UNIVERSITA' DEGLI STUDI DI PADOVA
DIPARTIMENTO DI BIOLOGIA

Corso di Laurea in Biologia Molecolare

ELABORATO DI LAUREA

SEQUENZIAMENTO COMPLETO DEL GENOMA DI *Cupriavidus necator* NH9 E RI-CLASSIFICAZIONE TASSONOMICA DEI CEPPI APPARTENENTI AI GENERI *Cupriavidus* E *Ralstonia*

Tutor: **Prof. Stefano Campanaro**
Dipartimento di Biologia

Laureando: **Filippo Fiorin**

A.A. 2021/2022

ABSTRACT

C. necator è un -proteobatterio Gram-negativo chemiolitotrofo facoltativo che ha suscitato negli ultimi decenni un discreto interesse da parte della comunità scientifica grazie alla grande varietà di processi metabolici che è in grado di svolgere e che lo rendono un organismo estremamente versatile in campo biotecnologico ed industriale.

C. necator fissando CO_2 produce naturalmente un biopolimero, il poli-(R)-idrossibutirrato, che viene stoccato in granuli citoplasmatici. L'interesse generale verso la possibilità di valorizzare commercialmente la CO_2 ha portato allo sviluppo di numerosi *tools* genetici per modifica ed incremento delle capacità metaboliche di *C. necator*, nonché all'ideazione di sistemi di *microbial electrosynthesis* (MES), che abbinano l'elettrolisi dell'acqua, ossidazione di H_2 e fissazione di CO_2 a fini produttivi.

L'elaborato prende in considerazione un articolo pubblicato circa il ceppo di *C. necator* noto come NH9, il quale trova impiego nella *bioremediation* in quanto in grado di catabolizzare composti notoriamente inquinanti e recalcitranti come i bifenili policlorinati (PCBs) e l'acido m-clorobenzoico. L'articolo mirava a caratterizzare geneticamente tramite sequenziamento *de novo,* al fine di definire meglio genomiche alla base delle sue attitudini metaboliche; per poi scoprire e quindi dipanare delle criticità tassonomiche circa la classificazione dei generi *Cupriavidus* e *Ralstonia.*

INDICE

1. INTRODUZIONE

1.1 *C. necator* OVERVIEW

Il genere *Cupriavidus* comprende microrganismi appartenenti alla famiglia *Burkholderiaceae,* parte di quelli che sono *β*-proteobatteri.
I *β*-proteobatteri sono un ordine di batteri Gram-negativi, a loro volta afferenti al Phylum *Pseudomonadota*. L'ordine dei *β*-protebatteri è in sé piuttosto variegato, con membri che possono essere eterotrofi, fotoeterotrofi, autotrofi e chemiolitotrofi. Nonostante esistano batteri di questa classe di notevole interesse clinico per l'uomo, come ad esempio *Neisseria meningitidis*, una certa frazione dei *β*-proteobatteri risulta d'interesse più biotecnologico.
In effetti, i membri di suddetta classe che hanno metabolismo autotrofo e chemiolitotrofo sono in grado di fissare una varietà di piccoli composti inorganici al fine di produrre tutti quei composti necessitano per le attività vitali. Questi composti inorganici possono anche essere composti inquinanti di cui si vuole limitare presenza e permanenza nell'ambiente e nel contempo, se possibile, valorizzarli. Tra i *beta*-proteobatteri di maggior interesse in tal senso vi sono quelli dei generi *Ralstonia* e, appunto, *Cupriavidus*.

C. necator fu individuato nel 1987 da Makkar e Casida, che hanno in effetti proposto il nome del genere, come un batterio naturalmente presente nel suolo e che risulta particolarmente favorito da elevate concentrazioni di rame nel mezzo di crescita (da cui il nome). Dal punto di vista morfologico, le cellule di *Cupriavidus* sono nell'ordine di pochi micrometri.
Makkar e Casida determinarono che il batterio identificato fosse per certo un Gram-negativo, aerobico, mesofilico (optimum a 27°C) e bastoncellare in grado di riprodursi per fissione binaria. Inoltre, tale batterio è risultato essere positivo ai test per catalasi e ossidasi, non emolitico.
L'osservazione al microscopio elettronico indica la presenza da 2 a 10 flagelli peritrichi che gli conferiscono motilità, grazie alla quale è in grado di esibire comportamenti predatori verso batteri sia Gram-positivi che negativi. Quando non preda, le richieste nutrizionali di *C. necator* risultano essere piuttosto semplici: il ceppo identificato da Makkar e Casida, noto come N-1, non utilizza il glucosio come fonte di carbonio preferenziale ma piuttosto il fruttosio e gli aminoacidi. Già nel 1987 fu notata la produzione di PHB nel contesto di un'ampia gamma di reazioni biochimiche eseguibili da *C. necator* N-1.

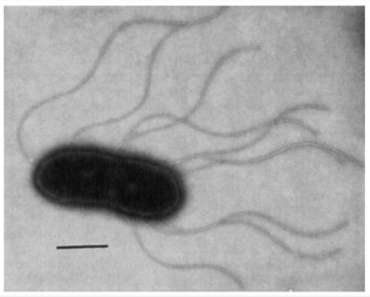

Fig. 1

Flagelli di C. necator N-1.
Tratta da Makkar e Casida.
Riferimento: 1um

Nel 2011 fu sequenziato completamente tramite *shotgun sequencing* il genoma dello strain N-1, per un totale di 8480857bp comprensivi di due cromosomi circolari (1 e 2) e di due plasmidi, pBB1 e pBB2, per un GC content del 57%. Il risultato fu quindi depositato in GenBank con i seguenti accession number:

- **Chr. 1:** CP002877
- **Chr. 2:** CP002878
- **pBB1:** CP002879
- **pBB2:** CP002880

Già con questo sequenziamento fu possibile, grazie alla similarità tra i genomi *C. necator* N-1 e *R. eutropha* H16 (ora classificata come *C. necator* H16), identificare in *C. necator* l'operone *cbb* deputato alla fissazione della CO_2 con anche diversi operoni codificanti per delle deidrogenasi. Furono annotati anche set di geni deputati al metabolismo dei poliidrossialcanoati, al catabolismo dei benzoati, fenoli e composti aromatici policlorinati.

1.2 IL QUESITO BIOLOGICO

L'articolo in esame prende in considerazione il ceppo della specie *Cupriavidus necator* NH9.
Già prima del 2011 era nota l'elevata similarità genetico-funzionale tra i generi *Cupriavidus* e *Ralstonia*. Le capacità metaboliche di degradazione di molti composti che risultano inquinanti recalcitranti e la possibilità di un metabolismo chemiolitotrofo facoltativo di questi generi li hanno resi nel tempo d'elevato interesse per la comunità scientifica, che ha portato all'identificazione di numerosi ceppi e allo sviluppo di tool d'ingegnerizzazione. Nella fattispecie, fu presto chiara la possibilità di utilizzare *C. necator* come bio-reattore per la produzione di bioplastiche (producendo naturalmente il PHB), e il successivo sviluppo delle competenze ha consentito di direzionare il suo metabolismo verso la produzione di bio-carburanti, con conseguente valorizzazione della CO_2.

La caratterizzazione del genoma degli strain di *C. necator* risulta pertanto fondamentale in un panorama ampio di applicabilità industriale e ambientale di suddetto batterio, come anche è fondamentale l'identificazione il più possibile univoca dei rapporti filogenetici esistenti tra le specie in grado di sostenere i processi metabolici d'interesse.
In effetti fu nel 2004 che Vandamme e Coenye si accorsero dell'elevata similarità esistente tra la sequenza 16S rRNA di un ceppo di *C. necator* (LMG 8453) e quella di diversi *β*-proteobatteri, e proposero una ri-classificazione dei generi allora esistenti secondo le evidenze raccolte. Nell'articolo viene esplicitamente detto che "*nonostante la rinominazione e la conseguente riclassificazione delle specie batteriche causi confusione e, non ultima, irritazione nella grande comunità microbiologica; aderire alle regole della nomenclatura è essenziale al fine di stabilire una vera tassonomia sistematica*".

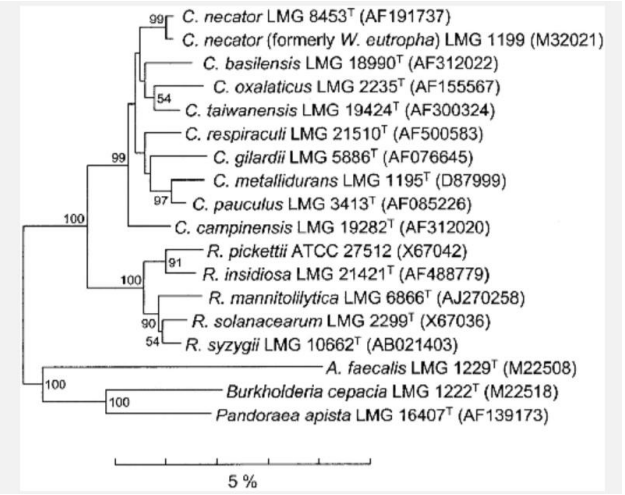

Fig. 2

Albero filogenetico non radicato basato sull'analisi del 16S rRNA, utilizzando A. faecalis come outgroup.
I numeri ai nodi rappresentano il supporto di bootstrap agli stessi generato con 1000 campionamenti con re-immissione.
Tratta da Vandamme e Coenye.

Nell'articolo in esame in effetti l'obiettivo iniziale era quello di caratterizzare dal punto di vista genomico i loci e gli operoni responsabili della particolare capacità dello strain di *C. necator* NH9 di catabolizzare acidi benzoici, inquinanti noti per recalcitranza ambientale.
C. necator NH9 fu isolato per la prima volta dal suolo in Giappone e per la sua identificazione si è optato per un sequenziamento genomico completo utilizzando la tecnologia SMRT PacBio RSII system, unita a sequenziamento con MiSeq Illumina. Per la validazione della scoperta, al sequenziamento fu inclusa un'analisi del 16S rRNA di *C. necator* NH9 essendo questo locus disovente utilizzato per le classificazioni tassonomiche. Nel corso dell'investigazione, furono notate delle inconsistenze nella classificazione proposta degli strain afferenti ai generi *Ralstonia* e *Cupriavidus*.

2. METODI

2.1 SMRT SEQUENCING, ASSEMBLY ED ANNOTAZIONE

Il sequenziamento "*Single Molecule Real Time*" (SMRT) è la tecnica di sequenziamento di terza generazione che, come suggerisce il nome, non richiede l'amplificazione delle molecole di acidi nucleici in fase di sequenziamento tramite PCR o sue varianti. I principali sequenziatori SMRT sono i sequenziatori PacBio e Oxford Nanopore.

In particolare, i sequenziatori SMRT dell'azienda Pacific Biosciences offrono una lunghezza delle sequenze generate (reads) maggiore di quella fornita dalla gran parte dei sequenziatori di seconda generazione, caratteristica che li rende adatti alla generazione di assembly di alta qualità di genomi e dei trascrittomi. Questa versatilità si riflette in particolar modo sullo studio delle regioni genomiche tipicamente difficili da risolvere, come quelle omopolimeriche e di DNA altamente ripetitivo, frequenti nei genomi degli eucarioti superiori. I sequenziatori PacBio si prestano molto bene anche alla caratterizzazione dei genomi microbici poiché la compattezza di quest'ultimi ne consente una fedele caratterizzazione in virtù della lunghezza delle sequenze generate. I dati generati con PacBio sono sovente, e anche nel caso dell'articolo in esame, accorpati a dati generati con sequenziatori di seconda generazione come Illumina, che di fatto garantiscono un'accuratezza superiore a livello di singola base e consentono di rifinire le bozze delle sequenze che si generano con le fasi iniziali dell'assembly (*draft sequences*).

In effetti, i sequenziatori PacBio è vero che generano sequenze di lunghezza superiore ed in tempi più brevi rispetto a quelle generate dai sequenziatori di seconda generazione, ma queste sono tipicamente soggette ad un tasso d'errore più alto. Il rendimento di uno dei più efficienti sequenziatori PacBio è più basso dei sequenziatori Illumina, che generano diverse migliaia di miliardi di paia di basi di sequenza tramite le tecniche paired-end più recenti. I tempi di generazione di queste sequenze sono però più alti nei sequenziatori Illumina, e possono richiedere più giorni.

Fig. 3

SMRTbell con adattatori circolarizzanti (verde). La polimerasi è ancorata alla ZMW. Tratta dal sito Oxford University Press.

I sequenziatori PacBio operano su dei templati chiamati SMRTbell, che sono essenzialmente molecole di DNA a doppio filamento alle cui

estremità sono legati degli adattatori a forcina, che vanno a “circolarizzare” la molecola, fornendo nel contempo una sequenza univocamente riconoscibile da una DNA-polimerasi. La DNA-polimerasi in questione è immobilizzata sul fondo di pozzetti chiamati ZMW (zero-mode waveguide).

Fig. 4

Dettaglio di una ZMW. Di 150000 ZMW presenti in una SMRT cell, 35000-75000 producono una sequenza in una run di 0,5-4h generando 0,5-1Gbp. Tratta dal sito Pacific Biosciences

Ogni cella ZMW è da intendersi come “unità di sequenziamento”, e ogni SMRT cell PacBio contiene 150000 ZMW. Una ZMW fornisce il minor volume possibile per la fluorescenza, e le polimerasi immobilizzate sul fondo di ciascuna ZMW sono in grado di legare gli adattatori di ogni SMRTbell ed avviare il sequenziamento per sintesi.
Dentro ogni ZMW infatti sono lasciati liberi di diffondere i quattro nucleotidi canonici marcati, la cui fluorescenza viene sempre misurata, a costituire un rumore di fondo. Qualora un nucleotide venisse incorporato dalla polimerasi nel filamento di nuova sintesi esso permarrebbe per un tempo superiore degli altri nella ZMW. Questo tempo di permanenza superiore alla media rende la fluorescenza del nucleotide univocamente marcato dal associato alla ZMW, così da determinare una nuova base della sequenza in analisi. Essendo il templato SMRTbell una molecola circolare, se la polimerasi è sufficientemente efficiente è possibile sequenziare entrambi i filamenti più volte in una singola run (ogni ciclo di sequenziamento di un filamento di templato è detta *pass*). Questo aumenta l’affidabilità delle sequenze generate poiché consente di generare delle sequenze consenso per ogni unità di sequenziamento. Ogni ZMW infatti fornisce quindi quello che viene chiamato “consenso circolare” (*Circular Consensus Sequence* o CCS). Se i sequenziatori PacBio generano sequenze con un tasso d’errore dell’11-15% per “*continuous long reads*” (CLR), la possibilità di eseguire numerosi passaggi è in grado di aumentare di molto l’accuratezza della sequenza generata (con 15 *passes* l’accuratezza è del 99%). In effetti, il numero di passaggi per SMRTbell e la lunghezza del CCS generato sono il risultato di un trade-off.

I dati generati dai sequenziatori PacBio ed Illumina nell'ambito del sequenziamento del genoma di *C. necator* NH9 sono descritti di seguito:

- PacBio RSII sequencing

  -prodotte 181370 "raw reads"

  -una volta filtrate dal software PreAssembler Filter v1 *per mantenere una qualità minima di 0,75 ed una lunghezza minima di 7,5Kbp si sono mantenute 86406 sequenze, con la mediana a 12.367bp e lunghezza massima a 41.609bp*

  -si sono quindi mantenute un totale di 1.050.061.719bp con circa 127-fold coverage

- Illumina MiSeq paired-end sequencing

  -prodotte 3.436.955 sequenze paired-end.

  -il filtraggio ("trimming") è stato eseguito su sequenze con Q-score inferiore a 15 e sulle sequenze degli adattatori

  -si sono quindi mantenute 2.305.131 sequenze paired-end per circa 138-fold coverage

N-fold coverage fa riferimento alla profondità di sequenziamento. Con esso si intende essenzialmente il numero di volte che le sequenze generate da un sequenziamento tendono a rappresentare lo stesso locus o, su scala genomica, il genoma stesso.

A seguito della generazione delle sequenze a partire da un processo di sequenziamento il cui obiettivo è caratterizzare un genoma prima solo parzialmente descritto, il naturale proseguimento delle operazioni è quello di assemblare il genoma *de novo.* L'assemblaggio dei genomi fa principalmente riferimento a tre strategie: la strategia "greedy", la strategia "string overlap" e quella basata sui grafi di de Bruijn.

Per l'assemblaggio del genoma di *C. necator* NH9 Moriuchi *et al.* hanno utilizzato un approccio descritto nel 2013 da Chin *et al.* definito HGAP, da Hierarchical Genome Assembly Process; che sfrutta la reiterazione di operazioni di assemblaggio eseguite software come Velvet che operano proprio coi grafi di de Bruijn.

Per grafo di de Bruijn si intende la rappresentazione grafica di un processo in cui le sequenze vengono indicizzate in *k-mers* di lunghezza *k*: da un set di sequenze è possibile generare un "nodo" per ogni *k-mer* che si presenta in suddette sequenze.

Per ogni *k-mer* si può quindi definire un suo suffisso ed un suo prefisso costituiti ciascuno da *k-1* elementi. Il prefisso mancherà dell'ultima posizione del *k-mer* ed il suffisso mancherà della prima.

Quindi, ogni *k-mer* potrà essere collegato direzionalmente ad un altro *k-mer* se quest'ultimo condivide il proprio prefisso col suffisso del *k-mer* iniziale. Create le connessioni si ricerca un percorso che copra tutti i nodi,

quindi tutti i *k-mer,* una singola volta. In alternativa, è possibile creare un grafo di de Bruijn in cui i vertici siano i prefissi ed i suffissi dei vari *k-mer*, così che siano i percorsi direzionati ("*edges*") a rappresentare i *k-mer* stessi: il percorso che prevede il passaggio una sola volta per ogni *edge* restituirà l'assembly. Questo elimina il vincolo posto dai passaggi per i nodi e abbassa il carico computazionale.
Risulta comunque evidente che la lunghezza dei *k-mer* è un parametro fondamentale per la bontà dell'assembly che un software operante tramite grafi di de Bruijn restituisce.

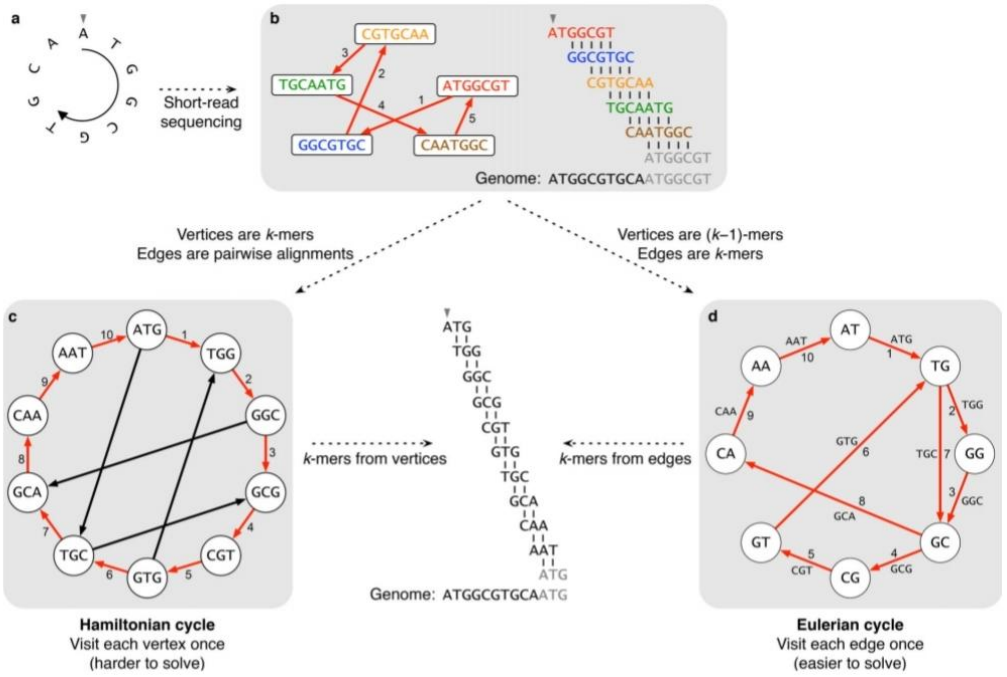

Fig. 5

Schematizzazione del processo di indicizzazione delle sequenze in k-mer e comparazione tra i processi di assembly tramite grafi di de Bruijn che ricercano cicli Hamiltoniani (passaggio per ogni nodo una singola volta) o Euleriani (passaggio per ogni edge una singola volta). Tratta da Compeau et al.

HGAP in particolare fa riferimento ad un processo *ad hoc* per le sequenze generate con PacBio, e che opera come segue:

- Crea un dataset con le sequenze CLR generate da sequenziatori PacBio
- Pre-assembla le sequenze filtrate per dimensione per generare sequenze "seme" tramite Velvet
- Allinea le sequenze più o meno lunghe contro le sequenze "seme"

- Crea dei consensi a partire dall'allineamento delle sequenze contro le sequenze "seme"
- Utilizzando il software Velvet si ri-crea un assembly usando le sequenze consenso, generando "*untigs"*
- Gli *"unitigs"* vengono infine utilizzati per generare un assembly finale pronto per essere rifinito usando software come Quiver o Arrow

Arrow è specificatamente il software utilizzato dall'articolo in analisi, e consiste di un programma che rifinisce gli assembly nel senso che incrementa l'accuratezza delle singole posizioni eseguendo multi-allineamenti tra l'assembly in input e le sequenze di partenza.
Infine, una volta generato l'assembly finale, esso è stato circolarizzato tramite Circlator 1.1.1 e allineato nuovamente con le sequenze filtrate ricavate dal sequenziamento Illumina, con errori identificati manualmente corretti.
I quattro genomi di *C. necator* NH9 così generati sono stati depositati in DDBI/ENA/GenBank con gli accession number CP017757-CP017760.

L'annotazione funzionale del genoma di *C. necator* NH9, da intendersi come l'identificazione delle sequenze relative a geni putativi e definizione della funzione dei relativi prodotti, è stata eseguita utilizzando numerosi software operanti sulla base delle informazioni inizialmente ricavate dagli strumenti NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP), Microbial Genome Annotation Pipeline (MiGAP) e Prokka 1.11.
Facendo riferimento al primo dei tre strumenti bioinformatici citati, PGAAP opera anzitutto l'annotazione strutturale dei genomi procariotici identificandone le ORF. Ciò è fatto comparando la sequenza genomica in questione con sequenze di geni rappresentati da *Hidden Markov Models* e sequenze di genomi con elevata qualità d'annotazione.
Inoltre, PGAAP inferisce la presenza di ORF non identificabili tramite allineamento grazie a predizioni *ab initio* delle regioni codificanti.
L'annotazione strutturale è contestuale all'annotazione funzionale, poiché una volta identificate le ORF il prosieguo del processo è la comparazione di queste con quelle di proteine note e ancora con sequenze note di domini conservati. Quest'ultimo processo è necessario per la classificazione dei geni putativi identificati con le operazioni precedenti, cosa che consente di inferire il metabolismo di una cellula ed in generale la sua biologia.
Questo step consente in ultima analisi di associare ad una consistente frazione di geni identificati nel genoma assemblato *de novo* una funzione.

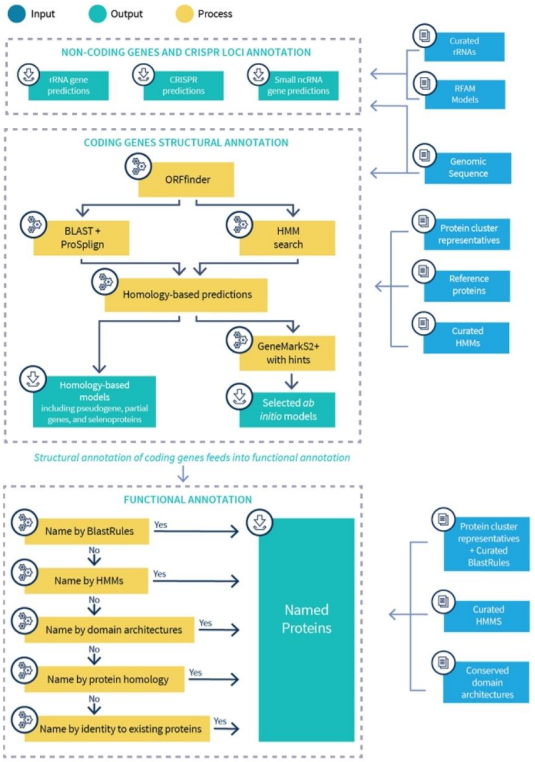

Fig. 6

Gerarchia operativa del processo di annotazione di genomi procariotici tramite PGAAP.
Tratta dal sito NCBI.

Tutte le proteine putative sono quindi classificate utilizzando gli strumenti forniti dal database COG, che le compara con tutti i genomi completamente sequenziati ed annotati di cui dispone al fine di inferirne l'ortologia con proteine note.
Infine, essendo uno dei principali obiettivi di *Moriuchi et al.* la caratterizzazione dei pathway degradativi riferiti ai composti aromatici, le informazioni ottenute dall'annotazione sono state utilizzate per la ricostruzione delle vie metaboliche tramite la Kyoto Enciclopedia of Genes and Genomes (KEGG).

2.2 ANALISI FILOGENETICA TRAMITE MLSA E 16S rRNA

L'analisi filogenetica dei ceppi di *C. necator* NH9 identificati col sequenziamento è stata fatta tramite:

- la tecnica standard dell'allineamento di sequenza tra i geni dell'rRNA 16S
- Multi-Locus Sequence Analysis (MLSA).

La determinazione dei rapporti di filogenesi tra ceppi batterici tramite il locus del 16S rRNA è una metodica nota da lungo tempo essendo questi geni altamente informativi . Questo principalmente perché i geni dell'RNA

ribosomale sono deputati alla formazione di ribonucleoproteine ribosomali e all'interazione con l'mRNA in corso di traduzione una volta che si è assemblato il ribosoma. In questo senso esistono necessariamente delle regioni degli RNA ribosomali che, in virtù della funzione che svolgono, sono soggette a pressione selettiva. Tale pressione selettiva tende ad impedire il fissarsi di mutazioni in specifiche posizioni e regioni dell'rRNA poiché particolarmente importanti nella costituzione e funzionalità dell'apparato di traduzione.
Noto il grado di divergenza tra due sequenze ed il tasso con cui esse specificatamente mutano è possibile inferire un tempo di divergenza e quindi i rapporti filogenetici tra due specie batteriche.
Note le sequenze del locus codificante il 16S rRNA di specie batteriche più o meno filogeneticamente affini a *C. necator* NH9 è stato quindi possibile per gli autori dell'articolo creare un albero filogenetico tramite ClustalW.
ClustalW in particolare opera generando un albero guida sulla base di un multiallineamento tra sequenze indicizzate e divise in *k-mer*. ClustalW infatti calcola inizialmente le distanze a coppie tra sequenze, utilizzando le informazioni note sul tasso di mutazione delle basi dei loci 16S rRNA; e genera contestualmente un albero filogenetico pesato e dotato di radice.
Il peso è applicato alle singole sequenze sulla base della distanza genetica che esse hanno con tutte le altre, e assegna il peso maggiore a quelle più distanti rispetto alle altre (*"outgroup"*).
Per il multiallineamento si sono utilizzate sequenze del locus 16S rRNA prive di gap di batteri appartenenti ai generi *Cupriavidus* e *Ralstonia* quali: *C. basilensis, C. gilardii, C. oxalaticus, C. pauculus, C. pinatubonensis, R. insidiosa* e *R. mannitolilytica*, e si è utilizzata la sequenza del medesimo locus di *Paraburkholderia xenovorans* come outgroup. La scelta di questa specie è legata alla sua appartenenza alla famiglia *Burkholderiaceae,* che la rende filogeneticamente legata ai generi Cupriavidus e Ralstonia, ma per l'appunto non ne condivide le classificazioni inferiori: nell'ambito di un'analisi filogenetica che mira a ridefinire i rapporti filogenetici intra-specifici è infatti necessario utilizzare come *outgroup* una specie che sia legata a quella in esame ma che garantisca comunque un certo grado di divergenza tra le sequenze utilizzate per l'analisi stessa.
L'albero generato è stato quindi valutato con un'analisi di bootstrap eseguita in 1000 repliche. L'analisi di bootstrap è un processo che valuta la significatività dei nodi generati in un dendogramma stimando quante volte essi compaiono nuovamente cambiando il set di dati in input. Trattasi di fatto di un campionamento multiplo con re-immissione: viene estratto casualmente dalle sequenze in input un subset di sequenze e viene eseguito un multiallineamento tra esse, generando così un albero guida

nuovo. Le sequenze estratte vengono quindi re-immesse nel set iniziale e si ri-campiona. Con le sequenze isolate dal ricampionamento si esegue nuovamente un multiallineamento, e si crea un terzo albero guida. L'operazione viene ripetuta ed infine si valuta quante volte, nonostante la casualità del ricampionamento, uno stesso nodo compare nell'albero guida generato da ClustalW. Maggiore il numero di volte in cui un nodo compare, maggiore è il suo *supporto di bootstrap*, e maggiore l'affidabilità che il nodo in questione rappresenti il percorso evolutivo più probabile.

L'analisi MLSA è un'analisi che richiama molto il significato biologico dell'analisi basata sul 16S rRNA con la differenza che viene eseguita su loci genici non tanto coinvolti in attività ribosomali quanto più in processi funzionali conservati.
La determinazione dei rapporti filogenetici tra specie batteriche tramite il locus 16S rRNA è complementata dall'analisi MLSA poiché quest'ultima fornisce maggiore di informazione circa il tempo di divergenza di due specie o ceppi strettamente legati in termini di filogenesi: il locus 16S rRNA non è particolarmente informativo quando il tempo di divergenza tra due taxa è relativamente piccolo poiché non avrà accumulato mutazione.
Negli studi MLSA le sequenze parziali di geni codificanti proteine a funzione conservata e di cui si conosce almeno un ortologo per specie in analisi vengono utilizzate per generare un albero filogenetico. È quindi chiaro come il punto critico di tutta l'analisi MLSA sia la selezione dei geni e delle relative regioni da analizzare e di cui determinare la divergenza tra specie. Tipicamente i geni selezionati per analisi MLSA non sono universalmente validi e vanno selezionati sulla base di criteri quali:

- devono essere presenti in singola copia per genoma
- devono essere tra loro omologhi e presenti in tutte le specie studiate
- devono codificare subunità specifiche di enzimi ubiquitari

Per lo studio in questione sono stati scelti i geni *atpD* (subunità *beta* dell'ATP sintasi)*, leuS* (gene della leucina-tRNA ligasi)*, rplB* (proteina ribosomale l2) e *gyrB* (subunità *beta* della DNA girasi).
L'albero filogenetico è stato ottenuto tramite multiallineamento con ClustalW, che lo ha operato utilizzando come sequenze input dei concatameri tra specifiche regioni dei geni succitati. Come outgroup si sono utilizzate le sequenze geniche della specie *P. xenovorans*.

2.3 COMPARAZIONI SU SCALA GENOMICA

Le comparazioni su scala genomica delle specie afferenti ai generi *Cupriavidus* e *Ralstonia* sono state fatte tramite analisi ANI (Average

Nucleotide Identity) e TNA (Tetra-Nucleotide Analysis) al fine di meglio delineare i rapporti tra i due generi.
L'analisi ANI è utile per verificare le identità tassonomiche tra batteri procariotici e confermare quelle già note. Il calcolo dell'identità media tra nucleotidi di diversi genomi batterici tipicamente viene eseguito frammentando *in silico* le sequenze note, allineando tra loro i frammenti ed identificando le regioni BBH (Bidirectional Best Hits), ossia le regioni a maggiore similarità tra loro rispetto ad ogni altra. Il calcolo del valore ANI è basato sul rapporto tra la sommatoria dei prodotti tra la percentuale di identità per ogni BBH e la lunghezza dell'allineamento con la lunghezza media delle regioni BBH identificate. Questo calcolo identifica una percentuale che può essere al di sopra o al di sotto di un valore soglia che definisce il limite inferiore di similarità per il quale due specie possono essere identificate come la stessa. Questo valore è perlopiù arbitrario sulla base dei dati di partenza, ma si tende a suggerire un valore soglia uguale o superiore al 95%. I risultati di un'analisi ANI possono essere riassunti tramite delle heat-map che identificano tramite un codice-colore le specie che tra loro hanno ANI più alto o più basso. L'analisi TNA è invece fatta comparando la frequenza di tutti i 256 tetranucleotidi osservabili in un genoma tra due genomi diversi. Fu osservato infatti da Pride et al. che ogni genoma ha un unico pattern di distribuzione dei vari tetranucleotidi, e che tendenzialmente più queste distribuzioni sono tra loro simili tanto più i genomi. TNA e ANI sono quindi evidentemente due ottimi strumenti per confermare un'analisi filogenetica basata sul 16S rRNA e su MLSA, raggruppando specie e relativi ceppi in heatmaps.

3. RISULTATI

3.1 PROPRIETA' DEL GENOMA DI *C. necator* NH9 E ABILITA' DEGRADATIVE

Il genoma di *C. necator* NH9 è stato osservato comprendere due cromosomi circolari Chr1 e Chr2 rispettivamente di 4,35Mbp e di 3,4Mbp. Inoltre, è stata confermata la presenza stabile di due plasmidi pENH91 e pENH92. L'analisi d omologia ha stimato un totale di 7290 sequenze codificanti.
L'analisi funzionale COG (Clusters of Orthologous Genes) ha evidenziato una certa polarizzazione funzionale del cromosoma 1 verso funzioni cellulari di base come la replicazione, traduzione, metabolismo aminoacidico e nucleotidico e del cromosoma 2 verso funzioni relative a trascrizione, motilità, metabolismo energetico ed in generale verso l'adattamento e la sopravvivenza. Questa osservazione non è assoluta dato che ad esempio i geni *ben*, fondamentali per la degradazione del

3-CB poiché codificanti le subunità dell'enzima benzoato diossigenasi, si localizzano nel Chr. 1. Il plasmide pENH92 presenta i loci *repB* (codificante la proteina iniziatrice della replicazione a cerchio rotante) e *parAB* fiancheggiati da elementi genetici mobili come integrasi e trasposasi. Questa evidenza, abbinata alla maggiore omologia della proteina repB di *C. necator* NH9 con quella di *Burkholdeira* rispetto a quella di *C. necator* N-1; suggerisce l'acquisizione di questa regione del megaplasmide tramite trasferimento genico orizzontale.
Per quanto concerne le abilità degradative di *C. necator* NH9, l'analisi dei pathway metabolici è stata eseguita usando il database KEGG. Si è osservato che le capacità degradative di questo ceppo sono simili a quelle di ceppi già caratterizzati, con però la peculiare abilità di catabolizzare il 3-clorobenzoato dovuta alla presenza dei geni *ben*. Inoltre, si è notata la mancanza di un operone comprendente tutti e tre i geni richiesti per il metabolismo del catecolo *catA, B* e *C* che è invece presente in molti Gram-negativi. Questi tre geni, essenziali per il completo catabolismo del 3-clorobenzoato, sono infatti dispersi nel genoma di *C. necator* tra il cromosoma 1 e 2 assieme al regolatore *catR*. Questa osservazione richiede ulteriori studi evoluzionistici per la comprensione dei meccanismi di acquisizione delle abilità degradative garantite dai suddetti geni.

3.2 INCERTEZZA TASSONOMICA

La necessità di risolvere le incertezze tassonomiche relative ai generi *Cupriavidus* e *Ralstonia* nasce sia dall'utilità, essendo funzionale ad una corretta annotazione dei genomi dal punto di vista delle capacità di degradazione, sia dalla nota sfida tassonomica che questi generi rappresentano da decenni (vedi Vandamme e Coenye, 2004). Al fine di ricercare differenze tra i generi *Cupriavidus* e *Ralstonia* dal punto di vista delle capacità degradative i ricercatori hanno eseguito analisi focalizzate su geni ortologhi deputati alla degradazione di composti aromatici. Si è osservato che, in modo non assoluto, il genere *Ralstonia* manca di diversi geni i cui prodotti sono necessari alla degradazione di suddetti composti che invece in molti ceppi di *Cupriavidus* sono presenti.
Il primo step dell'analisi filogenetica propriamente detta è in ogni caso stato eseguito tramite la sequenza del locus 16S rRNA. Una netta separazione tra i due generi è stata identificata dall'albero generato dal multiallineamento di queste sequenze, rivelando peraltro due ceppi prima classificati come afferenti al genere *Ralstonia* appartenenti in realtà al genere *Cupriavidus*. Il metodo basato sul 16S rRNA è ben consolidato, ma dato che il tasso di mutazione spesso non consente una risoluzione sufficiente per distinguere le specie batteriche afferenti ad uno stesso genere in modo preciso, si sono utilizzati approcci ulteriori per ampliare lo

studio. L'analisi MLSA approfondisce il metodo 16S rRNA utilizzando più loci diversi, fornendo nella fattispecie alberi dal supporto di bootstrap maggiore. L'analisi ANI, la tecnica TNA e l'approccio di valutazione della percentuale di proteine conservate suggerisce che 41 dei ceppi esaminati dai ricercatori andrebbero corretti, confermando peraltro la solidità di metodi ibridi che abbinino analisi filogenetiche a comparazioni su scala genomica.

4. CONCLUSIONI FINALI

La tesi si prefigge lo scopo di presentare in maniera adeguata e sintetica una serie di approcci utili alla caratterizzazione di specie microbiche in senso specie-specifico e ceppo-specifico, sottolineando come la diversità batterica dal punto di vista genetico e metabolico richieda numerosi sforzi e strumenti di nuova generazione al fine di essere descritta in maniera il più esaustiva possibile.

In particolare l'ampliamento delle conoscenze relative alle specifiche metaboliche dei vari ceppi della specie *Cupriavidus necator*, come il ceppo in esame NH9, non è solo in grado di fornire alla Scienza e alla tecnica degli strumenti applicativi ma apre anche delle finestre sulla complicata storia evolutiva di questa specie estremamente versatile.

Queste caratteristiche, come i ricercatori a più riprese nell'articolo hanno affermato, comprendono sia quelle relative ai genomi dei singoli ceppi che anche le relazioni filogenetiche tra specie che, di fatto, condividono nicchie ecologiche sovrapponibili. Questo perché gli studi genomici dei vari ceppi caratterizzano i geni chiave di vie degradative note ed ignote e potenzialmente sfruttabili, ne descrivono peculiarità ed esigenze nutrizionali ed in generale la biologia. L'approfondimento delle relazioni filogenetiche con approcci molecolari in questo senso risulta fondamentale per la creazione di una vera tassonomia sistematica che sia in grado di distinguere una specie da un'altra e di classificare proprietà d'interesse in modo specie-specifico.

Le tecniche di sequenziamento di nuova generazione e il sempre crescente pannello di strumenti informatici utilizzabili dai biologi sono ormai parte delle procedure standard nella definizione dei processi biologici, che quanto più vengono approfonditi tante più domande aprono. Ulteriori studi infatti potrebbero descrivere meglio come i vari ceppi di *C. necator* interagiscano tra loro e con altre specie nel creare comunità microbiche chemiolitotrofe complesse e come questi ceppi effettuino trasferimento genico orizzontale. Ancora, la definizione dei processi metabolici di un particolare ceppo richiede approfondimenti che descrivano dal punto di vista biochimico le proprietà di enzimi chiave nella degradazione di composti tossici ed inquinanti, come anche il modo in cui

questi composti vengono internalizzati da questi batteri e come ne alterino la fisiologia cellulare quando presenti.
In conclusione, l'articolo in esame esprime e conferma ulteriormente le potenzialità che ad oggi sono sfruttabili da chi esegue ricerca nel campo della Biologia, evidenziando come la conoscenza biologica necessiti di una continua complementazione via via che questi strumenti divengono più raffinati. Questa è la principale ragione che mi ha spinto alla scelta di questo articolo, abbinata alla consapevolezza del fatto che è solo con la conoscenza e l'approfondimento delle meccaniche che governano la vita a tutti i livelli, in particolare a quelli più "semplici", che si può riuscire a sovvertire la direzione presa dall'uomo che non può più dirsi inconsapevole e privo di strumenti.

5. BIBLIOGRAFIA

REFERENCES & RESOURCES

Moriuchi R, Dohra H, Kanesaki Y, Ogawa N. 2019. **Complete Genome Sequence of 3-Chlorobenzoate-Degrading Bacterium Cupriavidus necator NH9 and Reclassification of the Strains of the Genera Cupriavidus and Ralstonia Based on Phylogenetic and Whole-Genome Sequence Analyses.** Front. Microbiol, 10, 133. DOI: 10.3389/fmicb.2019.00133

Rhoads A, Au KF. 2015. **PacBio Sequencing and Its Applications.** Genomics Proteomics Bioinformatics, 13, 278-289. DOI: 10.1016/j.gpb.2015.08.002

Glaeser SP, Kampfer P. 2015. **Multilocus sequence analysis in prokaryotic taxonomy.** Systematic and Applied Microbiology, 38, 237-245. DOI: 10.1016/j.syapm.2015.03.007

Makkar N, Casipa L. 1987. **A Nonobligate Bacterial Predator of Bacteria in Soil?** International Journal of Systematic and Evolutionary Microbiology, 37, 323-326. DOI: 10.1099/00207713-37-4-323

Poehlein A, Kusian B, Friederich B et al. 2011. **Complete Genome Sequence of the Type Strain Cupriavidus necator N-1**. American Society for Microbiology, 5017. DOI: 10.1128/JB.05660-11

Compeau P. E. C, Pevzner A. P, Tesler G. 2011. **Why are de Bruijn graphs useful for genome assembly?** Nature Biotechnology, 29(11), 987-991. DOI: 10.1038/nbt.2023

Fichot E, Norman R. 2013. **Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform.** Microbiome, 1, 1-10. DOI: 10.1186/2049-2618-1-10

PacBio assembly with SMRT portal - ABRPI-Training (sepsis-omics.github.io)

HGAP — Applications 1.0 documentation (rhallpb.github.io)

Overview of PacBio SMRT sequencing: principles, workflow, and applications – CD Genomics (cd-genomics.com)