dell'area) della poissoniana, ovvero quelli per cui:

$$\int_0^{N^{(o)}} Pr_{N^{(e)}}[N']dN' \geq 0.95 \tag{2.6}$$

Chiamiamo d'ora in poi k-motivi questi particolari k-peptidi. I k-peptidi per cui invece si ha $N^{(o)} \approx N^{(e)}$ saranno chiamati k-peptidi *expected*.

Si vede che i k-motivi si presentano raramente da soli nelle proteine, ma sono quasi sempre clusterizzati o addirittura si sovrappongono (la fine di un k-motivo coincide con l'inizio di un altro), formando lunghi tratti statisticamente importanti (vedi figura 2.2).

>gi|14520241|ref|NP_125715.1| hypothetical MALTOSE
/MALTODEXTRIN TRANSPORT ATP-BINDING [Pyrococcus abyssi]

MVEVRLENLTKKFGNFTAVNKLNLTIKDGEFL**VLLGPSGCGKTTTLRMIAGLEE**PTE
GKIYFGDREVTYLPPRERNIS**MVFQSYA**VW**PHMTVY**D**NIAFPLK**IKKFPRDEIDKRV
RWAAELLQIEELLDRY**PAQLSGGQRQRVA**VARAIVVEPD**VLLMDEPLSNLDAKLR**VA
MRAEIKKLQQKLKV**TTIYVTHDQVEAMTMGDRIAVMN**RGQLLQVGPPTEVYLKPNSV
FVATFIGAPEMNIVEVSVGDGYLEGKGFKIELPQDIMELLRDYIGKTVLFGIRPEHM
TVEGVSELAHMKKTAKLNAKVDFVEALGTDTILHVKFGDELVKVKLPGHIPIEVGKE
VTIVIDLDMMHVFDKDTEKAII

Figura 2.2: In grassetto alcuni 6-motivi in una proteina dell'archaeon *P.Abyssi*. [6]

Il procedimento appena illustrato deve essere eseguito per ognuno dei proteomi delle N specie oggeto dell'analisi filogenetica.

2.1.2 I *k*-dizionari

Chiamiamo k-dizionaro e lo indichiamo con $Z_n(k)$ l'insieme costituito dai k-motivi che compaiono contemporaneamente in almeno n proteomi del set. Quindi $Z_1(k)$ indica l'insieme di tutti i k-motivi trovati, $Z_2(k)$ è l'insieme di quelli comuni almeno a due specie, mentre $Z_1(k) - Z_2(k)$ è l'insieme di quelli specifici: i k-motivi che si presentano in un solo proteoma. Si può vedere che se si riporta in un grafico il numero di elementi (entries) dei dizionari in funzione di k, normalizzati rispetto al numero totale di motivi espressi, al crescere di k, i k-motivi specifici superano in maniera consistente quelli condivisi (vedi figura 2.3).

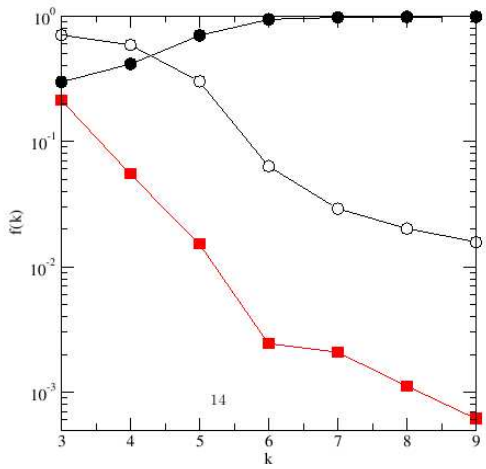

Figura 2.3: Frazione di *k*-motivi presenti nei diversi dizionari del set di proteomi elencati nel paragrafo 2.1.5. $Z_1(k) - Z_2(k)$ (cerchi pieni), $Z_2(k)$ (cerchi vuoti), $Z_6(k)$ (quadrati). [6]

2.1.3 Matrice di coespressione

Definiamo per ogni proteoma P una matrice di coespressione $A^{(P)}[Z_n(k)]$. I suoi elementi A_{ij} contano il numero di volte che l'i-esimo e il j-esimo k-motivo del dizionario $Z_n(k)$ si presentano contemporaneamente in una delle proteine di P.

Analizzando le matrici di coespressione di un dato set di proteomi di Archea e Batteri, si è visto che esse risultano tutt'altro che banali mostrando un considerevole gruppo di *k*-motivi coespressi parecchie volte in diverse proteine. Da ulteriori osservazioni sulle matrici si può dedurre che:

1. vi è un numero consistente di *k*-motivi comuni agli organismi dello stesso regno. Questi potrebbero costituire un dizionario che è stabile per tutte le unità tassonomiche;

2. vi è un un grosso numero di *k*-motivi specifici ($Z_1(k) - Z_2(k)$). Questi *k*-motivi probabilmente sono una manifestazione dell'evoluzione delle unità tassonomiche e ne costituiscono la "specificità linguistica";

3. vi è anche un numero consistente di *k*-motivi, quelli contenuti in $Z_2(k)$, che sono abbastanza specifici sebbene in comune ad alcune specie.

È plausibile che sia i *k*-motivi specifici che quelli "condivisi" interagiscano in qualche modo: la presenza di *k*-motivi comuni potrebbe essere in un certo senso modulata dai *k*-motivi specifici.

2.1.4 Una nuova definizione di distanza filogenetica

Si vuole a questo punto proporre la definizione di un nuovo concetto di distanza, utile ai fini di particolari analisi filogenetiche. $A^{(P)[Z_n(k)]}$ è una matrice simmetrica di dimensione n_k, dove n_k è il numero di *k*-motivi presenti in $Z_n(k)$. Possiamo "trasportare" la matrice di coespressione in un vettore di dimensione $\frac{n_k(n_k-1)}{2}$, essendo proprio $\frac{n_k(n_k-1)}{2}$ gli elementi distinti della matrice. Chiamiamo questo vettore *vettore di coespressione* e lo indichiamo con $V^{(P)}[Z_n(k)]$; usiamo l'accortezza di disporvi gli elementi della matrice di coespressione ordinati per riga, quindi:

$$V_s^{(P)}[Z_n(k)] = A_{ij}^{(P)}[Z_n(k)] \quad con\ j \geq i\ e\ s = 1,...,\frac{n_k(n_k-1)}{2} \tag{2.7}$$

Definiamo la distanza filogenetica $d_{P'P''}(k)$ tra i proteomi P' e P'' come il prodotto scalare tra i rispettivi vettori di coespressione costruiti sul dizionario $Z_j(k)$. In questo modo la distanza dipende sia da j che da k. Quindi:

$$d_{P'P''}(j,k) = 1 - \frac{\sum_s \{V_s^{(P')}[Z_n(k)] \cdot V_s^{(P'')}[Z_n(k)]\}}{\{|\mathbf{V}^{(\mathbf{P'})}| \cdot |\mathbf{V}^{(\mathbf{P''})}|\}} \tag{2.8}$$

2.1.5 Una particolare applicazione

Ferraro et al. [6], ideatori del metodo illustrato, hanno utilizzato il concetto di distanza appena introdotto per la costruzione dell'albero filogenetico di un insieme di 18 proteomi relativi ad alcuni procarioti (10 Archea e 8 Batteri) [2].

I dizionari $Z_n(k)$ sono stati costruiti per questo gruppo di specie seguendo il procedimento che abbiamo descritto nei paragrafi precedenti (vedi figura 2.2). Per la costruzione del vettore di coespressione (2.7), necessario per la definizione della distanza (2.8) e quindi per la costruzione dell'albero filogenetico, è stato utilizzato il dizionario $Z_2(6)$ (insieme dei 6-motivi presenti in almeno due proteomi del set). Illustriamo le ragioni per cui si è scelto di utilizzare

[2] I proteomi sono stati prelevati da Genbank. I 10 Archea sono: *Aeropyrum pernix, Archaeoglobus fulgidus, Halobacterium spNRC1, Methanococcus jannaschii, Methanobacterium thermoautotrophicum, Pyrobaculum aerophilum, Pyrococcus abyssi, Pyrococcus furiosus, Sulfolobus solfataricus Thermoplasma acidophilum*. Gli 8 batteri sono: *Agrobacterium tumefaciens, Bacillus subtilis, Chlorobium tepidum, Deinococcus radiodurans, Escherichia coli K12, Synechocystis spPCC6803, Thermotoga maritima, Yersinia pestis CO92*.

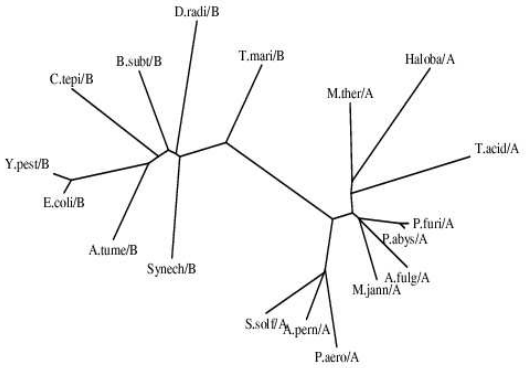

Figura 2.4: Albero unrooted costruito con il Neighbor-joining e utilizzando il concetto di distanza definito in questo capitolo. Dopo il nome delle specie, segue una A o una B: la A indica gli Archea mentre la B i batteri.

proprio $Z_2(6)$. Questo dizionario contiene 7712 entries, tra cui troviamo sia motivi specifici che comuni. Se si fosse utilizzato un dizionario $Z_n(k)$ con $n > 2$ si sarebbe condotta un'analisi filogenetica calcolando la distanza solo tra motivi fortemente conservati (presenti in più di 2 specie), senza quindi poter tener conto di quei motivi che modulano l'evoluzione di una particolare unità tassonomica. Inoltre i dizionari $Z_n(6)$ con $n > 2$ sono "piccoli". Nel nostro caso, per esempio, $Z_6(6)$ contiene solo 55 entries. Al contrario, invece, prendendo dizionari $Z_2(k)$ con $k < 6$ si hanno troppi elementi (nel nostro caso si hanno 161903 entries per $Z_2(5)$). L'inconveniente sta nel dover lavorare con matrici troppi grandi e in parte nella perdita di "significatività" dei *k*-motivi.

Si può eseguire un test statistico che permette di vedere quali *k*-motivi sono stabili al crescere di *k*, ovvero quali *k*-motivi non perdono di significatività. Se un *k*-motivo con *k* piccolo, significativo secondo la (2.6), è stabile, deve essere significativo anche uno dei dei 40 $(k+1)$-motivi che si possono formare aggiungendo una lettera all'inzio o alla fine della stringa di lunghezza *k*. Si vede che se *k* è piccolo, pochissimi *k*-motivi superano il $(k+1)$-test. Ciò vale a dire che il dizionario $Z_1(k+1)$ è molto più povero rispetto al dizionario $Z_1(k)$.

Se però *k*-cresce fino a 6, si raggiunge una certa stabilità, per cui si "perdono" solo pochi motivi applicando il $(k+1)$-test. Pertanto $Z_2(6)$ risulta il miglior compromesso per avere un numero adatto di entries senza perdere però i motivi proteoma-specifici.

A questo punto, utilizzando la definizione (2.8), è stata ricavata una matrice delle distanze

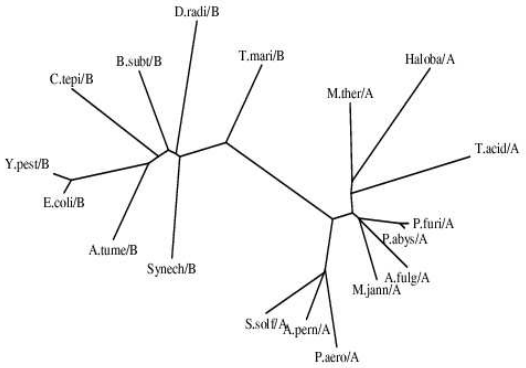

Figura 2.5: Albero filogenetico delle 18 specie considerate costruito mediante il confronto delle sequenze di SSU rRNA.

in modo da poter costruire un albero filogenetico con il metodo Neighbor-joining, così come descritto nel paragrafo 1.6.2. In questo modo è stato ricavato l'albero di figura 2.4. Come riferimento, invece, è stato preso l'albero filogenetico costruito con l'allineamento della SSU rRNA, mostrato in figura 2.5.

Il risultato potrebbe sembrare sconfortante, poiché i rami dei batteri e degli archea non sono ben risolti; verrebbe quindi a mancare la fondamentale tripartizione dell'albero della vita. Tuttavia essendo il metodo utilizzato basato su proprietà statistiche globali piuttosto che locali, a differenza dell'analisi condotta mediante confronto di SSU rRNA, le associazioni rivelate devono essere meglio investigate. Infatti, gli alberi costruiti applicando il metodo descritto mostrano la stessa topologia di quelli ottenuti mediante la *whole-genome analyses*, l'analisi condotta utilizzando l'intero genoma [9]. Questo tipo di analisi sta dando risultati interessanti, portando avanti l'ipotesi che afferma che gli eucarioti si siano in realtà formati dalla fusione di genomi procariotici preesistenti. Ciò metterebbe in discussione il concetto di albero della vita, sostituendolo con l'idea del *ring of life* (vedi figura 2.6). I risultati ottenuti sembrerebbero supportare quest'ultima congettura.

2.2 Secondo metodo

La possibilità di accedere, mediante banche dati, agli interi genomi e proteomi di un'enorme quantità di organismi ha messo in luce la riduttività di un'analisi filogenetica condotta esclu-

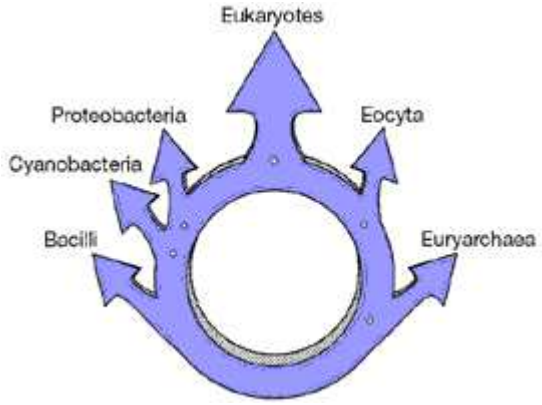

Figura 2.6: Immagine che schematizza l'ipotesi del *ring of life*. [10]

sivamente su particolari geni e proteine. Come già accennato all'inizio di questo capitolo e alla fine del paragrafo precedente, la possibilità di analisi su interi genomi ha avanzato dei dubbi sulla veridicità dell'analisi filogenetica basata sul confronto della SSU rRNA [8] e sulla tripartizione dell'albero della vita in Archaea, Bacteria e Eukarya.

All'inizio di questo capitolo abbiamo sottolineato le cause di errore in un'analisi filogenetica che fa riferimeno esclusivamente al confronto di alcuni geni, discutendo dell'importanza di uno studio che prenda in considerazione tutto il corredo genetico di una specie. Tuttavia è impossibile trattare e analizzare interi genomi o proteomi con i metodi tradizionali. Non ha senso cercare di allineare due genomi o due proteomi appartenenti ad organismi diversi, visto che ogni organismo ha un proprio patrimonio genetico e i geni sono disposti inmaniera diversa. Un altra difficoltà è dovuta alla differente lunghezza dei genomi. Si rende dunque necessario lo studio e l'implementazione di nuovi metodi che utilizzano altri espedienti per estrarre *contenuto informativo* dai dati in oggetto.

Il metodo che presentiamo adesso [7] possiede molti tratti in comune con il metodo descritto in paragrafo 2.1. Anche questo metodo è incentrato sul conteggio di peptidi di lunghezza k in una "collezione" di sequenze proteiche (il proteoma, o parte di esso). Il vantaggio di un'analisi di questo tipo sta nel non avere "parametri liberi" da fissare. Non ci sono, ad esempio, multiallineamenti che, implicitamente, dipenderebbero dalla matrice di sostituzione scelta (vedi Glossario e l'Appendice A).

2.2.1 Frequenza delle stringhe di lunghezza k

Data una sequenza di DNA o di aminoacidi di lunghezza L, si vuole contare il numero delle volte che compare una certa stringa di lunghezza k. Il massimo numero di stringhe diverse che possiamo trovare è 4^k per le sequenze nucleotidiche, 20^k per le sequenze aminoacidiche. Denotiamo la frequenza della k-stringa $\alpha_1\alpha_2...\alpha_k$, dove α_i è uno dei simboli dell'alfabeto in questione, con $f(\alpha_1\alpha_2...\alpha_k)$. La frequenza divisa per il numero totale di stringhe di lunghezza k $(L-k+1)$ ha il significato di probabilità di occorrenza della stringa $\alpha_1\alpha_2...\alpha_k$ nella sequenza in questione. La indichiamo con $p(\alpha_1\alpha_2...\alpha_k)$ e si ha:

$$p(\alpha_1\alpha_2...\alpha_k) = \frac{f(\alpha_1\alpha_2...\alpha_k)}{(L-k+1)} \tag{2.9}$$

Dopo aver calcolato in questo modo la probabilità di occorrenza della stringa $\alpha_1\alpha_2...\alpha_k$, è necessario sottrarre quello che è stato chiamato *random background*. A livello molecolare vi sono mutazioni dovute al caso. Queste mutazioni costituiscono una sorta di "rumore di fondo", il random background, che si sovrappone a quelle dovute alla pressione evolutiva e mantenute dalla selezione naturale. Questo background va sottratto al semplice conteggio delle occorrenze delle stringhe.

La probabilità di occorrenza della k-stringa viene predetta utilizzando un modello di Markov. Per fare ciò è necessario aver "contato" direttamente le occorrenze delle stringhe di lunghezza $(k-1)$ e $(k-2)$. In tal modo la probabilità predetta per la stringa $\alpha_1\alpha_2...\alpha_k$ è:

$$p^0(\alpha_1\alpha_2...\alpha_k) = \frac{p(\alpha_1\alpha_2...\alpha_{k-1})p(\alpha_2\alpha_3...\alpha_k)}{p(\alpha_2\alpha_3...\alpha_{k-1})} \tag{2.10}$$

L'apice 0 in p^0 sta ad indicare che si tratta di una quantità predetta.
Il ruolo della selezione naturale è espresso proprio dalla differenza tra il conteggio p e la probabilità predetta p^0. A questo punto per la nostra analisi è utile definire, per ogni possibile stringa $\alpha_1\alpha_2...\alpha_k$, la quantità $a(\alpha_1\alpha_2...\alpha_k)$ in questo modo:

$$a(\alpha_1\alpha_2...\alpha_k) = \begin{cases} \frac{p(\alpha_1\alpha_2...\alpha_k)-p^0(\alpha_1\alpha_2...\alpha_k)}{p^0(\alpha_1\alpha_2...\alpha_k)} & se\ p^0 \neq 0 \\ 0 & se\ p^0 = 0 \end{cases} \tag{2.11}$$

Semplifichiamo la notazione indicando le $a(\alpha_1\alpha_2...\alpha_k)$ con a_i, con $i = 1,...,N$, dove $N = 20^k$ nel caso di sequenze aminoacidiche e $N = 4^k$ nel caso di sequenze nucleotidiche. Tutte le possibili k-stringhe sono infatti 20^k nel primo caso e 4^k nel secondo caso.

2.2.2 Scelta delle sequenze aminoacidiche

Per ognuna delle unità tassonomiche del set di cui si vuole costruire l'albero filogenetico si definisce un vettore di composizione A nel modo seguente:

$$A = (a_1, a_2, ...a_N) \tag{2.12}$$

I vettori di composizione si possono costruire in tre modi diversi:

1. usando l'intera sequenza genomica della specie;
2. usando soltanto le sequenze codificanti del genoma della specie;
3. usando la traduzione aminoacidica delle sequenze codificanti del genoma della specie.

Poiché le sequenze codificanti mostrano un rate di mutazioni minore e meno casuale delle sequenze non codificanti, il modo migliore per costruire il vettore di composizione è intuitivamente il terzo. La validità della scelta può essere mostrata utilizzando un particolare criterio di consistenza. È importante infatti che al crescere di k la topologia dei vari alberi costruiti (mediante il concetto di distanza che definiremo in seguito) converga. Applicando tutti e tre i modi di costruire i vettori di composizione prima elencati, si è visto che la migliore consistenza si realizza utilizzando le sequenze aminoacidiche. D'ora in poi, quindi, ci riferiremo sempre e solo a sequenze proteiche.

2.2.3 Correlazione e matrice delle distanze

Siano A e B due specie i cui rispettivi vettori di composizione sono:

$$A = (a_1, a_2, ..., a_N) \quad e \quad B = (b_1, b_2, ..., b_N).$$

Calcoliamo la correlazione $C(A,B)$ tra le due specie utilizzando il coseno tra i due vettori N-dimensionali a e B:

$$C(A,B) = \frac{\sum_{i=1}^{N}(a_i \times b_i)}{(\sum_{i=1}^{N} a_i^2 \times \sum_{i=1}^{N} b_i^2)^{\frac{1}{2}}} \tag{2.13}$$

Definiamo la distanza $d(A,B)$ tra le due specie A e B nel modo seguente:

$$d(A,B) = \frac{1 - C(A,B)}{2} \tag{2.14}$$

Poiché $C(A,B)$ varia tra -1 e 1, la distanza $d(A,B)$ può assumere i valori da 0 a 1. Definita in questo modo la distanza tra coppie di specie, è immediato introdurre la matrice delle distanze D, simmetrica e di dimensione *n*, dove *n* è il numero di specie coinvolte nell'analisi. I coefficienti della matrice sono le distanze $d(i,j)$, con $i = 1,...,n$ e $j = 1,...,n$, dove i e j indicano i vettori di composizione delle *n* specie.

2.2.4 Un'applicazione

Qi et al. [7], ideatori del metodo appena illustrato, hanno utilizzato la distanza (2.14) per costruire l'albero filogenetico di un set di 109 specie diverse, tra cui 103 procarioti e 6 eucarioti. L'albero filogenetico è stato costruito utilizzando il metodo Neighbor-joining (vedi paragrafo 1.6.2) del pacchetto PHYLIP. Gli autori inoltre hanno effettuato il test statistico del bootstrap (vedi paragrafo 1.7) che ha mostrato risultati confortanti. Gli autori hanno eseguito un bootstrap che ha estratto per ogni specie un certo numero di sequenze proteiche, con possibilità di ripetizione. In questo modo alcune sequenze sono state estratte più volte, altre mai. Il risultato ha mostrato che non è necessario avere il proteoma completo per ricostruire l'albero filogenetico: è sufficiente analizzare la maggioranza delle proteine.

In figura 2.7 mostriamo l'albero filogenetico ottenuto. A differenza del metodo illustrato nel paragrafo 2.1, la tripartizione nei tre domini di Archaea, Bacteria e Eukarya è perfetta. Ciò è particolarmente entusiasmante visto che, pur utilizzando strategie completamente diverse, l'analisi condotta con questo metodo ricalca in maniera ineccepibile i risultati ottenuti dal confronto della SSU rRNA.

La maggiore differenza con il metodo descritto nel paragrafo 2.1 sta nell'aver utilizzato un modello di Markov per sottrarre il *random background*. Probabilmente i risultati diversi in termini di albero filogenetico derivano da questa scelta e non dall'aver utilizzato un set diverso di proteomi.

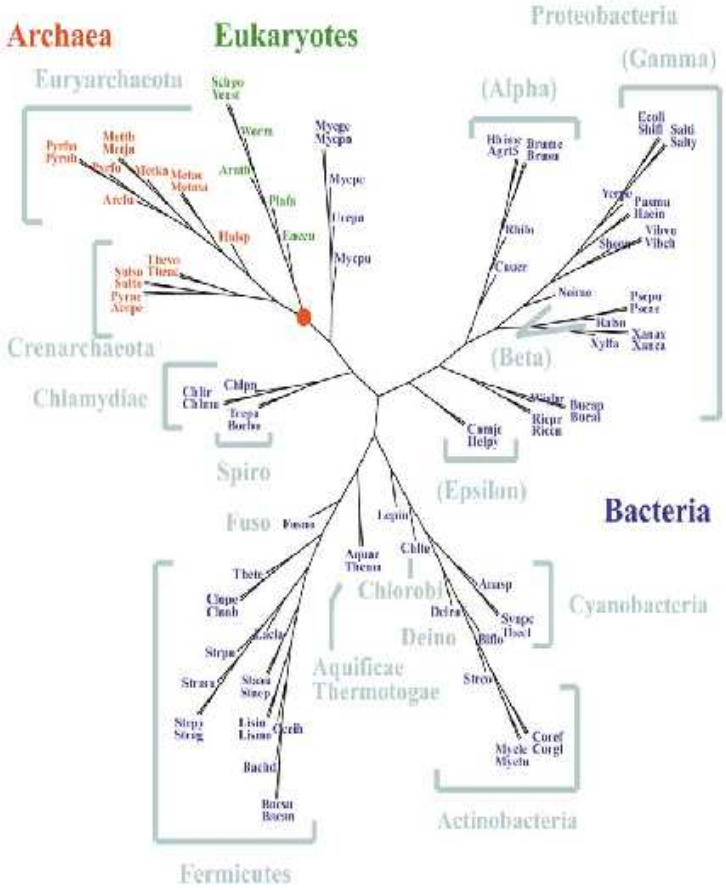

Figura 2.7: Albero filogenetico costruito utilizzando il valore $k = 6$ per le 109 specie utilizzate da Qi et al. [7]. Il punto rosso rappresenta la triforcazione dell'albero nei tre domini. Gli Archaea sono rappresentati in rosso, i Bacteria in blu e gli Eukarya in verde.

Appendice A

Matrici dei costi per gli aminoacidi

Tutti gli algoritmi per comparare le sequenze di proteine fanno riferimento ad alcuni schemi che assegnano costi (o pesi) in corrispondenza di ognuna delle 210 possibili coppie di aminoacidi (190 coppie di differenti aminoacidi più 20 coppie di aminoacidi identici). In generale, queste tabelle dei costi sono rappresentate come matrici 20x20 di similarità, dove aminoacidi identici e quelli con caratteristiche simili hanno un peso più alto rispetto a quell con maggiori differenze.
L'assegnazione dei pesi è molto importante nel confronto tra sequenze, infatti questi possono largamente influenzare gli esiti dei confronti. Idealmente, i pesi dovrebbero riflettere i fenomeni biologici che l'allineamento cerca di mostrare.
Introduciamo alcuni schemi di assegnamento:

- *Lo schema dei costi d'identità*. Le coppie di aminoacidi vengono classificati in due specie: *identiche* e *non identiche*.
- *Gli schemi dei costi per similarità fisiochimiche*. Vengono assegnati maggiori pesi agli allineamenti di aminoacidi con proprietà fisico-chimiche similari.
- *Gli schemi dei costi per sostituzioni osservate*. Sono i più usati attualmente e sono derivati dall'analisi delle frequenze delle sostituzioni osservate negli allineamenti tra sequenze.

I due schemi per l'assegnamento dei costi per sostituzioni osservate più importanti sono:

- Le matrici *n*PAM, ideate da M. Dayhoff,

- Le matrici BLOSUM*n*, ideate da Steven e Jorja Henikoff.

Le differenze principali tra queste due famiglie di matrici sono:

1. Le matrici PAM si basano su un esplicito modello evolutivo (le sostituzioni sono contate sui rami di un albero filogenetico), mentre le matrici BLOSUM sono basate su un modello implicito di evoluzione che non viene epresso formalmente.

2. Le matrici PAM sono basate su mutazioni osservate attraverso allineamenti globali, questo include sia regioni altamente conservate sia regioni altamente mutabili; le matrici BLOSUM invece, sono basate solamente su regioni altamente conservate con allineamenti locali che non consentano gaps.

3. La procedura BLOSUM utilizza gruppi di sequenze nelle quali non tutte le mutazioni vengono contate allo stesso modo: il calcolo considera se la mutazione interviene tra due sequenze appartenenti alla stessa famiglia o no.

Tuttavia questi tipi differenti di matrici possono essere comparati utilizzando una misura di quantità di informazione media per coppie di aminoacidi, in unità bit, della “entropia relativa”. A seguito di tale comparazione, si può concludere che per il confronto tra sequenze strettamente correlate occorre utilizzare le matrici BLOSUM con valori elevati o le matrici PAM con valori bassi; mentre per la comparazione tra sequenze relazionate lontanamente si utilizzano le BLOSUM con valori bassi e le PAM con valori alti.
I moderni programmi per il confronto tra sequenze (FASTA e BLAST) permettono di scegliere tra varie matrici PAM e BLOSUM in modo tale da permettere ai biologi di scegliere la matrice che possa “pesare” meglio gli allineamenti sottoposti. In generale comunque, le matrici maggiormente utilizzate in bioinformatica sono la 120PAM e al BLOSUM62. [4]

Appendice B

Il codice genetico

Il DNA dispone di un alfabeto di quattro lettere per specificare di circa 20 aminoacidi da cui possono essere costituite, secondo un preciso ordine di successione, le proteine: utilizzando gruppi di tre lettere si possono formare 64 parole; sperimentalmente si è verificato che, in effetti, il codice per un dato aminoacido è formato dalla successione di tre basi azotate, che viene detta tripletta o codone; inoltre, l'ordine in cui si susseguono le triplette corrisponde all'ordine in cui le rispettive molecole di aminoacidi si dispongono nella catena della proteina.

Il codice genetico è sorprendentemente universale e ugualmente valido sia per gli animali

	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	fine	UGA	fine	A
	UUG	Leu	UCG	Ser	UAG	fine	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Figura B.1: Tabella di conversione da codoni ad aminoacidi

superiori (uomo compreso), sia per i batteri e i virus. Questo fatto sembra confermare che tutti gli organismi viventi, dalle piante agli animali, dai virus all'uomo, abbiano avuto un progeni-

tore comune con un codice genetico che si è preservato durante tutta l'evoluzione biologica. Inoltre il codice genetico risulta ridondante, nel senso che, disponendo di 64 codoni per 20 aminoacidi, uno stesso aminoacido è codificato da più codoni; vi sono inoltre alcuni codoni ai quali non corrisponde alcun aminoacido, ma servono per segnalare la terminazione e l'inizio della catena proteica.

Glossario

ALLINEAMENTO: È il prodotto della procedura di confronto tra due o più sequenze in base all'ordine dei caratteri nelle sequenze in esame. Un allineamento può essere locale o globale, generalmente l'allineamento locale è il più utile dal punto di vista del biologo. Il migliore allineamento far due sequenze è quello cui viene associato il più alto punteggio di similarità così come determinato in ai parametri scelti.

AMINOACIDO: Uno dei 20 diversi tipi di molecole che sono uniti da legami pepidici per formare una proteina. All'interno di una proteina gli aminoacidi sono anche detti "residui aminoacidici" o semplicemente "residui".

BIOSEQUENZA: Successione di nucleotidi nel DNA e nel RNA o aminoacidi nelle proteine, tenuti assieme da legami chimici e ordinati secondo regole che determinano l'espletamento di funzioni biologiche.

BLOSUM: Una matrice di sostituzione. Confronta matrici di sostituzione.

CODICE GENETICO: La tabella di corrispondenza tra triplette di nucleotidi ed aminoacidi: le 64 (4^3) possibili triplette codificano i 20 aminoacidi che si trovano nelle proteine biologiche oltre ai codoni di terminazione (che segnalano all'apparato di traduzione la fine della catena proteica). Il codice genetico è ridondante e degenerato.

CODONE: Una sequenza di tre nucleotidi adiacenti che identifica uno specifico aminoacido o un segnale di inizio o terminazione della traduzione.

COMPLESSITÀ DI UNA BIOSEQUENZA: Misura del contenuto informazionale di una

biosequenza. Per esempio, biosequenze a bassa complessità mostrano una composizione dei nucleotidi o degli aminoacidi fortemente sbilanciata o ripetitiva.

EUKARYA: Eucarioti.

FENOTIPO: Ogni caratteristica visibile di un determinato carattere genetico.

GAP: Disallineamento tra due biosequenze causato da un'inserzione o delezione in una delle due sequenze. un gap può essere lungo da 1 a n nucleotidi o aminoacidi.

GENE: Frammento di DNA che codifica per uno specifico prodotto funzionale insieme a tutte le sequenza necessarie per la regolazione e il controllo della sua esposizione.

GENOMA: Tutto il materiale genetico di un organismo.

GENOTIPO: Il corredo genetico di un individuo.

LEGAME PEPTIDICO: Legame covalente formato tra due aminoacidi. Il legame peptidico è un legame forte.

MATRICE DI SOSTITUZIONE: Matrice 20x20 contenente i coefficienti di similarità far coppie di aminoacidi. Esistono differenti matrici di sostituzione, tra cui PAM e BLOSUM, generate con algoritmi che stimano la probabilità di conversione di un aminoacido in un altro o la probabilità di conservazione di un aminoacido a partire da allineamenti multipli di sequenze proteiche evolutivamente correlate. tali matrici vengono utilizzate per associare dei punteggi agli allineamenti tra sequenze proteiche.

MOTIVO: Elemento conservato in un allineamento di sequenze proteiche che disolito si associa ad unadetermianta funzione.

MUTAZIONE: Alterazione genetica ereditabile o acquisibile dall'individuo nell'arco della sua vita che include: mutazioni puntiformi o alterazioni su più larga scala.

NUCLEOTIDE: Una molecola di zucchero con 5 atomi di carbonio legata covalentemente a un gruppo fosfato e a una base azotata. I nucleotidi sono i costituenti fondamentali del DNA.

OMOLOGIA: Due geni o proteine si definiscono omologhi se si sono evoluti da un progenitore comune. Talvolta, e non correttamente, si confondono i concetti di similarità e di omologia.

ORTOLOGIA: Una coppia di geni appartenenti a due specie diverse si dice ortologa quando si presume che i due geni abbiano cominciato a divergere in seguito al processo di speciazione delle specie considerate. Proteine ortologhe svolgono generalmente al stessa funzione nei due organismi

PAM: Una matrice di sostituzione. Confronta matrici di sostituzione.

PARALOGIA: Due geni si dicono paraloghi se derivano da un evento di duplicazione genica. Mentre geni ortologhi hanno spesso la stessa funzione, geni paraloghi in genere svolgono funzioni diverse, anche se spesso correlate.

PEPTIDE: Corta sequenza di residui aminoacidici unita da legami peptidici.

PIRIMIDINA: Composto azotato dotato di un singolo anello. Timina (T) e citosina (C) sono pirimidine.

PURINA: Composto azotato con una strutttura a doppio anello. Adenina (A) e guanina (G) sono doppi anelli.

PROTEINA: Macromolecola composta da una catena di aminoacidi uniti da legami peptidici. Le proteine naturali sono composte da 20 aminoacidi diversi.

PROTEOMA: L'intero insieme delle proteine di un organismo.

PROTEOMICA: Lo studio del proteoma.

REPLICAZIONE: La sintesi di una macromolecola identica ad una data (per esempio DNA).

RIBOSOMA: Organello cellulare composto da RNA, detto appunto *ribosomiale*, e proteine; nel ribosoma avviene la traduzione in proteina dell'informazione contenuta nell'mRNA.

RNA: (Acido Ribo-Nucleico) Acido nucleico composto di ribosio, fosfato e di 4 nucleotidi: citosina, uracile, guanina e adenina. I principali tipi di RNA sono: l'RNA messaggero (mRNA), l'RNA transfer (tRNA) e l'RNA ribosomiale (rRNA).

SEQUENZIAMENTO: è il processo mediante il quale si ottiene la sequenza lineare di nucleotidi, nel caso di DNA e RNA, o di aminoacidi, nel caso delle proteine. Le sequenze vengono poi rappresentate come lunghe stringhe di caratteri appartenenti ad un alfabeto di 4 lettere nel caso di nucleotidi, di 20 lettere nel caso di aminoacidi.

SIMILARITÀ: Misura che si può associare ad un allineamento tra sequenz eproteiche e che si ottiene sommando i valori associati nella matrice di sostituzione prescelta alle coppie di residui allineati.

TRADUZIONE: Il processo di tarduzione dell'informazione contenuta nell'RNA messaggero in proteina. La traduzione viene operata nel ribosoma.

TRASCRITTO: La molecola di RNA che viene copiata da un gene.

TRASCRIZIONE: Il processo di copiatura di un gene in una molecola di RNA.

UNITÀ TASSONOMICA: La sequenza, o il gruppo di sequenze, facente parte del set per un'analisi filogenetica.

Bibliografia

[1] G. Valle, M. Helmer Citterich, M. Attimonelli, G. Pesole, *Introduzione alla bioinformatica*, Zanichelli, 2003.

[2] Slides del corso di *Genomica e Proteomica umana*, tenuto dal Prof. D.Condorelli presso la Scuola Superiore di Catania nell'A.A. 2004-05.

[3] http://parasitology.informatik.uni-wuerzburg.de/login/n/h/j_003-0900-z_mediaobjects_s00436-003-0900-zflb4.gif

[4] www.mat.uniroma3.it/didatticacds/studenti/laureati/fanuli/fanuli.ps

[5] users.unimi.it/ camelot/Didattica/LucaG/EvolMolec_parte2.pdf

[6] L. Ferraro, A. Giansanti, G. Giuliano, V. Rosato, *Co-expression of statistically over-represented peptides in proteomes: a key to phylogeny?* arXiv:q-bio.MN/0410011 v2 at http://lanl.arXiv.org

[7] J. Qi, Bin Wang, Bai-Iin Hao, *Whole proteome Prokaryote Phylogeny Without Sequence Alignment: a K-string Composition Approach*, J. Mol. Evol. **58**: 1-11 (2004).

[8] W.F. Doolittle, *Phylogenetic Classification and the Universal Tree*, Science **284**:2124-2128 (1999).

[9] C.H.House and S.T.Fitz-Gibbon, *Using Homolog Groups to Create a Whole-Genomic Tree of Free-living Organisms: An update*. J. Mol. Evol. **54** 539-547 (2002).

[10] M.C. Rivera and J.A.Lake, *The ring of life provides evidence for a genome fusion origin of eukaryotes*. Nature **431**:152-155(2004).

9