Biotecnologia applicata

La Bioinformatica nello studio delle trasmissioni infettive virali

D. Lombardi

CdL in Biotecnologie indirizzo Medico diagnostico, Università degli Studi di Firenze

Introduzione

Grazie all'acquisizione di sempre maggiori conoscenze nel campo della biologia molecolare, e all'enorme sviluppo che stanno avendo le tecnologie e le tecniche ad essa correlata, è possibile ad oggi studiare non solo la struttura e le principali funzioni correlate ad un dato virus, ma si possono individuare anche molti "piccoli dettagli" legati agli stati patologici che tale virus determina, il loro decorso da un punto di vista molecolare e persino tempi e modalità con cui l'infezione è stata contratta. Le analisi filogenetiche, ossia lo studio delle relazioni evolutive tra i vari organismi, sono sicuramente, e non solo, applicabili alla comprensione delle relazioni che intercorrono tra due o più agenti patogeni, e quindi tra due o più infezioni in diversi pazienti. Un ruolo chiave in tali analisi filogenetiche, ma potremmo dire in tutta la biologia molecolare, è sempre più svolto dalle analisi bioinformatiche che hanno notevolmente accelerato, e continueranno sempre più ad accelerare, i tempi di analisi e l'accuratezza dei risultati e dei responsi richiesti.
Scopo di questo breve contributo è quello di revisionare alcune applicazioni della bioinformatica all'analisi filogenetica per individuare, mediante alcuni semplici algoritmi e programmi costruiti su tali algoritmi, una possibile concatenazione tra una o più infezioni al fine di individuare eventuali correlazioni tra queste in pazienti diversi.
Sarà usato come esempio un caso di trasmissione del virus dell'epatite C (HCV) che coinvolge una madre, positiva all'infezione da tempo, e sua figlia, che divenuta a sua volta positiva, ha necessitato sapere se aveva contratto l'infezione per diffusione intrafamiliare, oppure se si dovevano ricercare cause esterne.
L'analisi filogenetica e la bioinformatica possono aiutarci nella soluzione di tale quesito permettendoci di ricercare e confrontare i virus che infettano entrambe al fine di stabilire eventuali correlazioni tra le due infezioni.
In questo caso l'analisi è stata effettuata sulla proteina E1/E2, codificata da un gene ad elevato grado di conservazione, situato nella regione codificante le proteine dell'envelope virale, le quali sono coinvolte nell'iterazione virus-cellula ospite e sono perciò fondamentali per assicurare la buona riuscita dell'infezione, e che proprio per questo mantengono il loro alto grado di conservazione. Questo ci permette di tracciare sviluppo ed evoluzione virale all'interno dello stesso individuo nonché tra individui diversi infetti da HCV, per stabilire se il virus abbia o meno la medesima origine.
In tale studio furono effettuati ben 27 campionamenti, attuati con regolarità in un ampio intervallo temporale, della proteina E1/E2 estratta dal virus da cui era affetta la madre, permettendoci così di conoscere l'evoluzione del virus durante il decorso infettivo. Ciò permette inoltre di ottenere gli spunti necessari per cercare di individuare il periodo in cui il virus può eventualmente essere stato trasmesso dalla madre alla figlia, andando in particolare ad individuare le analogie tra le sequenze geniche presenti in un dato momento dell'infezione materna, ma soprattutto le divergenze che si sono create dopo l'eventuale contagio della figlia. Alla figlia furono attuati due soli campionamenti di tale proteina, ma che furono sufficienti ad effettuare l'analisi filogenetica, la quale è stata completata aggiungendo altre 31 sequenze proteiche provenienti da vari ceppi generici di HCV, con la finalità di visualizzare, qualora la figlia non abbia contratto l'infezione dalla madre, la distanza relativa tra il virus materno e quello contratto invece dalla figlia.

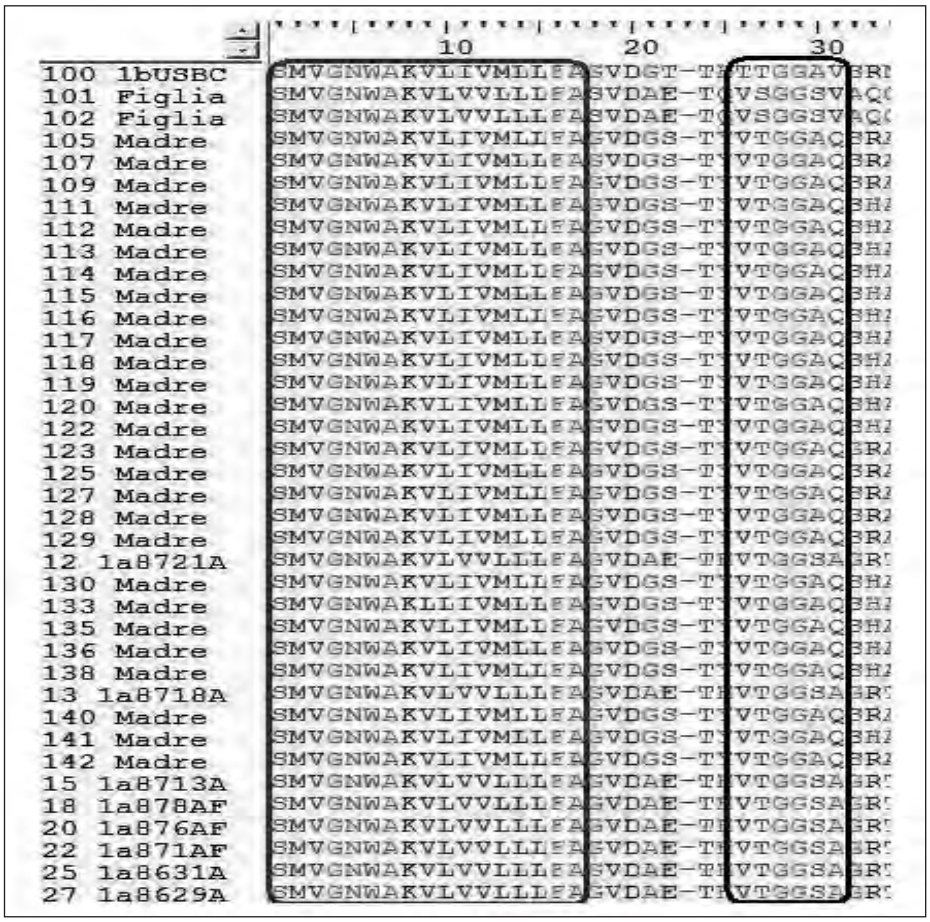

Fig. 1 - Sono evidenziate due regioni che in tutte le sequenze prese in esame sono altamente conservate.

Il caso

Il primo passo per un'analisi filogenetica è quello di eseguire un multi sequences alignment (MSA), ossia un allineamento tra le varie sequenze (in questo caso proteiche) prese in esame, al fine di appaiare i tratti omologhi delle diverse proteine e di individuare le divergenze tra le varie sequenze. È stato usato come algoritmo di allineamento ClustalW avvalendoci del programma di uso freeware BioEdit, il quale permette non solo l'allineamento, ma anche una facile visualizzazione dei risultati, in quanto le varie sequenze sono suddivise in righe ed allineate in base all'omologia dei vari tratti, come è possibile notare in Figura 1 ove sono evidenziate due regioni che in tutte le sequenze prese in analisi mantengono gli stessi aminoacidi.
I risultati del multiallineamento sono poi stati salvati in formato FASTA, ossia un formato analogo a quello txt con cui convenzionalmente sono salvate sequenze di DNA/RNA o proteine. Su tale file FASTA è stata poi eseguita l'analisi filogenetica mediante un altro programma sempre di uso freeware, Mega 4.1, il quale permette la costruzione di alberi filogenetici e la conversione del file di tipo FASTA in file di tipo .meg, ossia il formato che tale programma richiede per poter lavorare su tali sequenze. Una volta eseguita la conversione, che rappresenta sicuramente il primo step, il programma richiede su che tipo di sequenze si sta lavorando, ossia acido nucleico (in tal caso chiede anche se sia codificante o meno, in quanto le mutazioni in sequenze codificanti, che sono quindi lette in codoni, non hanno tutte lo stesso significato: se muta la prima base spesso cambia aminoacido, se muta l'ultima base spesso non varia l'aminoacido a causa della degenerazione del codice genetico: gli aminoacidi proteogenici sono 20 in totale, ma essendo il DNA costituito da 4 possibili basi azotate e letto per la traduzione in sequenze di tre basi azotate, otteniamo con 4^3=64 possibili combinazioni di basi azotate, ed è quindi chiaro che essendo 64 le combinazioni ma 20 gli aminoacidi avremo una certa ridondanza di codifica, e che combinazioni di basi diverse codificheranno per lo stesso aminoacido) o proteiche. Dopodiché è possibile costruire l'albero filogenetico, in questo caso utilizzando un metodo basato sulla distanza, il quale crea una matrice di dissimilarità tra

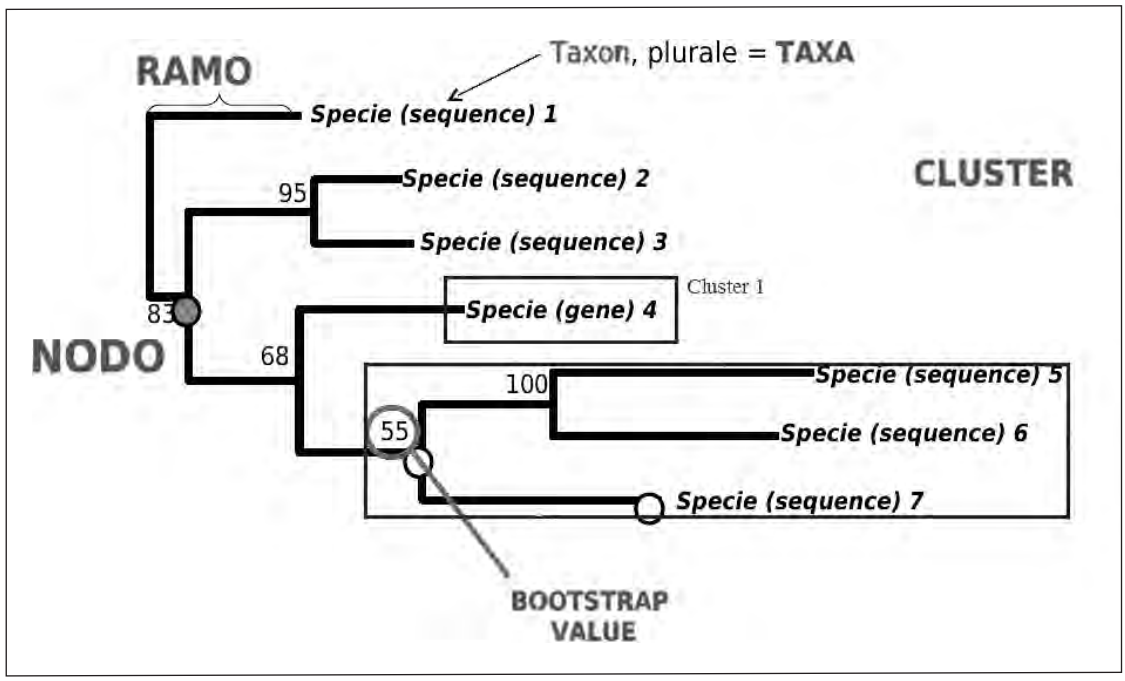

Fig. 2 - Schematizzazione di albero filogenetico per esplicazione visiva dei riferimenti testuali.

le varie sequenze facendo in particolare riferimento alla modalità neighbor-joining. Il programma prevede la possibilità di scegliere un qualsiasi numero di bootstrap, ossia il numero di replicazioni (confronti) che saranno eseguite tra le varie sequenze (è chiaro che maggiore sarà il numero di replicazioni impostato, più le sequenze saranno confrontate minuziosamente e, conseguentemente, l'albero sarà più accurato).
Prima di passare ai risultati è però necessario introdurre alcuni concetti sugli alberi filogenetici. In Figura 2 possiamo notare quelli che sono i parametri fondamentali per l'analisi di un albero filogenetico: sarà definito ramo qualsiasi segmento orizzontale che conduce ad un taxon (ossia un raggruppamento di organismi reali, distinguibili morfologicamente e geneticamente da altri e riconoscibili come unità sistematica, molto spesso si parla di taxon come di specie) o ad un nodo; nodo è invece il punto in cui un albero si divide dicotomicamente, mentre il valore di bootstrap rende conto dei valori statistici inerenti l'affidabilità dell'albero, i quali vanno da 1 a 100; 100 significa, ad esempio, che confrontando più e più volte quelle due date sequenze si è ottenuto sempre lo stesso allineamento e che perciò sono identiche o comunque strettamente correlate. Se ad esempio confrontiamo il valore di bootstrap evidenziato in Figura 2 notiamo che il cluster 1 è al 55% separato dal cluster 2. Inoltre è opportuno ricordare che la lunghezza dei rami è proporzionale alla divergenza tra le sequenze che sono prese in analisi: se il ramo è molto lungo vi è una notevole distanza tra le due sequenze che sono confrontate.

Conclusione

Nel caso qui riportato a scopo didattico è facile notare dall'albero filogeneticamente ricostruito (in Fig. 3), che la divergenza tra le sequenze del HCV associato alla figlia e di quello associato alla madre è considerevolmente alta ed è quindi altamente probabile che la figlia abbia contratto l'Epatite C tramite altra via che quella intra-famigliare (parentale). Al contrario i vari campionamenti fatti alla madre nel tempo si raggruppano in un solo cluster e questo ci conferma la correttezza dell'analisi filogenetica eseguita poiché tutti i virioni deriveranno da uno o pochi virus che hanno dato luogo all'infezione primaria nella madre, e che perciò non divergeranno particolarmente in un breve lasso di tempo, tantoché le distanze filogenetiche delle proteine appartenenti allo stesso virus modificatesi in tale breve intervallo temporale (l'arco vitale dell'organismo ospite) sono molto ridotte. Viceversa le distanze filogenetiche tra le sequenze appartenenti al HCV prelevato dalla figlia nei confronti di quelle prelevate alla madre, sono troppo grandi per ipotizzare che nell'arco di un tempo ristretto all'età della figlia, le due sequenze proteiche possano aver accumu-

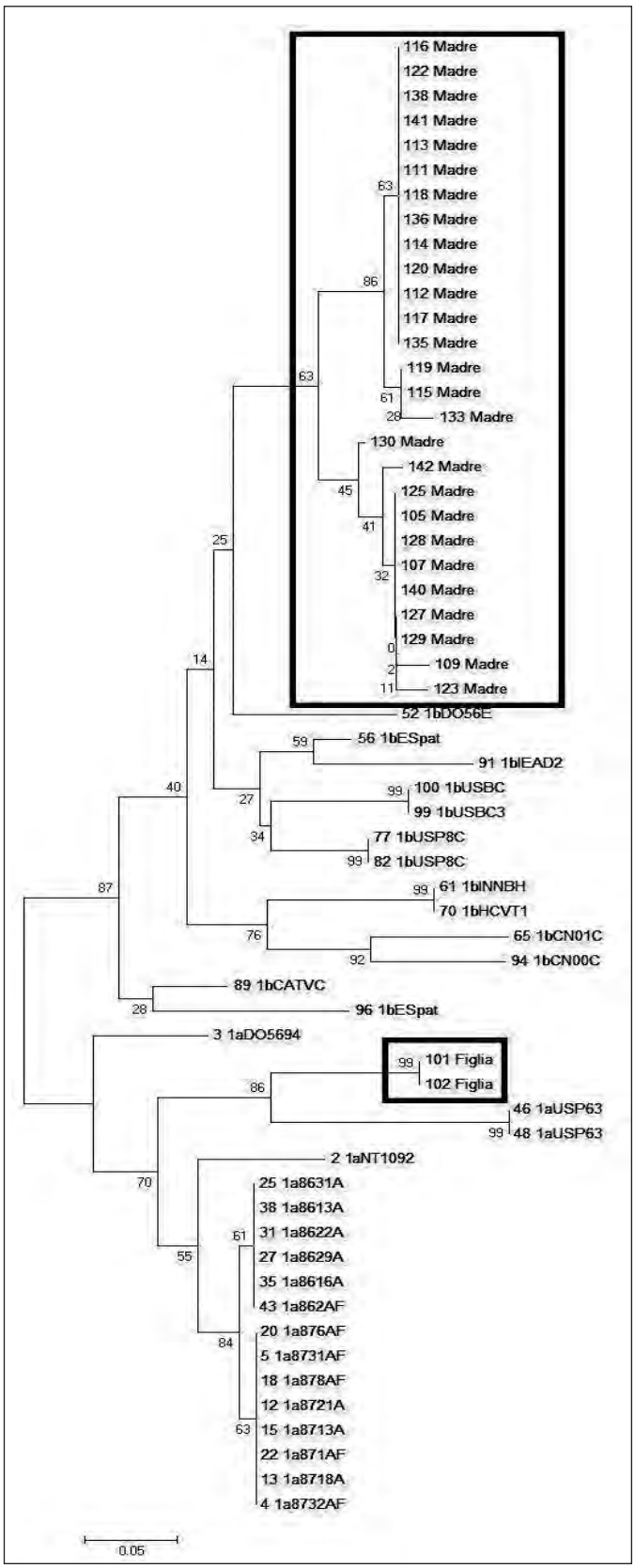

Fig. 3 - Sono evidenziati i cluster in cui si trovano le sequenze dell'HCV della madre e della figlia, le quali sono notevolmente distanti filogeneticamente.

lato una divergenza tale; è quindi chiaro che le cause dell'infezione della figlia sono perciò da ricercarsi in altri contesti diversi da quello di ambito familiare.
Come in questo caso grazie alla analisi biomolecolare ed alle sue applicazioni bioinformatiche si è potuti risolvere un quesito indaginoso e delicato, ed appare evidente come tale tipo di analisi possa trovare vaste applicazioni epidemiologiche in ambito sanitario financo ad essere usata per scopi medico legali e perché no assicurativi in contenziosi che sino a pochi anni orsono non vedevano l'equa soluzione e non solo per i pazienti.

Indirizzo degli Autori:
Duccio Lombardi
Via B. Scala 23
50126 Firenze
lombarduccio@alice.it

Letture consigliate

1. Cristina J, Colina R. Evidence of Structural genomic Region recombination in Hepatitis C virus. Virol J 2006; 3: 53.
2. Bermúdez-Aguirre AD, Padilla-Noriega L, Zenteno E, Reyes-Leyva J. Identification of amino acid variants in the hepatitis C virus non-structural protein 4A. Tohoku J Exp Med 2009 Jul; 218(3): 165-75.
3. Fani R, Fondi M. Bioinformatica, Firenze, Università degli studi di Firenze. (http://www1.unifi.it/dblemm/upload/sub/Lezione_4_2009.pdf
4. http://www1.unifi.it/dblemm/upload/sub/Lezione_5_2009.pdf)