Marco Galardini*, Alessio Mengoni*, Marco Bazzicalupo*

Produzione e analisi delle sequenze metagenomiche

introduzione

Si definisce qui metagenomica lo studio delle sequenze di DNA, per un totale di almeno 100 Mbp (milioni di coppie di basi), ottenute con un metodo di tipo "shotgun" cioè non ordinate, da un campione ambientale. Tipicamente l'analisi metagenomica mette insieme sequenze di DNA provenienti da molti microrganismi diversi, di cui non è richiesta la coltivabilità e quindi includendo anche i microrganismi difficili o impossibili da coltivare; la molteplicità delle informazioni di sequenza ricavate dall'analisi metagenomica consente infine di affrontare le problematiche connesse alle interazioni tra microrganismi presenti nella comunità microbica. L'analisi metagenomica quindi richiede di produrre e analizzare un numero molto grande di sequenze di DNA.

Questo breve capitolo cercherà di illustrare i problemi connessi a queste due fasi; in particolare, considerando la potenza di produzione di sequenze per mezzo delle apparecchiature di ultima generazione (*next generation sequencing technologies*), sembra evidente che lo sforzo maggiore vada concentrato sull'analisi dei dati, la quale spinge a sua volta la richiesta di potenti e sempre nuove applicazioni bioinformatiche.

In questa comunicazione cercheremo di affrontare le varie fasi dell'analisi metagenomica del suolo seguendo il loro sviluppo temporale, dalla progettazione del lavoro alla sua conclusione.

Le fasi che prenderemo in considerazione, e che costituiscono l'indice del presente capitolo, sono quindi: 1) Considerazioni pre-sequenziamento; 2) Campionamento e raccolta dati; 3) Sequenziamento; 4) Analisi dei dati,

* *Dipartimento di Biologia Evoluzionistica, Università degli Studi di Firenze*

quest'ultimo a sua volta organizzato in: a) Assemblaggio, b) Annotazione (predizione della funzione dei geni), c) Assegnazione tassonomica (*binning*), d) Composizione della comunità.

CONSIDERAZIONI PRE-SEQUENZIAMENTO

Prima di avviare il lavoro di raccolta del campione e di sequenziamento, è necessario stabilire i termini e le finalità del progetto. L'analisi metagenomica, e quindi i risultati che si raggiungono, infatti possono essere fortemente influenzati dalla costituzione della comunità microbica che si va ad analizzare ma anche dalla potenza di sequenziamento utilizzabile (quante basi si possono sequenziare) e infine dalla potenza di analisi disponibile (programmi bioinformatici e componenti hardware disponibili).

La prima domanda che ci si può porre è cosa includere nell'analisi: in un campione di suolo sono presenti procarioti (batteri e archea), eucarioti microscopici quali funghi e protozoi, ma anche piccoli animali quali nematodi e artropodi e infine una grande varietà di virus legati agli organismi suddetti. È chiaro che includere tutti i componenti del suolo è una scelta necessaria alla completezza dell'indagine, in particolare tenendo conto dello stretto legame ecologico, e spesso simbiotico, che c'è tra procarioti ed eucarioti. Includere tutto, in particolare la componente eucariota, significa d'altra parte appesantire enormemente la quantità di dati da analizzare poiché è noto che negli eucarioti le dimensioni molto grandi del genoma si accompagnano spesso a una molto bassa densità dei geni. Rimuovere il DNA eucariotico dal campione che si va a sequenziare può comunque essere complicato e può portare a introdurre altri tipi di errori di campionamento. La scelta quindi va calibrata sulla potenza di sequenziamento e di calcolo che si è in grado di mettere in campo e come è ovvio sugli obiettivi specifici della ricerca.

Un secondo aspetto che va considerato preliminarmente al campionamento e al sequenziamento è la "complessità" della comunità microbica da analizzare. La complessità genetica di una comunità è proporzionale al numero di specie (sequenze diverse) presenti, un parametro definito spesso come *richness*, e alla abbondanza relativa delle specie stesse, definita anche come *eveness*. Quindi una comunità che ha un alto numero di specie diverse, come è quasi sempre quella del suolo, e che presenta una abbondanza relativa delle specie ben distribuita, è molto complessa; al contrario una comunità con poche specie e che tra queste ne abbia solo qualcuna molto rappresentata, è una comunità poco complessa. La maggiore complessità di una co-

munità microbica si riflette nella minore probabilità che due sequenze qualsiasi siano contigue (appartengano cioè allo stesso frammento di DNA). La conseguenza di ciò è che l'analisi metagenomica di una comunità complessa produrrà un numero molto alto di piccoli frammenti non assemblabili, detti contig (il contig è una porzione di DNA genomico formato da sequenze sovrapposte tra loro), mentre una comunità poco complessa tenderà a produrre un numero minore di contig di dimensioni molto grandi, perché in questo caso le sequenze ottenute derivano da pochi genomi. Sapere quindi che la comunità microbica del suolo che si va ad analizzare contiene una o poche specie dominanti è un potente aiuto nell'organizzare i passi successivi del progetto metagenomico. Più avanti, quando parleremo di assemblaggio, vedremo come la grandezza dei contig sia un fattore determinante nel completare con successo un'analisi metagenomica. Quanto più sono grandi i contig, tanto più è possibile ricostruire porzioni significative dei genomi che compongono la comunità. Allo stato attuale dello sviluppo tecnologico, possiamo quantificare questo importante aspetto affermando che le comunità poco complesse, con poche specie dominanti, producono alla fine del processo di assemblaggio contig che vanno da 10 a 100 e più kbp (migliaia di coppie di basi). Le comunità complesse producono contig di dimensioni inferiori alle 10 kbp e spesso molto più piccoli.

Altri elementi da prendere in considerazione nella fase pre-sequenziamento riguardano la tecnologia di sequenziamento che si vuole adottare e la quantità di DNA da sequenziare. Questi due aspetti sono ovviamente legati tra loro e a loro volta legati all'impegno di spesa disponibile perché qualunque metodo di sequenziamento costa in proporzione alle basi sequenziate. Metodi di sequenziamento di ultima generazione (vedi più avanti) hanno un costo per base molto basso e sequenziano in breve tempo un numero di basi molto alto, tuttavia hanno lo svantaggio di produrre frammenti di sequenza molto piccoli che sono poi più difficili da assemblare. Il clonaggio del DNA in vettori di vario tipo d'altra parte, rallenta la velocità di sequenziamento e ne aumenta i costi, ma favorisce enormemente il lavoro di assemblaggio.

Per quanto riguarda la quantità di sequenze da produrre e analizzare non esiste un valore massimo, specialmente per il metagenoma del suolo è improbabile che qualunque progetto, per quanto ambizioso, riesca veramente a comprendere "tutte" le sequenze presenti in un certo suolo; quindi più sequenze si producono maggiore sarà il successo del progetto, sempre che si abbia un apparato di calcolo in grado di gestire tutti i dati ottenuti. Per analizzare una comunità microbica del suolo poco complessa e accontentandosi di mettere in evidenza solo le specie più rappresentative, si può ragionevolmente

stabilire una soglia minima di sequenze da determinare intorno a un valore tra i 100 e i 500 Mbp (milioni di coppie di basi).

CAMPIONAMENTO

Una volta scelto il sito in cui campionare il suolo per l'estrazione del DNA, ci sono almeno quattro aspetti molto rilevanti che vanno presi in considerazione prima di passare al campionamento vero e proprio.

Il primo aspetto riguarda il metodo di estrazione e purificazione del DNA che è strettamente correlato alla qualità del suolo, ma che è anche necessario mettere in relazione con la strategia di approccio al metagenoma. I tre punti critici sono la rappresentatività del campione rispetto alla composizione reale della comunità, la quantità di DNA necessaria all'intero progetto e la qualità intesa come dimensione dei frammenti ottenuti. Questi tre punti, insieme a molti altri correlati, sono ampiamente trattati in un altro capitolo di questa raccolta, per cui saranno adesso solamente accennati gli elementi che legano l'estrazione del DNA alle analisi successive.

Se si tratta di un progetto che prevede la produzione di una libreria di cloni bisogna privilegiare tecniche che non frammentano eccessivamente il DNA stesso, anche a scapito della resa; se invece il progetto è finalizzato al sequenziamento del DNA così come estratto, specialmente se si applicano tecniche di *next generation sequencing* il DNA può anche essere "maltrattato" e ridotto a piccoli frammenti, con un notevole incremento di resa. Nel caso di suoli molto poveri di materia organica, la resa del DNA può essere così bassa da rendere necessaria una amplificazione totale del DNA per raggiungere le quantità richieste per il sequenziamento. Questa procedura ha lo svantaggio di tendere ad alterare la rappresentatività relativa delle diverse sequenze, ma allo stesso tempo consente di recuperare le sequenze a singolo filamento, cosa essenziale se nel metagenoma si vuole includere anche quello virale.

Un altro accorgimento che va considerato è quello di prelevare campioni supplementari da utilizzare in parallelo per eventuali ulteriori analisi che si rendessero utili durante lo studio del metagenoma. La ragione di ciò sta nel fatto che se, durante lo studio, si rendesse necessario disporre di ulteriori campioni, il ricampionamento in un tempo successivo non garantisce l'equivalenza dei campioni stessi. Le analisi supplementari che potrebbero essere utili a migliorare i dati metagenomici sono ad esempio quelle sul metatrascrittoma, sul metaproteoma, l'ibridazione con sonde fluorescenti (FISH).

Il terzo aspetto da considerare preliminarmente al campionamento, è la

raccolta dei cosiddetti "metadati". Questi sono un complemento necessario alla descrizione del metagenoma e spesso ne consentono una corretta chiave interpretativa. Si tratta principalmente di dati collaterali di tipo geografico quali le coordinate geografiche del sito, la temperatura, l'umidità, la data di raccolta e le condizioni climatiche generali. Altri metadati sono la composizione chimica-fisica del suolo (ad esempio tessitura, contenuto in sostanza organica, C, N, P, salinità, presenza di composti organici, metalli pesanti ecc.) la profondità del campione, il tipo di copertura vegetale e la storia dei trattamenti subiti. Infine vanno anche considerate e incluse nella descrizione del campione le principali attività biochimiche riscontrate nel campione. Tra i metadati si sono aggiunti recentemente quelli che riguardano l'analisi delle capacità metaboliche del suolo dovute alla comunità microbica attiva in esso. Questa analisi, nota anche come metafenomica è descritta in dettaglio in un altro capitolo di questa raccolta e non sarà ulteriormente trattata qui.

Infine il quarto aspetto che va affrontato in vista del campionamento è la raccolta di informazioni preliminari sulla comunità microbica residente. Avere un profilo della comunità microbica è un'informazione importantissima prima di iniziare il progetto metagenomico vero e proprio. Usando marcatori tradizionali quali la sequenza del rRNA 16S o di altri geni metabolicamente rilevanti per quel particolare tipo di suolo, si può ricostruire un quadro della comunità che potrà aiutare a indirizzare il lavoro successivo. In particolare questa indagine può risultare conveniente a determinare l'abbondanza e la rappresentatività delle specie nel metagenoma e la possibile presenza di specie dominanti, che, come discusso in precedenza, può essere una informazione determinante per la strategia di sequenziamento.

IL SEQUENZIAMENTO

Il sequenziamento dei frammenti di DNA metagenomico è la fase cruciale del progetto, anche se non è forse quella più impegnativa in termini di tempo. Grazie alle nuove tecnologie *high-throughput* o *next-generation* (chiamate così perché si considera il metodo di sequenziamento di Sanger come prima generazione) l'ottenimento di miliardi di basi di sequenze è alla portata anche di piccoli laboratori con investimenti relativamente modesti. Non è negli scopi di questa relazione entrare nel dettaglio tecnico delle varie piattaforme di sequenziamento disponibili attualmente, si vuole solo dare un quadro generale delle possibilità e dei vantaggi legati a esse. Le tecnologie attualmente disponibili per il sequenziamento ad alta resa sono riportate nella tabella 1.

PIATTAFORMA	LUNGHEZZA DEL FRAMMENTO SEQUENZIATO	DURATA DEL PROCESSO (GIORNI)	BASI SEQUENZIATE PER PROCESSO (GBP)	COSTO DELLA MACCHINA (US$)	PRO	CONTRO
Roche 454 - Titanium	300-400	0.35	0.45	500000	Sequenze abbastanza lunghe; processo veloce	Alto costo del processo; errori di lettura
Illumina - Solexa	70-100	4-9	18-35	540000	La piattafroma attualmente più usata per metagenomica	Produzione degli stampi; tempi lunghi di processo; sequenze brevi
Life/APG SOLiD	50	7-14	30-50	600000	L'uso di dinucletidi riduce gli errori	Tempi di processo molto lunghi; sequenze molto brevi

Tab. 1 *Alcune delle piattaforme di sequenziamento di nuova generazione*

Tutti questi sistemi sono applicabili a frammenti di DNA estratti direttamente dal suolo e non necessitano del clonaggio in un vettore. Naturalmente questi stessi metodi si possono applicare a DNA metagenomico precedentemente clonato in un vettore quali plasmidi, cosmidi o fosmidi quando il progetto metagenomico preveda questa procedura per aumentare la possibilità di sequenziare interi geni o anche interi operoni. È da tener presente che le tecnologie di sequenziamento sono in continua e rapida evoluzione ed è quindi possibile che fra pochi anni anche quelle elencate nella tabella saranno diventate obsolete e saranno state sostituite da altre più potenti o più economiche. In ogni caso, nei grandi progetti metagenomici realizzati finora, generalmente sono state applicate in parallelo due tecniche di sequenziamento che hanno aumentato l'accuratezza delle letture e la possibilità di assemblare i frammenti sequenziati.

Alla produzione delle sequenze deve seguire la loro analisi che deve essere a sua volta preceduta da una fase detta di preprocessamento, spesso non attentamente considerata ma essenziale. In questa fase le brevi sequenze prodotte dalle macchine sequenziatrici sono controllate per l'attribuzione delle basi a partire dai grafici o cromatogrammi, in modo da associare a ciascuna base un punteggio che ne indichi il livello di qualità, cioè indichi la probabilità che una certa base sia effettivamente quella indicata. Nel caso si sia usata per il

sequenziamento una libreria di cloni plasmidici, sarà anche necessario eliminare tutte le sequenze appartenenti al vettore plasmidico stesso o all'ospite del clonaggio, in genere l'*Escherichia coli*. Queste operazioni sono svolte in maniera automatica attraverso l'impiego di programmi bioinformatici dedicati.

ASSEMBLAGGIO

Come si è accennato in precedenza, l'assemblaggio è una fase cruciale nell'analisi metagenomica e dalla sua qualità può dipendere la buona riuscita di un progetto metagenomico. L'assemblaggio è il processo attraverso cui le singole sequenze lette dalla macchina sequenziatrice vengono messe insieme in catene contigue dette *contig* sulla base della loro similarità. In genere un contig è prodotto dalla sovrapposizione di molti frammenti; il numero di frammenti che si sovrappongono in ogni data sequenza costituisce la "copertura" (*coverage*) o "profondità" della sequenza stessa; quanto più è alta la copertura tanto più è affidabile la sequenza che si sta considerando. Da un primo assemblaggio che consente di ottenere un numero grande di contig, una analisi successiva, la finitura, può consentire di mettere in evidenza una ulteriore contiguità tra contig diversi e assemblarli a loro volta in strutture più grandi e meno numerose dette "impalcature" (*scaffold*). L'assemblaggio delle sequenze in contig e scaffold è un'operazione impegnativa anche nei progetti genomici relativi a una singola specie a causa della presenza di zone ripetute e di possibili parti del genoma, anche piccole, che, per motivi strutturali, non riescono a essere sequenziate; queste difficoltà sono enormemente ampliate nell'assemblaggio delle sequenze metagenomiche, dove l'alto numero di specie e la presenza di sequenze simili appartenenti a genomi diversi rende arduo trovare le corrette sovrapposizioni. I due elementi che di più contribuiscono alla qualità dell'assemblaggio sono la lunghezza delle sequenze lette e la complessità della comunità con la presenza o meno di specie dominanti. Per quanto riguarda le lunghezza dei frammenti questa si è andata sempre più riducendo con l'utilizzo di macchine sequenziatrici di nuova generazione (tab. 1), per quanto riguarda la complessità della comunità questa è sempre molto alta nel metagenoma del suolo e il risultato è spesso una qualità di assemblaggio molto bassa.

Un cattivo assemblaggio può manifestarsi principalmente in due forme: una è costituita da errori di assemblaggio (*misassembly*) che nel metagenoma del suolo sono spesso dovuti alla formazione di contig "chimerici" che contengono parti di sequenza proveniente da specie diverse; l'altra è la dimensione ridotta e il conseguente numero alto di contig.

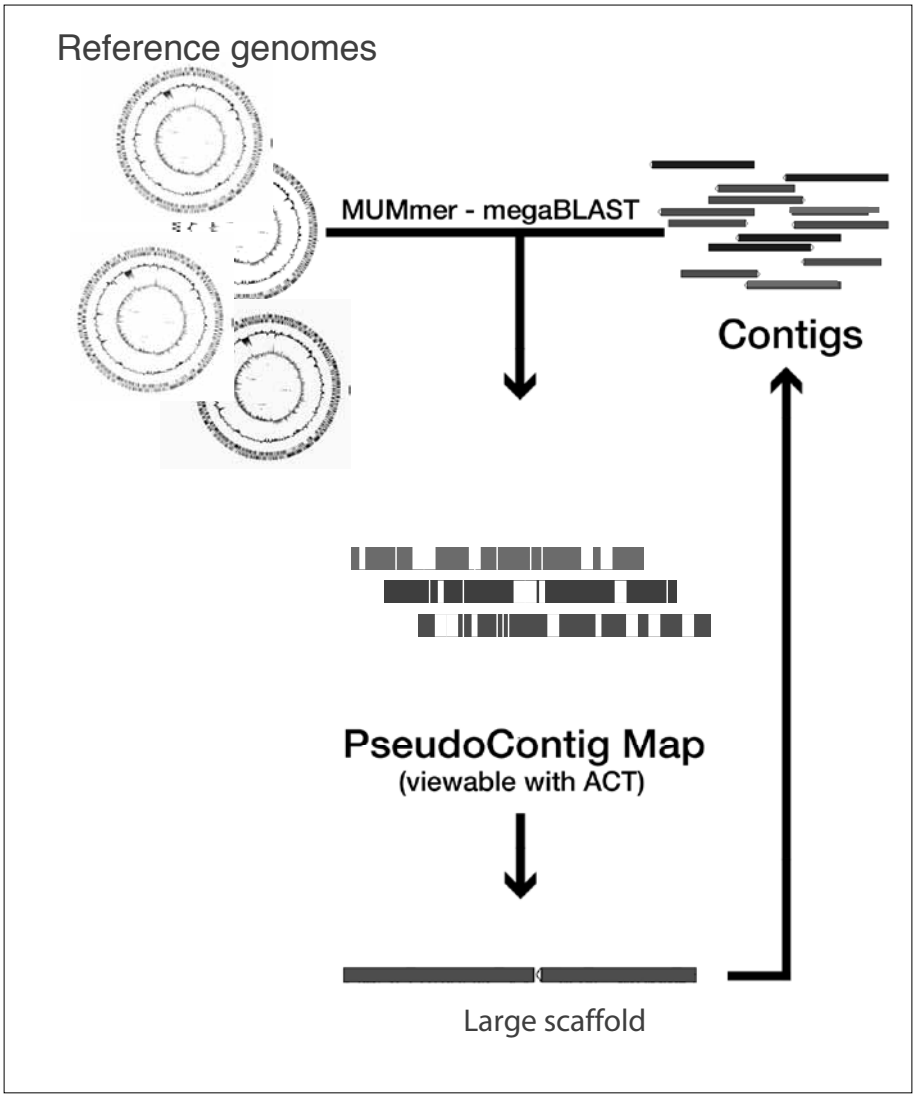

Fig. 1 *Schema di assemblaggio comparativo delle sequenze metagenomiche; i contig ottenuti vengono allineati contro genomi noti (Reference genomes) allo scopo di generare una mappa in cui due o più contig possono risultare vicini e per cui sia poi possibile progettare eventuali reazioni di PCR per unire questi contig. Il processo può essere quindi nuovamente iterato fino a esaurimento*

L'assemblaggio delle sequenze è assistito da una varietà di programmi bioinformatici che sono stati però spesso progettati per l'assemblaggio di singoli genomi e non sempre danno buona prova con dati metagenomici.

Un potente aiuto all'assemblaggio di sequenze genomiche, ma anche metagenomiche, è l'assemblaggio comparativo. Questo si basa sull'uso delle sequenze genomiche complete presenti in banca dati alle quali allineare i contig ottenuti dal sequenziamento, in modo da sfruttare i genomi già conosciuti come una guida per trovare sovrapposizioni tra contig e mapparli. Un esempio di schema di assemblaggio comparativo messo a punto nel nostro laboratorio è illustrato nella figura 1.

È utile ricordare che l'assemblaggio finale, quello che porta a unire tra loro tutti i contig e tutti gli scaffold e che costituisce l'obbiettivo di tutti i progetti genomici relativi a singole specie, non è comunque possibile nel caso del metagemoma del suolo, dove l'alto numero di specie e di ceppi fa sì che il risultato sia in ogni caso un insieme di sequenze continue molto più piccole dei singoli genomi da cui sono derivate. Come vedremo nel prossimo paragrafo questo può non essere un impedimento decisivo nell'analisi dei dati per disegnare un quadro anche preciso della comunità microbica che si sta indagando.

ANNOTAZIONE

L'annotazione è la fase dell'analisi metagenomica in cui si cerca di attribuire un significato alle sequenze prodotte. In pratica con questa operazione si identificano i geni codificati dalle sequenze di DNA e possibilmente le loro funzioni. Anche in questo caso l'approccio iniziale si basa sull'uso di programmi bioinformatici che sono stati sviluppati nella genomica tradizionale per annotare i genomi di singole specie. L'annotazione si può fare su qualunque tipo di sequenza, a partire dalle corte sequenze lette dalle macchine sequenziatrici ai contig assembalti e fino agli scaffold generati dal processo di finitura.

In genere l'annotazione si può realizzare con due diversi approcci. Il primo consiste nel cercare le omologie tra ciascuna sequenza del metagenoma e quelle presenti già annotate nelle banche dati. Questo metodo, basato sulla prova di somiglianza (*evidence-based*) si avvale dei numerosi programmi di ricerca di omologia disponibili tra cui va ricordato "BLAST" (Basic Local Alignment Search Tool) che è anche il più diffuso. Grazie all'enorme massa di sequenze annotate attualmente disponibili in banca dati questo approccio si rivela estremamente proficuo e in genere anche piuttosto veloce. L'altro approccio, che si basa sulle proprietà generali dei geni e non sulla conoscenza di sequenze specifiche, un metodo definibile come *ab initio*, parte dalle carat-

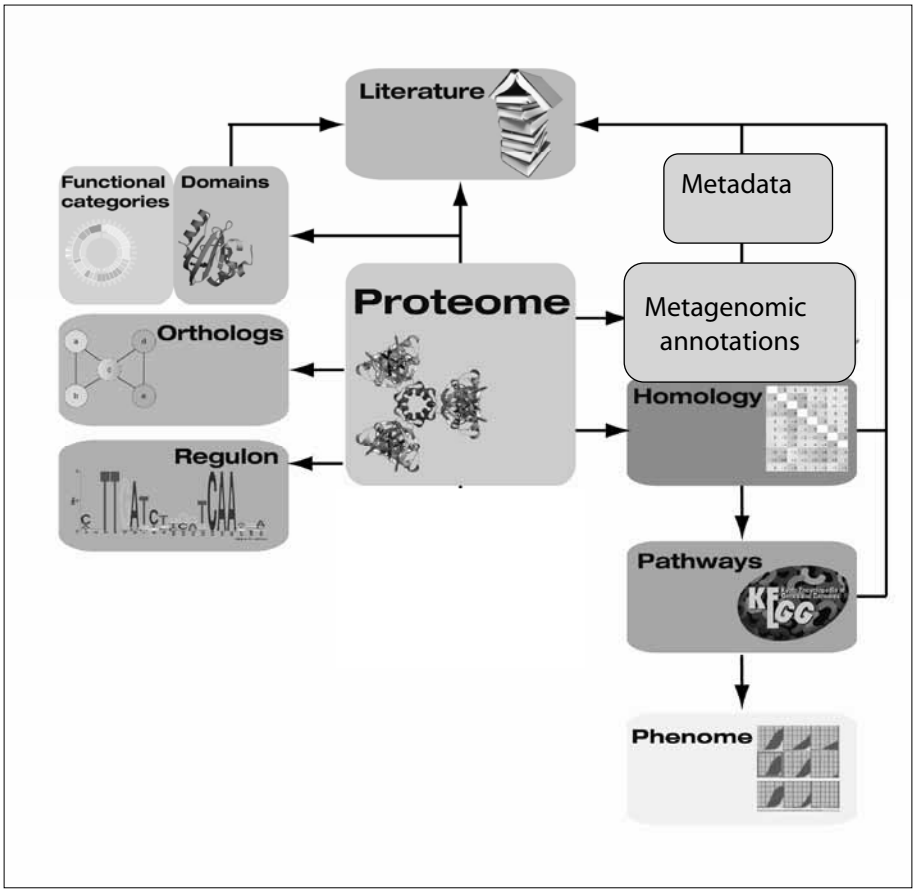

Fig. 2 *Schema di database per organizzare tutte le annotazioni disponibili riguardo al metagenoma in esame; le frecce indicano una relazione diretta fra le varie sezioni e la possibilità di incrociare i dati afferenti alle diverse sezioni*

teristiche intrinseche delle sequenze di DNA per individuare le possibili regioni codificanti, indipendentemente dal fatto che abbiano o meno omologia con sequenze in banca dati. Questo approccio consente quindi di individuare anche geni finora sconosciuti. Una corretta analisi metagenomica deve poter fare uso di entrambi questi approcci e integrarli con approssimazioni successive verso l'annotazione finale di ciascuna sequenza.

La qualità del lavoro di annotazione dipende in primo luogo dalla qualità e affidabilità del sequenziamento e dalla dimensione dei contig, se questi sono di buona qualità anche l'annotazione che ne segue sarà proporzionalmente più attendibile.

Un ulteriore aiuto all'accuratezza dell'annotazione viene infine dalle cosiddette informazioni di contesto. Queste riguardano dati sulla vicinanza genomica tra geni, su geni derivati da eventi di fusione, sui profili filogenetici, su dati di coespressione e altre informazioni disponibili riguardanti gli specifici geni individuati.

La grande quantità di dati che deriva dal lavoro di annotazione del metagenoma può risultare di difficile gestione dal punto di vista interpretativo e sembra quindi utile provare a sviluppare strumenti bioinformatici adatti a muoversi tra questi dati e capaci di indicare la loro connessione. Uno strumento simile è stato implementato presso il nostro laboratorio allo scopo di legare insieme, in un unico ambiente, tutte le informazioni genomiche (o metagenomiche) con quelle relative alla omologia di sequenza, alle reti regolative, ai dati della letteratura scientifica corrispondente, alla descrizione delle vie metaboliche e alle informazioni fenotipiche e infine ai metadati relativi al campionamento. Uno schema di questo programma è riportato nella figura 2.

ASSEGNAZIONE TASSONOMICA

Con l'assegnazione tassonomica delle sequenze, l'analisi del metagenoma si allontana definitivamente dalle analisi comuni anche al sequenziamento di genomi di singole specie. Il metagenoma è composto dall'insieme di molte specie diverse e quindi un primo obiettivo dell'analisi deve essere quello di identificare e quantificare il maggior numero possibile delle specie presenti nel campione. Questa operazione si sovrappone e completa l'analisi della comunità fatta in fase di pre-sequenziamento e si basa sull'identificazione di sequenze di marcatori filogenetici conservati tra i quali in primo luogo gli rRNA (RNA ribosomali) 16S e 23S, ma anche geni quali *recA* (codicante la proteina della ricombinazione e della riparazione del DNA), la proteina EF-Tu (fattore di allungamento della sintesi proteica), HSP70 (proteina indotta dallo stress di calore), il gene *rpoB* (subunità della RNA polimerasi). Tutte le sequenze di geni presenti nel metagenoma, possono essere allineate con quelle corrispondenti nelle banche dati e utilizzate per costruire degli alberi filogenetici che consentono di indicare la specie o il genere da cui la sequenza proviene.

Le maggiori limitazioni a quest'operazione sono dovute alla incompletezza e parzialità della banca dati anche per quanto riguarda questi marcatori filogenetici, alla frammentazione delle sequenze presenti nel dataset metagenomico

che rende arduo l'allineamento necessario alla produzione di alberi filogenetici, al fatto che solo una piccola parte del dataset metagenomico contiene i marcatori filogenetici menzionati, escludendo di fatto dall'assegnazione le specie rappresentate da altri geni.

L'operazione che cerca di classificare le sequenze metagenomiche come appartenenti allo stesso genoma, e quindi allo stesso ceppo o alla stessa specie, viene definita "inscatolamento" (*binning*) e costituisce il complemento all'assegnazione tassonomica nella descrizione della comunità metagenomica.

Gli approcci al binning sono di due tipi: un primo metodo si basa ancora una volta sulla omologia delle sequenze metagenomiche con quelle delle banche dati. In questo caso i problemi sono simili a quelli che si sono ricordati prima per i marcatori tassonomici. Un secondo approccio è invece basato sulla ricerca di caratteristiche particolari delle sequenze metagenomiche che le associno tra loro e a un gruppo tassonomico. Queste caratteristiche della composizione delle sequenze producono degli speciali contrassegni nelle sequenze stesse (*signature*) che le fanno distinguere come appartenenti a un particolare genoma. In particolare il contrassegno di solito più utilizzabile è la frequenza con cui si presentano determinati oligonucleotidi.

In generale il successo del binning dipende in gran parte dalla qualità dell'assemblaggio; infatti, se ogni contig deriva da un singolo genoma, in un grande contig è più probabile trovare sequenze di geni conservati o sequenze contrassegno con le quali attribuire l'identità all'intero frammento.

Mettendo insieme i risultati dell'annotazione, dell'assegnazione tassonomica basata sui marcatori filogenetici, e quelli del binning, si può cercare di dare un quadro della composizione della comunità microbica del campione metagenomico, definire la presenza di specie dominanti, e fare supposizioni ragionevoli sulla natura dei processi metabolici svolti dalla comunità stessa.

A conclusione di questo paragrafo va sottolineato come in una comunità microbica complessa come quella del suolo la quantità di ceppi, specie e generi differenti sia enorme e solo una altrettanto grande potenza di calcolo può essere in grado di trattare convenientemente i dati di sequenza per arrivare a una descrizione adeguata della comunità e della sua funzionalità.

ANALISI DEI SINGOLI GENI

Come accennato alla fine del paragrafo precedente, la complessità delle comunità microbiche del suolo può precludere, almeno in alcuni casi, una de-

scrizione che non sia approssimativa della comunità e quindi rischiare di non raggiungere le conclusioni desiderate.

Tuttavia, la natura dei genomi microbici, con la loro alta densità di geni, fa sì che anche in un metagenoma molto frammentato con un basso livello di assemblaggio, ogni singolo contig, per quanto piccolo, può contenere abbastanza informazione da riuscire a identificare il gene da esso codificato. Questo consente un tipo di analisi diverso, ma altrettanto utile ai fini della comprensione del funzionamento della comunità: l'analisi dei singoli geni. In questo tipo di analisi le funzioni putative, svolte dai singoli geni identificati sulle sequenze mediante le operazioni di annotazione, costituiscono la base della descrizione di una comunità, indipendentemente dalla possibilità o meno di attribuire il gene a un determinato gruppo tassonomico. In qualche modo in questo approccio la comunità è vista come un insieme, un aggregato, all'interno del quale non conta l'identità dei singoli componenti ma solo la loro funzionalità. La frequenza con cui una particolare funzione genica viene riscontrata nella comunità è direttamente un'indicazione della rilevanza ecologica che quella funzione ha nella comunità. Determinati progetti metagenomici, tesi a studiare certe specifiche funzioni come per esempio il ciclo dell'azoto o quello della zolfo, possono essere focalizzati solamente a evidenziare i geni relativi a quelle funzioni e in questo caso l'approccio sui singoli geni è l'unico applicato. Quest'approccio si presta particolarmente allo studio comparativo tra comunità, per mettere in evidenza differenze funzionali dovute a specifiche situazioni ambientali.

CONCLUSIONE

La produzione e l'analisi delle sequenze metagenomiche è oggi una realtà molto importante e costituisce un pezzo significativo di tutto lo sforzo di sequenziamento svolto a livello internazionale. Il metagenoma del suolo rappresenta, a causa della complessità delle sue comunità microbiche, la sfida più grande non solo dal punto di vista della produzione delle sequenze ma anche e soprattutto per quanto riguarda la possibilità di un'analisi soddisfacente. Solo con il contributo di importanti progressi a livello bioinformatico, adeguati e paragonabili a quelli già realizzati a livello di produzione delle sequenze, si riuscirà nel futuro a dominare questo campo di ricerca fondamentale per lo studio del suolo e per l'agricoltura in generale.

RIASSUNTO

La metagenomica sta diventando un approccio di studio sempre più utilizzato anche in laboratori di ricerca medio-piccoli e la metagenomica del suolo, sicuramente l'ambiente microbico più complesso, può essere oggi oggetto di progetti di ricerca. Può essere perciò importante in questa fase cercare di illustrare quali siano le procedure necessarie al corretto svolgimento di un progetto metagenomico.

La prima fase è quella di pre-sequenziamento, nella quale sono definiti gli obbiettivi del progetto e la sua estensione, entrambe correlate alla potenza di sequenziamento e di analisi computazionale che si può mettere in gioco. Nel pre-sequenziamento deve anche essere compresa una valutazione della complessità della comunità microbica che si studia, perché a questa sono connesse anche le decisioni riguardanti tutte le fasi successive del progetto.

La seconda fase è quella del campionamento durante il quale deve essere compiuto il massimo sforzo per preservare la qualità del DNA; in questa fase si raccolgono anche i cosiddetti "metadati" e si procede a collezionare campioni supplemetari di suolo per ulteriori possibili analisi.

La fase del sequenziamento è quella nella quale si producono effettivamente la sequenze di DNA, questa dipende in modo determinante dalla piattaforma tecnologica usata per il sequenziamento, in particolare di quale sistema di ultima generazione dispone il progetto.

La fase successiva prevede l'assemblaggio delle sequenze ottenute in insiemi continui più possibile lunghi detti contig. La qualità dell'assemblaggio può risultare decisiva in alcuni progetti metagenomici e può richiedere l'impiego di una notevole potenza di calcolo.

La fase dell'annotazione è quella nella quale si cerca di attribuire un nome alle sequenze assemblate, cioè se ne indica la funzione presunta; a questo punto il lavoro metagenomico comincia a mostrare i suoi frutti in termini di descrizione della comunità microbica.

Oltre all'annotazione la metagenomica del suolo richiede anche una fase in cui le sequenze vengono attribuite a specifici gruppi tassonomici; l'assegnazione tassonomica si basa sulla presenza di marcatori filogenetici e fornisce un quadro delle specie che svolgono un ruolo primario nella comunità. Infine, per il completamento dell'analisi metagenomica, si deve anche procedere all'analisi di singoli geni; questa volta, in modo indipendente dalla loro apparternenza a un particolare gruppo tassonomico, l'identificazione di singoli geni contribuisce a dare importanti informazioni sui processi che si svolgono nella comunità del suolo.

I progetti di metagenomica del suolo sono in realtà una sfida notevole, tuttavia grazie agli eccezionali sviluppi della tecnologia e della bioinformatica, possono essere oggi affrontati e dare un grande contributo alla scienza del suolo.

ABSTRACT

Metagenomic projects are becoming increasingly accessible also to medium and small size laboratories and even the metagenomics of soil, that is the more complex of the microbial environments, is currently under investigation. It is important at this point to

present a short account of the procedures involved in the metagenomic project with the aim to give a general picture of the most relevant steps necessary to proceed through in order to set up a successful research work.

Steps, in chronological order, are: Pre-sequencing issues involving the extent and the scope of the project which in turn are linked with the availability of sequencing and analysis power; this step also includes the evaluation of the complexity of the soil community to be investigated and its connections with the following steps of the project.

The second step is the sampling procedure, where much of the effort should be put in the quality of the DNA to be extracted and purified, but including also the gathering of metadata referred to the sampling site and the collection of supplementary samples for securing the possibility of further analysis.

Sequencing is the step when DNA sequences are actually generated, for which the approach is strongly related to the type of technological platform available, particularly which of the next-generation sequencing machines will be used.

After sequencing the obtained sequences should be assembled into continuous stretches, as long as possible, called contigs. The quality of assembly could be critical for the success of certain metagenomic projects and could require a very powerful computational effort.

Annotation is the step when the assembled sequences are named, that is they are attributed to some function. This is when results finally come out from the sequencing work and the microbial communities start to be described.

Beyond annotation the soil metagenomic project should also include the step of taxonomic attribution of the sequences, that is based on the occurrence of phylogenetic markers and give a picture of the species playing a major role in the community functioning. The completion of the metagenomic analysis finally includes the recognition and analysis of single genes that could be unrelated to the taxonomic position of the species to which the gene belongs, but nevertheless can give important information on the processes of the soil community.

Any soil metagenomic project is indeed a tremendous challenge, however, thanks to the astounding progresses in technology and bioinformatics, it can be afforded and greatly contribute to the soil science.

LETTURE CONSIGLIATE

Metzker M. L. (2010): *Sequencing technologies - the next generation*, «Nature Review Genetics», 11, pp. 31-46.

Huson D.H., Auch A.F., Qi J., Schuster S.C. (2007): *MEGAN analysis of metagenomic data*, «Genome Research», 17, pp. 377-386.

Kakirde K.S., Parsley L.C., Liles M.R. (2010): *Size does matter: Application-driven approaches for soil metagenomics*, «Soil Biology & Biochemistry», 42, pp. 1911-1923.

Kunin V., Copeland A., Lapidus A., Mavromatis K., Hugenholtz P. (2008): *A bioinformatician's guide to metagenomics*, «Microbiology and Molecular Biology Reviews», 72, pp. 557-578.

Mocali S., Benedetti A. (2010): *Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology*, «Research in Microbiology», 161, pp. 497-505.