UNIVERSITÀ DEGLI STUDI DI PADOVA

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN
INGEGNERIA BIOMEDICA

Studio di reti di interazione microbiche attraverso il metodo SparCC

Relatore:
CHIAR.MA PROF.SSA BARBARA DI CAMILLO

Laurenando:
JACOPO BALLIN
1219571

Correlatore:
DOTT. MARCO CAPPELLATO

Anno Accademico 2021/2022
Data 16/11/2022

Abstract

Le tecniche basate sul sequenziamento ad alto rendimento del gene 16S rRNA hanno il potenziale di chiarire quelli che sono i complessi meccanismi interni delle comunità microbiche naturali. Un passaggio fondamentale per comprendere tali dati è l'identificazione delle dipendenze tra i membri di queste comunità che viene ottenuta mediante l'analisi delle correlazioni. Tuttavia questi dati sono soggetti a delle problematiche matematiche come ad esempio il fatto che assumono la forma di frazioni relative di geni o specie piuttosto che assumere le loro abbondanze assolute. Questo porta ad avere dei risultati inaffidabili che comportano delle correlazioni errate tra i taxa ad esempio all'interno del microbioma umano. Analizzando dei dati simulati e reali si nota come tali effetti possono essere diffusi e gravi. Le correlazioni tra i taxa possono essere artefatte e le vere correlazioni possono anche essere di segno opposto. Per superare queste difficoltà hanno sviluppato un nuovo approccio, chiamato SparCC che è una nuova procedura adattata alle proprietà dei dati dell'indagine genomica che consente la derivazione di correlazioni tra geni o specie. Utilizzando infatti la rete SparCC come riferimento, si stima che l'approccio standard comporti una mancanza del 60% delle interazioni vere del microbioma umano. SparCC è quindi utilizzato per chiarire le reti di interazione tra le specie microbiche che vivono nel o sul corpo umano.

Indice

Capitolo 1

Dati di sequenziamento per il monitoraggio del microbiota

1.1 Cos'é il microbiota

Il microbiota è l'insieme dei microrganismi che vivono e caratterizzano uno specifico ambiente ed ecosistema, come l'intestino umano la bocca la gola, l'acqua o il suolo, senza danneggiarlo.
Il microbioma invece è la somma di tutto il patrimonio genetico di tutti i microrganismi presenti in un determinato organismo superiore. È quindi la totalità del patrimonio genetico che viene espresso dal microbiota.
I microrganismi quali funghi unicellulari, batteri, virus, protisti e archei vivono come delle comunità in ambienti complessi stabilendo delle interazioni tra di loro ma anche con l'ambiente che li ospita. Queste diverse comunità microbiche hanno ruoli fondamentali nell'ambiente e nella salute umana. Esse sono una componente essenziale in diversi habitat, quali aria, acqua suolo ma anche l'intestino di organismi sia semplici che complessi. Svolgono ruoli cruciali nei processi metabolici come il riciclaggio e la degradazione dei minerali, modulazione delle risposte immunitarie dell'ospite oppure produzione di vitamine.
Un passaggio fondamentale è quindi quello di comprendere le dipendenze tra i membri di queste comunità. Conoscere le interazioni microbiche e la loro composizione ci permette di comprendere gli effetti ecologici, ambientali e funzionali di un sistema microbico. Sebbene ci sia un apprezzamento sempre più crescente del ruolo di queste comunità all'interno dell'ambiente in cui abitano, possediamo una comprensione limitata delle loro interazioni e di come queste interazioni si combinano per generare dei comportamenti a livello di comunità.

La comprensione delle relazioni che si stabiliscono tra i vari membri ci permette di determinare e capire le cause dell'organizzazione di una determinata comunità. Vengono principalmente osservate e studiate due tipologie di relazione: quella microbica e quella ecologica. Grazie alle nuove tecnologie di Next-Generation-Sequencing siamo in grado di ricostruire l'intera composizione interna di una comunità batterica. I dati di sequenziamento che se ne ricavano possono quindi essere analizzati e studiati tramite metodi statistici e computazionali per dedurre dipendenze ed interazioni all'interno di una comunità. Le interazioni microbiche sono una serie di diverse tipi di relazioni che si stabiliscono tra i vari taxa. Un taxon (al plurale taxa) è un raggruppamento ordinato di organismi, che sono distinguibili morfologicamente dagli altri grazie a delle caratteristiche comuni.
L'interazione microbica che si sviluppa tra due taxa può produrre degli effetti ai rispettivi taxon che risulta essere:

- Positiva (+);
- Negativa (-);
- Neutrale (0);

Si possono quindi venire a formare 6 tipi di relazioni diverse:

- Interazione (+,+) : in questo caso si parla di mutualismo, una forma di simbiosi tra due specie che risulta essere vantaggiosa per entrambe. Ciò può avvenire grazie ad uno scambio di prodotti metabolici o ad una reciproca cooperazione nello svolgere una precisa funzione.
- Interazione (+,-) : in questo caso siamo di fronte a parassitismo o predazione, in cui un organismo è il parassita e l'altro l'ospite. In questo tipo di interazione il parassita trae vantaggio dall'ospite creandogli un danno.
- Interazione (+,0) : commensalismo, ovvero un'interazione in cui uno dei due organismi trae beneficio dall'altro senza provocarne alcun danno o aiuto.
- Interazione (-,0) : Amensalismo, ovvero una relazione in cui uno dei due organismi danneggia l'altro senza però trarne alcun beneficio o danno per se stesso.
- Interazione (-,-) : Competizione. Essa si verifica quando due specie si contendono una risorsa comune. È importante sottolineare che se questa risorsa è limitata

la specie dominante prevarrà nei confronti di quella più debole portandola all'estinzione o alla graduale spostamento di quest'ultima verso una nuova nicchia ecologica.

- Interazione (0,0) : Neutralismo; assenza di interazione.

L'interazione ecologica invece si verifica tra i taxa e l'ambiente. Un esempio importante di cui ho già accennato è ciò che si verifica nell'intestino umano in cui le cellule ospite vivono in simbiosi con il microbiota. Altri fattori determinanti sono l'età, lo stile di vita e la dieta che influenzano l'ambiente locale modificando così l'ambiente dove vivono i batteri. [1]
La ricerca microbica ha subito grandi cambiamenti e grandi rivoluzioni negli ultimi anni. Gli attuali metodi di coltura microbica in laboratorio riescono a replicare quelli che sono gli aspetti essenziali come il pH la temperatura, nutrienti e condizioni osmotiche che possono supportare la crescita di una piccola frazione della diversità microbica totale ma la maggior parte di essa rimane non coltivabile. Le tecniche microbiologiche convenzionali di isolamento e di arricchimento non riescono a supportare la crescita di tutti i microbi presenti in un campione. A causa di ciò, gli approcci microbiologici tradizionali di isolamento e caratterizzazione di nuovi microbi da una fonte ambientale sono ormai passati in secondo piano. I ricercatori non sono più molto interessati a comprendere una piccola percentuale della diversità microbica quanto più a comprendere e profilare l'intera e completa diversità. Il loro crescente interesse insieme ad un progresso degli strumenti e delle tecniche ha portato ad avviare molti progetti sia su scala locale e globale.
Un esempio molto importante è il progetto "Human Microbiome Project" avviato dal National Institutes of Health con l'intento di identificare e caratterizzare il microbioma umano, per apprezzare le diversità e la complessità delle comunità microbiche e il loro rapporto con lo stato di salute e di malattia dell'uomo. Per comprendere la diversità genetica e fisiologica umana si deve caratterizzare il microbioma e i fattori che vanno ad influenzare la distribuzione e l'evoluzione dei microrganismi costituenti il microbiota. Il risultato di queste scoperte ci fornisce un'ipotetica prospettiva sull'evoluzione umana contemporanea e quanto la trasformazione degli stili di vita umani e della biosfera, influenzi la "microevoluzione" degli esseri umani e quindi la predisposizione a varie malattie. Se conoscessimo le cause del perché una comunità microbica viene modificata si potrebbe agire sulla rete per mezzo di probiotici per ripristinare la corretta composizione della comunità. Comprenderne quindi le interazioni offre delle premesse per sintetizzare prodotti che contribuiscano al miglioramento della salute planetaria umana e anche animale. [6] [4]

Per analizzare e comprendere le comunità microbiche vengono utilizzati comunemente dei grafi che ci permettono di rappresentare relazioni di qualunque tipo, dove i nodi sono indicativi dei differenti membri che compongono la comunità e gli archi, che uniscono coppie di nodi, sono rappresentative dell'interazione che intercorre tra due nodi, nel nostro caso i vari taxa che compongono la rete. [1]
Di seguito andrò ad introdurre le tecniche di sequenziamento più utilizzate per ricavare i dati di abbondanza necessari a ricostruire le reti. Approfondirò poi un metodo di inferenza delle reti chiamato SparCC per la lettura e la ricostruzione delle reti di interazione microbiche

1.2 Tecnologie di sequenziamento

Grazie all'utilizzo delle tecniche di sequenziamento di nuova generazione l'analisi del microbioma di diverse specie è notevolmente migliorata. Grazie ad esse infatti riusciamo a comprendere meglio il ruolo metabolico, fisiologico e ecologico di molti microrganismi.
Il sequenziamento di nuova generazione (next-generation sequencing, NGS) ha aiutato ad identificare con precisione le specie microbiche e le vie metaboliche ad esse associate. Ha immensamente migliorato negli ultimi anni le aree di nuova previsione del genoma, di associazioni genetiche e di identificazione di agenti patogeni. Tuttavia ci sono ancora delle problematiche associate ai dati di sequenziamento che possono essere ad esempio la gestione errata del campione, la scelta dei metodi di estrazione del DNA e le analisi computazionali che potrebbero addirittura portare a dei risultati incoerenti o non confrontabili.
Le metodologie che vengono più comunemente utilizzate per l'identificazione microbica e la genotipizzazione ovvero il processo che permette di determinare quelle che sono le differenze nel corredo genetico o nel genotipo di un individuo sono principalmente basate:

- Su dei geni ampliconi/marcatori del gene;
- Shotgun sequencing

1.2.1 Shotgun sequencing

Con metagenomica ci si riferisce all'analisi genetica diretta dei genomi ottenuti da diversi ambienti. Permette di analizzare l'intero genoma dell'intera comunità di microrganismi presente in un campione. Essa cataloga in modo completo tutti i microrga-

nismi sia che essi sia coltivabili che non, in campioni ambientali complessi, riuscendo a profilare l'intera composizione tassonimica recuperando intere sequenze del genoma. Alcuni risultati degni di nota avuti grazie alla metagenomica sono ad esempio l'identificazione di phyla batterici ambientali con comportamento endosimbiotico, la scoperta della presenza diffusa di geni antibiotici nei batteri commensali intestinali o il tracciamento dei patogeni dell'epidemia umana.
Uno studio di "shotgun sequencing" si basa non sulla diversità di un singolo gene, ma sull'intero patrimonio genetico del campione e comprende le seguenti fasi principali (Figura 2.2): [3]

1. Disegno dello studio e del protocollo sperimentale. Questo passaggio viene solitamente sottovalutato ma è di vitale importanza;
2. Raccolta, elaborazione e sequenziamento dei campioni;
3. Pre-elaborazione computazionale delle letture di sequenziamento. Questo passaggio si pone come obiettivo quello di ridurre al minimo gli errori o gli artefatti di sequenziamento;
4. Analisi della sequenza per profilare le caratteristiche tassonomiche, funzionali e genomiche del microbioma;
5. Post-elaborazione statistica e biologica al fine di interpretare al meglio i dati;
6. Validazione e convalida.

L' obiettivo principale è quello di raccogliere una biomassa microbica sufficiente per il sequenziamento e ridurre al minimo la contaminazione dei campioni. Ad esempio fattori come la differenza tra il tempo di raccolta del campione e il tempo di congelamento/sgelo può influenzare in maniera determinante i profili della comunità microbica. La metodologia di estrazione del DNA deve essere efficace per i diversi taxa microbici o i risultati del sequenziamento possono risultare inaffidabili. Di seguito riporto quelle che sono le principali problematiche e sfide nella metagenomica: [3]

- "Entry-level access". È ancora molto costoso sequenziare e analizzare un gran numero di metagenomi senza un accesso a strutture di sequenziamento computazionali;
- Completezza dei cataloghi del genoma. Ovvero che gli strumenti computazionali metagenomici si basano sui genomi disponibili di facile coltivazione e risultano quindi essere influenzati;

- Distorsioni nella profilazione funzionale. La profilazione delle classi funzionali è impedita dalla mancanza di una convalidata letteratura per la maggior parte dei geni;
- "Materia oscura" microbica. Alcuni membri del microbioma potrebbero non essere stati caratterizzati con metodi basati su colture o metagenomica;
- Dilemma "vivo o morto". Dato che il DNA persiste nell'ambiente anche dopa la morte della cellula ospite i dati del sequenziamento possono non essere rappresentativi della comunità microbica attiva;
- "Maledizione della composizionalità". Le quantità di caratteristiche metagenomiche sono riportate come valori frazionari senza legami con la reale concentrazione assoluta.
- Sequenziamento del microbioma associato alla mucosa. I tessuti della mucosa umana sono delle interfacce cruciali tra i microbi e il sistema immunitario ma sequenziare il microbioma della mucosa con la shotgun sequencing è molto impegnativo in quanto è presente una frazione estremamente elevata di DNA umano e una frazione estremamente bassa della biomassa microbica

Ma la shotgun sequencing presenta anche numerose opportunità come ad esempio:

- Virome shotgun sequencing. È possibile rilevare organismi virali grazie alla shotgun sequencing;
- Progettazione di studi longitudinali.
- Validazione del microbioma dei biomarcatori. Dato che il microbioma dei biomarcatori è fortemente dipendente dallo studio svolto è fondamentale convalidare i biomarcaroti per migliorarne la riproducibilità;
- Condivisione dei dati e riproducibilità dell'analisi. La metagenomica deve ancora raggiungere il livello di standardizzazione caratteristico delle altre tecniche più consolidate.

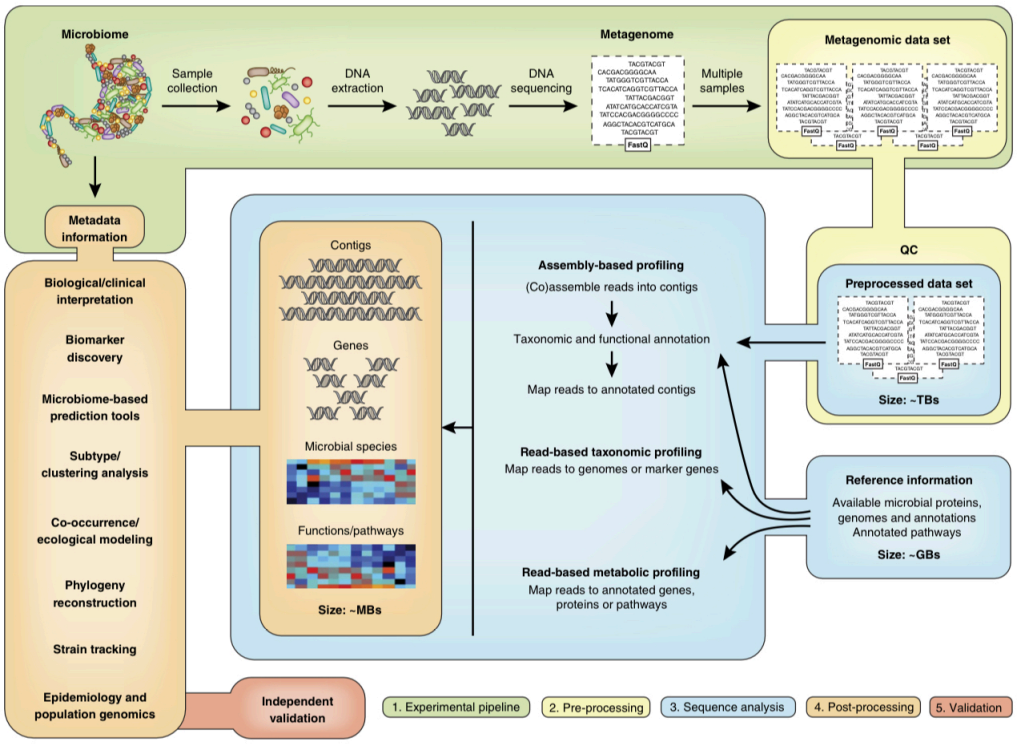

Figura 1.1: Shotgun sequencing

1.2.2 Sequenziamento dell'amplicone genico (16S rRNA)

Un amplicone è un frammento di DNA o RNA che è il prodotto di reazioni di amplificazione o di replicazione. Negli ultimi 25 anni, il sequenziamento dell'amplicone genico è stata la tecnica per analizzare e studiare la tassonomia di microbi complessi che erano in precedenza molto difficili da caratterizzare. Per batteri, funghi e archei vengono identificati geni marcatori che sono utilizzati per il sequenziamento degli ampliconi. Di seguito riporto alcuni parametri fondamentali da considerare durante la raccolta e la gestione del campione: [6]

- Contaminazione;
- Trasporto;
- Stoccaggio e sicurezza.

Una volta raccolto il campione il passaggio successivo riguarda l'estrazione dell'acido nucleico. È importante sottolineare che il metodo di estrazione deve catturare efficacemente tutti i tipi di microbi presi in esame. Ci sono due principali metodologie di estrazione:

- Lisi meccanica;
- Lisi chimica.

Isolare il DNA è un passo importante in quanto è il materiale di partenza che decide la qualità del risultato dei processi a valle. L'isolamento di una buona qualità e quantità di DNA è importantissimo per procedere coi processi di sequenziamento e analisi. Ci sono 3 passaggi chiave in un processo di isolamento:

- Corretta raccolta di campioni;
- Isolare il DNA intatto;
- Isolare il DNA privo di contaminanti.

La modalità con cui viene viene isolato il DNA deve essere scelta in modo che sia evitato un eccessivo taglio o degradazione del DNA. È preferibile una resa di DNA ad alto peso molecolare, in quanto può aumentare le dimensioni della popolazione in analisi. La resa a sua volta dipende dalla lisi cellulare scelta. Nella lisi diretta le cellule vengono lisate nel campione stesso per poi recuperare il DNA mentre in quella indiretta le cellule vengono prima separate dal campione e poi viene estratto il DNA. La maggiore purezza si ottiene con il metodo di lisi indiretta ma la modalità più comunemente utilizzata è quella di lisi diretta. Un altro punto fondamentale è quello di individuare il DNA proveniente da altri organismi che non siano quelli di interesse. Se ad esempio consideriamo un campione di pelle umana esso conterrà un piccola percentuale di DNA batterico e una grande quantità di DNA umano. Quindi il passo successivo sarà quello di eseguire una PCR per amplificare il gene bersaglio la quale comporterà delle difficoltà nella rimozione dei dati del DNA del gene di non interesse.
Nella stragrande maggioranza degli studi sulle analisi delle diversità delle popolazioni microbiche viene ampiamente utilizzato il sequenziamento del gene 16S rRNA. Esso contiene delle regioni condivise da tutte le specie e regioni ipervariabili che sono caratteristiche della singola specie, utilizzate quindi per indicare i diversi membri di una comunità microbica. Il gene 16S rRNA è presente in tutti i procarioti e ha più sottoregioni, chiamate V1-V9 che possono essere utilizzate per l'identificazione dei vari procarioti. Insieme a queste regioni ipervariabili, sono presenti delle regioni condivise dai procarioti che vengono utilizzate come siti di legame per i primer durante la fasi di amplificazione. La scelta del primer è fondamentale per la caratterizzazione della comunità batterica. Queste proprietà lo rendono utile per la classificazione e la separazione tassonomica. Per l'analisi della diversità microbica non è necessario sequenziare

l'intero gene ma è possibile utilizzare una singola o una combinazione di diverse regioni variabili. Ad esempio per l'analisi delle sequenze batteriche, le regioni V1-V4 sono quelle che hanno i tassi di errore più bassi e che mostrano i migliori risultati. A causa però dell'universalità della sequenza genica, l'amplificazione del gene 16S rRNA può portare a degli errori se sono presenti ad esempio contaminanti, amplificando anche prodotti indesiderati. In più nella maggior parte dei batteri ci sono più copie del gene 16S rRNA per cellule, e ciò può portare alla non comprensione delle proporzioni di microbi presenti in un campione.
Questa amplificazione mirata produce un gran numero numero di frammenti di 16S rRNA, gli ampliconi appunto, che vengono poi sequenziati usando piattaforme di sequenziamento di nuova generazione come Illumina oppure Ion Torrent, producendo una gran quantità di dati in maniera economica e veloce. Queste milioni di brevi sequenze, chiamate letture, vengono utilizzate per rilevare la presenza e l'abbondanza di diversi taxa nella popolazione originale. Successivamente le letture ottenute vengono pre-elaborate usando differenti software come QIIME, Mothur, i quali tramite delle tecniche di filtraggio riescono a scartare sequenze di coppie di base di breve e bassa qualità che non soddisfano i requisiti richiesti.
Per il sequenziamento di 16S rRNA letture fino a 200-250 coppie di basi sono sufficienti per la classificazione tassonomica. La lunghezza di lettura svolge un ruolo fondamentale negli studi di metagenomica perché identificare i livelli delle specie può essere difficoltoso con delle letture più brevi. Un altro parametro molto importante è il numero di volte che il genoma è stato sequenziato ovvero la Depth of Sequencing. Essa viene espressa in termini di milioni di letture da campionare. Con il sequenziamento 16S rRNA per identificare i diversi livelli di specie sono raccomandate circa 1000 milioni di letture. Successivamente diversi algoritmi eseguono il clustering delle letture nelle Unità Tassonomiche Operative (OTU), cluster di organismi raggruppati per somiglianza della sequenza del DNA. Il risultato finale dell'intero processo è una tabella contente i diversi OTU, in cui ogni elemento contiene il numero di volte in cui una lettura proviene da un particolare OTU, e la relativa tassonomia che descrive e cartterizza ogni OTU. Le matrici ottenute sono legate a caratteristiche biologiche e tecniche. I dati infatti non riflettono l'abbondanza assoluta ma piuttosto delle abbondanze relative. Concludo riportando che l'analisi NGS basata sul gene 16S rRNA ha contribuito ad identificare i cambiamenti nelle strutture della comunità e le alterazione delle funzioni associate alla comunità e ci ha permesso di acquisire una comprensione più profonda di diverse malattie associate all'intestino. Tuttavia rimangono ancora molte sfide aperte dovute soprattutto ad errori nella gestione dei campioni ed errori sperimen-

tali a valle. I fattori che influenzano la pre-elaborazione, così come la preparazione del campione e del sequenziamento dovrebbero essere catalogati con precisione in modo da migliorare l'affidabilità dei risultati. [6]

Capitolo 2

Il metodo SparCC per l'inferenza delle reti microbiche

2.1 Le reti di interazione microbica

Grazie alle tecniche di sequenziamento di nuova generazione descritte al capitolo 1 siamo in grado di caratterizzare efficacemente le comunità microbiche, sequenziando precisi bersagli molecolari come gli ampliconi del gene dell'RNA ribosomiale 16S per i batteri o utilizzando la shotgun metagenomics per i virus.
Gli approcci basati sulle reti si sono rivelati fondamentali per studiare le complesse interazioni che si sviluppano fra i vari microbi. Si trovano principalmente due categorie principali di studi sul microbioma: studi trasversali e studi longitudinali.
Gli studi trasversali coinvolgono più campioni in un momento specifico. Quella che si ottiene è una instantanea della comunità. Le reti che vengono dedotte da questi dati rappresentano relazioni a coppie tra le abbondanze di taxa. Gli studi longitudinali invece studiano le evoluzioni delle interazioni batteriche, anche correlate con i fattori esterni che perturbano il sistema, le reti che si deducono dunque rappresentano relazioni causa-effetto in una finestra temporale.
Ci sono tre principali problematiche legate alla decodifica delle relazioni di co-occorenza microbica:

- I dati del microbioma sono dati composizionali;
- La sparsità del set di dati può portare a false associazioni tra microrganismi
- È complesso distinguere tra associazioni indirette e dirette.

I primi metodi di inferenza delle reti non sono parametrici, in quanto non fanno alcuna ipotesi sulla distribuzione dei dati. I metodi basati sulle correlazioni, come ad esempio la correlazione di Pearson o Spearman, sono tra i metodi più comunemente utilizzati per studiare le interazioni microbiche. In questo approccio è importante il segno della correlazione (correlazione positiva o negativa). Un altro importante parametro è il Mutual Information (MI), un numero adimensionale che quantifica la dipendenza reciproca di due variabili casuali. Esse però non sono affidabili poiché producono delle associazioni artefatte e spurie, in quanto non riescono a tenere conto della composizionalità. Questo problema viene superato grazie ad una trasformazione del rapporto dei dati la quale assicura che il rapporto delle abbondanze tra due taxa sia lo stesso indipendentemente dal fatto che si considerano i dati come conteggi assoluti o proporzionali. Il rapporto quindi tra due taxa risulta essere indipendente dagli altri taxa. SparCC è un metodo molto popolare che impiega questa strategia.
Un'alternativa ai metodi di correlazione è quella basata sulla costruzione di modelli di regressione lineare regolarizzata in cui l'abbondanza di ogni taxon è modellata come una variabile di risposta utilizzando l'abbondanza di tutti gli altri taxa come variabili esplicative. Il coefficiente di ciascun taxon serve come misura lineare della forza di interazione tra due taxa [7]. Il problema di questi modelli è che sono inclini all'overfitting. Per ovviare a ciò si introduce un termine di penalità producendo modelli di regressione regolarizzati. Metodi che si basano su questi approcci sono ad esempio il metodo CCLasso o REBECCA
Il metodo Meta-Network, invece della regolarizzazione, utilizza un'estrazione avanzata delle regole di associazione per rilevare correlazioni complesse (Sia indirette che non lineari). Esso genera come prima cosa matrici di assenza-presenza per ogni campione. Successivamente, le frequenza di co-occorenza delle coppie di taxa sono composte producendo una matrice di probabilità di co-occorenza. Questa matrice viene viene poi utilizzata per costruire una rete con una probabilità di co-occorenza dell'80% (soglia definita dal metodo). Successivamente esso sfrutta due algoritmi per dedurre le associazioni non lineari e le relazioni indirette. Questi due metodi sono capaci entrambi in maniera indipendente, di catturare quei nodi e archi in grado di rappresentare la complessa natura della relazioni microbiche. [7]
I metodi basati sulla correlazione non riescono a distinguere tra associazioni dirette e indirette. Per ovviare a ciò si utilizza una correlazione parziale, che ci permette di ottenere un grafico nella quale gli archi implicano una condizione di dipendenza tra due taxa.
È importante sottolineare che la maggior parte dei metodi costruisce una sola rete di co-

occorenza non considerando le condizioni dello studio come la malattia, il trattamento, che in molti casi sono proprio le differenze tra quest'ultime l'oggetto di maggiore interesse. Per tenere conto di questi fattori sono stati sviluppati alcuni metodi per l'analisi differenziale della rete come ad esempio il Microbiome Differential Network Estimation (MIDNE). Il metodo EnDED si pone come obbiettivo quello di differenziare le associazioni dirette e indirette basate su fattori ambientali quali temperatura, salinità o i nutrienti, i quali possono influenzare l'intera dinamica dell'ecosistema. L'analisi delle reti in generale permette di ricavare preziose informazioni sulle reti di interazioni microbiche ma i metodi che sono attualmente disponibili e più utilizzati, non sono in grado di superare la moltitudine di sfide associate ai dati del microbioma, come il bias di composizione, la sovradispersione e le interazione trans-regno.
Il metodo che vado ad approfondire prende il nome di Sparse Correlations for Compositional data (SparCC). È un metodo molto popolare con applicazioni che vanno dallo studio del microbioma intestinale umano, a studi ambientali. Esso si basa su un approccio di approssimazione iterativo e utilizza dei dati log-trasformati per dedurre le correlazioni tra componenti. Queste trasformazioni di rapporto dei dati garantiscono che i rapporti tra due caratteristiche siano gli stessi, indipendentemente dal fatto che i conteggi siano assoluti o proporzionali.
In generale l'analisi delle reti fornisce preziosissime informazioni sulle reti di interazioni microbiche. Tuttavia i metodi attualmente disponibili non sono in grado di superare le sfide associate ai dati del microbioma come la composizione, la sovradispersione o le interazioni trans-regno. Devono essere condotti ulteriori studi per convalidare questi metodi utilizzando set di dati di riferimento universale.

2.2 Il metodo SparCC

2.2.1 Razionale di SparCC

Lo sviluppo di metodi di analisi dei dati ottenuti dalle tecnologie di sequenziamento ad alto rendimento come la profilazione del gene16S rRNA è ancora in corso. Le sfide principali riguardano in primis l'ottenimento di dati affidabili e informativi dalle sequenze geniche 16S rRNA, filtrando le letture e raggruppandole in maniera significativa. L'obiettivo è quello di identificare le correlazioni tra i taxa all'interno delle comunità ecologiche. Le sfide associate ai dati dell'indagine genomica (GDS) derivano principalmente dal fatto che esse sono una misura relativa, piuttosto che assoluta dell'abbondanza dei componenti della comunità. L'analisi dei dati ad esempio presi

dalle letture del gene 16S rRNA inizia in genere con una normalizzazione dei conteggi osservati per il numero totale di conteggi. Le frazioni ottenute da ciò rientrano quindi in una classe di dati chiusa e compositiva. Le frazioni poiché devono sommarsi a 1 non sono indipendenti e tendono ad avere correlazioni negative indipendentemente dalle vere correlazioni tra le abbondanze assolute. Pertanto le stime delle correlazioni riflettono quella che è la natura compositiva dei dati e non sono indicative dei processi biologici.
La composizione di ogni unità è descritta in termini di unità tassonomiche operative (OTU) e dato che sono disponibili solo abbondanze relative per ogni OTU, questi dati sono soggetti a potenziali errori come ho già sottolineato. Se si applicano infatti metodi basati sulle tecniche di correlazione standard ai dati dell'indagine del Human Microbiome Project (HMP), come ad esempio le correlazioni di Pearson per dedurre le reti, si nota che è presente un OTU correlato negativamente con altri OTU. Nonostante si provi ad attribuire un significato biologico a queste correlazioni esse derivano da un processo di normalizzazione. Il meccanismo alla base di queste correlazioni errate è dovuto al fatto che quando un OTU ha un'abbondanza media molto elevata, una variazione della sua abbondanza relativa ha un forte effetto sulle abbondanze del resto della comunità in quanto esse devono soddisfare il requisito che le abbondanze relative di tutte le OTU si sommino a 1. Quindi quando un OTU con una abbondanza molto elevata varia le abbondanze relative degli altri OTU variano anch'esse generando correlazioni artificiali negative con essa e correlazioni artificiali positive tra di loro. Gli errori dovuti a questi effetti composizionali possono essere limitati aumentando la diversità della comunità. Meno OTU compongono la comunità, peggiori sono gli effetti composizionali. Inoltre è importante sottolineare che anche se c'è una grande diversità ma con pochi OTU che dominano la comunità questi effetti possono risultare comunque molto gravi. Questo termine di diversità può essere quantificato utilizzando il numero effettivo di Shannon, il quale varia da 1 quando la comunità è dominata completamente da un singolo OTU al numero di OTU della comunità quando essi hanno uguale abbondanza. Per comprendere al meglio l'effetto della diversità sugli artefatti composizionali notiamo come in Figura 2 le correlazioni vere (che sono la Figura 2.1A-B-C) vengono recuperate solo quando la comunità è molto diversificata (Figura 2.1F). Quando invece il numero effettivo di Shannon si avvicina a 1 notiamo che le connessioni sono dominate da correlazioni negative con l'OTU più abbondante e sono presenti anche correlazioni positive i restanti OTU (Figura 2.1D-E). Questo effetto è talmente forte che trasforma le correlazioni negative tra OTU 4 e OTU 3 e 5 in una apparente correlazione positiva (Figura 2.1D) [8].

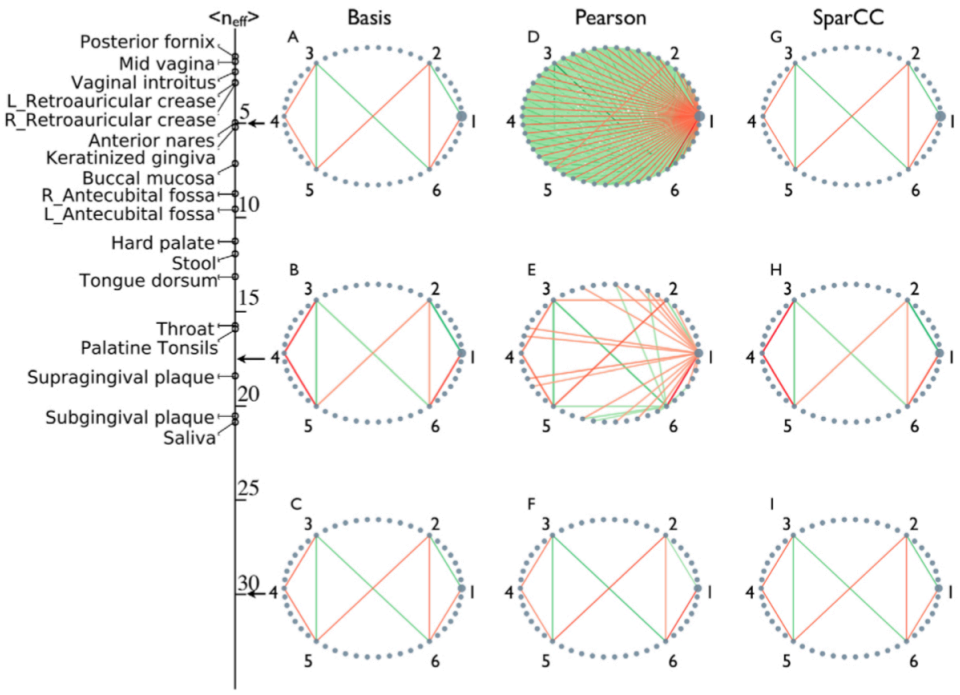

Figura 2.1: SparCC e Person a confronto

SparCC (Sparse Correlations for Compositional data) è una tecnica per dedurre le correlazioni dai dati composizionali. Stima le correlazioni lineari di Pearson tra i componenti log-trasformati. SparCC utilizza delle approssimazioni basate su :

- Il numero di diversi componenti è grande;
- La vera rete di correlazione è ”sparsa” (ovvero che la maggior parte dei componenti non è fortemente correlata)

SparCC non si basa su alcuna particolare distribuzione delle variabili di base, cioè le vere abbondanze della comunità possono seguire qualsiasi distribuzione. Il metodo può essere utilizzato non solo nel contesto dei dati del gene 16S rRNA, dove i componenti sono le OTU e le variabili di base sono le loro reali abbondanze nella comunità, ma può essere applicato a qualsiasi dato composizionale per il quale è valida la sua approssimazione.
SparCC si basa, come molte tecniche di analisi dei dati composizionali nella trasformazione:

$$y_{ij} = log(\frac{x_i}{x_j}) = logx_i - logx_j \tag{2.1}$$

dove x_i è la frazione di OTU i. Questa trasformazione porta dei vantaggi come:

- Le nuove variabili y_{ij} contengono le vere abbondanze di OTU, perché il rapporto delle frazioni è uguale al rapporto delle reali abbondanze;
- Il rapporto delle frazioni di due OTU è indipendente dalle altre OTU che sono inclusi nell'analisi
- Le nuove variabili y_{ij} inoltre sono libere di assumere qualsiasi valore reale, e utilizzando il logaritmo si rimuove il vincolo di positività introducendo (anti) simmetria nel trattamento delle variabili.

Per descrivere le dipendenze in un set di dati composizionale, si utilizza la teoria di Aitchison, il quale definisce la varianza del rapporto logaritmico come una metrica per quantificare la dipendenza tra due variabili compositive. Si introduce quindi la quantità:

$$t_{ij} \equiv Var\big[log\frac{x_i}{x_j}\big] = Var[y_{ij}] \tag{2.2}$$

dove la varianza è prelevata da tutti i campioni. $t_{ij} = 0$ vuol dire che tutte le OTU sono perfettamente correlate e il loro rapporto è costante, mentre il rapporto di OTU non correlate varia e il rispettivo t_{ij} è grande. Anche se t_{ij} contiene informazioni sulla dipendenza tra OTU, è una valore difficile da interpretare in quanto privo di una scala. Non è ancora chiaro cosa possa costituire un valore grande o piccolo (ad esempio $t_{ij} = 0,1$ può indicare una dipendenza forte, debole o nessuna dipendenza?). t_{ij} viene quindi messo in relazione con la nostra quantità di interesse, la correlazione tra le vere abbondanze delle OTU. La relazione viene data da:

$$t_{ij} = \omega_i^2 + \omega_j^2 - 2\rho_{ij}\omega_i\omega_j \tag{2.3}$$

dove ω_i^2 e ω_j^2 sono le varianze delle abbondanze di base log-trasformate delle OTU i e j e ρ_{ij} è la correlazione tra loro. Si può comprendere ora come t_{ij} può essere interpretato solo in relazione alle varianze delle abbondanze di base.

- $t_{ij} < \omega_i^2 + \omega_j^2$ indica una correlazione positiva
- $t_{ij} > \omega_i^2 + \omega_j^2$ indica una correlazione negativa

Noi vorremmo idealmente risolvere tutte 3 le equazioni descritte sopra per tutte le coppie di OTU e dedurre sia le varianze di base che le correlazioni. Tuttavia sono presenti più variabili sconosciute che equazioni e quindi questo il più delle volte non è possibile. Si può tuttavia ottenere una buona approssimazione delle varianze se, in media, le OTU sono poco correlate. Ottenute le varianze di base, queste possono essere inserite nelle 3 equazioni, per dedurre le correlazioni tra ogni coppia di OTU, che non

devono essere piccole a differenza delle correlazioni medie. Per ottenere una stima più accurata si itera questo procedimento. Ad ogni iterazione la coppia di OTU con una correlazione più forte identificata nell'iterazione precedente viene esclusa dalla stima della varianza di base. Questo rafforza la sparsità tra le coppie rimanenti producendo una migliore stima della varianza e della correlazione.
Per applicare SparCC le frazioni delle OTU devono essere stimate dai conteggi osservati. La normalizzazione di ogni OTU dai conteggi totali nel campione non è però affidabile per le OTU più rare perché sopravvaluta il numero di frazioni zero. Questo può dare origine ad artefatti che sono guidati dalla profondità del sequenziamento. Questi artefatti ti portano a sottocampionare i dati in modo che tutti i campioni abbiano gli stessi conteggi totali, e tuttavia il sottocampionamento non elimina o riduce gli artefatti dovuti agli effetti composizionali e richiede anche l'eliminazione di una parte sostanziale di dati disponibili. Pertanto si utilizza un approccio Bayesiano per stimare le frazioni dei componenti, che ci consente di una valutazione della robustezza dell'analisi a valle e l'assegnazione di valori affidabili.
Come si può notare ancora in Figura 2.1G-I SparCC è molto accurata nel dedurre correlazioni anche quando i dati composizionali sono altamente problematici perché dominati da una singola OTU. Per valutare al meglio SparCC vediamo come essa si comporta con più set di dati di diversa diversità e densità. Per farlo misuriamo la densità come la media delle correlazioni di Pearson tra le OTU, in modo tale che i set di dati più densi abbiano OTU più fortemente correlati, venendo meno all'ipotesi di sparsità utilizzata da SparCC. Per ogni combinazione di densità e diversità, più reti di vere correlazioni sono state assegnate e i corrispondenti dati sono stati campionati. Le reti dedotte da SparCC o dalle correlazioni standard sono state valutate utilizzando la radice dell'errore quadratico medio (RMSE Root Mean Square Errore) (Figura 2.2). Le tecniche standard hanno fornito stime ragionevoli solo per reti molto diverse e sparse (RMSE $\sim$ 0,02), mentre per reti con diversità simili a quelle osservate dall'HMP, l'RMSE era troppo elevato raggiungendo valori di 0,5. Al contrario le prestazioni di SparCC hanno dato risultati migliori per tutti i valori dei parametri, anche per reti molto dense in cui è violata l'ipotesi di sparsità. La peggiore precisione raggiunta (RMSE 0,02) è paragonabile alla migliore precisione raggiunta utilizzando le correlazioni standard su campioni altamente diversi. Utilizzando SparCC si scopre come in media i 3/4 delle coppie correlate di OTU usando Pearson sono false e che i 2/3 delle reali coppie di OTU correlate non sono state identificate. Inoltre usando SparCC si osserva una maggiore probabilità tra taxa filogeneticamente correlati, una scoperta che sembra sostenere un ruolo nelle dinamiche comunitarie neutrale, poiché è probabile

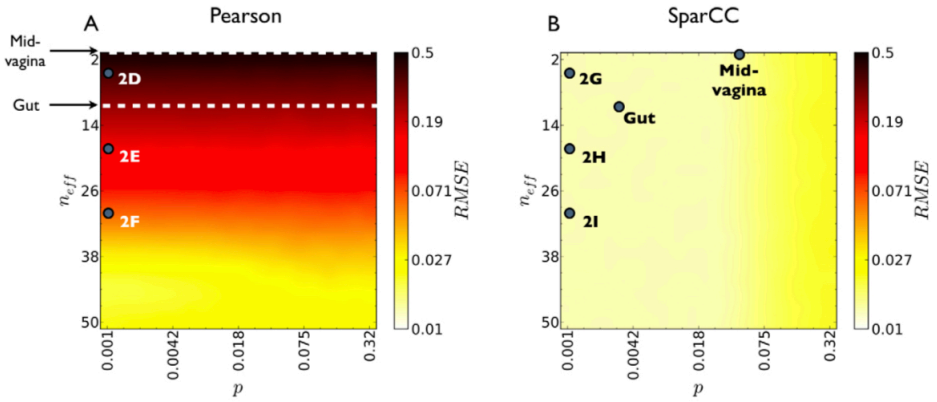

Figura 2.2: SparCC performa meglio dei metodi di inferenza standard.

che organismi correlati abitino nicchie simili, ma non sembra ci sia una dominanza di competizioni esclusive. Per riassumere quindi la diversità delle specie e la densità delle interazioni sono i due fattori più importanti che determinano la gravità degli effetti composizionali sulle stime delle correlazioni. I dati con un'alta densità e una bassa diversità sono i dati più complicati da comprendere utilizzando i metodi standard. Un altro problema che si riscontra è legato alla preponderanza dei valori zero. Questi zeri possono rappresentare dei componenti che sono realmente assenti nella comunità, o componenti rari che non erano presenti nel campione estratto dalla comunità. Queste due opzioni sono indistinguibili, e sta al ricercatore l'interpretazione di tali dati e la scelta del metodo di analisi. [8]
I limiti principali legati al metodo SparCC sono: Il fatto che esso si basa sull'avere un numero di componenti affidabile;
Le correlazioni stimate misurano la relazione lineare tra le abbondanze log-trasformate. Metodi per dedurre dipendenze più generali tra i componenti, come correlazioni di rango e informazioni per i dati non composizionali non sono ancora stati sviluppati;
Per ultimo SparCC non è in grado di mettere in relazione i modelli rilevati all'interno di una comunità coi fattori esterni, rilevando modelli temporali all'interno e tra le comunità.

2.2.2 L'algoritmo SparCC

Come abbiamo già visto la quantità:

$$t_{ij} \equiv Var\left[log\frac{x_i}{x_j}\right] \tag{2.4}$$

contiene le informazioni sulle dipendenze tra i componenti i e j e può essere correlata con le correlazioni di base tramite la relazione:

$$\begin{aligned} t_{ij} &\equiv Var\left[log\frac{x_i}{x_j}\right] = Var\left[log\frac{w_i}{w_j}\right] = Var[log(w_i) - log(w_j)] \\ &= Var[log(w_i)] + Var[log(w_j)] - 2Cov[logw_1, logw_j] \\ &\equiv \omega_i^2 + \omega_j^2 + 2\rho_{ij}\omega_i\omega_j \end{aligned} \tag{2.5}$$

in cui ω_i^2 e ω_j^2 sono le varianze delle variabili di base i e j log trasformate e ρ_{ij} la correlazione tra loro. Ciò che si deve fare è ora utilizzare l'equazione 2.5 per riuscire a dedurre la matrice di covarianza dalle variabili di base log trasformate Ω, dalla matrice di Aitchison $\mathbf{T}$, i cui elementi sono t_{ij}. Poiché però le varianze di base sono sconosciute a priori, non è possibile risolvere il problema in quanto sottodeterminato. Sono necessarie almeno 4 componenti per rilevare le deviazione dalla completa indipendenza tra tutte le componenti. SparCC per ottenere una risoluzione al problema utilizza un'approssimazione che è valida quando si hanno molti componenti scarsamente correlati tra di loro. L'equazione 2.5 può essere riscritta per ottenere:

$$\rho_{ij} = \frac{\omega_i^2 + \omega_j^2 - t_{ij}}{2\omega_i\omega_j} \tag{2.6}$$

che può essere risolta per ottenere le correlazione di base avendo le varianze di base. Per stimare le varianze di base si procede utilizzando le seguenti approssimazioni. Come prima cosa si calcola la variazione del componente i come:

$$\begin{aligned} t_i &\equiv \sum_{j=1}^{D} t_{ij} + \sum_{j\neq i} \omega_j^2 - 2\sum_{j\neq i} \rho_{ij}\omega_i\omega_j \\ &= d\omega_i^2\left[1 + \frac{1}{d}\sum_{j\neq i}\frac{\omega_j^2}{\omega_i^2} - 2\frac{1}{d}\sum_{j\neq i}\rho_{ij}\frac{\omega_j}{\omega_i}\right] \\ &\equiv d\omega_i^2\left[1 + \langle\left(\frac{\omega_j}{\omega_i}\right)^2\rangle_i - 2\langle\rho_{ij}\frac{\omega_j}{\omega_i}\rangle_i\right] \end{aligned} \tag{2.7}$$

dove $d \equiv D-1$ (D varianza) e $\langle . \rangle_i$ invece rappresenta la media su tutte le coppie che coinvolgono il componente i. Ora si assume che i termini di correlazione nell'equazione 2.7 siano piccoli, ovvero:

$$1 + \langle\left(\frac{\omega_j}{\omega_i}\right)^2\rangle_i >> 2\langle\rho_{ij}\frac{\omega_j}{\omega_i}\rangle_i \tag{2.8}$$

in modo così da trascurarli, avendo quindi ottenuto un insieme approssimativo di equazioni:

$$t_i \simeq d\omega_i^2 + \sum_{j \neq i} \omega_j^2 \quad , i = 1, 2, ..., D \tag{2.9}$$

infine, risolvendo l'equazione 2.9 si ottiene un'approssimazione delle varianze di base da inserire nell'equazione 2.6 ottenendo i valori delle correlazioni di base. Se si considera il caso in cui tutte le variabili di base hanno la stessa varianza l'equazione 2.8 si semplifica in:

$$1 >> \langle \rho_{ij} \rangle_i \tag{2.10}$$

ovvero si suppone che siano piccola la media della correlazioni, piuttosto che richiedere che una particolare correlazione sia piccola. Utilizzando queste approssimazioni la procedura di inferenza di base si sviluppa nei seguenti 5 punti:

1. Stimare le frazioni dei componenti per tutti i campioni per ottenere la matrice delle frazioni **X**
2. Calcolare la matrice di variazione **T**
3. Calcolare le variazione dei componenti $\{t_i\}$
4. Risolvere l'equazione 2.9 per ottenere un valore per tutte le varianza di base $\{\omega_i\}$
5. Inserire le varianze di base nell'equazione 2.5 per ottenere le correlazioni di base $\{\rho_{ij}\}$

2.2.3 La versione iterativa dell'algoritmo

La procedura di inferenza di base può essere migliorata utilizzando il seguente schema iterativo:

1. Stimare le correlazione di base come descritto sopra
2. Identificare la coppia più fortemente correlata. Se il suo valore supera una un determinato valore di soglia, aggiungere tale coppia all'insieme delle coppie escluse. Altrimenti terminare la stima
3. Identificare i componenti che formano solo coppie escluse ed escluderle dall'analisi. Se tutti i componenti sono stati esclusi tranne 3 terminare la procedura perché l'ipotesi di sparsità viene violata
4. Se sono stati esclusi componenti ristimare le frazioni dei componenti rimanenti.

5. Calcolare la variazione dei componenti $t_1^{(n)}c$, escludendo tutte le coppie fortemente correlate. Ovvero se $c_i^{(n)}$, è l'insieme degli indici dei componenti identificati con una correlazione forte con i alla precedente, n^{th} iterazione allora:

$$t_i^{(n+1)} = \sum_{j \notin c_i^{(n)}} t_{ij} \tag{2.11}$$

6. Utilizzare le variazioni appena trovate per calcolare le correlazioni di base con nei punti 4 e 5 della procedura di inferenza di base descritta sopra

7. Ripetere i passaggi da 2 a 6 per un numero di iterazioni o fino a che non vengono identificate nuove coppie fortemente correlate.

2.2.4 Diagramma di flusso

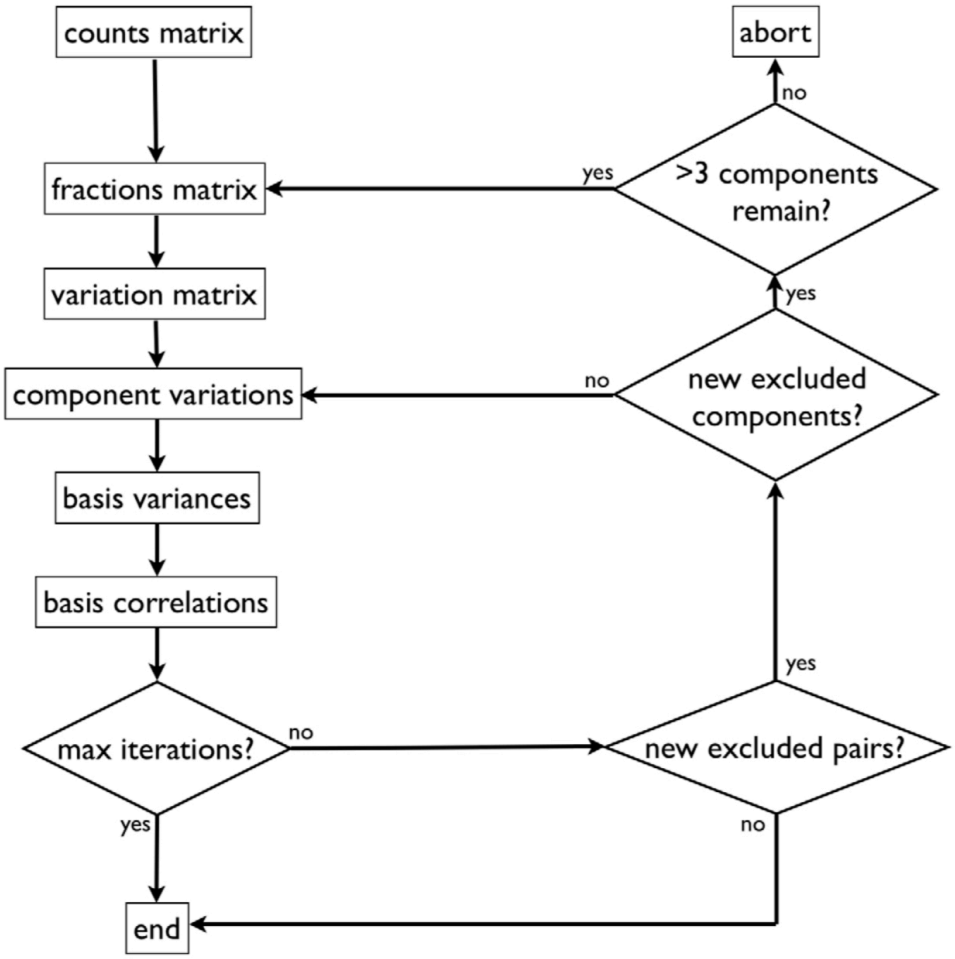

Figura 2.3: Diagramma di flusso [8]

Capitolo 3

Codice e Risultati

3.1 Analisi dei dati

Ho utilizzato SparCC in dei dati presi da un dataset reale. In particolare i dati sono stati presi da *Schubert et al* [9] articolo riguardante la disbiosi che si osserva in pazienti affetti da infezione da Clostridium difficile. Nell'articolo si sottolinea come l'utilizzo di antibiotici ha sì salvato molti pazienti da infezioni potenzialmente letali uccidendo l'agente patogeno di interesse, portando però anche, il più delle volte, a degli squilibri nelle comunità microbiche che si trovano ad esempio nel nostro tratto intestinale. Un microbioma intestinale intatto e privo di squilibri è fondamentale per fornire resistenza e protezione alla colonizzazione di *Clostridium difficile*. L'utilizzo di antibiotici e l'avanzare dell'età, sono entrambi fattori di rischio per una possibile infezione da *Clostridium difficile*. Analizzare quindi le differenze nei microbiomi dei soggetti con e senza *C. difficile* è fondamentale per comprendere i cambiamenti associati al microbioma con l'infezione batterica [9]. Tuttavia non è scopo di questa tesi approfondire le differenze che vengono a crearsi nelle diverse comunità, mi sono limitato a ricreare due grafi rappresentati ciascuno la rete di interazione della comunità microbica presente nell'intestino dei soggetti sani e dei soggetti malati in modo da poterle visivamente confrontare.
Il DNA batterico è stato estratto da ogni campione di feci utilizzando un kit di isolamento del DNA. La regione V35 del gene 16S rRNA è stata amplificata e sequenziata utilizzando la piattaforma di pirosequenziamento 454 GS FLX. Le sequenze sono state raggruppate in unità tassonimiche operative (OTU). I dati ottenuti presi da *Schubert et al* [9] sono rappresentativi di 246 soggetti di cui 92 malati e 154 sani e per ognuno ne è stata specificata l'età, il genere, il peso, l'etnia e la dieta. Abbiamo poi una matrice con le abbondanze di ogni batterio per soggetto contenente un totale di 166 batteri, matri-

ce da dare in input al metodo SparCC per ottenere i valori di correlazioni tra batteri. Infine una matrice delle tassonomie che per ogni taxa descrive la rispettiva tassonomia ovvero i livelli tassonomici ad esso associate.
Per prima cosa ho dovuto capire che tipo di dato dovessi dare in input al metodo SparCC. Esso richiedeva in input una matrice coi vari soggetti nelle righe e con i valori delle rispettive abbondanze dei batteri nelle colonne, e richiedeva che questi dati fossero delle stringhe convertibili in numero. Allora ho proceduto a fare un primo codice (3.1) che andasse a rimuovere per ogni dato di abbondanza i caratteri che non fossero numeri.

```
import numpy as np
in_file = open('otu_table.txt', 'r')
matrix = []
for line in in_file:
    numbers = line.split()
    chars = '\"s_tax'
    for i in range(len(numbers)):
        if numbers[i] =='\"':
            numbers[i] = 0
        else:
            for x in range(len(chars)):
                numbers[i] = numbers[i].replace(chars[x],"")
        numbers[i] = int(numbers[i])
    matrix.append(numbers)
in_file.close()
a = np.array(matrix,dtype=int)
np.savetxt('matrix.txt',a,fmt='%d')
```

Listing 3.1: Primo codice

Ottenuto una matrice con i rispettivi dati di abbondanza dei batteri leggibile dal codice ho proceduto a quantificare le correlazioni utilizzando SparCC al sito: [10]. Una volta terminata la procedura di stima delle correlazioni ho calcolato gli pseudo p-values tramite una procedura di bootstrap che mi ha permesso di ottenere un centinaio di set di dati mescolati e per ciascuno ho successivamente eseguito SparCC. Ottenuti i valori di tutte le correlazioni calcolate dai set di dati mischiati, sono stato in grado di ottenere gli pseudo p-valori. Ho tenuto conto così solo delle coppie di Taxa che avevano valori di p che soddisfacevano il test di ipotesi. Il livello di significatività è stato impostato a $p = 0,05$, scartando quindi tutte le Taxa che avevano un valore di p minore o uguale.
Ho proceduto infine alla costruzione delle reti rispettivamente dei sani e dei malati. Le reti di interazioni sono state ottenute collegando tutte le coppie di Taxa che avevano una grandezza di correlazione maggiore di una determinata soglia. È stato utilizzato

un valore di soglia pari a $0,3$. La rete è ricostruita mediante l'utilizzo di una matrice di adiacenza in cui un 1 è indicativo di una correlazione e quindi della presenza di un arco tra i rispettivi Taxa mentre uno 0 è indicativo di una non correlazione tra i due Taxa. In questo modo ho scritto un codice in cui tramite la matrice delle correlazioni ho costruito una matrice di adiacenza in cui un valore pari a 1 rispecchia un valore di correlazione maggiore o uguale a $0,3$, altrimenti 0 (per comodità riporto solo il codice che mi ha permesso di ricavare la matrice di adiacenza per la ricostruzione delle rete dei sani, il codice per la rete dei malati è analogo) (codice:3.2).

```
import numpy as np
import matplotlib.pyplot as plt
import networkx as nx
file_correlazioni = open('cor_sparcc_sani_.txt','r',)
matrice_di_adiacenza = []
for line in file_correlazioni:
    correlazioni = line.split(',')
    righe_adiacenza = []
    x = 0
    for j in range(1,len(correlazioni)):
        if abs(float(correlazioni[j]))<=0.3:
            righe_adiacenza.append(0)
        else:
           righe_adiacenza.append(1)
    matrice_di_adiacenza.append(righe_adiacenza)
file_correlazioni.close()
matrice_di_adiacenza.pop(0)
matrice_sani = np.array(matrice_di_adiacenza,dtype=int)
np.savetxt('matrice_di_adiacenza_sani.txt',matrice_sani,fmt='%d')
```

Listing 3.2: Matrice di adiacenza

Ottenuta la matrice di adiacenza tramite la libreria di Python "Networkx" [11] sono stato in grado di ricostruire le rispettive reti le quali sono dei grafi in cui ogni nodo è rappresentativo di un batterio e la presenza di un arco è indicativo del fatto che i due batteri siano correlati. (codice 3.3)

```
G = nx.from_numpy_matrix(matrice_sani, create_using=nx.Graph)
G.remove_edges_from(nx.selfloop_edges(G))
nx.draw(G,pos=nx.circular_layout(G),node_size=40,node_color='
    green')
plt.savefig("rete_sani.png")
```

Listing 3.3: Realizzazione del grafo

Ho così ottenuto le due reti di interazioni microbiche. La figura 3.1 è rappresentativa della rete di interazioni dei soggetti sani mentre la figura 3.2 è rappresentativa della rete di interazione dei soggetti malati.

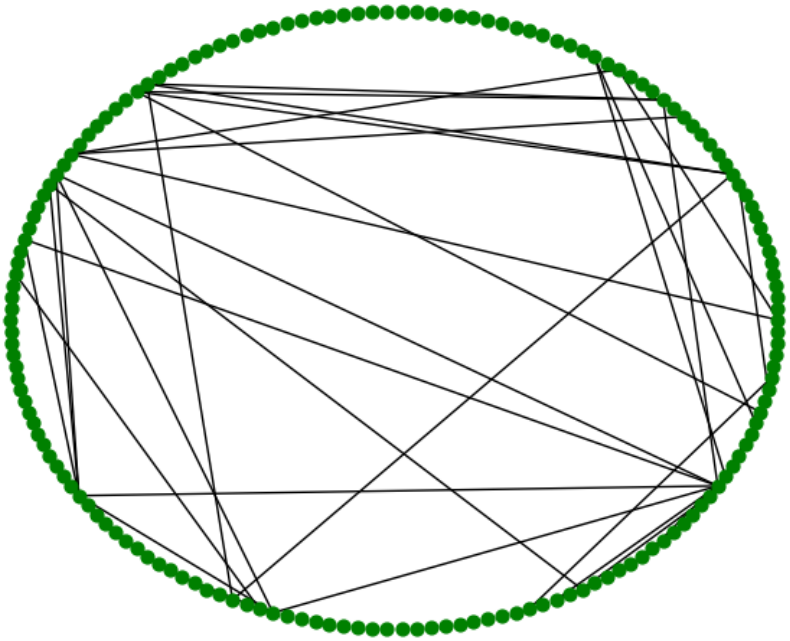

Figura 3.1: Rete soggetti sani

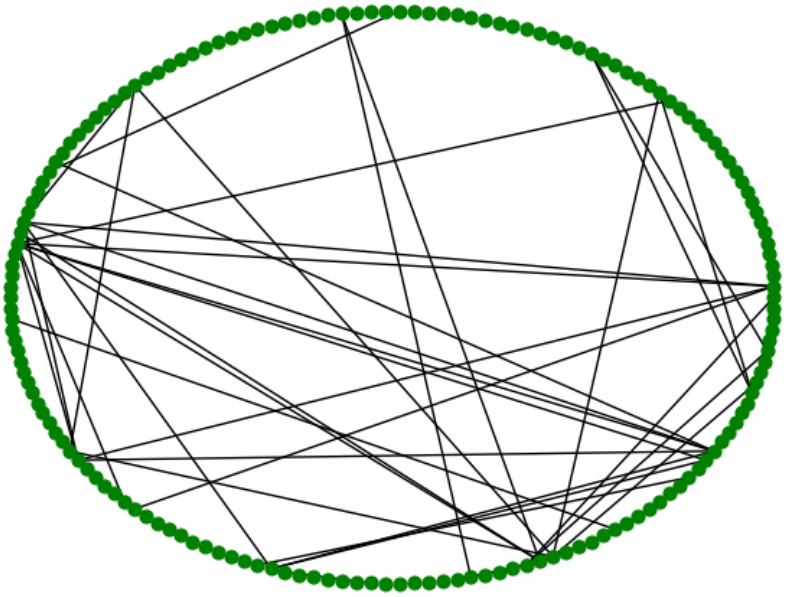

Figura 3.2: Rete soggetti malati

3.2 Considerazioni

Ottenute i due grafi essi possono essere analizzati con varie metriche, le quali sono indicatori numerici rappresentativi di una proprietà delle rete quali ad esempio la densità di un grafo che misura il rapporto tra il numero di archi del grafo rispetto al numero di coppie di nodi o più semplicemente quanti nodi o archi le due reti hanno in comune (Tabella 3.1). Queste metriche sono importanti per comprendere le diversità e le somiglianze tra le reti in modo tale da poter condurre degli studi per intervenire su esse. Si può ad esempio condurre un studio sulle reti in modo tale da fornire un potente strumento predittivo e terapeutico nel campo della salute umana. In questo caso si potrebbe ad esempio agire sulla rete per mezzo di probiotici per andare a ripristinare la corretta composizione della comunità. Abbiamo infatti capito come la moltitudine di batteri che si trovano nel tratto gastrointestinale siano responsabili del controllo della colonizzazione di agenti patogeni come infezioni da *Clostridium difficile*, e grazie alla teoria delle reti e alla modellazione statistica studi futuri saranno in grado di progettare terapie probiotiche adatte al mantenimento all'interno della comunità.

Grafo	Totale Archi	Archi in comune	Densità
Rete dei soggetti sani	68	16	0,00248
Rete dei soggetti malati (s)	92	16	0,00335

Tabella 3.1: Metriche di analisi

3.3 Conclusioni

Il lavoro di questa tesi è stato quello di sottolineare come i microrganismi abbiano costruito ecosistemi molto complessi in molti ambienti che spaziano dal suolo all'acqua fino a vari organi del corpo umano e di come comprendere la natura delle co-occorenze microbiche e i modelli delle correlazioni all'interno della comunità ci possano fornire informazioni su alcune malattie complesse. Tuttavia per poterle comprendere è necessario che si sviluppino modelli computazionali adeguati in modo da poter inferire le reti di interazioni microbiche e successivamente studiarla per comprendere il ruolo dei microrganismi e loro interazioni rispetto agli ambienti esterni. La scoperta delle interazioni microbiche è di particolare interesse soprattutto nella ricerca medica,

le comunità microbiche infatti sono dei bersagli farmacologici chiave nella medicina preventiva. Gli approcci basati sulle reti sono fondamentali per modellare e studiare queste relazioni. Abbiamo compreso quali sono i gravi effetti dovuti al bias composizionale e quali i suoi errori nella ricostruzione delle reti. Abbiamo capito quali sono i fattori chiave che influenzano la gravità di questi effetti, e di come una bassa diversità accompagnata a dei dati con un'alta densità siano i dati più complessi per i quali determinare le correlazioni con l'utilizzo dei metodi standard. Si è visto come SparCC non si basi su un'alta diversità e di come esso sia molto robusto anche quando l'ipotesi di sparsità viene violata. La scelta del metodo di analisi deve essere estremamente accurata e ponderata in base al set di dati che si ha a disposizione. Sono necessari degli sforzi maggiori per garantire affidabilità nella deduzione delle interazioni microbiche. È necessario avere un quadro di simulazione solido e affidabile infatti la mancanza di un'osservazione diretta di quella che è la verità biologica, complica la valutazione delle interazioni che vengono a svolgersi all'interno di una comunità. [1] Devono essere condotti un numero maggiore di studi per riuscire a convalidare i metodi di analisi delle reti utilizzando dei set di dati di riferimento universali. È necessario avere delle linee guida complete in modo da valutare sistematicamente le prestazioni dei modelli di rete esistenti. Raggiunto questo obiettivo saremo in grado di, tramite una conoscenza delle interazioni, manipolare e controllare il complesso mondo del microbiota.

Bibliografia

[1] Marco Cappellato, Giacomo Baruzzo, Ilaria Patuzzi, Barbara Di Camillo *Modeling Microbial Community Networks: Methods and Tools*, Current Genomics, 2020.

[2] David Lovell, Warren Müller, Jen Taylor, Alec Zwart, Chris HelliwellCaution! Compositions!Can constraints on omics datalead analyses astray, 2010

[3] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, Nicola Segata *Shotgun metagenomics, from sampling to analysis*, Nature Biotecnology, 2017

[4] Richa Bharti, Dominik G. Grimm *Current challenges and best-practice for microbiome analysis*, Briefings in Bioinformatics, 2019.

[5] Yili Qian, Freeman Lan, Ophelia S. Venturelli *Towards a deeper understanding of microbial communities integrating experimental data with dynamic models*.

[6] Asmita Kamble, Shriya Sawant, Harinder Singh *16S Ribosomal RNA Gene-Based Metagenomics: A Review*, Departement of Biological Sciences, Sunandan School of Science, 2020.

[7] Monica Steffi Matchado, Micheal 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, 2021.

[8] Jonathan Friedman, Eric J. Alm *Inferring Correlation Networks from Genomic Survey Data*, Plos Computational Biology, 2012.

[9] Alyxandria M. Schubert, Mary A. M. Rogers, Cathrin Ring, Jill Mogle, Joseph P. Petrosino, Vincent B. Young, David M. Aronoff, Patrick D. Schloss *Microbiome Data Distinguish Patients with Clostridium difficile Infection and Non-C. difficile-Associated Diatthea from Healty Controls*

[10] Codice SparCC: https://github.com/dlegor/SparCC.

[11] Libreria Netwotkx: https://networkx.org.

[12] Definizioni: https://www.treccani.it/vocabolario/

Ringraziamenti

Un ringraziamento speciale ai miei genitori, mia mamma Roberta e mio papà Michele, che mi hanno dato la possibilità di affrontare questo percorso, per avermi sostenuto, sopportato e supportato in questi tre anni.
Ringrazio i miei fratelli Giovanni e Marco per avermi ascoltato,e per avermi dato la forza nei momenti più difficili.

Ringrazio la Prof.ssa Di Camillo che mi ha concesso la possibilità e la disponibilità per questo lavoro di tesi.
Un grazie di cuore al Dottor Marco Cappellato per essermi sempre stato vicino in questo lavoro, per la sua disponibilità e soprattutto nell'aiuto a risolvere i problemi incoraggiandomi sempre a fare del mio meglio.

Un grazie al mio migliore amico Alberto che mi ha ascoltato e consolato nei periodi più bui, dandomi sempre la forza di andare avanti e ricordandomi sempre quanto sia importante il valore dell'amicizia.

Un grazie a Sofia per la compagnia, per i momenti felici ma anche i momenti brutti passati insieme, momenti che mi hanno fatto crescere e maturare interiormente, portandomi ad avere una maggiore consapevolezza di me.

Un grazie al mio gruppo di amici di WLJD, la mia seconda famiglia, per avermi alleggerito momenti di ansia e stress, per tutti i momenti felici e di festa passati insieme e per essermi stati sempre vicini anche se distanti e per aver condiviso anche questo importante momento.

Un grazie infine ai miei compagni di università Alice, Linda, Alessandro e Andrea per il sostegno e l'aiuto durante lo studio, per avere reso le lezioni più leggere e divertenti e per tutto l'ottimismo trasmesso.