UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

ANALISI DEL TOOL META-NETWORK PER LO STUDIO DI COMUNITÀ MICROBICHE

Relatore: Prof. ssa Barbara Di Camillo
Correlatore: Dott. Marco Cappellato

Laureanda: Erica Benfatto

ANNO ACCADEMICO 2021– 2022
Data di laurea 16-11-2022

Abstract:

I microorganismi come batteri, funghi e virus vivono in comunità che caratterizzano determinati ambienti. Analizzare il microbioma usando le tecniche di New Generation Sequencing apre grandi opportunità per una migliore comprensione del ruolo metabolico, fisiologico ed ecologico che il microbiota svolge all'interno dell'ambiente che lo ospita. Allo stesso tempo il campo è ancora poco esplorato e propone costantemente nuove sfide nelle tecniche di sequenziamento e negli strumenti computazionali. I metodi basati sulle reti si ripropongono di ricostruire la rete batterica partendo dalle matrici di abbondanza ricavate dal sequenziamento del microbioma. I metodi di co-occorrenza sono vari: vanno da quelli basati sulla correlazione a quelli basati su modelli grafici complessi. Nella panoramica dei tool proposti per tale scopo emerge Meta-network, che descrive le interazioni batteriche attraverso una rete di co-occorrenza in grado di identificare non solo le relazioni dirette, ma anche quelle indirette e non lineari.

Indice

1 Introduzione

Il microbiota è l'insieme di microrganismi come batteri, funghi, virus e archei che costituiscono una componente essenziale di specifici ecosistemi e habitat, tra cui la saliva e l'intestino di organismi semplici e complessi, ma anche l'aria, il suolo e l'acqua [1]. Lo studio del microbiota è di estremo interesse da quando è stato provato che i microrganismi hanno un'ampia influenza sui meccanismi ecologici dei loro habitat. Questo micromondo complesso è caratterizzato da diversi tipi di interazioni; comprendere queste relazioni premette di sviluppare strumenti utili per esplorare cause ed effetti dell'organizzazione della comunità [2]. Questa ricerca avviene analizzando il patrimonio genetico del microbiota, che prende il nome di microbioma. Gli studi sul microbioma sono stati potenziati dai progressi del Next Generation Sequencing (NGS) che ha aiutato a identificare precisamente le specie microbiche e i percorsi metabolici ad esse associati [1]. Possiamo osservare due tipi di interazioni batteriche: microbiche ed ecologiche. Le prime si instaurano tra diversi taxa; le seconde tra i taxa e l'ambiente, dove le cellule dell'organismo ospitante influenzano l'ecosistema microbico e dove, viceversa, i batteri promuovono numerose funzioni fisiologiche. Lo studio delle interazioni microbo-microbo e ambiente-microbo è di estrema importanza per comprendere l'organizzazione della comunità in relazione ai fattori che determinano la biodiversità. In aggiunta le reti microbiche possono svolgere un importante ruolo di prevenzione e di strumento terapeutico nel campo della salute dell'uomo. Le informazioni su come si modifichi la comunità in rapporto a un determinato stimolo permettono ad esempio di agire per mezzo di probiotici per restaurare la corretta composizione del microbioma [2]. Nelle interazioni microbiche studiate, si è notato che i microrganismi che condividono lo stesso ambiente ospitante, sono in costante competizione, in questa circostanza alcuni organismi sviluppano relazioni simbiotiche in cui cooperano tra loro per ottenere dei vantaggi che possono o meno essere positive per l'organismo ospitante [3]. Gli approcci analitici basati sulla rete si sono dimostrati utili per studiare sistemi con interazioni complesse e infatti rappresentano un potente strumento nella biologia dei sistemi per dedurre la regolazione genica. La teoria delle reti si è dimostrata uno strumento ottimo per indagare il vasto panorama delle comunità batteriche. I grafi sono usati frequentemente nella biologia molecolare per rappresentare le relazioni tra le entità, qui i nodi sono i diversi membri del microbioma mentre gli archi rappresentano la loro relazione che può essere eventualmente descritta anche tramite una direzione e un peso dell'arco. La presenza di

una relazione tra taxa viene dedotta dai valori di abbondanza tramite approcci di reverse engineering [2].

2 Il sequenziamento del microbioma di nuova generazione

Il sequenziamento di nuova generazione (NGS) è una tecnologia di sequenziamento massivo parallelo che offre un'elevatissima produttività, scalabilità e velocità. Questa tecnologia viene utilizzata per determinare l'ordine dei nucleotidi in interi genomi o in regioni mirate di DNA o RNA. Il NGS ha rivoluzionato le scienze biologiche, consentendo un'ampia varietà di applicazioni e permettendo lo studio dei sistemi biologici a un livello mai raggiunto prima. Le complesse domande di genomica di oggi richiedono una profondità di informazioni che va oltre la capacità delle tradizionali tecnologie di sequenziamento del DNA. Il NGS ha colmato questa lacuna ed è diventata uno strumento di uso quotidiano per rispondere alle domande poste dallo stato di avanzamento della ricerca nel campo del microbioma [4]. Queste tecniche e altri progetti hanno aiutato a migliorare enormemente il campo della previsione genomica, dell'associazione di geni, lo studio del microbioma umano, l'identificazione di patogeni e le diagnosi cliniche [1].

Dopo l'introduzione di nuove tecniche molecolari dell'ultimo decennio, la ricerca microbica ha visto una grande rivoluzione. L'attuale metodo di coltura microbica su terreni standard è in grado di riprodurre fedelmente gli aspetti essenziali come pH, temperatura, nutrienti e condizioni osmotiche. Tuttavia, questa tecnica può supportare solo la crescita di una piccola frazione della diversità microbica totale, mentre la maggior parte di essa rimane non coltivabile. I dati relativi al Next Generation Sequencing hanno dimostrato che l'enorme estensione del mondo microbico non coltivato rimane inesplorato dalle tecniche convenzionali. Questo perché i metodi standard di isolamento e arricchimento microbiologico, non sono in grado di supportare la crescita di tutti i microbi presenti in un campione. A causa di questo ostacolo, l'approccio microbiologico tradizionale per l'isolamento e la caratterizzazione di microbi da campioni ambientali è passato in secondo piano. Oggi i ricercatori sono più interessati a catturare e profilare la diversità microbica presente in un determinato campione, piuttosto che una piccola percentuale di essa, come si può fare con i metodi convenzionali. Insieme al miglioramento della strumentazione e delle tecniche di sequenziamento, la fonte dei campioni si è ampliata enormemente, i campioni ora provengono da una serie di fonti naturali o artificiali, come il suolo agricolo o ambientale, l'habitat acquatico, la flora presente all'interno di altri organismi, come gli animali domestici o il corpo umano. L'accresciuto interesse ha dato il via a numerosi progetti di ricerca sul microbioma su

scala locale e globale. Un famoso esempio è il Progetto Microbioma Umano (HMP), avviato dai National Institutes of Health (NIH) e lanciato nel 2005. Oppure Salute (NIH), lanciato nel 2008, che ha come obiettivo l'identificazione del microbioma umano completo e sano per apprezzare la diversità e complessità delle comunità microbiche [5].

I due metodi NGS più comunemente utilizzati per l'identificazione del microbioma e del suo genotipo sono le tecniche basate sul gene ampicon (es:16SrRNA) o sul shotgun Metagenomics. (Figura 1)

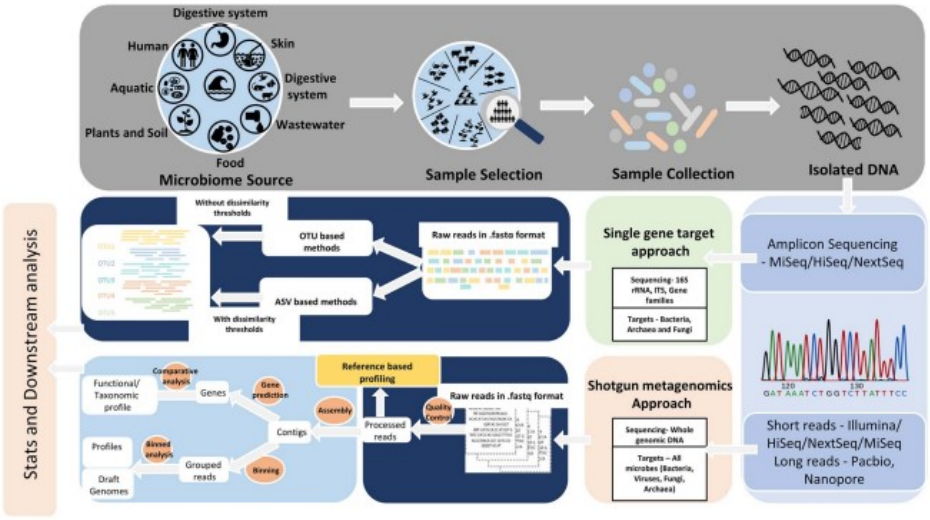

Figura 1: Illustrazione degli approcci di sequenziamento mirato amplicon e metagenomic. Una panoramica schematica che illustra i diversi tipi di campioni, le piattaforme di sequenziamento comunemente utilizzate, nonché le fasi di elaborazione dei dati.

2.1 Sequenziamento basato sul gene amplicon

È stata la tecnica primaria degli ultimi 25 anni per studi di filogenesi e tassonomia di microbi complessi, difficili da catalogare. Include diversi passaggi come la raccolta dei campioni, l'isolamento del DNA, l'amplificazione PCR del 16SrRNA, l'analisi delle sequenze utilizzando una variegata strumentazione computazionale [5].

2.1.1 Raccolta dei campioni e isolamento del DNA

Durante la raccolta dei campioni è indispensabile che la dimensione della biomassa sia sufficiente rispetto all'ambiente che deve rappresentare (es: la pelle contiene meno massa microbica rispetto all'intestino; sarà quindi sufficiente un campione ridotto). Inoltre, bisogna sempre tenere in considerazione alcuni parametri. Innanzitutto, evitare la contaminazione con altri campioni ambientali, con l'addetto alla raccolta e nel laboratorio; secondo, prestare attenzione alle condizioni e alla durata del transito dal momento della raccolta del campione dall'ambiente: refrigerare immediatamente è

considerato indispensabile; infine, controllare le condizioni di conservazione ottimali in laboratorio, che possono cambiare la composizione dei campioni e alterare i risultati finali, come alcuni studi hanno dimostrato [1].

Per quanto riguarda la scelta di metodi di isolamento del DNA è importante che questi catturino effettivamente tutti i tipi di microbi della porzione ambientale sotto esame. Esistono due metodi principali di estrazione del DNA: la lisi meccanica e quella chimica. Quella meccanica da rese superiori se eseguita precisamente, ma è da evitare una lisi meccanica vigorosa perché rischia di tagliare gli acidi nucleici [1].

2.1.2 Amplificazione PCR del gene 16sRNA

Il sequenziamento basato sul gene amplicone si basa sull'identificazione di un gene target: il più comunemente usato è il 16s rRNA che codifica la piccola subunità procariotica 30s appartenente al complesso ribosomiale 70s, nella maggior parte dei batteri e archei [1]. Il ruolo cruciale del gene 16s rRNA consiste nella caratteristica combinazione di zone altamente conservate, che permettono il riconoscimento del gene target, intervallate da zone ipervariabili che permettono la precisa classificazione genomica (che può essere di conosciuta o sconosciuta tassonomia) [2]. La sequenza del gene è quindi costituita dalle zone altamente conservate che sono usate come siti di ancoraggio del primer e 9 regioni ipervariabili (V1-V9) che servono a identificare e catalogare i profili microbici [1].

2.1.3 Preparazione degli acidi nucleici

Prima del sequenziamento avviene la fase di preparazione degli acidi nucleici: si procede nell'amplificazione con coppie di primer e nella purificazione del DNA, utilizzato poi per la preparazione delle librerie. Per ogni piattaforma di sequenziamento sono presenti diversi pacchetti che si occupano di questi passaggi [1]. L'amplificazione produce un grande numero di frammenti di 16s rRNA che vengono poi sequenziati attraverso piattaforme di NGS come Illumina MiSeq, che si dimostra essere la migliore per quanto riguarda le cosiddette letture brevi (short-reads). Illumina fornisce un output limitato (15 Gb) ed è utilizzato principalmente per il sequenziamento amplicon in quanto fornisce letture più lunghe (2× 300 bp) con un costo di sequenziamento molto più basso rispetto ad altri sequenziatori ad alta velocità. È interessante notare che Illumina genera letture brevi fino a 1,5 Tb per run. Sfortunatamente, le tecniche NGS basate su letture brevi hanno applicazioni limitate nell'analisi dei genomi poliploidi a causa della esclusiva applicabilità del loro algoritmo ai dati di metagenomica. In questo contesto, le piattaforme

di sequenziamento di terza generazione come Pacific Biosciences RS II/Sequel e Oxford Nanopore MinION si sono dimostrate più efficienti grazie alle dimensioni delle letture più lunghe, alla risoluzione più elevata e all'assenza di bias dovuti all'amplificazione del DNA [1].

2.2 Sequenziamento basato sul shotgun metagenomics

La metagenomica si riferisce all'analisi genetica del genoma ottenuto direttamente dall'ambiente, ha come obiettivo la catalogazione di tutti i microorganismi presenti nel campione in complessi campioni ambientali. L'approccio "shotgun sequencing" si differenzia dal sequenziamento con 16s rRNA perché invece di marcare un singolo gene target analizza l'intero genoma. Infatti, non si basa sulla diversità di un singolo gene, ma sull'intero patrimonio genetico del campione: questo permette di risalire alla composizione genetica delle comunità microbiche ottenendo una migliore risoluzione tassonomica e una migliore informazione genica generale. Il metodo aiuta inoltre ad associare particolari funzioni a comportamenti filogenetici e a creare profili evolutivi; permette infine di indentificare nuovi virus. L'aspetto negativo del metodo è che sono numerosi i fattori che possono introdurre errori ed alterare quindi la classificazione finale.[1]

2.3 Le fasi di preprocessing e clustering

I risultati del sequenziamento di tipo NGS producono milioni letture che possono essere impiegate per trovare la presenza e l'abbondanza di diversi taxa nella popolazione originale. Sono diversi i software sviluppati per la fase di preprocessing, che incorporano metodi per il denoising e per il filtraggio qualitativo, in grado di scartare sequenza troppo brevi o di bassa qualità. Inoltre, molti algoritmi si occupano di clusterizzare le letture in Operational Taxonomic Units(OTUs): raggruppamenti di organismi in base alla somiglianza di DNA. Metodi di clustering più recenti permettono la ricostruzione delle Amplicon Sequence Variants (ASVs), una versione a più alta risoluzione delle classiche OTUs. ASV può distinguere le sequenze che variano anche solo per un singolo nucleotide, senza necessitare di una soglia arbitraria di somiglianza com'è per il clustering in OTUs. L'ultimo passaggio della fase di preprocessing è l'assegnamento di un'annotazione tassonomica ad ogni OTU/ASV da parte di un classificatore addestrato su un database di riferimento. L'output dell'intero processo è la tabella OTU/ASV, deve ogni elemento contiene il numero di volte che la lettura emersa da un dato campione è stata attribuita ad un particolare OTU/ASV, e la relativa tassonomia che descrive e

caratterizza ogni OTU/ASV nel modo più approfondito possibile. Le matrici di conteggio delle sequenze risultanti hanno delle caratteristiche peculiari. Innanzitutto, la limitata profondità di sequenziamento rende le matrici sparse che significa che il 70-95% dei valori è zero. In secondo luogo, i dati non riflettono l'abbondanza assoluta, ma piuttosto una porzione dell'intera sequenza, che riflette la proporzione degli individui appartenenti a uno specifico gruppo tassonomico. In questo modo, si generano degli artefatti, chiamati bias composizionali. Per correggere i bias, normalizzare e tener conto della variabilità tecnica sono stati proposti numerosi metodi, come, ad esempio, la trasformazione in rapporti logaritmici [2].

3 Le reti di interazione microbica

Gli approcci basati sulle reti si stanno affermando come uno degli strumenti più utili per l'analisi della struttura delle comunità microbiche. Offrono nuovi spunti metodologici e biologici per indagare le interazioni tra le specie. Molti microrganismi coesistono interagendo tra loro ed esercitano efficacemente diverse funzioni. A causa dell'attuale insufficiente comprensione della struttura della comunità, il volume crescente di dati metagenomici rende difficile per la tradizionale analisi di rete recuperare le reali relazioni nella comunità batterica [6]. Progetti di sequenziamento recenti hanno quantificato l'abbondanza di centinaia di taxa microbici contando i geni marcatori (di solito 16S rRNA) sequenziati in un gran numero di campioni. Questi grandi numeri di campioni aprono la possibilità di svelare le complesse relazioni tra i microrganismi [7]. Gli approcci di analisi e di ricostruzione delle reti si sono dimostrati efficaci per decifrare i pattern di co-occorrenza microbica a partire dai risultati del sequenziamento dei batteri della comunità microbica sotto esame [3]. L'analisi della co-occorrenza è una tipica istanza nell'inferenza di rete, in grado di prevedere le relazioni tra gli oggetti partendo da misurazioni ripetute della presenza o dell'abbondanza degli oggetti [7]. L'idea di base è che due taxa co-occorrono quando hanno valori di abbondanza simili secondo la metrica usata, in tal caso è presente un arco tra i due [3]. Sono stati sviluppati diversi metodi di co-occorrenza che vanno dalla correlazione ai modelli grafici complessi.

3.1 I metodi di co-occorrenza per le reti microbiche

Le tecniche correlation-based, come la correlazione di Pearson e quella di Spearman, sono tra i metodi più popolari per studiare le interazioni microbiche nell'intestino dell'uomo, del cavo orale e del suolo. La correlazione indica la tendenza che hanno due variabili a variare insieme, ovvero a covariare. La correlazione lineare indica la forza

della relazione lineare tra due variabili. Il coefficiente di correlazione di Pearson è un valore che varia tra –1 e 1: se negativo indica che la correlazione è negativa, ovvero quando una variabile aumenta, l'altra diminuisce; quando si avvicina a zero indica che la correlazione è debole, se è zero che è inesistente; i valori 1 e –1 indicano una relazione lineare perfetta. Per la correlazione di Pearson è necessario fare l'ipotesi sulla distribuzione normale delle variabili, se la condizione per l'ipotesi non sussiste si può ricorrere alla correlazione di Spearman, una misura non parametrica di correlazione, che possiamo considerare un caso particolare di Pearson, nel quale prima del calcolo del coefficiente, si convertono i valori in ranghi. Il valore di correlazione si calcola sulla matrice OTU/ASV e si definisce il valore p come la probabilità che il valore di correzione osservato per ogni coppia di taxa sia maggiore di quello ottenuto a caso. Molti metodi di correlazione hanno implementato delle varianti rispetto a Pearson o Spearman per stimare le interazioni tra coppie di taxa. Tuttavia, queste misure, oltre a fornire dati artefatti o spuri nelle componenti microbiche meno abbondanti, non tengono conto della composizionalità, infatti, ad esempio, un aumento dell'abbondanza assoluta di un solo taxon è seguito da una diminuzione delle abbondanze relative di tutti gli altri taxa anche se la loro abbondanza assoluta non cambia. Questo inconveniente può essere mitigato dalla trasformazione dei dati in rapporti. Le trasformazioni di rapporto assicurano che i rapporti tra due caratteristiche siano gli stessi, sia che i dati siano conteggi assoluti sia che siano proporzioni. Se poi si prosegue facendo il logaritmo di questi conteggi, i dati risultano ulteriormente simmetrici e linearmente correlati. I coefficienti di correlazione risultanti sono quindi coerenti dal punto di vista compositivo, cioè il rapporto logico di due taxa è completamente indipendente dagli altri taxa [3].

I metodi basati sulla correlazione di solito non riescono a distinguere tra associazioni dirette e indirette. Per tenerne conto, sono stati sviluppati numerosi metodi che sfruttano la dipendenza condizionale. Solitamente presentano una complessità computazionale e tempi di esecuzione più elevati rispetto ai metodi basati sulla correlazione. La correlazione parziale e i relativi approcci vengono qui utilizzati per distinguere tra interazioni dirette e indirette, ottenendo un grafo pesato indiretto in cui gli spigoli implicano la dipendenza condizionale tra due taxa [3].

Infine, association rule minig è un metodo ibrido basato sulla correlazione di Pearson e sui metodi grafici in grado di identificare le relazioni dirette e indirette, lineari e non lineari.

3.2 Meta-network

Meta-network è il tool proposto da Yang et al per studiare le comunità microbiche. L'obiettivo è quello di affrontare le nuove sfide introdotte dall'avvento della grossa mole di dati metagenomici costruendo una rete di co-occorrenza che sia in grado di individuare non solo le relazioni dirette e lineari tra le coppie di batteri, ma anche quelle indirette e non lineari. Per farlo introducono prima il metodo loose definition e poi association rule mining.

Meta-Network genera matrici di indicatori di presenza-assenza per ogni campione. Successivamente, vengono calcolate le frequenze di co-occorrenza delle coppie di taxa, ottenendo una matrice di probabilità di co-occorrenza. Questa matrice viene utilizzata per costruire una rete con una probabilità di co-occorrenza dell'80% (soglia predefinita in Meta-Network). Seguendo questa loose definition Meta-Network utilizza un metodo basato sul grafo, il peso di somiglianza funzionale (Functional Similarity Weight, FS-Weight) per individuare le relazioni indirette e il metodo PCA-PMI (Partial Mutual Information filtrato con Path Consistency Algorithm) per dedurre associazioni non lineari. Questi due metodi (FS-Weight e PCA-PMI) sono in grado di catturare in modo indipendente molti nodi e archi, il che, secondo gli autori, indica che entrambi sono in grado di descrivere la natura complessa delle relazioni microbiche. [3]

3.2.1 Loose definition

Per misurare il numero di campioni nei quali coppie di specie coesistono, si introduce la probabilità di co-occorrenza. Per essere precisi, si converte prima la matrice di abbondanza originale in una matrice di presenza-assenza. Poi, per ogni coppia, si calcola il rapporto tra i campioni nei quali i due elementi sono entrambi presenti rispetto ai campioni totali: questa è la probabilità di co-occorrenza per ogni coppia. Quando la probabilità di co-occorrenza raggiunge una soglia definita dall'utente, viene calcolata la correlazione per la coppia (l'80% è utilizzato come valore predefinito).

3.2.2 FS-Weight per le relazioni indirette

I vicini di livello 1 interagiscono tra loro ed è probabile che partecipino ad alcuni percorsi comuni. Quindi hanno una maggiore probabilità di condividere alcune funzioni. Questa è la rilevanza biologica sottostante alla relazione diretta. Il concetto di relazione indiretta è diverso, ma non per questo meno intuitivo. I vicini di livello 2 interagiscono con alcuni elementi comuni. Quindi possono condividere alcune caratteristiche fisiche o biochimiche che consentono loro di legarsi a questi elementi. Più elementi comuni con

cui interagiscono, maggiore è la possibilità che condividano alcune funzioni. [8] Mentre le relazioni dirette sono state riportate in molti lavori esistenti nel contesto della comunità batterica intricata, è possibile rilevare le relazioni indirette utilizzando association rule mining.

Per questo motivo, nel nostro contesto, il metodo FS-Weight è applicato per rilevare le correlazioni indirette (Fig. 1, a.2). Il metodo FS-Weight misura la sovrapposizione tra le coppie, ed è originariamente progettato per stimare l'associazione tra correlazioni dirette e indirette basandosi sulla struttura della rete [6].

Per ogni coppia, il valore di FS-Weight viene calcolato in due fasi. In primo luogo, si calcolano le correlazioni dirette utilizzando una soglia definita dall'utente per il valore dell'arco. In questo lavoro, è stata scelta la correlazione del coefficiente di Pearson. In secondo luogo, si applica la rete costruita da FS-Weight per filtrare le correlazioni meno affidabili e per aggiungere relazioni indirette significative [6].

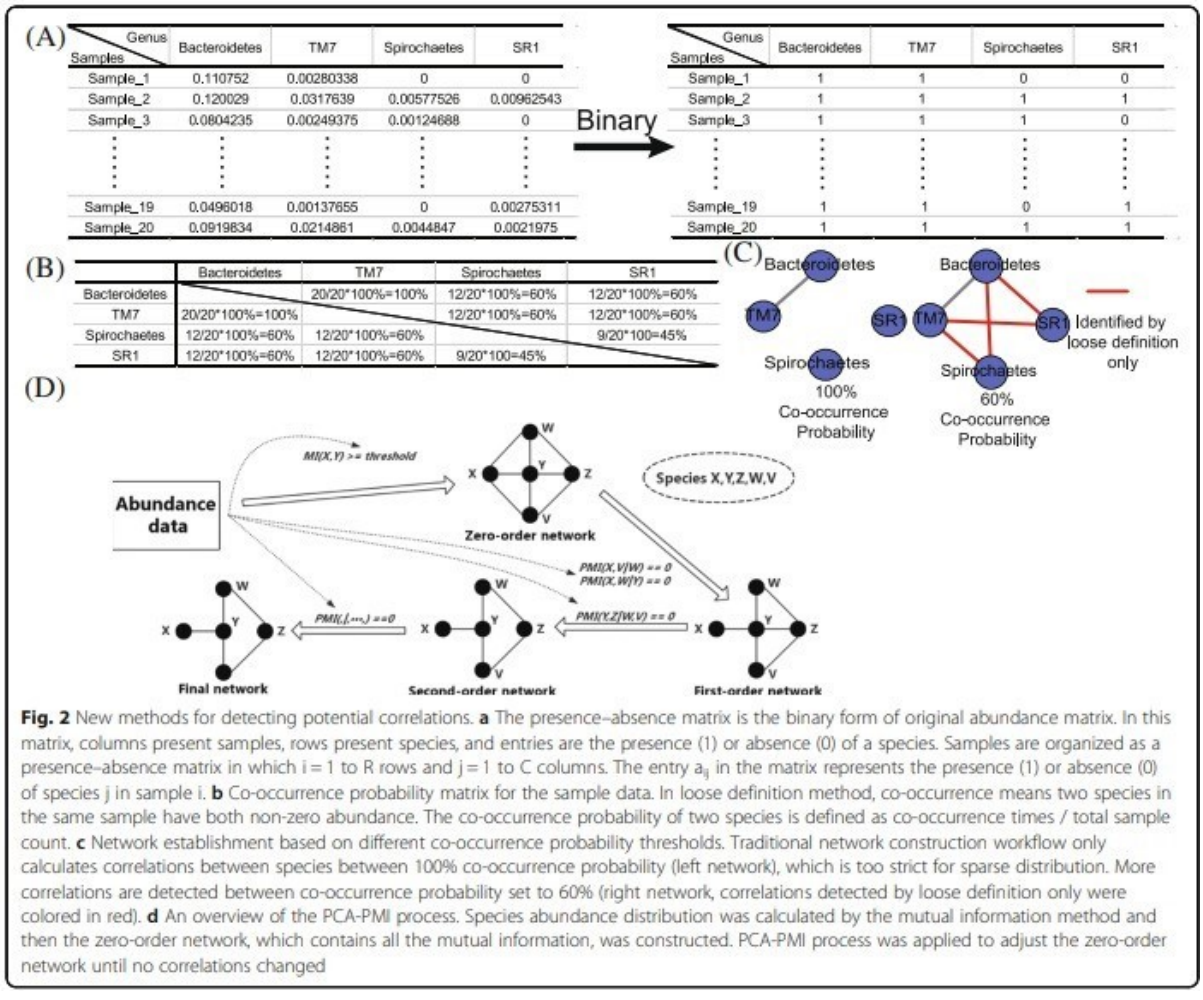

Fig. 2 New methods for detecting potential correlations. **a** The presence–absence matrix is the binary form of original abundance matrix. In this matrix, columns present samples, rows present species, and entries are the presence (1) or absence (0) of a species. Samples are organized as a presence–absence matrix in which i = 1 to R rows and j = 1 to C columns. The entry a_{ij} in the matrix represents the presence (1) or absence (0) of species j in sample i. **b** Co-occurrence probability matrix for the sample data. In loose definition method, co-occurrence means two species in the same sample have both non-zero abundance. The co-occurrence probability of two species is defined as co-occurrence times / total sample count. **c** Network establishment based on different co-occurrence probability thresholds. Traditional network construction workflow only calculates correlations between species between 100% co-occurrence probability (left network), which is too strict for sparse distribution. More correlations are detected between co-occurrence probability set to 60% (right network, correlations detected by loose definition only were colored in red). **d** An overview of the PCA-PMI process. Species abundance distribution was calculated by the mutual information method and then the zero-order network, which contains all the mutual information, was constructed. PCA-PMI process was applied to adjust the zero-order network until no correlations changed

Figura 2 Passaggi per passare dalla matrice delle abbondanze alla rete microbica finale

3.2.3 Il metodo PCA-PMI per le associazioni non lineari

Nelle comunità microbiche, anche le relazioni non lineari svolgono un ruolo importante. Pertanto, è stato adottato il metodo PCA-PMI (Path Consistency Algorithm-Part Mutual Information) per esplorare le correlazioni non lineari nei microbi. Il metodo PCA-PMI calcola l'informazione parziale per misurare la relazione lineare e non lineare per ogni correlazione a coppie nella comunità microbica. [6]

Il PMI(X,Y|Z) è il coefficiente di correlazione parziale e quantifica la dipendenza di X (valore target) dalla variabile di input Y che non è rappresentata dalla variabile di input Z. In altre parole, il PMI misura le informazioni aggiuntive che una nuova variabile fornisce a un modello di previsione preesistente; pertanto, il PMI è potenzialmente uno strumento adeguato a distinguere le variabili ridondanti.

Il PMI è definito nel modo seguente: assumendo che X e Y siano due variabili unidimensionali che rappresentano la distribuzione dell'abbondanza in tutti i campioni per le specie A e B rispettivamente, e Z una matrice n-2 dimensionale (n-2 > 0) che rappresenta la distribuzione dell'abbondanza delle altre specie in tutti i campioni, il PMI tra specie x e y dato il vicino indiretto z è definito come di seguito:

$$\mathrm{PMI}(x,\ y|z) = \sum_{x,y,z} p(x,y,z)\ \log\ \frac{p(x,y|z)}{p^*(x|z)p^*(y|z)} \quad (1)$$

Where the Part independence of species x and y given indirect neighbor z is defined as:

$$p^*(y|z) = \sum_x p(y|z,x)p(x) \quad (2)$$

$$p^*(x|z) = \sum_y p(x|z,y)p(y) \quad (3)$$

Nella PCA-PMI di ordine uno, una sola specie viene considerata come vicino indiretto tra due specie, e il Path Consistency Algorithm (PCA) è applicato per aggiustare la distribuzione delle correlazioni utilizzando una soglia di correlazione definita dall'utente (0,02 come valore predefinito). Questa soglia è stata impostata in base confronto tra diverse soglie. La matrice sparsa è definita come una matrice con un gran numero di nodi e una distribuzione rada delle specie, per cui è possibile utilizzare il metodo PCA per costruire la rete di co-occorrenze tra specie. Dopo aver calcolato tutte le correlazioni lineari e non lineari, la rete viene aggiornata rilevando gli ordini più elevati di PMI (più specie sono state impostate come vicine indirette). Gli ordini superiori di PMI vengono calcolati per controllare l'affidabilità delle correlazioni in modo iterativo fino a quando gli archi non si modificano più [6].

3.2.4 Identificazione di cluster e hub nella rete

L'individuazione di modelli di co-occorrenza e del loro significato biologico fornisce un importante materiale per comprendere la rete microbica. In primo luogo, si esegue il calcolo dei cluster e l'individuazione dei nodi hub sulla base del clustering di densità. Si applica l'algoritmo MCODE per individuare i potenziali cluster. L'allineamento dei membri del cluster rispetto a database di annotazioni tassonomiche o funzionali possono prevedere le potenziali funzioni dei cluster e le composizioni tassonomiche. Per identificare i nodi hub, calcoliamo i nodi più connessi come candidati per i nodi hub. Questi nodi sono selezionati per eseguire il test di Kruskal-Wallis per decidere se sono o meno nodi hub (verifica delle differenze nella distribuzione di rete prima e dopo l'eliminazione del nodo hub). Infine, si interpretano i cluster e i nodi hub sulla base di database tassonomici e funzionali e sulla base della letteratura [6].

4 Conclusioni

Il New Generation Sequencing costituisce la base per comprendere i ruoli svolti dai microrganismi nei loro ambienti. La matrice di abbondanza che fornisce in output il NGS è il punto di partenza per lo sviluppo di molti metodi di co-occorrenza, che permettono di ricostruire le interazioni microbiche tra gli elementi del microbioma. Tuttavia, i metodi ora disponibili non sono in grado di superare le sfide imposte dalla ricerca sul microbioma, come i bias composizionali. Meta-network è capace di individuare anche le relazioni indirette e non lineari e, una volta applicato al microbioma dell'intestino umano, ha permesso di scoprire la relazione nascosta tra due batteri intestinali, Syntrophomonas e Methanogens, che svolgono un importante ruolo nella trasformazione dell'acido grasso in metano ed energia. Meta-network però, non affronta i bias composizionali che seguono inevitabilmente dall'impiego di alcuni suoi metodi. Molte questioni biologiche rimangono irrisolte, ma appaiono promettenti le premesse e i traguardi raggiunti negli ultimi anni. Presto si potrà catturane profondamente la complessità delle interazioni microbiche.

Bibliografia:

[1] Richa Bharti, Dominik G Grimm, Current challenges and best-practice protocols for microbiome analysis, *Briefings in Bioinformatics*, Volume 22, Issue 1, January 2021, Pages 178–193, https://doi.org/10.1093/bib/bbz155

[2] Cappellato Marco, Baruzzo Giacomo, Patuzzi Ilaria and Di Camillo Barbara, Modeling Microbial Community Networks: Methods and Tools, Current Genomics 2021; 22(4) . https://dx.doi.org/10.2174/1389202921999200905133146

[3] Monica Steffi Matchado, Michael Lauber, Sandra Reitmeier, Tim Kacprowski, Jan Baumbach, Dirk Haller, Markus List,Network analysis methods for studying microbial communities: A mini review,Computational and Structural Biotechnology Journal,Volume 19, 2021, Pages 2687-2698, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2021.05.001.

[4] Illumina.com

[5] Asmita Kamble, Shriya Sawant, Harinder Singh, 16S ribosomal RNA gene-based metagenomics: A review Department of Biological Sciences, Sunandan Divatia School of Science, NMIMS Deemed to be University, Vile Parle (W), Mumbai, India https://www.brjnmims.org/article.asp

[6] Yang, P., Yu, S., Cheng, L. *et al.* Meta-network: optimized species-species network analysis for microbial communities. *BMC Genomics* 20 (Suppl 2), 187 (2019). https://doi.org/10.1186/s12864-019-5471-1

[7] Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape. F1000Research 2016;5:Oct. https://doi.org/10.12688/ f1000research.9050.2.

[8] Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics. 2006;22:1623–30.