RELAZIONE SCIENTIFICA FINALE

Programma Short Term Mobility 2013

Fruitore: Andrea Strazzulli

Istituto di afferenza: Istituto di Biochimica delle Proteine

Qualifica: Borsista

Titolo del programma: Analisi metagenomica di ambienti estremi per l'identificazione di nuovi biocatalizzatori di interesse biotecnologico.

INTRODUZIONE

Uno dei più ragguardevoli e recenti sviluppi nel campo delle scienze biologiche, ed in particolare della microbiologia ambientale e delle biotecnologie, è stato l'avvento e lo sviluppo della metagenomica. La metagenomica è definita come la analisi genetica diretta delle comunità microbiche (microbioma) presenti in un determinato ambiente in modo da aggirare i limiti imposti dalle difficoltà di coltivazione ed isolamento delle singole specie microbiche in laboratorio[1]. Questa tecnica consente di studiare in dettaglio la composizione genica di un microbioma e rappresenta un potente strumento di analisi filogenetica che supera qualitativamente l'analisi basata su di un singolo gene, (per esempio rRNA 16S), fornendo preziose informazioni per l'identificazione di nuovi biocatalizzatori e la correlazione genetica tra funzione e filogenia di organismi non coltivabili[2,3].
Le aree in cui è presente attività geotermale sono largamente diffuse in tutto il pianeta ed ospitano microorganismi estremofili archaeali e batterici. Un tipico esempio di questi ambienti estremi sono le fumarole e le pozze di fango bollente presenti in aree con attività vulcanica, come ad esempio l' Obsidian pool nel parco nazionale di Yellowstone nel Wyoming o nella Caldera Flegrea di Napoli. Queste zone possiedono caratteristiche fisiche, ambientali e chimiche tali da renderle habitat unici per determinati microorganismi estremofili, in particolare termofili ed ipertermofili, noti per essere fonte di numerose attività enzimatiche idrolitiche di interesse biotecnologico come lipasi, glicosidasi e proteasi[4–6]. Purtroppo, molti di questi microorganismi sono recalcitranti ai comuni protocolli di studio basati sull'isolamento e la coltivazione in laboratorio e, in questo scenario, l'analisi del DNA metagenomico (gDNA) rappresenta un valido approccio per lo studio filogenetico di questo microbioma e per la comprensione e la caratterizzazione del suo potenziale funzionale.

OBIETTIVO

Obiettivo del progetto è l'identificazione di nuovi enzimi ipertermofili, in particolare cellulasi ed emicellulasi, attraverso l'analisi bioinformatica di dataset metagenomici del microbioma

presente nella pozza solfatarica di Pisciarelli (Agnano, Napoli), arricchito successivamente su diversi substrati di tipo lignocellulosico.

ATTIVITÀ SVOLTA

Durante il suo soggiorno a Sydney presso l'University of New South Wales, il Dr. Strazzulli, sotto la guida del Dr. Federico M. Lauro, ha pianificato ed ottimizzato un workflow per l'analisi delle sequenze metagenomiche ottenute mediante Illumina® paired-end sequencing (Beijing Genomics Institute) del gDNA estratto da campioni provenienti da due pozze presenti nella solfatara Pisciarelli (40° 49' 43.5"N -14° 8' 45" E) e arricchiti su diversi materiali lignocellulosici per identificare e successivamente isolare e caratterizzare nuovi microorganismi e attività enzimatiche coinvolte nella degradazione di biomasse vegetali (Figura 1A).
Il workflow in silico ottimizzato (Figura 1B) è stato ottenuto mediante l'impiego di strumenti bioinformatici programmati ad hoc dal Dr. Lauro e di programmi di analisi metagenomica riportati in letteratura ed opportunamente adattati per l'analisi dei campioni d'interesse.

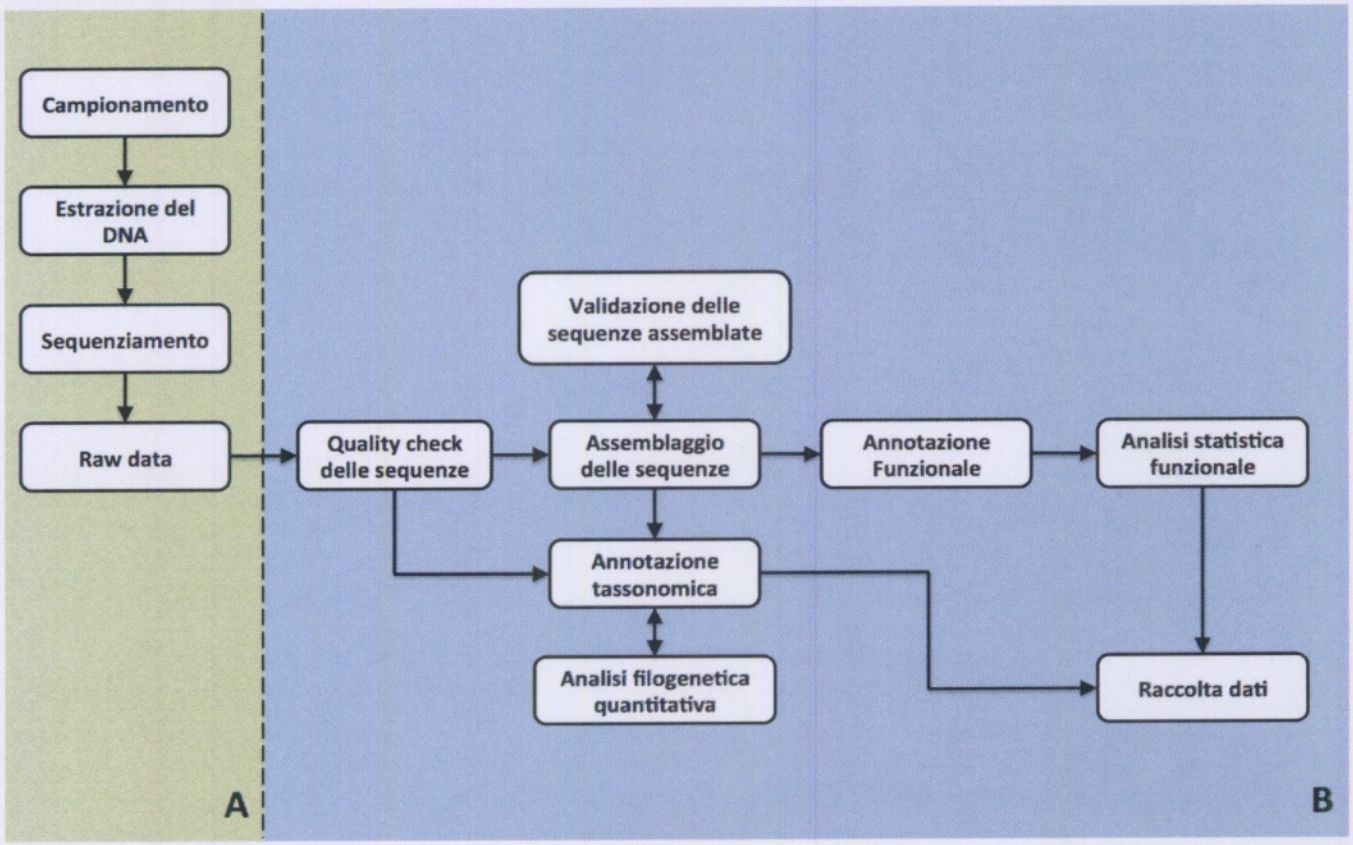

Figura 1: Workflow di analisi metagenomica. **Parte A)** Campionamento e sequenziamento. **Parte B)** Analisi in silico.

- **Quality check delle sequenze.**

Le sequenze ottenute dalla BGI (raw data) sono state controllate mediante l'utilizzo di FastQC[7], un software per analisi delle reads che consente di stabilire i parametri qualitativi delle reads ottenute, indispensabili per il successivo step di assemblaggio, e di stabilire la qualità del

sequenziamento ottenuto analizzando in particolare la lunghezza, la percentuale %GC ed il quality score per ogni sequenza. FastQC ha mostrato una buona qualità media delle sequenze, con una lunghezza di 90bp ciascuna compatibile con quanto atteso dal sequenziamento. Le reads sono quindi state sottoposte allo step di assemblaggio.

- **Assemblaggio e validazione**

Le sequenze paired-end validate sono state assemblate utilizzando SOAPdenovo2[8], un software per assemblaggio di sequenze di tipo de Bruijn. Sebbene questo tipo approccio richieda una potenza di calcolo superiore al metodo alternativo, basato sull'utilizzo di genomi di riferimento (reference-based assembly), esso è stato preferito in quanto consente di evitare artefatti dovuti principalmente alla variabilità genetica dei campioni analizzati rispetto a quelli riportati nelle banche dati disponibili, e alla presenza di sequenze completamente nuove che potrebbero non trovare un riferimento sufficientemente robusto per permettere l'assemblaggio.
Scelto un valore di k-mers corrispondente al numero di basi che il software andrà a cercare nelle singole reads, SOAPdenovo2 permette di ottenere i contigs dall'assemblaggio diretto delle reads e, successivamente, da questi estendersi alla costruzione di scaffolds. Al termine del processo viene restituita una statistica indicante i parametri qualitativi dell'assemblaggio eseguito.
I parametri d'interesse sono:

- **Numero dei contigs**
- **Lunghezza complessiva dei contigs**
- **N50:** numero dei contigs che hanno almeno il 50% della lunghezza complessiva di tutti i contigs.
- **N90**: numero dei contigs che hanno almeno il 90% della lunghezza complessiva di tutti i contigs.
- **Lunghezza massima dei contigs**.
- **Lunghezza minima dei contigs.**

Solo l'utilizzo di parametri di assemblaggio differenti, ed in particolare diversi valori di k-mers, consentiranno di stabilire quale sia il risultato di assemblaggio migliore. Al momento sono in corso test di assemblaggio per stabilire i parametri migliori da utilizzare per proseguire con l'annotazione funzionale.

- **Annotazione tassonomica e analisi filogenetica**

Uno dei principali obiettivi della metagenomica è l'analisi tassonomica. La quantizzazione relativa delle specie presenti in una determinata comunità microbica è di fondamentale importanza per paragonare campioni provenienti da diverse origini e che spesso sono solo stimati a livello di phylum mediante analisi convenzionale dell'rRNA 16S.
Per ottenere un'informazione completa delle specie presenti nel campione di partenza e dopo l'arricchimento selettivo è stato utilizzato il tool GASiC (Genome Abundance Similarity Correction)[9] che consente di stimare la quantità relativa delle specie a partire dalle reads ottenute. Per prima cosa è stato necessario scegliere dei genomi noti che sono diventati il dataset di riferimento. A questo scopo sono stati selezionati per ora 54 differenti genomi tra archaea, batteri e virus prevalentemente termofili (Tabella 1) ai quali sono state

successivamente allineate, mediante Bowtie[10], tutte le reads fornendo una stima preliminare delle abbondanze osservate.

Genbank ID	Nome	Genbank ID	Nome
NC_000853	*Thermotoga maritima MSB8*	NC_010154	*Acidianus filamentous virus 8*
NC_000854	*Aeropyrum pernix K1*	NC_010155	*Acidianus filamentous virus 3*
NC_002578	*Thermoplasma acidophilum DSM 1728*	NC_010175	*Chloroflexus aurantiacus J-10-fl*
NC_002689	*Thermoplasma volcanium GSS1*	NC_010537	*Acidianus filamentous virus 9*
NC_002754	*Sulfolobus solfataricus P2*	NC_011766	*Desulfurococcus kamchatkensis 1221n*
NC_003106	*Sulfolobus tokodaii str 7*	NC_012588	*Sulfolobus islandicus M1425*
NC_003364	*Pyrobaculum aerophilum str IM2*	NC_012589	*Sulfolobus islandicus LS215*
NC_004113	*Thermosynechococcus elongatus BP-1*	NC_012883	*Thermococcus sibiricus MM 739*
NC_005213	*Nanoarchaeum equitans Kin4-M*	NC_013205	*Alicyclobacillus acidocaldarius DSM 446*
NC_005830	*Acidianus filamentous virus 1*	NC_013501	*Rhodothermus marinus DSM 4252*
NC_006461	*Thermus thermophilus HB8*	NC_013769	*Sulfolobus islandicus LD85*
NC_007181	*Sulfolobus acidocaldarius DSM 639*	NC_013921	*Thermoanaerobacter italicus Ab9*
NC_007409	*Acidianus two-tailed virus*	NC_014205	*Staphylothermus hellenicus DSM 12710*
NC_008696	*Thermofilum pendens Hrk*	NC_014471	*Ignisphaera aggregans DSM 17230*
NC_008701	*Pyrobaculum islandicum DSM 4184*	NC_014658	*Methanothermus fervidus DSM 2088*
NC_009012	*Clostridium thermocellum ATCC 27405*	NC_015185	*Desulfurobacterium thermolithotrophum DSM 11699*
NC_009033	*Staphylothermus marinus F1*	NC_015216	*Methanobacterium sp AL-21*
NC_009328	*Geobacillus thermodenitrificans NG80-2*	NC_015518	*Acidianus hospitalis W1*
NC_009376	*Pyrobaculum arsenaticum DSM 13514*	NC_017274	*Sulfolobus solfataricus 98/2*
NC_009437	*Caldicellulosiruptor saccharolyticus DSM 8903*	NC_017275	*Sulfolobus islandicus HVE10/4*
NC_009440	*Metallosphaera sedula DSM 5348*	NC_017276	*Sulfolobus islandicus REY15A*
NC_009452	*Acidianus bottle-shaped virus*	NC_017312	*Enterococcus faecalis 62*
NC_009718	*Fervidobacterium nodosum Rt17-B1*	NC_017625	*Escherichia coli DH1*
NC_009884	*Acidianus filamentous virus 2*	NC_018001	*Desulfurococcus fermentans DSM 16532*
NC_009954	*Caldivirga maquilingensis IC-167*	NC_018092	*Pyrococcus furiosus COM1*
NC_010152	*Acidianus filamentous virus 6*	NC_021058	*Sulfolobus islandicus LAL14/1*
NC_010153	*Acidianus filamentous virus 7*	NC_021169	*Archaeoglobus sulfaticallidus PM70-1*

Tabella 1 Lista dei genomi di riferimento utilizzati per l'analisi di GASiC.

È stata quindi creata una matrice di similarità tra i 54 genomi scelti utilizzando ancora Bowtie e Mason[11], un emulatore di reads capace di restituire reads simulate tipiche di una tecnica di sequenziamento, in questo caso Illumina® Paired-end, di un genoma dato. La matrice di similarità e le abbondanze osservate sono quindi combinate in un sistema di equazioni lineari applicato ad una procedura a cicli di bootstrapping (ricampionamento con reimmissione) per calcolare l'abbondanza reale nel campione metagenomico delle specie selezionate come dataset di riferimento. L'analisi preliminare finora effettuata ha consentito di attribuire e quindi di correggere l'abbondanza delle reads disponibili solo ad un ridotto numero di specie tra quelle selezionate. Per ottenere un risultato maggiormente accurato e completo, sono stati mandati a sequenziare gli rRNA 16S ottenuti dai diversi campioni metagenomici. Il risutalto dell'analisi del 16S consentirà di identificare i phyla presenti nei diversi campioni e ciò permetterà di perfezionare il processo di annotazione tassonomica.

- **Annotazione e statistica funzionale**

Durante la STM sono state effettuate alcune analisi preliminari sui contigs e gli scaffolds finora assemblati Una volta che l'assemblaggio verrà ottimizzato, i contigs e gli scaffolds ottenuti saranno analizzati con BlastX, uno strumento di ricerca di allineamento locale usando come database di sequenza proteiche non ridondanti "nr". Ciò consentirà una prima identificazione dei geni codificanti per proteine presenti nel campione iniziale.
Inoltre, il Dott. Strazzulli ha acquisito il know how per il successivo utilizzo di glimmer-MG[12,13], un sistema per ricercare geni in sequenze di DNA metagenomiche che utilizza modelli Markov interpolati e che può essere addestrato ad hoc, in base al database refseq, per identificare regioni codificanti e distinguerle dal DNA non codificante. Una volta estrapolate le sequenze codificanti per proteine, queste verranno sottomesse a BlastP per la ricerca COGs (Clusters of Orthologous Groups) verso il database "myva" [14] e per la ricerca della loro versione archaeale arCOGs verso il database "ar120"[15]. Questa analisi consentirà di associare le sequenze codificanti a specifiche funzioni sia batteriche che archaeali e di raggrupparle in categorie funzionali per ricostruire informazioni relative a pathway metabolici delle comunità microbiche presenti nel campione di partenza.

Conclusioni e prospettive future

Durante il suo periodo di soggiorno presso l'University of New South Wales, grazie al supporto del Dott. Lauro, il Dott. Strazzulli ha acquisito un solido know how bioinformatico mirato in particolare ad analisi metagenomiche. Le conoscenze acquisite, inoltre, hanno permesso al Dott. Strazzulli di mettere a punto, presso il gruppo del CNR in cui lavora, una piattaforma bioinformatica simile a quella utilizzata dal Dott. Lauro, con il quale il gruppo resta in attiva collaborazione. I risultati preliminari ottenuti durante la permanenza a Sydney sono al momento in fase di completamento e ottimizzazione dal Dott. Strazzulli sulla piattaforma configurata presso il gruppo del CNR.

References

1. Kozubal, M. a *et al.* Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *The ISME journal* **7,** 622–34 (2013).
2. Gilbert, J. A. *et al.* Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PloS one* **3,** e3042 (2008).
3. Wilmes, P. & Bond, P. L. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends in microbiology* **14,** 92–7 (2006).
4. Streit, W. R., Daniel, R. & Jaeger, K.-E. Prospecting for biocatalysts and drugs in the genomes of non-cultured microorganisms. *Current opinion in biotechnology* **15,** 285–90 (2004).
5. Daniel, R. The soil metagenome--a rich resource for the discovery of novel natural products. *Current opinion in biotechnology* **15,** 199–204 (2004).
6. Handelsman, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Society* **68,** 669–685 (2004).
7. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)* **27,** 863–4 (2011).

8. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1,** 18 (2012).
9. Lindner, M. S. & Renard, B. Y. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic acids research* **41,** e10 (2013).
10. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10,** R25 (2009).
11. Holtgrewe, M. Mason – A Read Simulator for Second Generation Sequencing Data FACHBEREICH MATHEMATIK UND INFORMATIK SERIE B • INFORMATIK Abstract. (2010).
12. Delcher, A. L., Bratke, K. a, Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics (Oxford, England)* **23,** 673–9 (2007).
13. Kelley, D. R., Liu, B., Delcher, A. L., Pop, M. & Salzberg, S. L. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic acids research* **40,** e9 (2012).
14. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4,** 41 (2003).
15. Meereis, F. & Kaufmann, M. Extension of the COG and arCOG databases by amino acid and nucleotide sequences. *BMC bioinformatics* **9,** 479 (2008).

DATA 16/07/2013 ..... **IL FRUITORE**