UNIVERSITÀ DEGLI STUDI DI PADOVA
FACOLTÀ DI SCIENZE MM. FF. NN.

LAUREA SPECIALISTICA
IN BIOLOGIA EVOLUZIONISTICA

ELABORATO DI LAUREA

ANALISI GENOMICA DEL TASSO DI SOSTITUZIONE AMINOACIDICA IN BATTERI PIEZOFILI

RELATORE: Prof. Giorgio Valle
Dipartimento di Biologia

CORRELATORE: Dr. Alessandro Vezzi
Dipartimento di Biologia

LAUREANDA: Laura Treu

Anno Accademico 2006 – 2007

INDICE

PREMESSA

Il pianeta Terra è per il 75% ricoperto d'acqua e nonostante la gran parte raggiunga profondità molto elevate a tutt'oggi l'ambiente oceanico abissale rimane quasi completamente sconosciuto. Allo stesso modo sono poco noti gli organismi che lo popolano e gli adattamenti che mettono in atto per adeguarsi alle enormi pressioni, all'assenza di luce, alle basse temperature ed alla scarsità di nutrienti che lo caratterizzano.

Dopo il sequenziamento del genoma del batterio abissale *Photobacterium profundum* ceppo SS9, il laboratorio nel quale ho svolto il progetto di tesi si è occupato del *finishing* del genoma del batterio *Shewanella benthica* ceppo KT99, al quale ho partecipato. Questa procedura è fondamentale per poter completare la struttura di un genoma e per permettere le successive analisi bioinformatiche volte, per esempio, all'identificazione di geni, di promotori, della struttura degli operoni e così via.

Al fine di comprendere meglio come questi batteri si adattino alle condizioni estreme del loro ambiente ho quindi svolto un'analisi a livello genomico, confrontando i geni ortologhi di *P. profundum* con quelli di altri tre batteri non piezofili appartenenti alla famiglia delle Vibrionaceae e gli ortologhi di *S. benthica* con altre tre Shewanelle non piezofile. Lo scopo è quello di identificare quali classi geniche presentino un tasso di sostituzione non-sinonimo rispetto al sinonimo più elevato dell'atteso, per identificare quali geni sono sottoposti ad una pressione selettiva positiva, mirata a migliorare la funzionalità della proteina e ad aumentare la *fitness* del microorganismo.

Questo studio, svolto parallelamente su due famiglie batteriche differenti, ha consentito di identificare svariati geni e classi funzionali, che sembrano avere un ruolo nell'adattamento alle alte pressioni. Questi due elementi sono a volte comuni ai due gruppi di batteri, altre volte specifici ad un gruppo e permettono così di identificare sia meccanismi comuni che specifici adottati dai diversi organismi.

INTRODUZIONE

2.1 Generalità sui Batteri Marini

Il mondo subacqueo brulica di microscopiche forme di vita.
Le comunità di eubatteri, archea, protisti e funghi unicellulari costituiscono la maggior parte di questa biomassa oceanica e sono una componente essenziale dell'ecosistema, in quanto responsabili del 98% della produzione primaria e catalizzatori di tutti i cicli biogeochimici (Whitman *et al.*, 1998). L'oceano intero è un sistema vivente integrato in cui le trasformazioni energetiche sono regolate da processi fisici, chimici e biotici interdipendenti. Se da un lato la maggior parte dei principi chimici e fisici è ormai nota, solamente ora si stanno mettendo a punto degli approcci molecolari per descrivere ed interpretare i processi oltre che la diversità biologica. E' ormai indubbio comunque che sia necessario includere l'abbondanza microbica, la sua diversità, le sue dinamiche e la sua influenza sulla chimica oceanica nello sviluppo di modelli per una migliore comprensione dell'ecosistema marino, fondamentale nelle sue interazioni con l'atmosfera per la regolazione del clima mondiale (Azam *et al.*, 2004).

Le Comunità Microbiche Oceaniche.

I modelli di flusso delle sostanze organiche oceaniche sono il risultato di complicate interazioni tra diversi biota. I batteri sono la maggiore forza biologica del ciclo oceanico del carbonio in quanto la materia organica disciolta, *Dissolved Organic Matter* (DOM), viene da loro metabolizzata e resa disponibile a livelli trofici superiori, in quello che viene definito *microbial loop*.
Un efficace esempio di contesto in cui è evidente tale attività microbica è rappresentato dagli *hot spot* in cui strutture polimeriche colloidali, nano e micro gel interagiscono nel creare un *continuum* di materia organica, che è il substrato ideale per un'ampia biodiversità (Azam, 1998). La produttività oceanica non viene mantenuta attraverso un elevato apporto di sostanze organiche ma piuttosto da un loro rapido riciclo. Questo sistema può modificarsi per mezzo dell'accumulo di biomassa o di alcune sue esportazioni, sotto la spinta di condizioni fisiche variabili o dell'aumento modulato dei nutrienti stessi.
La produzione di materia organica viene regolata in funzione della più bassa concentrazione relativa di nutrienti presenti, necessaria per la crescita del batterioplancton. Ciò implica che non sono necessariamente nè il più alto tasso di richiesta di carbonio nè la presenza di tracce organiche nell'ambiente a controllare il tasso produttivo totale dell'ecosistema, bensì una bilanciata dinamica tra domanda e disponibilità di nutrienti.

Queste dinamiche sono a loro volta fortemente influenzate dai processi fisici, inclusi il clima, l'alta pressione e la scarsa luce (Arrigo, 2005).
Nella figura sotto riportata sono schematizzate le reti trofiche pelagiche con particolare interesse per il *microbial loop* e le sue interazioni con la *grazing food chain* (Azam, 1998).

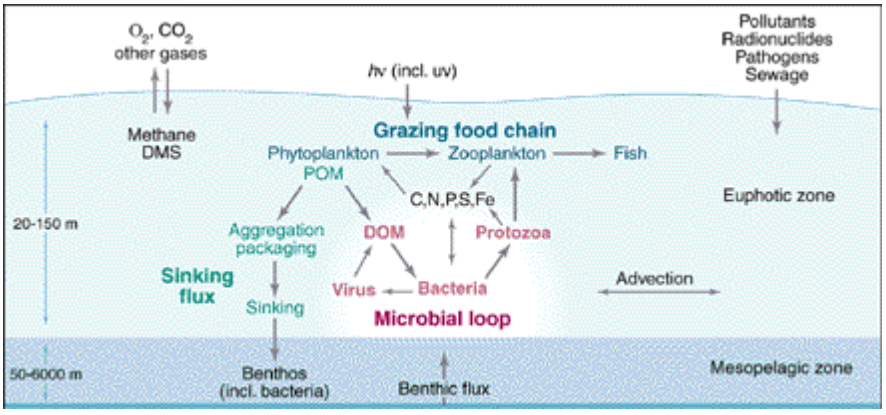

Fig. 1. Rappresentazione dei flussi di materia organica oceanici tratta da Azam, 1998. DOM indica la *dissolved organic matter* e DMS è il *dimethylsulfide*.

Attualmente è possibile studiare i singoli passaggi delle complesse dinamiche biologiche, coinvolte in tali processi, per mezzo delle più recenti metodiche di studio che portano ad identificare i geni e le vie metaboliche responsabili dell'adattamento degli organismi alle diverse condizioni ambientali. Il riconoscimento del batterioplancton quale componente fondamentale della rete trofica oceanica, ha portato all'avvio del sequenziamento di svariati genomi batterici, allo scopo di indagare più a fondo sulle strategie che permettono l'utilizzo della materia organica e dei composti inorganici come supplemento all'eterotrofia (DeLong *et al.*, 2005).

Progetti di Sequenziamento Genomico.

Grazie all'applicazione della genomica ai problemi di oceanografia microbica è possibile cercare di espandere la comprensione di metabolismo, ecologia ed evoluzione dei batteri. La determinazione della sequenza di un genoma fornisce una quantità e variabilità di dati, difficilmente ottenibili in altro modo, che non ha precedenti nella storia della biologia; il DNA infatti contiene tutta l'informazione che determina quelli che saranno la struttura e il metabolismo di un organismo.
Informazioni riguardanti geni specie specifici, proteine ipotetiche conservate in differenti taxa nonché la presenza di elementi ripetuti o di inserzioni possono essere estrapolate dall'analisi dell'intero genoma (Nelson *et al.*, 2000).

Inoltre la conoscenza della totalità della sequenza permette di compiere una più dettagliata ricostruzione dei *pathway* metabolici e delle complesse interazioni intracellulari ed intercellulari, indirizzando in qualche modo lo studio successivo di analisi funzionale di gruppi specifici di geni.
A partire dal 1995, anno in cui *The Institute for Genomic Research* (TIGR) ha pubblicato la prima sequenza genomica completa di un batterio, sono stati portati a termine altri 574 genomi di microrganismi (GenBank, *release* dell'1 settembre 2007) e molti altri sono in corso di completamento. Esemplificativa è la linea temporale di metodologie sviluppate che hanno influenzato la microbiologia marina degli ultimi vent'anni, come il sequenziamento *Whole Genome Shotgun* (WGS) e le applicazioni di PCR per l'individuazione di nuove specie e la stima della biomassa del batterioplancton, come riportato nella figura 2.

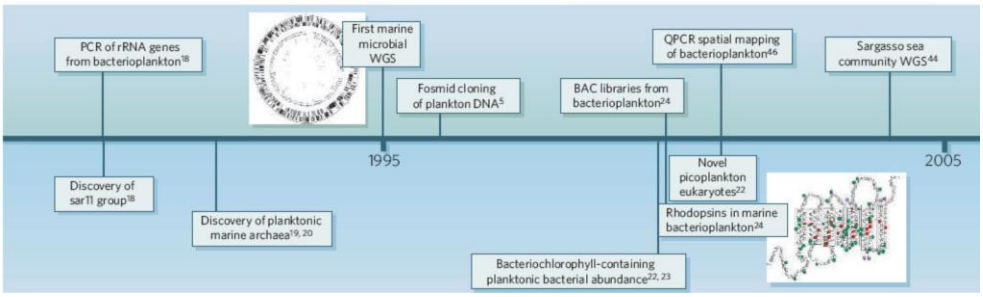

Fig. 2. Immagine tratta da DeLong *et al.*, 2005 con riportate le più importanti tecnologie degli ultimi anni utili per lo studio dei microorganismi marini.

Infine il sequenziamento massivo di metagenomi procariotici sta mettendo a disposizione un enorme mole di dati per l'identificazione di abilità metaboliche sconosciute e di metodi innovativi di concettualizzare e studiare la biodiversità procariotica. Per fare solo un esempio si consideri la grande importanza rivestita dalla scoperta di 782 nuovi fotorecettori simili alla rodopsina identificati nel progetto di metagenomica condotto da J. Craigh Venter su campioni prelevati nel Mar dei Sargassi (Venter *et al.*, 2004). La scoperta dell'enorme numero di queste fotorodopsine permette infatti di avanzare nuove ipotesi sull'accoppiamento tra la ricezione dell'energia luminosa ed il ciclo del carbonio attraverso un *pathway* non basato sulla clorofilla. Senza contare l'enorme quantità di nuovi dati apportati alla fotobiologia oceanica da questo progetto.

2.2 La Vita in Condizioni Estreme

E' l'adattamento alle caratteristiche uniche del loro ambiente ciò che definisce l'essenza delle specie microbiche marine. Considerato che l'oceano presenta una profondità media di 3800 m e quindi una pressione idrostatica media di 38 MPa, una larga parte della biosfera marina deve essere in grado di vivere ad alte pressioni. I batteri isolati in questo ambiente sono detti piezofili e presentano il loro optimum di crescita a pressioni maggiori di 40 MPa, quindi adattamenti specifici che permettono loro di sopravvivere a questa condizione estrema. Al contrario un batterio è definito piezotollerante se presenta una crescita ottimale a pressione inferiore di 40 MPa e cresce ugualmente a pressione atmosferica (DeLong *et al.*, 1997). La diversità microbica delle profondità comprende anche batteri mesofili, con optimum a pressione atmosferica, ma che sopravvivono comunque all'alta pressione con tasso di crescita ridotto. Questi ultimi sono interessanti per studiare in che modo possono rispondere alle variazioni di pressione idrostatica.

Meccanismi di Adattamento all'Alta Pressione.

Le basse temperature influenzano la velocità delle reazioni enzimatiche e del trasporto dei soluti, riducono la fluidità di membrana, provocano la formazione di cristalli di ghiaccio all'interno della cellula (Cavicchioli *et al.*, 2002).

Allo stesso modo l'aumento di pressione idrostatica impedisce la crescita degli organismi tramite l'inibizione di diversi processi cellulari. Ad esempio in *Escherichia coli* si è osservato come motilità cellulare, trasporto dei substrati, divisione e crescita cellulare, replicazione del DNA, trascrizione e traduzione siano sensibili e vengano ridotti a differenti valori di pressione (Bartlett, 2002). L'impedimento della divisione cellulare si manifesta fenotipicamente con la formazione di lunghe cellule filamentose, con la conseguenza di una variazione morfologica del batterio alquanto marcata.

Gli organismi che sono in grado di crescere a basse temperature ed alte pressioni devono quindi necessariamente presentare una serie di adattamenti che permettano loro di vivere in queste condizioni estreme.

In *P. profundum* ceppo SS9 è stato isolato un gene essenziale per la crescita ad alte pressioni con elevata similarità a *recD*, che codifica una proteina coinvolta nella ricombinazione omologa in *E. coli*. L'introduzione del gene *recD* di SS9 in mutanti *recD* di *E. coli* ne consente la crescita ad alta pressione, prevenendo la formazione di cellule filamentose (Bidle *et al.*, 1999). Il gene del ceppo SS9 presenta una porzione di 192 paia di basi che non si trova nell'omologo di *E. coli* e che potrebbe essere determinante per l'attività della proteina ad alte pressioni.

Acidi Grassi, Proteine di Membrana e Trasporto.

La fluidità delle membrane cellulari nei batteri delle profondità marine, sembra essere mantenuta tramite un fine controllo della percentuale di acidi grassi insaturi dei fosfolipidi di membrana (Yayanos, 1995). L'aumento della viscosità delle membrane causato dalla crescita degli organismi a basse temperature o ad alte pressioni sarebbe quindi compensato da un corrispondente aumento del grado di insaturazione degli acidi grassi. Tuttavia non sono ancora noti i recettori in grado di percepire le fluttuazioni delle condizioni ambientali, come questi si rapportino con gli elementi regolatori intracellulari ed alcuni dei *pathway* metabolici coinvolti nella sintesi degli acidi grassi insaturi (Allen *et al.*, 2002).

L'innalzamento della pressione idrostatica è simile ad un decremento della temperatura in termini di riduzione di fluidità di membrana e mobilità molecolare di fosfolipidi e proteine, questo effetto può determinarne una riduzione della funzionalità.

I trasportatori sono un gruppo rilevante di proteine di membrana che possono essere soggette a modificazioni di espressione negli organismi piezofili e non in seguito a cambiamenti nei parametri ambientali (Vezzi *et al.*, 2005; Iwahashia *et al.*, 2005). Questo sembra correlato all'influenza che la pressione esercita sulle reazioni bochimiche che prevedono variazioni del volume di attivazione, come ad esempio il trasporto di molecole attraverso la membrana (Abe *et al.*, 2000).

2.3 Shewanellaceae e Vibrionaceae

Relazioni Filogenetiche tra Specie Piezofile.

L'ambiente fisico delle profondità oceaniche è caratterizzato da assenza di luce, scarsità di nutrienti, alta pressione idrostatica e basse temperature. Nonostante l'enorme complessità della comunità microbica piezofila è interessante notare che una parte di questi microorganismi è filogeneticamente vicina.

Infatti da un confronto tra le sequenze di rRNA 16S di microorganismi piezofili e piezotolleranti appartenenti al philum dei γ-Proteobatteri vengono raggruppati sullo stesso ramo del genere Shewanella tutti i ceppi piezofili obbligati considerati (DB5501, DB6101, DB6906, DB172F e *Shewanella* sp. PT99) e alcuni moderatamente barofili (DSS12 e *S. benthica*), mentre altri piezotolleranti (*Shewanella* sp. SC2A, *Photobacterium* sp. SS9 e DSJ4) sono distribuiti in modo omogeneo nel gruppo dei γ-Proteobatteri (Kato *el al.*, 1996).

A questo proposito sono stati recentemente svolti ulteriori studi che hanno inserito nella filogenesi molecolare dei γ-Proteobatteri svariati ceppi con origini geografiche differenti.
Grazie ai nuovi dati a disposizione è stato infatti possibile effettuare un confronto tra i microorganismi isolati nelle fosse oceaniche e i ceppi a loro strettamente correlati, non adattati all'alta pressione (Lauro *et al.*, 2007).
Le relazioni filogenetiche sono descritte di seguito nell'albero costruito con il *maximum-likelihood* basandosi sulle sequenze dell'rRNA 16S.

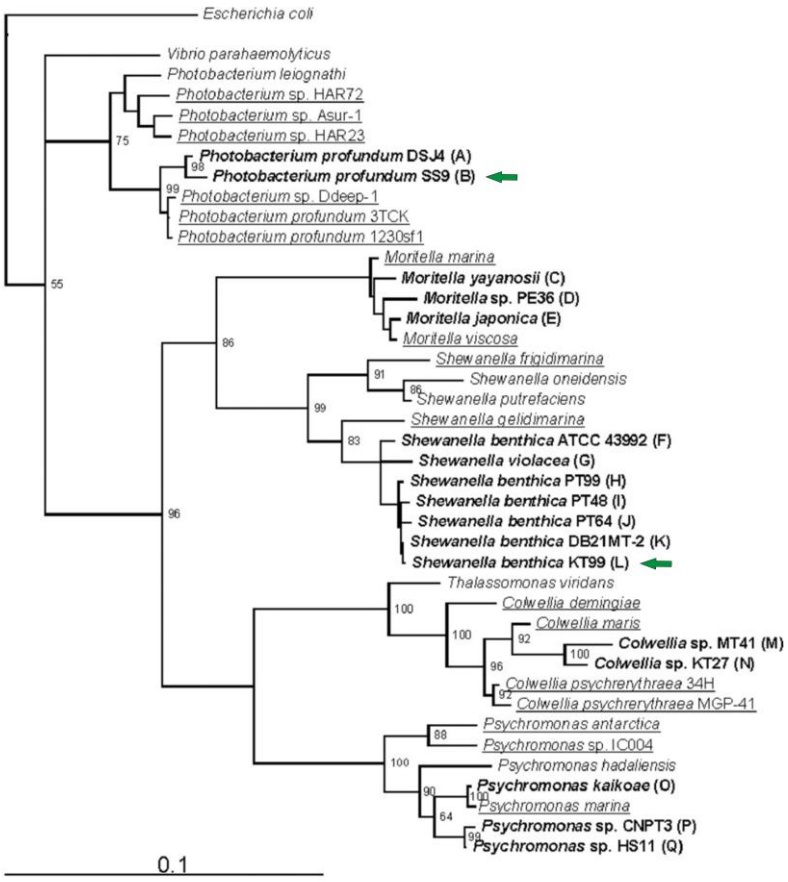

Fig. 3. Rappresentazione delle relazioni filogenetiche tra i batteri isolati ad elevate profondità e quelli di superficie a loro vicini. Il supporto di solidità dell'albero è dato da 1000 repliche di *bootstrap* e la scala rappresenta il numero medio di sostituzioni nucleotidiche per sito. I microorganismi piezofili sono indicati in grassetto, quelli psicrofili sono sottolineati; le lettere tra parentesi indicano i siti geografici d'isolamento riportate nell'articolo di Lauro *et al.*, 2007 da cui è tratta l'immagine.

Della filogenesi sopra riportata fanno parte anche i microorganismi che vengono analizzati in questo lavoro di tesi.

P. profundum ceppo SS9.

Il batterio *P. profundum* ceppo SS9 è stato isolato per la prima volta da un omogenato di anfipodi raccolti a 2551 m di profondità e ad una temperatura di circa 9°C nel mare Sulu, al largo delle coste delle Filippine. E' un batterio Gram negativo con morfologia a bastoncello e fa parte del phylum dei γ-Proteobatteri, a cui appartengono tutti gli eubatteri piezofili fino ad oggi isolati.

L'analisi filogenetica basata sulla sequenza dell'RNA ribosomale 16S ha permesso di collocare il ceppo SS9 nella famiglia delle Vibrionaceae. Le sue condizioni ottimali di crescita sono 28 MPa e 15°C, alle quali il batterio si duplica all'incirca ogni due ore; è in grado di crescere in un intervallo di temperatura compreso tra 2°C e oltre 20°C e ad una pressione compresa tra 0,1 MPa e circa 70 MPa.

La capacità di crescere a temperatura ambiente e a pressione atmosferica ha fatto di questo organismo un sistema modello, essendo facilmente coltivabile rispetto ad altre specie psicrofile o piezofile obbligate.

S. benthica ceppo KT99.

Il genere Shewanella include al suo interno specie che vivono in un ampio *range* ambientale e sono in grado di utilizzare diversi elementi come accettori di elettroni durante la respirazione anaerobia, inclusi alcuni ossido-metalli insolubili. Il ceppo di *S. benthica* KT99 è stato isolato da una colonia formatasi ad alta pressione a partire da un campione di sedimento oceanico raccolto a 8600 m di profondità e 1,8°C di temperatura nel Kermadec Trenc, nell'Oceano Pacifico. Questo batterio ha la peculiarità di produrre, come adattamento all'alta pressione, acidi grassi polinsaturi (PUFA) al posto dei normali fosfolipidi di membrana per mantenere la fluidità (DeLong *et al.*, 1997).

La sequenza parziale del genoma di *S. benhica* è stata ottenuta dal J. Craig Venter Institute (JCVI) a Rockville, Maryland, nell'ambito del progetto "Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing", tramite l'approccio di frammentazione casuale dell'intero genoma, il *whole genome shotgun* (WGS), e il successivo sequenziamento dei frammenti di DNA. Il progetto di sequenziamento di genomi microbici marini della Moore Foundation è iniziato nel 2004 con lo scopo di aumentare la conoscenza del batterioplancton ecologicamente importante, considerando i batteri selezionati da una commissione di autorevoli microbiologi marini. Si è quindi iniziato a lavorare sui genomi dei microorganimi scelti, grazie ad una donazione al J. Craig Venter Institute; le sequenze di DNA auto-annotate vengono sistematicamente depositate in GenBank, nel database del *National Institute of Health* (NIH).

Il *Finishing* per Ottenere la Sequenza Completa.

L'approccio WGS si basa fondamentalmente sul sequenziamento sistematico di frammenti casuali del DNA genomico di interesse, clonati in una libreria plasmidica ed una libreria fosmidica aventi dimensioni medie dell'inserto rispettivamente di 2000 bp e 34000 bp. Via via che le sequenze vengono prodotte aumenta il *genome coverage* o ridondanza media, che rappresenta il rapporto tra le basi di DNA complessivamente sequenziate e la lunghezza del genoma. Già con un *coverage* di 5 oltre il 99% del genoma è teoricamente coperto e la percentuale non cambia di molto per *coverage* più elevati. Per questi motivi nei progetti genomici la fase *shotgun* è condotta generalmente fino ad una copertura teorica che varia tra 6 e 10, per poi passare all'assemblaggio delle sequenze e al *finishing*. Le sequenze ottenute vengono confrontate tra loro da programmi di assemblaggio che ne determinano le possibili sovrapposizioni e creano dei contigui (*contig*), in cui insiemi di sequenze sono uniti tra loro a formare una sequenza consenso (*consensus*).
La fase più complessa e lunga dell'approccio WGS è certamente il *finishing* cioè il sequenziamento mirato, utile a chiudere i *gap* rimasti nella sequenza del genoma e a migliorare la qualità dei dati. Questa parte della tecnica richiede infatti un intervento manuale per poter identificare le regioni dell'assemblaggio da sottoporre ad ulteriore analisi, e costituisce la parte del progetto di questa tesi mirata al miglioramento della qualità finale della sequenza di *S. benthica*.

2.4 Come Evolve un Genoma

I processi generali che determinano l'evoluzione dei genomi microbici sono fondamentalmente le mutazioni puntiformi ed il *lateral gene transfer*. Mentre quest'ultimo meccanismo rappresenta per i batteri un metodo rapido che può portare all'acquisizione anche di grosse porzioni di DNA, le mutazioni determinano una lenta e graduale modificazione del genoma (Ochman *et al.*, 2000). Entrambi sono comunque essenziali per i batteri al fine di riuscire ad adattarsi a nuovi ambienti: il trasferimento genico laterale costituisce argomento di tesi del laureando R. Rosselli, mentre io mi sono occupata dell'aspetto riguardante le mutazioni sinonimo-non sinonimo.

L' Importanza del Tasso di Sostituzione Aminoacidica.

Un metodo molto diffuso per comprendere le dinamiche dell'evoluzione molecolare delle sequenze dei geni è la stima del tasso di sostituzione sinonimo e non sinonimo. Mentre le mutazioni sinonime sono invisibili alla selezione naturale, quelle non sinonime sono largamente sottoposte alla pressione selettiva che può essere purificatrice, per cui tende ad eliminarle, o positiva, quando le fissa nella sequenza.

Una comparazione della frequenza dei due tipi di mutazione può essere utile per comprendere i meccanismi di evoluzione delle sequenze di DNA e il livello di pressione selettiva a cui sono sottoposte (Yang *et al.*, 1997). Un'analisi di questi parametri individua come dN il numero di sostituzioni non-sinonime per sito non sinonimo e come dS il numero di sostituzioni sinonime per sito sinonimo. Rapporti dN/dS variabili tra i siti di un gene possono suggerire quali regioni si trovano sotto pressione selettiva positiva e le restrizioni funzionali che le caratterizzano. Questo rapporto tra dN e dS, definito come omega (ω), viene frequentemente utilizzato come indice della pressione selettiva a cui sono sottoposti i geni (Yang *et al.*, 1997).

Propositi dello Studio.

Le sequenze geniche sono soggette a mutazioni casuali che vengono selezionate positivamente o negativamente; questo processo porta ad ottimizzare la struttura proteica per consentirne la migliore funzionalità alle particolari condizioni chimico-fisiche dell'habitat in cui l'organismo vive. Quindi il mio progetto di tesi si basa sull'assunto che l'adattamento degli organismi alle condizioni estreme può essere valutato in funzione della variazione della composizione nucleotidica dei loro geni rispetto a quelle degli organismi non piezofili. L'approccio utilizzato in questa tesi si basa sul fatto che i valori di ω ottenuti dal confronto tra i geni ortologhi di organismi piezofili e non, rappresentano una misura della selezione. In particolare si andrà a selezionare una serie di geni ortologhi degli organismi i cui valori di ω relativi ai confronti piezofilo-non piezofili siano significamente più alti dei rispettivi ω relativi ai confronti tra non piezofili. Ciò che interessa maggiormente è vedere se la selezione opera in modo più marcato a carico dei geni coinvolti in processi particolari che si presume, dati studi precedenti, essere influenzati dall'alta pressione (Vezzi *et al.*, 2005). Dato il numero sempre maggiore di genomi batterici che vengono attualmente sequenziati è diventato possibile confrontare tutti i geni ortologhi di numersi organismi appartenenti ad una stessa famiglia. Questo permette di avere a disposizione i dati necessari per l'applicazione dell'analisi del tasso di sostituzione aminoacidica ai batteri piezofili che si è interessati a studiare. Infatti questo lavoro si inserisce in un più ampio progetto di ricerca sull'adattamento dei microorganismi all'ambiente abissale, nato dalla collaborazione del laboratorio del Prof. G. Valle con il gruppo del Prof. D. Bartlett dello *Scripps Instution of Oceanography*, San Diego. In precedenza era già stato completato il sequenziamento del genoma del batterio moderatamente piezofilo *P. profundum* ceppo SS9 mentre durante l'internato di laurea ho preso parte al *finishing* della sequenza di *S. benthica*. A partire dai risultati così ottenuti è stata possibile l'identificazione dei geni ortologhi e le successive analisi bioinformatiche.

MATERIALI E METODI

Data la diversità delle analisi svolte nel corso del progetto, questo capitolo è stato necessariamente suddiviso in due sezioni distinte.

3.1 Metodologie Biomolecolari Utilizzate.

Il materiale biologico di partenza è stato gentilmente fornito dal Dott. F. Lauro sotto forma di estratto di DNA genomico di *S. benthica* ceppo KT99 e di piastre contenenti i cloni della libreria fosmidica, trasformati nel ceppo batterico *E. coli* TransforMax™ EPI300™ (Epicentre Biotecnologies) con genotipo: F ¯ *mcrA Δ(mrr-hsdRMS-mcrBC) Φ80dlacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ ¯ rpsL nupG trfA dhfr*.

Procedure e Materiali di Base

Legenda di abbreviazioni e termini di uso comune:

abs = assoluto
AF = autoclavato e filtrato
bijoux = contenitore sterile da 8 ml
°C = gradi Celsius
dNTPs = dATP + dTTP + dCTP + dGTP
EDTA = etilendiammino-tetra-acetato
eppendorf = provetta di polipropilene da 1,5 ml o 2 ml
EtBr = bromuro d'etidio
EtOH = etanolo
EtOH abs = etanolo assoluto
falcon = provetta graduata di polipropilene
h = ore
H_2O mQ= acqua purificata mediante il sistema Milli RO 15 (Millipore) o simili
HCl = acido cloridrico
KAc = acetato di potassio
kbp = paia di chilobasi
KCl = cloruro di potassio
LB = loading buffer
min = minuti
Mpb= mega paia di basi (10^6 pb)
NaAc = acetato di sodio
NaCl = cloruro di sodio
NaOH = idrossido di sodio
O/N = tutta la notte (over night)
pb = paia di basi
rpm = rotazioni per minuto
SDS = sodio dodecil solfato
sec = secondi
TE = Tris-EDTA
Tris = Tris-idrossimetilamino-metano
w/v = peso/volume

Tamponi, soluzioni e terreni di crescita:

1. Tampone TAE 50X
   - 2 M Tris
   - 0.05 mM EDTA
   - 2 M acido acetico
   - portare a volume con H_2O deionizzata e autoclavare
2. Tampone Taq polimerasi 10X
   - 0.2 M Tris-HCl (pH 8.3)
   - 0.5 M KCl
   - 1% Tween 20
   - 20 mM $MgCl_2$
3. TE buffer
   - 10 mM Tris-HCl (pH 8.0)
   - 1 mM EDTA (pH 8.0)
4. Terreno di coltura LB
   - 1 % bacto-triptone
   - 0.5 % estratto di lievito
   - 1 % NaCl
   - pH 7 finale
5. Soluzione di risospensione (S1)
   - 50 mM Tris-HCl
   - 10 mM EDTA
   - 100 µg/ml Rnasi A
   - pH 8.0
6. Soluzione di lisi (S2)
   - 200 mM NaOH
   - 1% SDS
7. Soluzione di neutralizzazione (S3)
   - 2.8 M KAc
   - pH 5.1

Il marcatore di peso molecolare impiegato come confronto nelle corse elettroforetiche è Gene Ruler^{TM} 1 kbp DNA ladder (MBI Fermentas).

Protocolli per Estrazione ed Amplificazione del DNA

Il vettore fosmidico utilizzato al Venter Institute per la creazione delle librerie fosmidiche genomiche è pCC1FOS (*Epicentre Biotecnologies*). La peculiarità di questo vettore è che normalmente viene mantenuto a singola copia nella cellula ospite, ma può essere indotto a multicopia tramite aggiunta di L-arabinosio (0,01% w/v finale).

Estrazione di DNA fosmidico tramite lisi alcalina.

Il primo giorno il protocollo prevede il preinoculo del clone selezionato dalle piastre contenenti le librerie fosmidiche in 1,8 ml di LB + cloramfenicolo (12,5 µg/ml) e la sua crescita O/N a 37°C in incubatore orbitale (230rpm).
Il secondo giorno si arresta la crescita dei fosmidi e, al momento del re-inoculo, si aggiungono 0,2 ml di coltura a 2,8 ml di LB + cloramfenicolo (12,5 µg/ml), quindi circa 1/15 del volume finale sarà costituito dalla coltura batterica del primo giorno. Si incuba per un ora a 37°C e 230 rpm, per poi aggiungere in ciascun re-inoculo dell'L-arabinosio pari ad un 0,01% w/v finale. A questo punto si incuba O/N.
Il terzo giorno i batteri vengono trasferiti in provette eppendorf e centrifugati a massima velocità per qualche minuto a 4°C (centrifuga Eppendorf 5415R). Si risospende la coltura in 0,3 ml di soluzione S1 per poi compiere la lisi dei batteri aggiungendo 0,3 ml di soluzione S2 e mescolando delicatamente il tutto tramite 5-10 inversioni. Trascorsi un massimo di 5 min si neutralizza tramite l'aggiunta di 0,3 ml di soluzione S3; anche in questo caso si mescola delicatamente tramite 5-10 inversioni. Si collocano le eppendorf per qualche minuto in ghiaccio per poi centrifugarle per 15 min a 4°C a massima velocità (centrifuga Eppendorf 5415R).
Si raccoglie quindi il surnatante e lo si mette in una nuova eppendorf da 1,5 ml. Dopo aver aggiunto 0,7 ml di isopropanolo e aver mescolato per inversione, si centrifuga per 40 min a massima velocità a 4°C. Si compiono quindi due lavaggi successivi con etanolo 70%, ogni volta 15 min, a massima velocità e 4°C: la prima volta con 0,5 ml, la seconda con 0,16 ml. Si centrifuga a vuoto per rimuovere gli eventuali residui di etanolo.
Infine, dopo aver lasciato seccare il pellet, lo si risospende in 30 µl di acqua mQ AF preriscaldata a 65°C, temperatura utile ad eliminare eventuali DNasi.

Metodologia di amplificazione di DNA tramite reazione a catena della polimerasi (PCR).

Per poter ottenere un quantità di DNA sufficiente alla successiva reazione di sequenziamento, diverse regioni target del genoma di *S. benthica* ceppo KT99 sono state amplificate tramite reazione a catena della polimerasi. Ciascuna reazione ha previsto la preventiva progettazione da parte mia di una coppia di oligonucleotidi (che indicherò come *primer for* e *primer rev*) che fossero allo stesso tempo specifici e compatibili tra loro.

La reazione di PCR viene effettuata in tubini da PCR da 0,2 ml (STARLAB) in ciascuno dei quali è stata aliquotata la seguente miscela di amplificazione:

0,4 µl	DNA genomico (pari a 10ng)
0,4 µl	*primer for* (10 µM)
0,4 µl	*primer rev* (10 µM)
0,4 µl	dNTPs (10 mM)
0,6 µl	$MgCl_2$ (50 mM)
2 µl	tampone 10X Taq polimerasi
0,2 µl	Taq polimerasi (5U/µl) Polymed
15,6 µl	H_2O mQ AF
20 µl	

Tutte le reazioni sono state svolte nel termociclatore *Mastercycler ep gradient* (Eppendorf), programmato per eseguire il seguente ciclo di amplificazione:

1. 95°C 5 min (denaturazione del DNA stampo)

2. 30 cicli di:
   - 95°C 20 sec (denaturazione del DNA stampo)
   - 55°C 30 sec (ibridazione dei *primer* con il DNA stampo)
   - 72°C 3 min (estensione delle eliche di nuova sintesi)

3. 72°C 10 min (estensione finale delle eliche sintetizzate)

Elettroforesi e Sequenziamento

La metodica di elettroforesi su gel di agarosio è stata utilizzata per verificare il successo nell'estrazione del DNA fosmidico descritta in precedenza e per stimare le quantità di DNA ottenute. Allo stesso modo anche i prodotti di PCR, che sono frammenti di DNA abbastanza corti e quindi facilmente sequenziabili, vengono sottoposti a verifica tramite elettroforesi per stimare le dimensioni precise degli amplificati e il loro quantitativo. Per l'analisi dei risultati si prepara un gel di agarosio (allo 0,8% nel caso dell'estrazione dei fosmidi, all'1,5% nel caso dei prodotti di PCR) e per ciascun campione si caricano sul gel 3 µl. La stima della concentrazione del DNA dei campioni è necessaria in quanto la resa della reazione di sequenziamento è dipendente dalla quantità di templato utilizzato nella stessa. Come regola generale i prodotti di PCR sono stati sequenziati prelevando un numero di µl pari a 2 ng di DNA ogni 100 basi di amplificato che sono stati messi in un tubino da PCR da 0,2 ml (STARLAB) e seccati.

Il DNA estratto dai cloni fosmidici è invece di notevoli dimensioni (40-45 kpb), in quanto costituito da inserto più vettore e quindi ne è richiesta una quantità maggiore (750-1000ng) affinché la reazione di sequenziamento abbia successo. Anche in questo caso si preleva il volume necessario che si trasferisce in un tubino da PCR da 0,2 ml (STARLAB) e si secca.

Dato che la reazione di sequenziamento richiede l'impiego di un *primer*, specifico per la regione in *finishing*, si aggiungono inoltre 3,2 µl del medesimo (10 µM), in ciascuno dei campioni fosmidici prelevati. Invece nel caso dei templati di PCR si mettono 0,64µl di *primer* (10 µM) in tubini indipendenti in quanto gli amplificati, prima della reazione di sequenziamento, devono essere purificati; una reazione che porterebbe alla degradazione dello stesso *primer* di sequenziamento.
La razione di sequenziamento dei campioni di DNA è stata svolta dai colleghi del gruppo di *sequencing* presente nel laboratorio dove ho svolto il tirocinio. Questa metodica si basa sull'utilizzo, in una reazione a cicli (*cycle sequencing*) simile alla PCR, di una miscela nella quale, assieme ai deossinucleotidi, sono presenti in una adeguata percentuale dei dideossinucletidi. Questi ultimi, quando incorporati dalla polimerasi, bloccano l'allungamento della catena nascente. Il prodotto finale di una simile reazione è un insieme di molecole di lunghezza variabile dipendente dal momento in cui uno dei quattro dideossinucleotidi è stato incorporato.
Il risultato del sequenziamento sono degli elettroferogrammi relativi alla regione genomica che si voleva migliorare. Per ognuno sono state verificate qualità e specificità della sequenza e, quando tali parametri risultavano soddisfatti, si è potuto aggiungere le nuove sequenze all'assemblaggio.

3.2 Elaborazione Bioinformatica dei Dati.

La bioinformatica è una branca della biologia in rapida evoluzione altamente interdisciplinare in quanto usa tecniche e concetti che derivano da informatica, statistica, matematica, chimica, biochimica e fisica. Il *National Center for Biointecnology Information* (NCBI) definisce la bioinformatica come la scienza nella quale biologia, informatica e tecnologia dell'informazione si uniscono in un'unica disciplina. Esistono tre importanti applicazioni dell'informatica utili nel campo biologico:
- lo sviluppo di nuovi algoritmi e statistiche con i quali valutare le relazioni tra i membri di un ampio data set;
- l'analisi e l'interpretazione di vari tipi di dati che includono sequenze aminoacidiche e nucleotidiche, domini e strutture proteiche;
- lo sviluppo e l'implementazione di strumenti, o *tool*, che aumentano l'efficienza di accesso e gestione dei differenti tipi di informazione.

Assemblaggio delle Sequenze: *PhredPhrap* e *Consed*

Per allineare tra loro le sequenze, o *reads*, ottenute dal WGS sono stati utilizzati degli appositi *software* di assemblaggio. Questi valutano le possibili sovrapposizioni di sequenza in modo da determinare la migliore soluzione di allineamento, generano una sequenza consenso per ogni contiguo e attribuiscono un valore di affidabilità ad ogni base del consenso.

Per il progetto di sequenziamento di *S. benthica* ceppo KT99 è stato utilizzato il programma *Phrap* (*phragment assembly program*) sviluppato da Phil Green dell'University of Washington Genome Center di Seattle. Il programma, che è stato inizialmente creato per l'assemblaggio di sequenze *shotgun* di cloni BAC, è incorporato in una serie di procedure automatiche definite dallo *script phredPhrap* che lancia, in successione, i programmi di elaborazione delle sequenze (Ewing *et al.*, 1998 a-b).

Ciò che avviene ogni volta che viene lanciato lo *script phredPhrap* è:

- tutte le sequenze appena introdotte nella cartella *chromat_dir* sono analizzate dal programma *Phred* che assegna ad ogni picco dell'elettroferogramma, quindi ad ogni nucleotide, un valore che dipende dalla posizione dei picchi adiacenti, dall'area del picco in esame e dal rapporto tra l'intensità del picco e il rumore di fondo. I risultati ottenuti sono salvati in nuovi file che mantengono il nome delle sequenze originali con l'estensione ".phd"; tali file sono conservati nella cartella *phd_dir*.
- viene lanciato il programma *DetermineReadType* che estrae dal nome della sequenza i dati accessori relativi al campione (libreria di provenienza, *primer* e chimica di sequenziamento) e li inserisce in coda nel corrispondente file .phd.
- successivamente *Phrap* allinea le sequenze in modo da generare i *contig* tenendo conto della qualità delle singole basi attribuita da *Phred*, quindi per ciascuna posizione della sequenza consenso ottenuta viene calcolata la corrispondente probabilità di errore. In ciascun *contig* esisteranno così delle regioni ad alta e bassa qualità, in genere dovute ad una alta o bassa copertura di sequenza, che avranno una valutazione diversa nell'assemblaggio globale.

L'*output* dell'assemblaggio eseguito da *Phrap* viene salvato nella cartella *edit_dir* in un file con estensione ".ace". Quando sono prodotte altre *read* e viene lanciato un nuovo assemblaggio tutte le sequenze presenti in *chromat_dir* vengono riallineate, per cui si possono ottenere risultati differenti dall'assemblaggio precedente; *l'output* sarà salvato in un file ".ace" differente.

I dati di assemblaggio generati da *Phrap* sono salvati in un file di testo alquanto complesso che non può essere facilmente esaminato. Per risolvere questo problema è stato sviluppato il programma *Consed* (*consensus editor*) da David Gordon dell'University of Washington Genome Center di Seattle (Gordon *et al.*, 1998).

Consed è un'interfaccia grafica che permette di visualizzare ed eventualmente modificare manualmente i risultati di assemblaggio. Tutti i contigui generati da *Phrap*, le sequenze che li compongono e i singoli elettroferogrammi delle stesse possono essere quindi analizzati e valutati.
La qualità di ciascuna regione del *consensus*, così come quelle delle singole posizioni nucleotidiche, può essere quindi facilmente intuita in quanto a qualità (= probabilità di errore) diverse sono assegnate tonalità differenti di grigio. Anche la presenza di sequenze ripetute o di eventuali problemi in assemblaggio è evidenziata mediante l'utilizzo e la sovrapposizione di colori forti. Inoltre il programma è in grado di generare, se richiesto, una serie di finestre nelle quali viene riportata per ciascun contiguo la tipologia e la posizione nella sequenza consenso dei diversi punti problematici riscontrati in assemblaggio.
Tutto questo permette all'operatore di concentrare il lavoro sulle porzioni della sequenza a più bassa qualità o con possibili errori di allineamento, rendendo più agevole il lavoro di *finishing*. In realtà gli stessi contigui creati da *Phrap* devono essere analizzati in modo fine per poter individuare possibili regioni di allineamento sbagliato, infatti la presenza di sequenze ripetute complica il processo di assemblaggio per la possibile creazione di contigui chimerici, dove le due regioni ai lati della porzione ripetuta non riflettono la reale posizione delle stesse nel genoma.

Finishing del Genoma

Verranno ora descritti i metodi applicati per ridurre il numero di *contig* risolvendo gli assemblaggi sbagliati e identificando quando possibile l'esatta struttura della sequenza ripetuta o *repeat*.
Ciascun contiguo è costituito da un insieme di sequenze generate a partire dalle due estremità dell'inserto di DNA, le così dette *paired pairs*. Il dato che accompagna le *paired pairs* è duplice in quanto ciascuna fornisce informazione di sequenza nonché indicazione di posizione: le due estremità dei cloni plasmidici sono in genere ad una distanza di 2000 paia di basi l'una dall'altra, così come le due estremità dei cloni fosmidici sono in genere ad una distanza di 30000-35000 paia di basi. E' quindi l'analisi della corretta distanza ed orientamento tra le due sequenze ottenute dal medesimo clone che permette di creare una mappa fisica (*scaffolding*) del genoma e di individuare gli eventuali contigui chimerici. In base all'informazione di posizione fornita dalle *paired pairs* è possibile allestire diversi esperimenti utili sia a chiudere i *gap* tra i *contig* e migliorare la qualità del consenso sia a risolvere le regioni in *repeat*. Nel primo caso i cloni fosmidici, le cui sequenze collegano due differenti contigui, possono essere utilizzati come stampo in una reazione di sequenziamento svolta in presenza di un oligonucleotide interno di nuova sintesi.

E' quindi necessario selezionare gli eventuali *primer* che permettessero di chiudere i *gap* tra i contigui e, per ciascun esperimento, determinarne la sequenza e i cloni da utilizzare come templato. Inoltre, dato che la costruzione della mappa fisica ha richiesto l'esame di ciascun contiguo, sono stati individuati diversi esperimenti per risolvere le regioni ripetute e quelle a bassa qualità. In alcuni casi gli esperimenti di *finishing* hanno invece previsto l'allestimento di reazioni di PCR, per ciascuna delle quali è stata individuata una coppia di *primer* specifici per la regione genomica in analisi. I prodotti di PCR sono stati successivamente utilizzati nelle reazioni di sequenziamento. Invece per risolvere le *repeat* è stato necessario:

- recuperare le sequenze che cadono in una regione ripetuta, ma che presentano l'altra estremità del clone in una regione unica;
- integrare le *reads* recuperate con quelle che appartengono solo a parte della porzione ripetuta, come ad esempio x e y nella figura 4;
- creare il nuovo assemblaggio in esperimenti specifici, accompagnato dalla verifica della corretta distanza tra i *paired pairs.* Tutto questo permette spesso di ottenere il corretto consenso della regione ripetuta. Anche nel caso in cui la *repeat* non fosse stata risolta, la presenza dell'assemblaggio specifico permette di isolare e quindi di lavorare sulla singola unità ripetuta separandola dal contesto genomico.

Questa metodologia permette di assemblare in modo corretto le regioni genomiche identificando per inconsistenza di posizione i contigui chimerici. Un esempio di scorretto allineamento di sequenze è ripotato nella sezione C della figura 4.

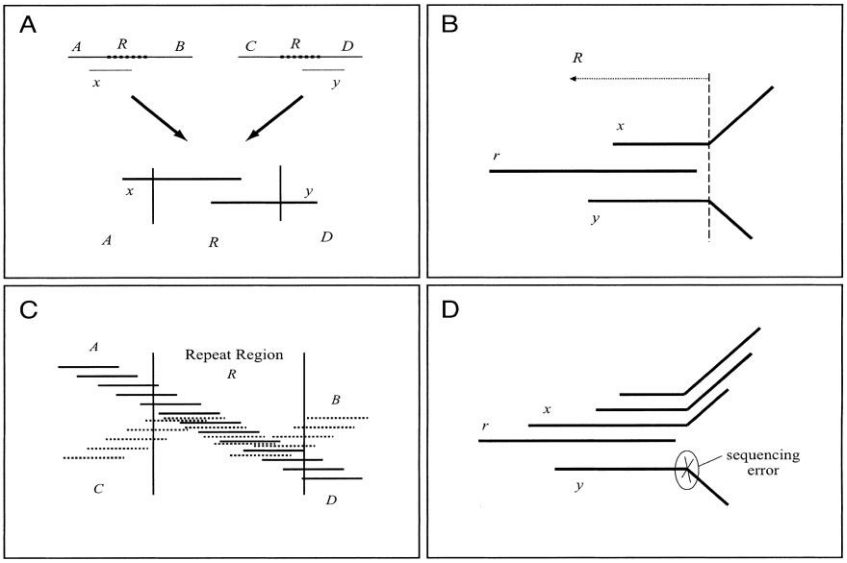

Fig. 4. Viene qui riportata un'immagine tratta da Batzoglou *et al.*, 2002 per chiarificare la procedura utilizzata per la risoluzione delle *repeat* (R). x e y rappresentano sequenze che sono solo in parte nella regione ripetuta, mentre r è una *read* completamente in *repeat,* quindi impossibile da collocare nel corretto contiguo. Nella sezione C viene schematizzata una regione ripetuta in due contigui diversi che collassano tra loro; le sequenze tratteggiate appartengono ad uno mentre quelle continue all'altro.

Organizzazione e Gestione Dati: MySQL e PERL

Lo *Structured Query Language* (SQL) è sicuramente l'interfaccia tra database ed utente più diffusa; è uno standard per l'accesso, la definizione e la gestione di database di cui esistono implementazioni come ORACLE e MySQL. Quest'ultimo viene generalmente utilizzato per la memorizzazione, l'indicizzazione e la conservazione delle informazioni biologiche. Perciò è stato sfruttato in questo lavoro di tesi per la progettazione di un database ad hoc, chiamato *Genomi*, per organizzare in modo ordinato tutti i dati relativi agli organismi oggetto dello studio.

I *record* rappresentano l'unità di memorizzazione del database e, all'interno di ognuno, ci sono parole chiave che permettono di individuare i diversi campi; in questo progetto di tesi è stato creato un record per ogni gene considerato. Esiste inoltre un campo (PID) che identifica univocamente ogni record del database ed è stato definito intenzionalmente in modo da essere identico al PID di riferimento dell'NSBI, per maggiore praticità.

La computational biology è la disciplina che si occupa dell'analisi dei dati tipo la ricerca di similarità tra sequenze, lo studio dei genomi, l'allineamento multiplo di sequenze e le analisi filogenetiche. Per fare tutto ciò è necessario possedere degli strumenti che permettano, di volta in volta, di adattare il proprio metodo d'indagine al tipo specifico di dati da trattare.

Il *Practical Extraction and Report Language* (PERL) è un linguaggio di programmazione nato negli anni '80 che svolge perfettamente questo ruolo. PERL per la sua praticità, potenza e compattezza, soprattutto nell'estrazione di informazioni da file, o *parserizzazione*, è stato largamente utilizzato per collegare tra loro le varie fasi del progetto. Sono stati infatti scritti più di 20 piccoli programmi, o *script*, al fine di interagire con il database Genomi, gestire la manipolazione delle migliaia di sequenze geniche nei numerosi passaggi intermedi delle analisi, parserizzare i file di *output* degli altri programmi e crearne i particolari file *input* richiesti.

Selezione dei Geni Ortologhi: BLAST.

Il programma *Basic Local Alignment Search Tool* (BLAST) compie ricerche di similarità tra una sequenza *query* e un database, utilizzando un metodo euristico che permette di abbreviare i tempi di esecuzione. Tale metodo prevede dapprima la ricerca di brevi pattern identici tra le due sequenze e, solo dopo, l'allineamento delle sequenze a partire da tali pattern (Altschul *et al.*, 1990).

Esistono diverse versioni del programma che consentono di allineare tra loro due o più sequenze sia a livello di acidi nucleici sia di poteine:
- blastn che cerca in un database di sequenze nucleotidiche a partire da una sequenza *query* di DNA;
- blastp che cerca in un database di sequenze aminoacidiche a partire da una sequenza *query* proteica;
- blastx che cerca in un database dati di sequenze aminoacidiche, a partire da una sequenza di nucleotidi, dopo averla automaticamente tradotta considerando i sei possibili *frame* di lettura;
- tblastn che cerca in un database di sequenze nuleotidiche tradotte a partire da una sequenza *query* proteica;
- tblastx che cerca in un database di sequenze nuleotidiche tradotte a partire da una sequenza *query* tradotta di nucleotidi;

BLAST è stato impiegato svariate volte nel corso del progetto sia per selezionare quali organismi batterici, tra quelli presenti all'NCBI, fossero più adatti per essere confrontati con i due batteri piezofili sia per selezionare i geni ortologhi all'interno delle famiglie di Vibrionaceae e Shewanellaceae.
La procedura per identificare quali organismi scegliere ha previsto l'uso di BLAST per confrontare a coppie le sequenze proteiche in formato FASTA del genoma dei batteri e la parserizzazione dei suoi risultati con i programmi scritti in PERL per estrarre il numero di geni in comune, considerando solo gli allineamenti significativi. Quindi, per quanto riguarda l'identificazione dei geni ortologhi, è stato utilizzato blastp per confrontare tutti i file FASTA dei genomi dei microorganismi, cercando le sequenze dei non-piezofili come *query* sul database creato con le sequenze dei piezofili e viceversa. L'*output* ottenuto per ogni coppia di batteri è stato quindi parserizzato con gli appositi programmi scritti in PERL per estrarre una lista di geni che rispettasse *e*-value, identità ed estensione dell'allineamento fissati come soglia in base alla letteratura (Tatusov *et al.*, 1997). Lo stesso procedimento di *parsing* è stato effettuato sia sugli *output* dei blastp sulle coppie con il genoma del batterio piezofilo come database che sui reciproci per eliminare dalla lista i geni paraloghi grazie alla selezione del *best reciprocal hit*. In base alle liste così ottenute sono stati prescelti per le analisi successive i batteri che avevano il maggior numero di ortologhi in comune rispettivamente con *S. benthica* e *P. profundum*. Infine un altro *script* compila i due elenchi definitivo dei geni ortologhi comuni, uno per le Shewanellaceae e uno per le Vibionaeae.

Calcolo di Sostituzioni Aminoacidiche e CAI

Questa parte dell'analisi è stata svolta parallelamente e allo stesso modo per entrambe le famiglie batteriche. Quindi per facilitare la descrizione della procedura si farà riferimento solamente ad una di esse.

Per poter calcolare il tasso di sostituzione aminoacidica sono state innanzitutto estratte dal database Genomi le sequenze dei geni ortologhi comuni, quindi è stato necessario allineare in file indipendenti per ogni gene le sequenze proteiche dei quattro organismi con *Clustal W*, un pacchetto di programmi per l'allineamento multiplo di sequenze (Thompson *et al.*, 1994). In seguito il programma *RevTrans* legge il set di sequenze peptidiche allineate correttamente da *Clustal W* e utilizza le corrispondenti sequenze di DNA fornitegli per costruire una versione inversamente tradotta dell'allineamento (Wernersson *et al.*, 2003). Questi passaggi, convenientemente uniti da adeguati *script* in PERL, sono stati necessari affinchè gli allineamenti rispettassero il frame di lettura delle proteine codificate e lo standard *input* del prossimo programma.

Per calcolare il tasso di sostituzione sinonimo (dS) e non sinonimo (dN) per ogni codone di ogni gene ortologo in confronti a coppie dei vari organismi considerati è stato impiegato il programma *yn00* del pacchetto per *Phylogenetic Analysis by Maximum Likelihood* (PAML) (Yang, 1997). Dall'*output* di *yn00* sono stati estratti, per ogni ortologo, i 6 valori di ω (= dN / dS) relativi alle possibili combinazioni del confronto delle sequenze dei 4 batteri.

Infine con il programma *CodonW* sono stati calcolati i valori di *codon adaptation index* (CAI) per tutti i geni di ogni microorganismo, in funzione di una lista selezionata ad hoc dei geni maggiormente utilizzati (Peden, 1999). *CodonW* è un *software* utilizzato per semplificare l'analisi multivariata, o analisi di corrispondenza, dell'utilizzo dei codoni e degli aminoacidi. Esso consente anche il calcolo indici standard di utilizzo dei codoni. Questa lista è stata ottenuta sulla base di una serie di esperimenti di *microarray* condotti su *P. profundum* SS9 nel laboratorio del Prof. Valle (Vezzi *et al.*, 2005; Campanaro *et al.*, 2005) o ottenuti dal database ArrayExpress (http://www.ebi.ac.uk/) per quel che riguarda *Vibrio cholerae* e *Shewanella oneidensis*. Da questi esperimenti sono stati selezionati solamente i geni aventi valori di espressione molto elevati, considerando una soglia minima di valore di fluorescenza pari a 10000.

Tutti i dati, appena sono stati ottenuti, sono stati sistematicamente inseriti nel database Genomi.

SAM e *GoMiner*

Significance Analysis of Microarrays (SAM) è un programma statistico che viene utilizzato per identificare i geni in repliche multiple indipendenti di esperimenti che coinvolgono grandi numeri di dati, come per esempio nell'analisi dei *microarrays* o nel caso di questo progetto per i valori di ω. Tuttavia si può usufruire di questo *software* in qualsiasi applicazione che preveda di testare un gran numero di ipotesi indipendenti (Tuscher *et al.*, 2001). La sua popolarità deriva dal fatto che può essere aggiunto al programma di calcolo EXCEL come componente aggiuntivo e, pertanto, risulta di semplice utilizzo.

SAM implementa un t-test modificato che consente di risolvere elegantemente e in maniera statisticamente rigorosa i problemi che si presentano quando si testano un gran numero di ipotesi indipendenti e vengono calcolati un numero enorme di valori di *p*. Come noto *p* rappresenta la probabilità di osservare un valore del test uguale o più estremo del valore ottenuto dal campione sotto l'ipotesi nulla (H_0); misura quindi l'evidenza fornita dai dati contro l'ipotesi nulla: minore è il valore di *p*, più forte è l'evidenza contro H_0.

Ogni volta che viene testata un'ipotesi nulla esiste la probabilità di rigettarla erroneamente. Maggiore è il numero di ipotesi testate più alta è la possibilità di rigettare erroneamente H_0 e quindi il numero di falsi positivi che si accettano. Pertanto se si testano K ipotesi indipendenti la probabilità che i test siano congiuntamente non significativi è data da:

$(1 - \alpha)^K$ dove α è il coefficiente di confidenza del test

Ne consegue che la probabilità di avere almeno un test significativo è:

$1 - (1 - \alpha)^K$.

Quindi risulta che, se si testano un gran numero di ipotesi diventa molto probabile di commettere degli errori *tipo I* rigettando erroneamente l'ipotesi nulla.

Nel nostro caso H_0 è che i valori di ω ottenuti dal confronto tra un batterio piezofilo ed uno non-piezofilo non siano significativamente differenti da quelli ottenuti nel confronto tra due batteri non piezofili.

Il programma utilizza la seguente strategia: calcola una statistica *t* per ogni gruppo di valori, nel nostro caso i valori di ω dei sei confronti, in cui una statistica *t* viene definita come il rapporto della differenza media del gruppo divisa per la deviazione standard stimata. Questi valori vengono chiamati statistiche *t* sperimentalmente osservate.

Poi vengono effettuate delle permutazioni e successivamente calcolate le statistiche *t* per ciascun gene, basate sulle permutazioni dei gruppi di campioni che nel nostro caso sono i valori di ω del confronto tra piezofilo contro non piezofilo e quelli dei confronti tra non piezofili. Queste vengono dette statistiche *t* calcolate.
Il programma genera così un grafico delle statistiche *t* osservate rispetto a quelle calcolate. Questo punto è il fulcro del programma SAM che confronta questi due set di statistiche ed identifica i geni realmente significativi selezionando quelli con una statistica *t* osservata più grande della statistica *t* calcolata.
Il programma permette quindi all'utilizzatore di definire un valore Δ che consente di definire quanto maggiore deve essere il valore ottenuto dalla statistica *t* osservata, in confronto alla statistica *t* calcolata, affinchè un gene possa essere definito significativo. Infine viene calcolato il *False Discovery Rate* (FDR) sulla base dei dati sperimentali originari, i dati permutati ed il valore di Δ definito. Il metodo consente infatti di determinare la probabilità di avere un'ipotesi nulla falsamente rigettata (cioè un falso positivo) data una lista di ipotesi nulle rigettate. In pratica un FDR del 5% indica che in una lista di geni identificati come significativi, il 5% sono falsi positivi. Il FDR viene calcolato per ogni gene "i", considerando il suo valore p_i:

$$FDR_i = [(p_i\ N)/K_i]$$

dove N e' il numero di geni ortologhi e K_i è il numero di geni che hanno un valore *p* minore di quello del gene in questione.
Vengono poi scelti i geni in base al loro FDR al posto del valore *p*. Il programma calcola quindi il valore di *q* per ogni gene che definisce il valore minimo di FDR a cui quel gene risulta significativo. Questo valore è quello che è stato considerato per la selezione dei risultati ottenuti dall'analisi e discussi nel paragrafo 4.4 ed è riportato nella tabella 3.

Viene di seguito descritta la procedura di identificazione delle categorie di Gene Ontology statisticamente arricchite, tramite il programma *GoMiner*™ (Zeeberg *et al.*, 2003). *GoMiner*™ è un *tool* per l'interpretazione biologica dei dati che risulta molto utile per le analisi della distribuzione di un gran numero di geni all'interno delle categorie di *Gene Ontology* (GO) (Ashburner *et al*, 2000). La GO organizza i geni in categorie gerarchiche sulla base del loro processo biologico, della funzione molecolare e della componente cellulare di cui fanno parte.
Gli esperimenti di genomica spesso generano delle liste di decine o centinaia di geni che nel loro insieme difficilmente possono essere interpretati nel loro significato biologico globale. Per questo sono essenziali degli strumenti informatici che ne facilitino l'interpretazione.

GoMiner interroga il database della GO per identificare i processi biologici, le funzioni molecolari e le componenti cellulari con cui sono annotati i geni nelle liste che gli vengono fornite. Il suo scopo infatti è quello di evitare l'analisi condotta gene per gene, classificandoli invece all'interno di categorie biologiche coerenti.
Il programma produce dei file di *output* di tipo quantitativo e statistico, e permette di visualizzare i dati in differenti maniere, utili ad una migliore comprensione del risultato dell'esperimento. Fornisce poi un'analisi statistica condotta su ogni classe di GO utilizzata nell'annotazione delle due liste dei geni che gli vengono date come *input* secondo la procedura di seguito descritta: il set totale dei geni, nel nostro caso tutti i geni ortologhi individuati per Shewanellaceae e Vibrionaceae, ed un *subset* di questo gruppo, nel nostro caso i geni identificati dall'analisi statistica effettuata con SAM, descritta precedentemente.
GoMiner effettua un test statistico, il test di Fisher a due code, per identificare le categorie di GO in cui i geni della sottolista in esame risultano sovra-rappresentati rispetto all'atteso, cioè rispetto ad una distribuzione casuale basata semplicemente sulla rappresentatività dei geni nelle differenti classi di GO. La statistica che viene eseguita produce un valore di *p* per ogni categoria di GO che si basa sul test condotto sull'ipotesi nulla (H_0). Nello studio sugli ortologhi H_0 indica che una certa categoria non è arricchita nè presenta una sotto-rappresentazione dei geni della sottolista rispetto a quanto atteso sulla base solamente del caso.

L'Ambiente R per le Analisi Statistiche

R è un pacchetto di *software* adatti alla manipolazione dei dati, all'analisi statistica ed alla visualizzazione grafica. Esso è costituito da differenti parti: strumenti per lo stoccaggio dei dati e la loro manipolazione, un insieme di operatori per effettuare calcoli sugli *array*, un ampio set di *tool* integrati per l'analisi, strumenti grafici per la visualizzazione dei dati e un linguaggio di programmazione detto "S".
Il termine "ambiente" con cui ci si riferisce a R è dovuto al fatto che esso è un sistema completamente pianificato e coerente. Il *software* statistico SAM, pur essendo una componente aggiuntiva di EXCEL, per funzionare richiede l'installazione di R. Lo stesso programma esiste comunque anche come "pacchetto" aggiuntivo per R che funziona in maniera indipendente da EXCEL.
L'utilizzo di questo ambiente si è reso necessario per quattro scopi principali descritti di seguito.

- R consente di generare degli istogrammi che riportano la frequenza o la numerosità di una particolare distribuzione di dati entro una serie di intervalli definiti, come nel caso della figura 7 nel paragrafo 4.3.

- E' stato inoltre usato per calcolare il coefficiente di correlazione di Spearman tra il valore di ω dei geni ortologhi ed il corrispondente valore di CAI. Nei calcoli statistici è stata utilizzata la correlazione di Spearman (ρ) poichè, a differenza di quella di Pearson, non richiede l'assunzione che la relazione tra le due variabili sia lineare.
ρ dà inoltre una misura non parametrica di correlazione, cioè consente di definire quanto una arbitraria funzione monotona è in grado di descrivere la relazione tra due variabili, senza fare nessun assunto a priori riguardo la distribuzione di frequenza delle variabili.
- R è stato utilizzato per l'analisi della regressione multipla che confronta il valore di ω di ogni gene, il corrispondente valore di CAI e l'appartenenza del gene ad una particolare classe di *Cluster of Orthologous Group* (COG) (Tatusov *et al*, 1997). Il proposito generale della regressione multipla è quello di comprendere la relazione esistente tra più variabili indipendenti e una singola variabile dipendente. I risultati relativi verranno discussi nel paragrafo 4.6.
- Infine R è stato impiegato per l'analisi statistica dell'arricchimento delle classi di COG, utilizzando le liste dei geni identificate dal *software* SAM, come descritto in precedenza. Questa analisi è stata fatta sulla base della distribuzione ipergeometrica.
In statistica essa rappresenta una distribuzione discreta di probabilità, che descrive il numero di successi in una sequenza di *n* campionamenti effettuati su una popolazione finita, senza reinserimento. In pratica su un insieme di N oggetti, in cui *m* sono in difetto, la distribuzione ipergeometrica descrive la probabilità che in un campione di *n* oggetti distinti, prelevati dall'insieme N, siano difettivi esattamente *k* oggetti.
In generale se una variabile casuale *X* segue la distribuzione ipergeometrica con parametri N, *m* ed *n*, allora la probabilità di avere esattamente *k* successi è data da:

$$f(k; N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

Ci sono $\binom{N}{n}$ possibili campioni, senza rimpiazzamenti, ci sono $\binom{m}{k}$ modi per ottenere *k* degli oggetti presenti in difetto e ci sono $\binom{N-m}{n-k}$ modi per prelevare il resto dei campioni non difettivi.

RISULTATI E DISCUSSIONE

Nel seguente capitolo verranno discussi quelli che sono i due obiettivi principali di questo progetto di tesi e i progressi certamente ottenuti per raggiungerli. Questi sono il completamento del genoma di *S. benthica* e l'analisi dei suoi geni per comprendere i meccanismi evolutivi nell'adattamento alle condizioni estreme abissali.

4.1 Progressi nell'Assemblaggio

Fondamentale per poter affrontare in modo accurato lo studio di un organismo è riuscire ad ottenere la sequenza dell'intero genoma con una precisione maggiore possibile. Con questo intento è stata condotta l'analisi per migliorare la qualità della sequenza di DNA del ceppo KT99 di *S. benthica*. Per una più immediata comprensione dei risultati ottenuti in questo ambito viene di seguito riportata la tabella 1.

	***Contig* Totali**	**Regioni in *Repeat***	**Sequenze Totali**	**Dimensioni *Contig* Maggiore**	**Sequenze *Contig* Maggiore**	**Errore *Contig* Maggiore**
Fasta.screen. ace.1	125	30	36937	231032 bp	1913	22.18 err/10 kbp
Fasta.screen. ace.2	73	12	37531	574092 bp	4047	1.27 err/10 kbp

Tab. 1. Descrizione delle caratteristiche dell'assemblaggio all'inizio (Fasta.screen.ace1) ed alla fine (Fasta.screen.ace.2) del mio lavoro di tesi. Nelle colonne sono riportati il numero di *contig* totale, le regioni ripetute da risolvere, il numero di sequenze totali presenti nell'assemblaggio, le dimensioni in paia di basi del *contig* più grande, il numero di sequenze che si assemblano a formarlo e la qualità della sequenza del *contig* più grande calcolata come numero medio di errori ogni 10000 paia di basi.

I progressi fatti nel *finishing* di questo genoma hanno portato ad un aumento medio delle dimensioni di tutti i *contig*, che da 125 si sono ridotti di numero fino a 73. Infatti nel primo file di assemblaggio, fasta.screen.ace.1, il maggiore tra i contigui aveva una dimensione di 231 kbp, mentre ora, dopo una fase preliminare di chiusura del genoma, la sequenza univoca più lunga è di 574 kbp. Questo *contig*, denominato 73 nel fasta.screen.ace.2, non è dato dall'aumento di dimensione del *contig* di 231 kb del primo assemblaggio ma è il risultato dell'unione di altri *contig* più piccoli.
Tutto ciò è stato possibile grazie ai 594 nuovi elettroferogrammi aggiunti, che, come descritto nel paragrafo 3.1 dei materiali e metodi, sono stati ottenuti per mezzo degli esperimenti di PCR e reazioni di sequenziamento a partire dai 600 oligonucleotidi sintetizzati ad hoc.

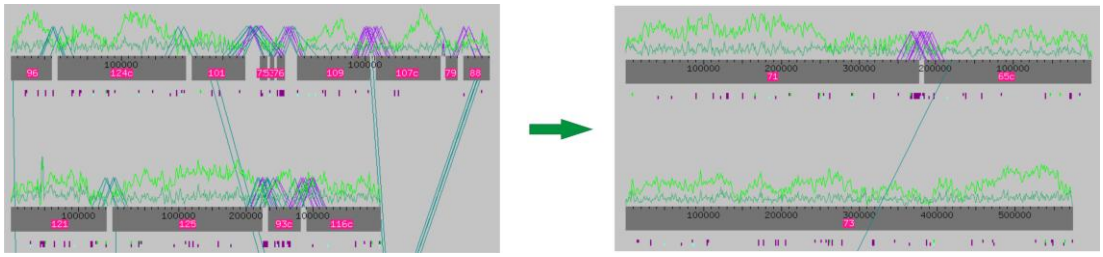

Fig. 5. Visualizzazione dei risultati dell'assemblaggio del genoma di *S. benthica* prima e dopo il lavoro di *finishin*g generato dal programma *Consed.*

Inoltre, date le 30 *repeat* che sono state individuate nel primo assemblaggio, è stato ottenuto il corretto consenso di 18 di queste regioni che ha contribuito all'esatto posizionamento di diversi contigui tra loro e ha favorito l'unione di alcuni di essi.
Infine, come evidenziato dalla tabella 1, il lavoro di *finishing* ha portato ad una forte diminuzione media dell'errore (calcolato su 10 kbp), che viene esemplificata dal confronto del valore di 22,18 err/10kbp del *contig* più grande del primo assemblaggio e il valore di 1,27 del *contig* 73 prima citato.
Questo risultato è fondamentale ai fini delle analisi svolte in questa tesi in quanto l'influenza del tasso d'errore sulla sequenza genomica può generare variazioni che vanno ad influenzare sia la predizione genica sia la sequenza stessa dei geni ortologhi del batterio, utilizzati in seguito. Infatti considerando il modello di stima dell'influenza degli errori di sequenziamento sui geni proposto da B. Dujon, si può facilmente capire quanto una variazione nell'accuratezza media da 99,9% a 99,99% implichi un aumento dell'esattezza delle sequenze dei geni dal 33% all'85% (Dujon, 1996).

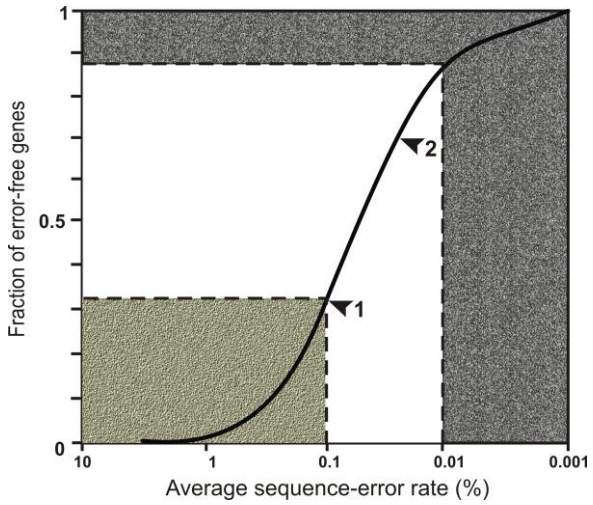

Fig. 6. Viene qui riportato il grafico tratto da Azam, 1998 relativo al rapporto dell'accuratezza media delle sequenze geniche in funzione del tasso medio di errore.

4.2 La Selezione dei Geni Ortologhi

Il punto cruciale di ogni progetto di studio è la scelta degli organismi oggetto dell'analisi, decisione che può essere presa in funzione di svariati parametri. Nel caso dei nostri batteri piezofili la scelta è stata guidata dal fatto che il *finishing* di entrambi è stato effettuato dal gruppo di ricerca del Prof. G. Valle. Inoltre un discreto numero di specie strettamente correlate filogeneticamente, e con la sequenza genomica completa, sono note sia per *S. benthica* che per *P. profundum*.
Il nostro approccio quindi ha previsto una selezione mirata ad identificare gli organismi che avessero il maggior numero possibile di geni ortologhi con i due piezofili. Sono state per questo scaricate dal sito ftp di GenBank tutte le sequenze dei genomi completi dei membri delle famiglie Shewanellaceae e Vibrionaceae. Seguendo quindi la procedura descritta nel paragrafo 3.2 (Selezione dei geni ortologhi: il programma BLAST), sono stati prescelti *V. fisheri*, *V. parahaemolyticus* e *V. vulnificus* per il confronto con *P. profundum* e *S. baltica*, *S. oneidensis* e *S. frigidimarina* per quello con *S. benthica*.
La procedura utilizzata per la selezione dei geni ortologhi prevede l'impiego di programmi scritti ad hoc in PERL, che hanno consentito di parserizzare gli *output* di BLAST. Il ceppo 3TCK di *P. profundum*, considerato inizialmente un buon candidato per le analisi, perchè piezosensibile, è stato in seguito escluso poichè avrebbe potuto sbilanciare le analisi essendo troppo vicino filogeneticamente.
L'utilizzo di tre organismi per il confronto con i piezofili è motivato dalla necessità di raggiungere un compromesso tra l'identificazione di un numero sufficientemente elevato di geni ortologhi comuni e un gruppo di organismi bastevole per conferire solidità statistica alle analisi. A questo punto è stato possibile creare una tabella MySQL in cui inserire rispettivamente i 2180 geni ortologhi delle Shewanellaceae e i 2174 geni delle Vibrionaceae. Il numero finale di geni così ottenuti nelle due famiglie risultava confacente alle aspettative in quanto simile, mentre si è deciso di fare una comune selezione dei geni solamente all'interno dei due gruppi separati. Questo perchè una ulteriore cernita di ortologhi estesa a tutti i microorganismi avrebbe potuto ridurre troppo il numero di geni finali.

Da qui in poi si è deciso di proseguire le analisi in modo indipendente sui due gruppi, in quanto era comunque possibile confrontare i risultati ottenuti alla fine dall'arricchimento dei geni all'interno delle classi funzionali di COG e GO. Il lavoro descritto nei prossimi paragrafi è stato svolto per entrambe le famiglie batteriche parallelamente e allo stesso modo di conseguenza per praticità si farà riferimento solamente ad uno dei due gruppi.

4.3 Tasso di Sostituzione Aminoacidica

Per poter effettuare i calcoli del tasso di sostituzione aminoacidica è stato necessario allineare le quattro sequenze di ogni gruppo di ortologhi, in modo da ottenere un *input* particolare. Infatti questo allineamento multiplo deve essere fatto in modo da mantenere necessariamente il *frame* di lettura del gene e rispettare i requisiti specifici del programma del pacchetto PAML. Per ottenere tale *input* sono stati implementati con *script* in PERL i programmi *Clustal W* e *RevTrans* come descritto nel paragrafo 3.2 (Calcolo delle sostituzioni aminoacidiche e CAI). La stima del tasso di sostituzione sinonimo e non-sinonimo sugli allineamenti multipli di tutti gli ortologhi e l'individuazione di una pressione selettiva sono quindi stati eseguiti dal programma *yn00* appartenente al pacchetto PAML.
Yn00 utilizza un'implementazione del metodo di Yang e Nielsen che calcola il tasso di sostituzioni non-sinonimo per sito non sinonimo (dN), il tasso di sostituzioni sinonimo per sito sinonimo (dS) ed il rapporto tra i due (ω) (Yang *et al.*, 2000). Sono stati quindi parserizzati i file di *output* per estrarre i valori riferiti ad ogni confronto fra le coppie di sequenze.
Per una migliore visualizzazione della distribuzione dei valori di ω, questi sono stati rappresentati in un istogramma; due esempi, riferiti rispettivamente ai confronti di un organismo piezofilo e uno non per famiglia batterica, sono riportati di seguito nella figura 7.

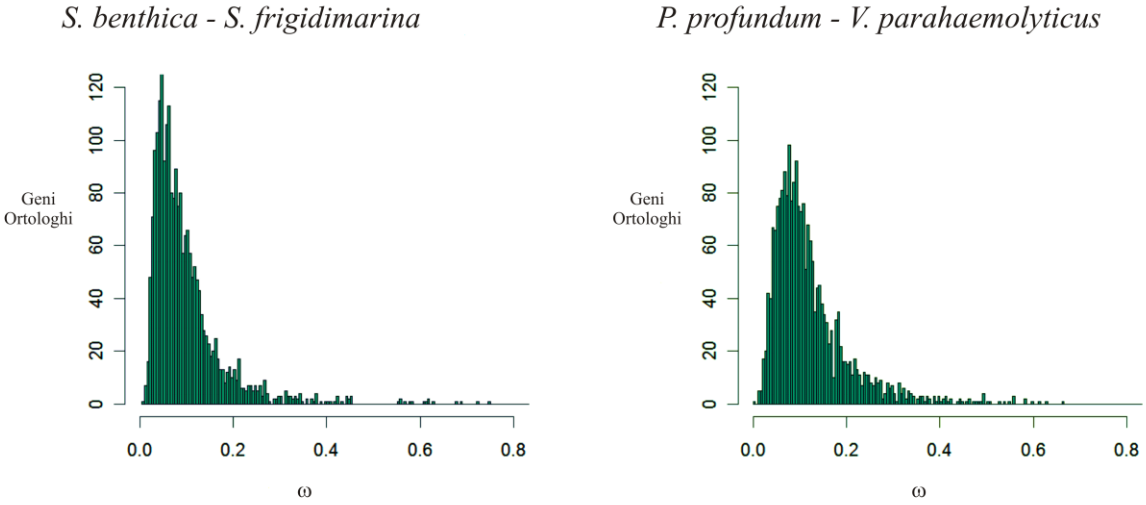

Fig. 7. Distribuzione del numero dei geni ortologhi in Shewanellaceae e Vibrionaceae, in funzione degli intervalli dei valori di ω (dN/dS) calcolati rispettivamente sui confronti *S. benthica - S. frigidimarina* e *P. profundum - V. parahaemolyticus*.

Anziché fissare una soglia per discriminare tra i valori, è stato impiegato il *software* statistico SAM. Questo ci ha permesso di selezionare quei geni che presentano valori di ω significativamente più alti nei confronti tra piezofili e non rispetto ai confronti tra i non piezofili. In bibliografia sono state fissate delle soglie teoriche di ω in cui i geni vengono considerati come sottoposti ad una pressione selettiva nulla (ω=1), purificatrice (ω<= 0,3) e positiva (ω>=3).
Tuttavia altri autori, basandosi su dati sperimentali, definiscono l'intervallo di valori di 0,05<ω<1 come indice di una pressione selettiva positiva che agisce in modo preferenziale solo su alcune porzioni delle proteine. Dato che la maggior parte dei geni ortologhi in analisi, come evidente nel grafico N, ricade in quest'ultimo intervallo, si è reso necessario selezionarli sulla base del *False Discovery Rate* (FDR) calcolato da SAM. Sono stati quindi fissati come limiti il 5% e 10% di FDR che hanno portato all'individuazione rispettivamente di un limitato numero di geni pari a 34 e 61 per le Shewanellaceae e a 65 e 213 per le Vibrionaceae. La lista dei geni significativi e i relativi valori sono nelle tabelle T2 e T3 dei materiali supplementari.
Era già stato ipotizzato, dati precedenti studi sul ceppo SS9 di *P. profundum*, che il numero di geni coinvolti nell'adattamento all'alta pressione non fosse particolarmente elevato. Infatti uno *screening* di mutanti pressione sensibili effettuato allo *Scripps* aveva permesso di individuare un gruppo di 30 geni (comunicazione personale del Dr. F. Lauro), mentre il profilo di espressione genica ottenuto a diverse pressioni per mezzo di *microarray* aveva rilevato circa 200 trascritti. Questi dati assieme alle percentuali di 2,8% e 9,8%, relative ai geni sotto pressione selettiva positiva rispetto al totale degli ortologhi di Shewanellaceae e Vibrionaceae, sembrano indicare che solamente un ristretto numero di proteine sono implicate nei processi chiave dell'evoluzione.

4.4 Arricchimento delle Categorie Funzionali

Si è voluto a questo punto seguire le indicazioni date dai risultati degli esperimenti di *microarray* condotti dal Dr. S. Campanaro che sembravano evidenziare come prevalente il ruolo di alcune classi funzionali rispetto alle altre. Pertanto è stata eseguita una valutazione dell'arricchimento dei geni selezionati con SAM all'interno delle classi funzionali dei *Cluster of Orthologous Groups* (COG) e *Gene Ontology* (GO), le prime più generiche, le seconde più specifiche riguardo all'annotazione dei geni. I calcoli statistici sul database della GO sono stati fatti con il *software GoMiner* mentre quelli sul COG con la distribuzione ipergeometrica, come descritto nel paragrafo 3.2. I geni identificati per mezzo del grado più basso di FDR e quelli appartenenti alla soglia meno stringente hanno portato a risultati similari, riportati nelle tabelle 2 e T1 (quest'ultima nei materiali supplementari).

FUNCTIONAL CATEGORIES	COGs	1°s	tot	*p* value	2°s	tot	*p* value	1°v	tot	*p* value	2°v	tot	*p* value
RNA processing and modification	A	0	1	0,015	0	1	0,027	0	1	0,029	0	1	0,100
Energy production and conversion	C	5	149	**0,024**	5	149	0,214	5	125	0,151	12	125	0,482
Cell cycle control, cell division, chromosome partitioning	D	1	28	0,068	1	28	0,175	0	23	0,493	2	23	0,407
Amino acid transport and metabolism	E	4	173	0,123	7	173	0,093	6	200	0,359	22	200	0,261
Nucleotide transport and metabolism	**F**	3	48	**0,006**	4	48	**0,009**	1	60	0,525	10	60	**0,032**
Carbohydrate transport and metabolism	G	0	50	0,543	1	50	0,396	4	91	0,121	9	91	0,424
Coenzyme transport and metabolism	H	2	114	0,255	2	114	0,606	2	107	0,608	4	107	0,987
Lipid transport and metabolism	I	0	82	0,726	0	82	0,899	4	65	**0,038**	6	65	0,476
Translation, ribosomal structure and biogenesis	J	3	124	0,120	5	124	0,117	3	125	0,495	9	125	0,817
Transcription	K	0	126	0,865	0	126	0,971	3	136	0,563	10	136	0,816
Replication, recombination and repair	L	1	107	0,496	3	107	0,330	1	100	0,796	5	100	0,945
Cell wall/membrane/envelope biogenesis	M	1	106	0,491	4	106	0,158	7	113	**0,015**	16	113	**0,052**
Cell motility	N	0	73	0,683	1	73	0,596	1	75	0,647	9	75	0,210
Posttranslational modification, protein turnover, chaperones	O	1	111	0,516	2	111	0,587	3	102	0,342	10	102	0,441
Inorganic ion transport and metabolism	**P**	2	72	0,097	4	72	**0,044**	8	107	**0,003**	16	107	**0,033**
Secondary metabolites biosynthesis, transport and catabolism	Q	0	34	0,412	0	34	0,610	0	28	0,563	1	28	0,786
General function prediction only	R	4	209	0,214	4	209	0,683	7	217	0,290	29	217	**0,035**
Function unknown	S	10	659	0,436	19	659	0,319	9	577	0,984	52	577	0,789
Signal transduction mechanisms	T	0	105	0,811	0	105	0,948	4	116	0,244	11	116	0,494
Intracellular trafficking, secretion, and vesicular transport	U	1	74	0,316	2	74	0,325	3	81	0,206	13	81	**0,027**
Defense mechanisms	V	0	25	0,323	3	25	**0,004**	1	36	0,280	2	36	0,713
totale classi		38	2470		67	2470		72	2485		213	2485	

Tab. 2. In questa tabella sono riportati i valori ottenuti dal calcolo dell'arricchimento delle categorie funzionali del COG ottenuti dalla distribuzione ipergeometrica. Nelle varie colonne si trovano rispettivamente: (1) la categorie funzionale, (2) il codice della categoria, (3) il numero di geni ortologhi delle Shewanellaceae di quella categoria identificati dal *software* SAM considerando un FDR del 5%, (4) il numero di totale di geni ortologhi di quella categoria, (5) il *p*-value determinato in base all'analisi della distribuzione ipergeometrica. Le colonne (6)-(7)-(8) corrispondono alle (3)-(4)-(5) ma i geni sono quelli identificati dal *software* SAM considerando un FDR del 10%. Le colonne da (9) a (14) corrispondono a quelle da (3) a (8) ma i valori sono relativi alle Vibrionaceae.

Il risultato delle analisi statistiche è una selezione di classi di COG e GO con un grado di significatività molto elevato (soglia *p*-value 0,05), che sono rispettivamente ***Nucleotide transport and metabolism*** e ***Inorganic ion transport and metabolism*** per il COG e ***Establishment of localization***, ***Localization***, ***Transport*** **e** ***Membrane*** per la GO. Tra le classi arricchite delle due annotazioni c'è una corrispondenza, comunque in parte attesa, che ne conferma l'attendibilità. Dall'arricchimento appare evidente che i trasportatori risultano estremamente frequenti tra i geni aventi un valore di ω tendenzialmente alto nel confronto tra piezofili e non.

Questo risultato, in parte atteso, indica che questa classe di proteine richiede un adattamento per poter svolgere efficientemente il proprio ruolo a pressione elevate. Studi presenti in letteratura hanno messo in evidenza come il trasporto sia uno dei processi più sensibili all'alta pressione, ed una delle dimostrazioni più chiare di ciò è stata fornita dagli esperimenti condotti sul gene *TAT2* di *Saccharomyces cerevisiae* (Abe *et al.*, 2000). La marcata sovraespressione di questo gene conferisce al lievito la capacità di crescere fino a 25 MPa; gli autori spiegano questo effetto sulla base dell'alto valore del volume di attivazione associato al trasporto del triptofano, un aminoacido molto voluminoso.

E' noto infatti che la pressione influenza le reazioni in misura proporzionale alla variazione di volume che le accompagnano. Analizzando la lista di proteine appartenenti alla classe transport della GO emergono alcuni ABC transporter per la traslocazione degli aminoacidi, tra cui per esempio il ***Putative ABC-type arginine transport system, permease component*** (SWISSPROT Q6LNL4). Quest'ultimo era già stato identificato in precedenti analisi trascrizionali condotte su *P. profundum* SS9 e la sua espressione risulta regolata in base alla variazione della pressione (Vezzi *et al.*, 2005).

Un punto interessante è l'individuazione nella classe *transport* di componenti di sistemi di traslocazione delle proteine attraverso la membrana come ad esempio la **P*reprotein translocase subunit* SecF** (Q6LU66) che, assieme a SecD, costituisce il sistema integrale di membrana, essenziale per la traslocazione delle proteine che devono essere secrete. E' noto che le basse temperature inibiscono fortemente questo processo in *E coli*, alcuni autori infatti suggeriscono l'esistenza di passaggi temperatura-sensibili in questo *pathway* (Pogliano *et al*., 1993).

E' stato inoltre dimostrato che alte pressioni e basse temperature hanno effetti simili sugli organismi in quanto entrambi determinano una riduzione della fluidità delle membrane biologiche (Royer, 1995). Questo effetto viene infatti contrastato da alcuni batteri abissali come *P. profundum* SS9 e *S. benthica* KT99 utilizzando differenti strategie, tra queste la sintesi di acidi grassi poliinsaturi Omega-3 (PUFA) (DeLong *et al.*, 1985). L'irrigimento delle membrane determina una riduzione nella mobilità dei fosfolipidi e delle proteine di membrana ed in ultima analisi una riduzione della funzionalità di queste ultime (Macdonald *et al.*, 1987). Non stupisce pertanto che almeno uno dei geni codificanti proteine implicate nel *pathway* di traslocazione delle proteine Sec sia sottoposto a pressione selettiva.

L'importanza della struttura delle membrane risulta evidente anche dalla numerosità delle proteine identificate da SAM, 42 nelle Vibrionaceae e 13 nelle Shewanellaceae, ed appartenenti alla componente cellulare ***Membrane***.

Vale la pena notare che esistono quindi dei processi comuni nell'adattamento alle alte pressione in batteri abissali differenti. In ultima analisi infatti *P. profundum* SS9 e *S. benthica* KT99 hanno differenti *optimum* di crescita alle alte pressioni: il primo infatti è piezotollerante, cresce comunque bene anche a pressione atmosferica ed ha una temperatura ottimale di crescita di 9°C (determinata dalle condizioni ambientali particolari del Mar di Sulu), il secondo invece è un piezofilo abissale obbligato, non cresce a meno di 60 MPa. Le differenti caratteristiche potrebbero comunque rendere conto del fatto che alcune risultano caratteristiche di uno solo dei due batteri analizzati. La specificità di alcune classi di COG e GO potrebbe quindi riflettere adattamenti ristretti ad uno solo dei due piezofili considerati.

4.5 Geni Comuni nella Selezione Finale

La presenza di categorie di geni appartenenti alla stesse classi di GO e di COG individuate dall'analisi dell'arricchimento, mi ha indotto a chiedermi se e quanti geni identificati dall'analisi di SAM (soglia del 10% di FDR) fossero ortologhi negli organismi in analisi.
Dalla ricerca fatta in database ne sono emersi 12, riportati nella tabella 3, che in percentuale rappresentano il 18% ed il 5,6% del totale dei geni identificati da SAM per le Shewanellaceae e le Vibrionaceae. La percentuale abbastanza esigua sembrerebbe indicare che l'adattamento nei due gruppi di batteri si verifica a livello di classi funzionali piuttosto che di geni; in altre parole, batteri diversi potrebbero adattarsi alla medesima condizione ambientale tramite modifiche a carico della stessa classe funzionale di geni, mentre non è detto che la pressione selettiva agisca esattamente sulle stesse proteine. Naturalmente questo risultato potrebbe essere parzialmente influenzato dalla scelta degli organismi che formano i due gruppi o dalla difficoltà di definire esattamente la soglia da utilizzare per la scelta dei geni nelle analisi statistiche.

Anche nel gruppo dei geni ortologhi comuni alle due famiglie di batteri sono stati identificati dei trasportatori ma sono emerse anche alcune altre proteine interessanti appartenenti ad altre classi funzionali. Una di questa è **TonB** (Q6LQC5; Q8EFY6) che fa parte del sistema di trasporto dei siderofori nel periplasma ed è essenziale per il trasporto del ferro in numerosi batteri. Pur non essendo chiaro se il ruolo svolto da questa proteina sia lo stesso anche nei piezofili in esame è comunque noto che l'accumulo del ferro nei batteri oceanici è un processo limitante (Church *et al.*, 2000). Ad esempio in *V. cholerae* è così complesso che viene svolto da almeno cinque sistemi differenti a seconda dell'ambiente di crescita (Wyckoff *et al.*, 2007).

Un'altra proteina interessante è la ***Primosomal replication protein N''*** (Q6LTF5, Q8EFZ7), che assieme alla *primosomal proteins N* agisce nell'assemblaggio del primosoma, un complesso di replicazione multiproteica che opera sul filamento *lagging* della forca replicativa (Zavitz *et al.*, 1991). Questa proteina è dotata di attività elicasica come un'altra proteina di *P. profundum* SS9, RecD, nota per la sua importanza nell'adattamento alle alte pressioni e coinvolta nella riparazione al danno del DNA, in grado di conferire resistenza alle alte pressioni ed impedire la formazione del fenotipo filamentoso in mutanti *recD* di *E. coli* (Bidle *et al.*, 1999). Forse l'elevato valore di ω di questo gene va ricollegato al fatto che la pressione agisce sui batteri mesofili inibendo la biosintesi di DNA e RNA (Yayanos *et al.*, 1969) ma ancor più inibendo la divisone cellulare ed inducendo la formazione di cellule altamente filamentose (ZoBell *et al.*, 1962).

ID Geni Ortologhi	TrEMBL	COG	n°AA	*q*-val (%)	n°	Function
SHEWANELLACEAE						
shewa62	Q8E946	**4149P**	226	10,0	**1**	Molybdenum ABC transporter, permease protein
shewa269	Q8EJX6	-	484	0,0	**2**	Hypothetical protein
shewa642	Q8EBG4	**0041F**	163	0,0	**3**	Phosphoribosylaminoimidazole carboxylase, catalytic subunit
shewa881	Q8EDK9	0132H	231	4,9	**4**	Dithiobiotin synthetase
shewa986	Q8ED85	2233F	412	0,0	**5**	Uracil permease
shewa1414	Q8EFY6	0810M	207	10,0	**6**	TonB2 protein
shewa1422	Q8EFZ7	-	222	7,0	**7**	Primosomal replication protein N'', putative
shewa1478	Q8ECL1	0350L	167	7,0	**8**	Methylated-DNA--protein-cysteine methyltransferase
shewa1565	Q8EBJ7	1252C	429	0,0	**9**	NADH dehydrogenase
shewa1668	Q8EHE8	1012C	482	0,0	**10**	Succinate-semialdehyde dehydrogenase
shewa2015	Q8E9Z7	1450NU	560	10,0	**11**	MSHA biogenesis protein MshL
shewa2141	Q8EBL9	0841V	1047	10,0	**12**	RND multidrug efflux transporter MexF
VIBRIONACEAE						
vibrio2101	Q6LSC3	**0581P**	296	4,3	**1**	Putative phosphate ABC transporter, permease protein
vibrio595	Q6LQK3	-	522	4,8	**2**	Hypothetical protein
vibrio1730	Q6LLJ8	**0041F**	201	0,0	**3**	Hypothetical phosphoribosylaminoimidazole carboxylase, catalytic subunit
vibrio716	Q6LPR5	0132H	220	8,3	**4**	Dithiobiotin synthetase
vibrio95	Q6LVQ0	2233F	459	8,3	**5**	Putative xanthine/uracil permease
vibrio1864	Q6LQC5	0810M	259	4,4	**6**	Hypothetical TonB protein
vibrio658	Q6LTF5	-	183	8,3	**7**	Hypothetical primosomal replication protein N''
vibrio1783	Q6LLC3	0350L	154	4,3	**8**	Hypothetical methylated DNA-protein cysteinemethyltransferase
vibrio2015	Q6LS87	1251C	849	8,3	**9**	Putative nitrite reductase (NAD(P)H), large subunit
vibrio1584	Q6LVE5	1012C	485	3,3	**10**	Putative succinylglutamate 5-semialdehyde dehydrogenase
vibrio1931	Q6LP98	4964U	461	3,3	**11**	Hypothetical Flp pilus assembly protein
vibrio617	Q6LNM7	0841V	1043	4,3	**12**	Putative multidrug resistance protein

Tab. 3. Geni ortologhi comuni tra quelli identificati indipendentemente nelle Shewanellaceae e nelle Vibrionaceae dal *software* SAM. Nelle colonne sono riportati rispettivamente: (1) codice ID del database MySQL, (2) codice del databse TrEMBL, (3) codice della categoria di COG, (4) numero di aminoacidi della proteina, (5) valore di *q* ottenuto dall'analisi con il *software* SAM, (7) annotazione del gene.

Infine è stata fatta una ricerca per verificare se alcuni dei geni di questo lavoro erano precedentemente stati identificati in esperimenti di *microarray* condotti sull'adattamento alle alte pressioni e nella mutagenesi su larga scala effettuati in *P. profundum* SS9 (Vezzi *et al.*, 2004; F. Lauro comunicazione personale). Sono stati identificati 14 geni comuni al primo studio e 2 comuni al secondo aggiungendo quindi delle nuove informazioni di carattere adattativo a quelle funzionali già identificate.

ID Geni Ortologhi	Locus_tag	TrEMBL	COG	Function	28 MPa	4 °C	45 MPa
vibrio1363	PBPRA2914	Q6LN69	**0659P**	Hypothetical sulfate permease family protein	-0,80		
vibrio907	PBPRA1773	Q6LR96	0764I	3-hydroxydecanoyl-ACP dehydratase	0,78		
vibrio1953	PBPRA2759	Q6LNJ5	1979C	Putative iron-containing alcohol dehydrogenase	-1,60	-1,40	-0,8
vibrio1907	PBPRB0074	Q6LLD4	0282C	Acetate/propionate kinase	-0,80		-1
vibrio1280	PBPRA0158	Q6LVS8	2271G	Putative glycerol-3-phosphate transporter	0,60	0,90	
vibrio1337	PBPRA0525	Q6LUS7	0747E	Putative peptide ABC transporter, periplasmicpeptide-binding protein	1,20	1,60	
vibrio1884	PBPRA3547	Q6LLL1	3314S	Hypothetical protein	0,60		
vibrio449	PBPRA2941	Q6LN51	**2011P**	Putative ABC-type metal ion transport system,permease component	-	-	-
vibrio655	PBPRA2740	Q6LNL4	4215E	Putative ABC-type arginine transport system, permease component	-1,20		
vibrio530	PBPRA1027	Q6LTD8	0367E	Asparagine synthetase B	-0,90		
vibrio299	PBPRA0552	Q6LUQ0	0460E,0527E	Bifunctional aspartokinase I/homeserine dehydrogenase I	-0,90		
vibrio970	PBPRA2341	Q6LPQ0	1448E	Aspartate aminotransferase	-0,80		
vibrio770	PBPRA2115	Q6LQB3	4135R	Hypohetical ABC transporter, permeaseprotein		-1,20	
vibrio910	PBPRA1769	Q6LRA0	1092R,0116L	Hypothetical N6-adenine-specific DNA methylase		-0,80	

Tab. 4. Geni identificati nelle Vibrionaceae dall'analisi con il *software* SAM e presenti anche tra quelli evidenziati negli esperimenti di *microarray* effettuati in *P. profundum* (Vezzi *et al.*, 2005; Campanaro *et al.*, 2005). Nelle colonne sono riportati rispettivamente: (1) codice ID del database MySQL, (2) Locus tag dell'NCBI, (3) codice del databse TrEMBL, (4) codice della categoria di COG, (5) annotazione del gene, (6) valore del $\log_2$ del rapporto di espressione del gene a 28 MPa rispetto a 0,1 MPa ottenuto tramite i *microarray*, (7) valore del $\log_2$ del rapporto di espressione del gene a 4°C rispetto a 16°C, (8) valore del $\log_2$ del rapporto di espressione del gene a 45MPa rispetto a 28MPa

4.6 Relazione tra variabili

E’ noto dalla letteratura che negli organismi unicellulari esiste una forte correlazione tra il livello di espressione di un gene e il *codon bias*, utilizzo preferenziale di alcuni codoni. E' stato quindi ritenuto necessario considerare nelle analisi il *Codon Adaptation Index* (CAI), che rappresenta una misura dell'adattamento relativo dell'utilizzo dei codoni di un gene rispetto all'utilizzo dei codoni dei geni altamente espressi (Sharp *et al.*, 1987). Per questo sono stati utilizzati i valori di ω ottenuti dai confronti tra tutti gli organismi considerati per valutare il grado di correlazione con i rispettivi CAI, calcolati per ogni batterio, come descritto nel paragrafo 3.2 (Calcolo di sostituzioni aminoacidiche e CAI). La distribuzione di questi valori viene riportata nella figura 8 sia per le Shewanellaceae che per le Vibrionaceae.
Già i grafici evidenziano il rapporto inverso esistente tra i due parametri, confermato ulteriormente dal valore di ρ calcolato usando la correlazione di Spearman. I dati così ottenuti collimano con quelli già presenti in letteratura, calcolati per altri batteri (Rocha *et al.*, 2004). Nel lavoro a cui si fa riferimento sono stati inoltre considerati i valori di CAI, classi funzionali, essenzialità dei geni e costo metabolico per valutare quanto ognuno di questi influenzi l'evoluzione delle sequenze. Il nostro interesse invece si è focalizzato sul contributo dato da CAI e dalle classi di COG come verrà descritto più avanti.
Gli ortologhi aventi una significativa differenza nel valore di ω, ottenuta dal confronto tra batteri piezofili e non piezofili, dimostrano chiaramente di appartenere a delle classi ben definite di COG, come già detto nel paragrafo 4.4. Questo ha fatto sorgere la domanda se alcune classi di geni potessero contribuire in modo significativo a determinare il valore di ω o, in altri termini, se potesse esistere una correlazione tra la classe di COG ed il valore di ω. Calcolando il valore medio di ω per tutti i geni appartenenti ad una certa classe di COG non sono state trovate delle differenze molto evidenti tra il valore medio di ω delle varie classi per il confronto tra i piezofili/non-piezofili e tra non-piezofili, dati non riportati. Tuttavia si è comunque deciso di indagare meglio su questo punto utilizzando un’analisi di regressione multipla ed effettuando un’analisi statisticamente più corretta.
E’ dato dalla letteratura che il valore di ω è positivamente correlato con vari parametri biologici, come per esempio il costo metabolico degli aminoacidi; questo è stato dimostrato ad esempio in *E. coli* ad in *B. subtilis*. Per quanto riguarda la suddivisione in categorie funzionali, la differenza tra il valore di ω medio delle varie classi sembra essere meno evidente e dipendente dalla modalità di raggruppamento dei geni in classi (Rocha *et al.*, 2004). Non ci risulta comunque dalla letteratura che sia stata condotta un’analisi di correlazione suddividendo i geni sulla base delle classi di COG.

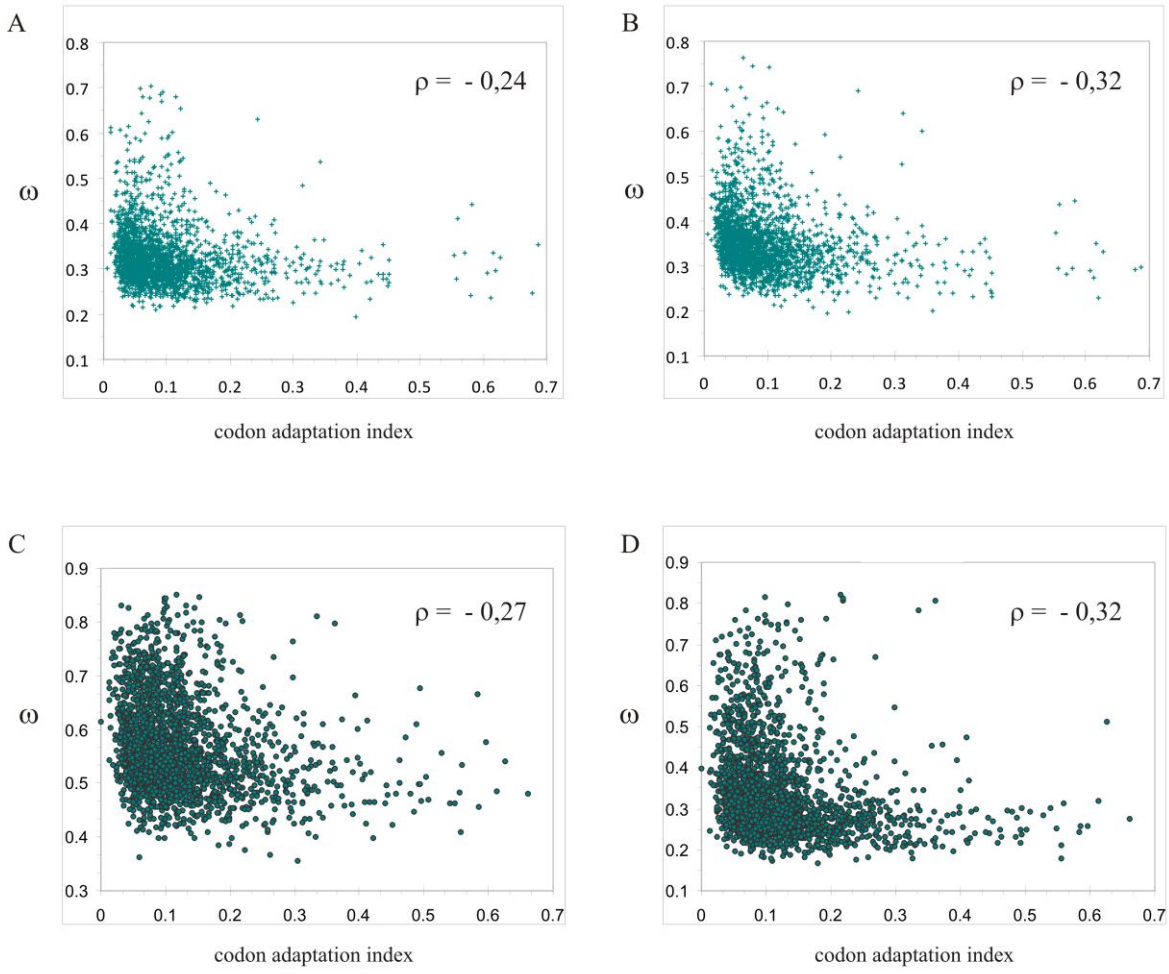

Fig. 8. Distribuzioni dei valori di ω (dN/dS) ottenuti dal confronto dei geni ortologhi di *S. benthica* e *S. frigidimarina* in funzione del codon adaptation index (CAI) calcolato su *S. benthica* (A) e su *S. frigidimarina* (B).
Distribuzioni dei valori di ω (dN/dS) ottenuti dal confronto dei geni ortologhi di *P. profundum* e *V. parahaemolyticus* in funzione del codon adaptation index (CAI) calcolato su *P. profundum* (C) e su *V. parahaemolyticus* (D).
ρ indica il coefficiente di correlazione di Spearman calcolato tra le due serie di dati.

Il grado di correlazione esistente tra ω e tutti questi parametri biologici è comunque reso complesso dal fatto che tutte queste variabili (ω, CAI, costo metabolico, classe funzionale ed appartenenza dei geni al gruppo di quelli essenziali per l'organismo) sono correlate tra loro. Ad esempio è noto dalla letteratura che il costo metabolico degli aminoacidi è correlato con il *codon usage* (Akashi *et al.*, 2002).
Data la complessità di questa correlazione, il metodo migliore è quello di effettuare un'analisi di regressione multipla sui dati dei batteri del gruppo di Vibrionaceae e delle Shewanellaceae per indagare la relazione tra i parametri in nostro possesso: ω, CAI ed appartenenza ad una specifica classe di COG. Non si è considerata la suddivisione in classi di *Gene Ontology* in quanto il numero di queste classi è estremamente elevato ed il raggruppamento dei geni in un numero minore di classi risulta complesso nonostante l'esistenza dei *GO Slims*.
I valori di ω sono stati trasformati in logaritmi al fine di ottenere un miglior risultato di correlazione (Draper *et al.*, 1998). Le basi statistiche dell'analisi della varianza (ANOVA) e delle tecniche di regressione sono le stesse: ANOVA viene utilizzato per analizzare il ruolo di variabili nominali/ordinali e la regressione è utilizzata per analizzare variabili continue.
In questo caso abbiamo utilizzato il pacchetto di analisi statistica R per effettuare una regressione multipla su un insieme di variabili continue (CAI) e discrete (le classi di COG) contemporaneamente. Le variabili discrete vengono codificate sotto forma di *dummy variables*, procedura essenziale per l'analisi dell'importanza relativa delle variabili nel modello di regressione. Il modello che viene generato da questa analisi consente di identificare quali variabili contribuiscono maggiormente alla regressione.
Come atteso, il ruolo maggiore viene rivestito dal valore di CAI in quanto il suo parametro β ottenuto è il più alto, compreso tra -0.54 e -0.72 per le Vibrionaceae, ed è inversamente correlato con il valore di ω, come riportato nella figura S1 dei materiali supplementari. Un ruolo minore viene rivestito dalle classi di COG, ma è comunque evidente che in tutti i casi alcune di queste classi (J, M, P, U per Vibrionaceae e M, S per le Shewanellaceae) hanno una correlazione leggermente più alta delle altre con il valore di ω.

Tuttavia ci si poteva aspettare che il contributo delle singole classi potesse variare notevolmente nel confronto tra piezofili e non piezofili; in altre parole ci si poteva attendere che la pressione selettiva media su specifiche classi di COG fosse differente tra piezofili e non piezofili. Al contrario il risultato ottenuto sembrerebbe indicare che l'analisi condotta globalmente sui geni appartenenti ad una classe di COG non rivela una pressione selettiva maggiore a carico di una classe specifica nell'adattamento all'ambiente abissale.

Questo risultato comunque non fa che supportare la precedente analisi condotta con il *software* SAM che ha consentito di identificare un numero abbastanza ristretto di geni specificamente sottoposti alla pressione selettiva nell'adattamento all'alta pressione e che difficilmente possono essere identificati da un'analisi come quella descritta in questo paragrafo e condotta sulle classi di COG nella loro globalità.
Rimane comunque di notevole interesse notare come questa analisi consenta di identificare delle particolari classi di COG aventi una correlazione maggiore con il valore di ω:

J *Translation, ribosomal structure and biogenesis*
M *Cell wall/membrane/envelope biogenesis*
P *Inorganic ion transport and metabolism*
U *Intracellular trafficking, secretion, and vesicular transport*
S *Function unknown*

Si può vedere inoltre dal grafico nella figura S1 dei materiali aggiuntivi che, come atteso, l'andamento del valore di correlazione è abbastanza simile per l'analisi condotta sulle Shewanellaceae e sulle Vibrionaceae, quindi la correlazione tra la classe di COG ed il valore di ω è simile per entrambe le famiglie batteriche.

CONCLUSIONI

L'interesse del laboratorio in cui ho svolto questo lavoro di tesi per lo studio dell'adattamento all'alta pressione è nato ormai quattro anni fa dalla collaborazione con il Prof. D. Bartlett dello *Scripps Institution of Oceanography*, durante il sequenziamento del genoma di *P. profundum* ceppo SS9. Dopo questa prima fase condotta al CRIBI Biotecnology Centre dell'Università di Padova sotto la supervisione del Dr. A. Vezzi, il progetto è continuato con lo sviluppo di tool bioinformatici da parte del Dr. N. Vitulo e con le analisi dell'espressione genica per mezzo della tecnologia dei *microarray* effettuate dal Dr. S. Campanaro. La sequenza del genoma ha notevolmente facilitato lo *screening* sistematico dei mutanti piezosensibili condotta dal Dr. F. Lauro, presso lo *Scripps*.

Recentemente, grazie all'interesse della Moore Foundation per la microbiologia marina che ha portato al sequenziamento del genoma di *S. benthica*, ne sono state avviate le analisi di *finishing*. Come già detto in più occasioni, arrivare ad avere la sequenza completa e priva di errori del genoma di ogni organismo è basilare per l'accuratezza di qualsiasi altro studio successivo, sia per assegnare una corretta posizione relativa dei geni, sia per la loro esatta identificazione.
Avendo così a disposizione i genomi di due microorganismi piezofili ed un cospicuo numero di batteri non piezofili a questi strettamente correlati filogeneticamente, è diventato possibile sia realizzare analisi sul tasso di sostituzioni dN/dS con un buon supporto statistico, sia verificare l'esistenza di eventuali convergenze adattative evolutesi nelle due famiglie. Sono stati quindi sviluppati un cospicuo numero di *tool* bioinformatici per automatizzare i passaggi e gestire l'elevata quantità di dati, dovuta al numero di organismi considerati mai così alto prima per un progetto di analisi del tasso di sostituzioni nucleotidiche sui batteri.

Il risultato del lavoro è un numero relativamente piccolo ma significativo di geni, utilizzato in seguito per i calcoli di arricchimento delle categorie funzionali. La presenza di un numero di geni superiore all'atteso nelle categorie relative a Trasporto e Membrana della *Gene Ontology* e del COG ha confermato l'importanza di queste classi nell'adattamento all'alta pressione sia delle Shewanellaceae che delle Vibrionaeae. Tuttavia ci sono anche altre indicazioni dell'importanza relativa di alcune classi differenzialmente nell'una o nell'altra famiglia, che potrebbero costituire adattamenti specifici.

Alcuni dei risultati raggiunti in questo progetto, volto alla comprensione delle strategie adattative dei batteri abissali *P. profundum* ceppo SS9 e *S. benthica* ceppo KT99, sono la presenza di proteine sottoposte ad una notevole pressione selettiva tra i trasportatori e nel *Sec pathway*, e la conferma dell'importanza di proteine che svolgono un ruolo essenziale alle alte pressioni, sulla base dei dati di espressione e dell'analisi di mutanti piezosensibili.

Sequenziamento, analisi di espressione genica, mutanti ed identificazione dei geni sottoposti a pressione selettiva stanno, così, componendo un quadro genomico sempre più particolareggiato dei batteri abissali e forniscono informazioni sempre più precise su quali possono essere i geni migliori da utilizzare come candidati per studi mirati di tipo funzionale, di modellizzazione o mutagenesi.

BIBLIOGRAFIA

- Abe, F. and Horikoshi, K., 2000. Tryptophan Permease Gene *TAT2* Confers High-Pressure Growth in *Saccharomyces cerevisiae*. Molecular and Cellular Biology 20(21), 8093–8102

- Akashi, H. and Gojobori, T., 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and Bacillus subtilis. Proc Natl Acad Sci USA 19, 99(6), 3695-700

- Allen, E.E. and Bartlett, D.H., 2002. Structure and regulation of the omega-3 polyunsaturated fatty acid synthase genes from the deep-sea bacterium *Photobacterium profundum* strain SS9. Microbiology 148, 1903-1913.

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool, Journal of Molecular Biology 5, 215(3), 403-410.

- Arrigo, K.R., 2005. Marine microorganisms and global nutrient cycles. Nature 437, 349-355, Review.

- Ashburner, M., 2000. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25-29.

- Azam, F., 1998. Microbial Control of Oceanic Carbon Flux: The Plot Thickens. Science 280, 694-696.

- Azam, F. and Worden, A.Z., 2004. Microbes, molecules, and marine ecosystems. Science 303, 1622-1624.

- Bartlett DH., 2002. Pressure effects on in vivo microbial processes. Biochim Biophys Acta. 25, 1595(1-2), 367-381.

- Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P. and Lander, E.S., 2002. ARACHNE: a whole-genome shotgun assembler. Genome Res. 12(1), 177-189.

- Bidle, K.A. and Bartlett, D.H., 1999. RecD Function Is Required for High-Pressure Growth of a Deep-Sea Bacterium. Journal of Bacteriology 181 (8), 2330–2337.

- Campanaro, S., Vezzi, A., Vitulo, N., Lauro, F.M., D'Angelo, M., Simonato, F., Cestaro, A., Malacrida, G., Bertoloni, G., Valle and G., Bartlett, D.H., 2005. Laterally transferred elements and high pressure adaptation in *Photobacterium profundum* strains. BMC Genomics. 14, 6, 122.

- Cavicchioli, R., Siddiqui, K.S., Andrews, D. and Sowers, K.R., 2002. Low-temperature extremophiles and their applications. Curr Opin Biotechnol. 13(3), 253-261.

- Church, M.J., Hutchins, D.A. and Ducklow, H.W., 2000. Limitation of bacterial growth by dissolved organic matter and iron in the Southern ocean. Appl Environ Microbiol 66, 455-466.

- DeLong, E.F., 2005. Microbial community genomics in the ocean. Nature 3, 459-469, 336-342, Review.

- DeLong, E.F. and Karl, D.M., 2005. Genomic perspectives in microbial oceanography. Nature 437, 5, Review.

- DeLong, E.F., Franks, D.G. and Yayanos, A.A.,1997. Evolutionary Relationships of Cultivated Psychrophilic and Barophilic Deep-Sea Bacteria. Appl. Environ. Microbiol. 63, 2105 – 2108.

- DeLong, E.F. and Yayanos, A.A., 1985. Adaptation of the membrane lipids of a deep-sea bacterium to changes in hydrostatic pressure. Science. 228(4703), 1101-1103.

- Draper, N.R. and Smith, H., 1998. Applied regression analysis. J. Wiley & Sons, New York.

- Dujon, B., 1996. The yeast genome project: what did we learn? Trends Genet. 12(7), 263-270.

- (a) Ewing, B., Hillier, L., Wendl, M.C. and Green, P, .1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8(3), 175-185.

- (b) Ewing, B. and Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8(3), 186-194.

- Gordon, D., Abajian, C. and Green, P., 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8(3), 195-202.

- Iwahashia, H., Odanib, M., Ishidoua, E. and Kitagawaa E., 2005. Adaptation of *Saccharomyces cerevisiae* to high hydrostatic pressure causing growth inhibition FEBS Letters 579, 2847–2852.

- Lauro, F.M., Chastain, R.A., Blankenship, L.E., Yayanos, A.A. and Bartlett, D.H., 2007. The unique 16S rRNA genes of piezophiles reflect both phylogeny and adaptation. Appl Environ Microbiol. 73(3), 838-845.

- Kato, C., Inoue, A., Horikoshi, K., 1996. Isolating and characterizing deep-sea marine microorganisms. Trends Biotehnol. 14(1), 6-12.

- Macdonald, A. G. 1987. The role of membrane fluidity in complex processes under high pressure, p. 207–223. In R. E. Marquis, A. M. Zimmerman, and H. W. Jannasch (ed.), Current perspectives in high pressure biology. Academic Press, London, England.

- Nelson, K.E., Paulsen, I.T., Heidelberg, J.F. and Fraser, C.M., 2000. Status of genome projects for nonpathogenic bacteria and archaea. Nat Biotechnol. 18(10), 1049-1054.

- Ochman, H., Lawrence, J.G. and Groisman, E.A., 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 18, 405(6784), 299-304.

- Peden, J.F., 1999. Analysis of codon usage. PHD Thesis, Department of Genetics University of Nottingham.

- Pogliano, K.J. and Beckwith, J., 1993. The Cs Sec Mutants of *Escherichia coli* Reflect the Cold Sensitivity of Protein Export Itself Genetics. 133(4), 763–773.

- Rocha, E.P.C. and Danchin, A., 2004. An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. Mol. Biol. Evol. 21(1), 108-116.

- Royer, C.A., 1995. Application of pressure to biochemical equilibria: the other thermodynamic variable. Methods Enzymol. 259, 357–377.

- Sharp, P.M. and Li, W.H., 1987. The Codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15(3), 1281-1295.

- Tatusov, R.L., Koonin, E.V. and Lipman, D.J., 1997. A genomic perspective on protein families. Science 24, 278(5338), 631-637.

- Thompson, J.D., Higgins, D.G. and Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22,4673-4680.

- Tusher, V.G., Tibshirani, R. and Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA .24, 98(9), 5116-5121.

- Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H. and Smith, H.O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304(5667), 66-74.

- Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comp. Appl. BioSc. 13, 555-556.

- Yang, Z. and R., Nielsen, 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17,**32-43.

- Yayanos, AA., 1995. Microbiology to 10,500 meters in the deep sea. Annu Rev Microbiol. 49, 777-805.

- Yayanos, A.A., and E. C. Pollard. 1969. A study of the effects of hydrostatic pressure on macromolecular synthesis in *Escherichia coli*. Biophysics 9, 1464-1482.

- Vezzi, A., Campanaro, S., D'Angelo, M., Simonato, F., Vitulo, N., Lauro, F.M., Cestaro, A., Malacrida, G., Simionati, B., Cannata, N., Romualdi, C., Bartlett, D.H. and Valle, G., 2005. Life at depth: *Photobacterium profundum* genome sequence and expression analysis. Science 307(5714), 1459-1461.

- Whitman, W.B., Coleman, D.C. and Wiebe, W.J., 1998. Prokaryotes: The unseen majority. PNAS 95, 6578-6583. Review.

- Wernersson, R. and Pedersen, A.G., 2003. RevTrans - Constucting alignments of coding DNA from aligned amino acid sequences. Nul. Acids. Res. 31(13), 3537-3539.

- Wyckoff, E.E., Mey, A.R., Payne, 2007. Iron acquisition in *Vibrio cholerae*. SM Biometals. 20(3-4), 405-416.

- Zavitz, K.H., DiGate, R.J. and Marians, K.J., 1991. The priB and priC replication proteins of *Escherichia coli*. Genes, DNA sequence, overexpression, and purification. J Biol Chem. 25, 266(21), 13988-13995.

- Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C. and Weinstein, J.N., 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 4(4), R28.

- ZoBell, C. E., and A. B. Cobet. 1962. Growth, reproduction, and death rates of *Escherichia coli* at increased hydrostatic pressures. J. Bacteriol. 84,1228–1236.

Siti internet:

- ftp://ftp.ncbi.nih.gov/genomes/Bacteria/
- www.ncbi.nlm.gov/Genebank/
- www.sanger.ac.uk/
- www.tigr.org/
- www.ebi.ac.uk/

MATERIALI SUPPLEMENTARI

GENE ONTOLOGY CATEGORIES *

Bio Proc	GO:	1°v	tot	*p* value	2°v	tot	*p* value	1°s	tot	*p* value	2°s	tot	*p* value
cell motility; ciliary or flagellar motility	0006928; 0001539	-	-	-	5	22	0,055	-	-	-	-	-	-
diaminopimelate biosynthesis	0019877	-	-	-	-	-	-	1	2	0,029	-	-	-
establishment of localization; localization	**0051234; 0051179**	16	315	0,016	42	315	0,013	-	-	-	12	217	0,004
gas transport; oxygen transport	0015669; 0015671	-	-	-	-	-	-	1	2	0,029	1	2	0,049
localization of cell; locomotion	0051674; 0040011	-	-	-	5	22	0,055	-	-	-	-	-	-
membrane lipid biosynthesis	0046467	2	7	0,017	-	-	-	-	-	-	-	-	-
membrane lipid metabolism; phospholipid metabolism	0006643; 0006644	2	8	0,022	-	-	-	-	-	-	-	-	-
molybdate ion transport	0015689	-	-	-	-	-	-	-	-	-	1	2	0,049
nicotinamide metabolism	0006769	2	9	0,027	-	-	-	-	-	-	-	-	-
nucleoside monophosphate biosynthesis; metabolism	0009124; 0009123	-	-	-	-	-	-	-	-	-	2	10	0,024
nutrient import	0009935	-	-	-	-	-	-	-	-	-	1	2	0,049
oxidoreduction coenzyme metabolism	0006733	3	15	0,009	4	15	0,050	-	-	-	-	-	-
phospholipid biosynthesis	0008654	2	7	0,017	-	-	-	-	-	-	-	-	-
protein folding	0006457	-	-	-	5	22	0,055	-	-	-	-	-	-
protein transport	0015031	-	-	-	6	27	0,041	-	-	-	-	-	-
purine nucleoside monophosphate biosynthesis; metabolism	0009127; 0009126	-	-	-	-	-	-	-	-	-	2	7	0,012
purine ribonucleoside monophosphate biosynthesis; metabolism	0009168; 0009167	-	-	-	-	-	-	-	-	-	2	7	0,012
pyridine nucleotide metabolism	0019362	2	9	0,027	-	-	-	-	-	-	-	-	-
pyruvate biosynthesis	0042866	1	1	0,030	-	-	-	-	-	-	-	-	-
ribonucleoside monophosphate biosynthesis; metabolism	0009124; 0009123	-	-	-	-	-	-	-	-	-	2	9	0,019
serine family amino acid biosynthesis	0009070	-	-	-	-	-	-	-	-	-	2	14	0,045
serine family amino acid metabolism	0009069	-	-	-	-	-	-	-	-	-	3	22	0,016
siderophore transport	0015891	-	-	-	-	-	-	1	1	0,015	1	1	0,025
transport	**0006810**	14	264	0,018	-	-	-	-	-	-	12	171	0,001
UDP-glucose metabolism	0006011	1	1	0,030	-	-	-	-	-	-	-	-	-

Cell Comp	GO:	1°v	tot	*p* value	2°v	tot	*p* value	1°s	tot	*p* value	2°s	tot	*p* value
envelope	0031975	5	62	0,034	-	-	-	-	-	-	-	-	-
external encapsulating structure	0030312	5	63	0,036	-	-	-	-	-	-	-	-	-
external encapsulating structure part; cell envelope	0044462; 0030313	5	60	0,030	-	-	-	-	-	-	-	-	-
flagellum; cell projection	0019861; 0042995	-	-	-	6	29	0,060	-	-	-	-	-	-
glycine cleavage complex	0005960	-	-	-	-	-	-	-	-	-	1	2	0,049
glycine dehydrogenase complex (decarboxylating)	0005961	-	-	-	-	-	-	-	-	-	1	1	0,025
integral to membrane; intrinsic to membrane	0016021; 0031224	11	149	0,003	-	-	-	-	-	-	-	-	-

membrane	**0016020**	17	377	0,039	-	-	-	-	-	-	13	300	0,023
membrane part	0044425	11	184	0,016	-	-	-	-	-	-	-	-	-
phosphoribosylaminoimidazole carboxylase complex	**0009320**	1	1	0,030	-	-	-	1	1	0,015	1	1	0,025
pyruvate dehydrogenase complex	0045254	-	-	-	-	-	-	1	1	0,015	1	1	0,025

Molec Function	**GO:**	**1°v**	**tot**	***p* value**	**2°v**	**tot**	***p* value**	**1°s**	**tot**	***p* value**	**2°s**	**tot**	***p* value**
2-hydroxy reductase activity; 3-hydroxy dehydrogenase activity	0008679; 0008442	1	1	0,030	-	-	-	-	-	-	-	-	-
3-hydroxyacyl dehydratase activity; hydroxydecanoyl	0019171; 0008693	1	1	0,030	-	-	-	-	-	-	-	-	-
adenosine deaminase activity	0004000	-	-	-	-	-	-	-	-	-	1	1	0,025
adenylylsulfate kinase activity	0004020	-	-	-	-	-	-	-	-	-	1	1	0,025
aspartate carbamoyltransferase activity	0004070	-	-	-	-	-	-	1	1	0,015	1	1	0,025
carboxyl- and carbamoyltransferase activity	0016743	-	-	-	-	-	-	1	2	0,029	1	2	0,049
cation:amino acid symporter activity	0005416	-	-	-	-	-	-	-	-	-	1	2	0,049
CDP-alcohol transferase activity; CDP-diacyl phosphate activity	0017169; 0008444	1	1	0,030	-	-	-	-	-	-	-	-	-
chaperone regulator activity	0030188	-	-	-	2	2	0,009	-	-	-	-	-	-
chorismate pyruvate lyase activity	0008813	1	1	0,030	-	-	-	-	-	-	-	-	-
cis-trans isomerase activity	0016859	-	-	-	2	3	0,027	-	-	-	-	-	-
cyclo-ligase activity	**0016882**	-	-	-	2	2	0,009	1	2	0,029	1	2	0,049
cyclopropane-fatty-acyl-phospholipid synthase activity	0008825	1	1	0,030	-	-	-	-	-	-	-	-	-
dethiobiotin synthase activity	0004141	-	-	-	-	-	-	1	1	0,015	1	1	0,025
di-, tri-valent inorganic cation transporter activity	0015082	-	-	-	-	-	-	-	-	-	2	7	0,012
dihydrodipicolinate synthase activity	0008840	-	-	-	-	-	-	1	1	0,015	1	1	0,025
dihydrolip. acyltransferase activity; dihydrolip. Transferase activity	0030523; 0004742	-	-	-	-	-	-	1	1	0,015	1	1	0,025
disulfide oxidoreductase activity	0015036	-	-	-	-	-	-	2	20	0,033	-	-	-
glycine dehydrogenase (decarboxylating) activity	0004375	-	-	-	-	-	-	-	-	-	1	1	0,025
GTPase activity	0003924	-	-	-	2	4	0,050	-	-	-	-	-	-
inositol or phosphatidylinositol phosphatase activity	0004437	-	-	-	-	-	-	-	-	-	1	2	0,049
ion transporter activity	0015075	-	-	-	-	-	-	-	-	-	4	52	0,037
iron ion transporter activity	0005381	-	-	-	-	-	-	-	-	-	2	4	0,004
metal ion transporter activity	0046873	-	-	-	-	-	-	-	-	-	2	7	0,012
methylated-DNA-[protein]-cysteine S-methyltransferase activity	0003908	-	-	-	-	-	-	-	-	-	1	2	0,049
molybdate ion transporter activity	0015098	-	-	-	-	-	-	-	-	-	1	2	0,049
nucleobase transporter activity; nucleoside transporter activity	0015932; 0005337	-	-	-	-	-	-	-	-	-	1	1	0,025
nucleoside kinase activity	0019206	-	-	-	2	2	0,009	-	-	-	-	-	-
oxidoreductase activity, acting on paired donors	0016706	-	-	-	-	-	-	-	-	-	1	1	0,025
oxidoreductase activity, acting on the CH-NH2 group of donors	0016642	-	-	-	-	-	-	-	-	-	1	1	0,025

oxidoreductase activity. acting on superoxide radicals as acceptor	0016721	1	1	0,030	-	-	-	-	-	-	-	-	-
peptidyl-prolyl cis-trans isomerase activity	0003755	-	-	-	2	3	0,027	-	-	-	-	-	-
phosphoribosyl-AMP cyclohydrolase activity; ATP diphosphatase	0004635; 0004636	1	1	0,030	-	-	-	-	-	-	-	-	-
phosphoribosylaminoimidazole carboxylase activity	**0004638**	1	1	0,030	-	-	-	1	1	0,015	1	1	0,025
pseudouridine synthase activity; pseudouridylate synthase activity	0009982; 0004730	2	12	0,047	-	-	-	-	-	-	-	-	-
purine-nucleoside phosphorylase activity	0004731	-	-	-	2	2	0,009	-	-	-	-	-	-
quinolinate synthetase A activity	0008987	1	1	0,030	-	-	-	-	-	-	-	-	-
S-acetyltransferase activity	0016418	-	-	-	-	-	-	1	1	0,015	1	1	0,025
S-acyltransferase activity	0016417	-	-	-	-	-	-	1	3	0,043	-	-	-
siderophore transporter activity; siderophore-iron transporter activity	0042927; 0015343	-	-	-	-	-	-	1	1	0,015	1	1	0,025
sodium:amino acid symporter activity	0005283	-	-	-	-	-	-	-	-	-	1	2	0,049
superoxide dismutase activity	0004784	1	1	0,030	-	-	-	-	-	-	-	-	-
transition metal ion transporter activity	0046915	-	-	-	-	-	-	-	-	-	2	7	0,012
transporter activity	0005215	-	-	-	-	-	-	-	-	-	12	169	0,001
tRNA isopentenyltransferase activity	0004811	-	-	-	-	-	-	-	-	-	1	1	0,025
UDP-sugar pyrophosphorylase activity; UTP:glucose-1-phosphate	0051748; 0003983	1	1	0,030	-	-	-	-	-	-	-	-	-

Tab. T1. In questa tabella sono riportati i valori ottenuti dal calcolo dell'arricchimento delle categorie funzionali della GO ottenuti dalla *software GoMiner*. Nelle varie colonne si trovano rispettivamente: (1) la categorie funzionale, (2) il codice della categoria, (3) il numero di geni ortologhi delle Shewanellaceae di quella categoria identificati dal *software* SAM considerando un FDR del 5%, (4) il numero di totale di geni ortologhi di quella categoria, (5) il *p*-value determinato in dal *software GoMiner*. Le colonne (6)-(7)-(8) corrispondono alle (3)-(4)-(5) ma i geni sono quelli identificati dal *software* SAM considerando un FDR del 10%. Le colonne da (9) a (14) corrispondono a quelle da (3) a (8) ma i valori sono relativi alle Vibrionaceae.

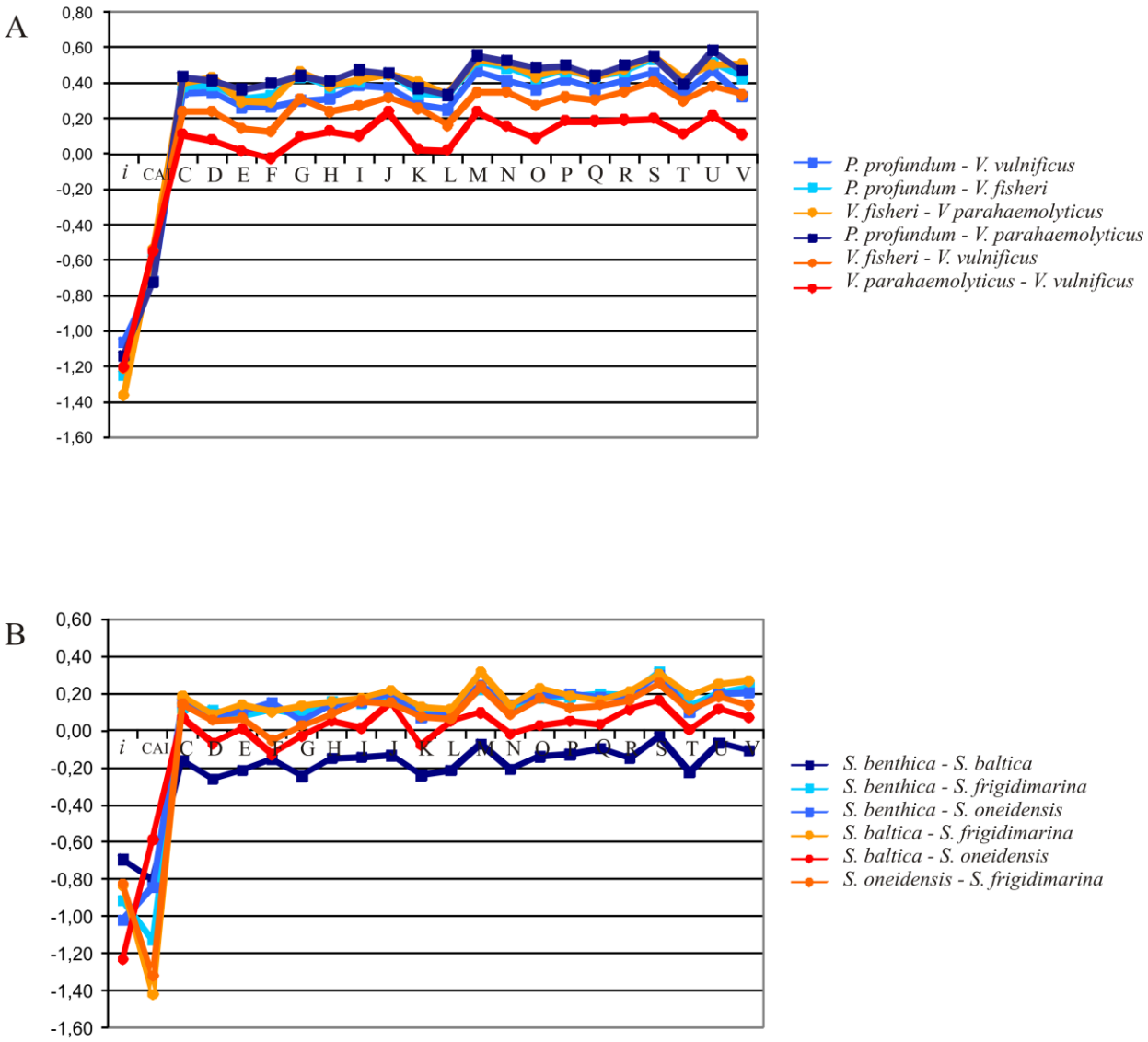

Fig. S1. In questi grafici vengono riportati i valori dell'intercetta (*i*) e dei parametri β ottenuti con la Regressione Multipla utilizzata per calcolare l'influenza sul valore di ω di variabili continue, il CAI, e discrete, le classi funzionali del COG (lettere dalla C alla V).
Il grafico A è relativo all'analisi condotta sulle Vibrionaceae, il B a quella sulle Shewanellaceae.

ID Geni Ortologhi	*q*-val (%)	COG	Locus Tag	Trembl	*S.onei* *S.bent*	*S.frigi* *S.bent*	*S.balt* *S.bent*	*S.frigi* *S.onei*	*S.balt* *S.onei*	*S.balt* *S.frigi*	CAI *S.bent*	CAI *S.onei*	CAI *S.frigi*	CAI *S.balt*
shewa30	0,0	-	SO_0039	Q8EKQ1	0,270	0,288	0,274	0,066	0,037	0,068	0,321	0,302	0,362	0,289
shewa642	0,0	**0041F**	SO_3554	Q8EBG4	0,472	0,333	0,455	0,023	0,039	0,043	0,364	0,327	0,351	0,347
shewa351	0,0	0419L 1196D	SO_4008	Q8EAA0	0,245	0,243	0,251	0,045	0,019	0,025	0,375	0,318	0,385	0,334
shewa352	0,0	-	SO_4007	Q8EAA1	0,657	0,627	0,661	0,070	0,041	0,052	0,325	0,267	0,331	0,290
shewa353	0,0	-	SO_4006	Q8EAA2	0,705	0,686	0,474	0,073	0,018	0,087	0,353	0,255	0,298	0,305
shewa1668	0,0	1012C	SO_1275	Q8EHE8	0,251	0,269	0,255	0,094	0,065	0,111	0,310	0,265	0,365	0,310
shewa1650	0,0	**0540F**	SO_1301	Q8EHC7	0,382	0,377	0,442	0,041	0,027	0,040	0,299	0,275	0,309	0,312
shewa1565	0,0	1252C	SO_3517	Q8EBJ7	0,125	0,125	0,122	0,037	0,033	0,042	0,267	0,253	0,347	0,305
shewa986	0,0	**2233F**	SO_2879	Q8ED85	0,820	0,678	0,536	0,085	0,019	0,068	0,246	0,241	0,292	0,225
shewa285	0,0	0075E	SO_4343	Q8E9E1	0,240	0,235	0,235	0,063	0,037	0,076	0,327	0,262	0,373	0,300
shewa2032	0,0	0508C	SO_0425	Q8EJN8	0,254	0,266	0,210	0,069	0,082	0,110	0,397	0,367	0,418	0,391
shewa269	0,0	-	SO_0330	Q8EJX6	0,149	0,158	0,156	0,039	0,055	0,042	0,297	0,238	0,326	0,278
shewa1942	0,0	2358R	SO_0456	Q8EJK7	0,572	0,380	0,534	0,087	0,031	0,058	0,316	0,281	0,360	0,294
shewa2033	0,0	2609C	SO_0424	Q8EJN9	0,140	0,137	0,140	0,037	0,034	0,016	0,470	0,468	0,479	0,486
shewa952	4,9	1671S	SO_2894	Q8ED72	0,211	0,214	0,261	0,059	0,054	0,120	0,271	0,229	0,273	0,285
shewa2062	4,9	-	SO_4229	Q8E9N8	0,337	0,321	0,380	0,210	0,151	0,208	0,299	0,229	0,230	0,276
shewa1297	4,9	0042J	SO_2006	Q8EFG7	0,099	0,101	0,124	0,034	0,030	0,032	0,270	0,246	0,300	0,280
shewa1417	4,9	0811U	SO_1825	Q8EFY9	0,563	0,312	0,500	0,072	0,064	0,034	0,310	0,457	0,527	0,544
shewa1705	4,9	4232OC	SO_3659	Q8EB73	0,308	0,443	0,426	0,115	0,086	0,183	0,278	0,217	0,260	0,292
shewa1837	4,9	0111HE	SO_0585	Q8EJ83	0,227	0,185	0,189	0,097	0,102	0,097	0,262	0,210	0,328	0,279
shewa1735	4,9	-	SO_3725	Q8EB11	0,542	0,612	0,407	0,271	0,151	0,136	0,236	0,169	0,273	0,228
shewa2162	4,9	**1629P**	SO_4743	Q8E8C4	0,223	0,167	0,174	0,072	0,024	0,066	0,357	0,318	0,379	0,347
shewa1679	4,9	2252R	SO_1236	Q8EHI5	0,146	0,134	0,147	0,069	0,034	0,049	0,250	0,242	0,261	0,275
shewa881	4,9	0132H	SO_2737	Q8EDK9	0,195	0,168	0,170	0,092	0,097	0,080	0,282	0,229	0,309	0,232
shewa100	4,9	0244J	SO_0222	Q8EK76	0,207	0,214	0,234	0,105	0,107	0,134	0,430	0,507	0,543	0,512
shewa344	4,9	-	SO_0973	Q8EI74	0,133	0,158	0,172	0,053	0,061	0,052	0,247	0,186	0,359	0,229
shewa188	4,9	-	SO_4560	Q8E8U6	0,376	0,423	0,422	0,117	0,034	0,206	0,324	0,236	0,332	0,274
shewa407	5,4	-	SO_0755	Q8EIT0	0,182	0,133	0,196	0,061	0,065	0,069	0,358	0,403	0,428	0,385
shewa398	5,4	**3301P**	SO_4568	Q8E8T8	0,163	0,159	0,188	0,071	0,035	0,084	0,234	0,250	0,276	0,236
shewa1377	5,4	2755E	SO_2928	Q8ED42	0,187	0,148	0,163	0,080	0,077	0,090	0,260	0,274	0,255	0,275
shewa536	5,4	-	SO_1372	Q8EH58	0,184	0,304	0,258	0,088	0,024	0,068	0,347	0,329	0,404	0,374
shewa1797	5,4	3529R	SO_0886	Q8EIF3	0,121	0,168	0,176	0,028	0,043	0,060	0,335	0,338	0,355	0,439
shewa1028	5,4	0329EM	SO_1879	Q8EFT7	0,113	0,106	0,097	0,047	0,040	0,051	0,316	0,299	0,376	0,303
shewa120	5,4	0094J	SO_0243	Q8EK57	0,113	0,103	0,099	0,047	0,054	0,040	0,501	0,529	0,482	0,560
shewa1628	5,4	0730R	SO_1333	Q8EH96	0,189	0,212	0,210	0,132	0,097	0,116	0,248	0,237	0,243	0,216
shewa1734	7,0	**0529P**	SO_3723	Q8EB13	0,154	0,214	0,277	0,055	0,064	0,059	0,280	0,231	0,358	0,272
shewa938	7,0	3751O	SO_3563	Q8EBF5	0,325	0,204	0,206	0,076	0,039	0,070	0,350	0,251	0,306	0,301
shewa1478	7,0	0350L	SO_3126	Q8ECL1	0,170	0,208	0,199	0,125	0,110	0,106	0,304	0,241	0,279	0,231
shewa1111	7,0	0534V	SO_2295	Q8EES3	0,139	0,122	0,138	0,038	0,024	0,070	0,284	0,219	0,260	0,263
shewa1422	7,0	-	SO_1817	Q8EFZ7	0,206	0,156	0,172	0,078	0,072	0,103	0,262	0,322	0,367	0,271
shewa36	8,2	**1816F**	SO_4731	Q8E8D4	0,480	0,212	0,519	0,054	0,016	0,073	0,371	0,280	0,364	0,345
shewa499	8,2	-	SO_0808	Q8EIM9	0,276	0,314	0,257	0,177	0,077	0,147	0,485	0,448	0,639	0,490
shewa275	8,2	0251J	SO_0337	Q8EJW9	0,166	0,160	0,172	0,082	0,034	0,094	0,332	0,281	0,268	0,322
shewa445	10,0	-	SO_1159	Q8EHQ8	0,240	0,321	0,256	0,158	0,178	0,160	0,300	0,215	0,375	0,291
shewa1897	10,0	3248M	SO_4131	Q8E9Y1	0,264	0,141	0,251	0,090	0,031	0,075	0,330	0,287	0,385	0,367
shewa585	10,0	0084L	SO_1213	Q8EHK8	0,159	0,191	0,187	0,104	0,061	0,109	0,294	0,267	0,296	0,291
shewa2015	10,0	1450NU	SO_4112	Q8E9Z7	0,079	0,104	0,098	0,039	0,047	0,038	0,269	0,244	0,292	0,265
shewa586	10,0	3248M	SO_1215	Q8EHK6	0,159	0,191	0,187	0,104	0,061	0,109	0,266	0,532	0,593	0,561
shewa62	10,0	**4149P**	SO_4447	Q8E946	0,117	0,127	0,108	0,075	0,055	0,056	0,237	0,201	0,223	0,214
shewa2141	10,0	0841V	SO_3492	Q8EBL9	0,149	0,323	0,320	0,071	0,036	0,076	0,269	0,219	0,304	0,256
shewa1826	10,0	0324J	SO_0602	Q8CX50	0,080	0,084	0,090	0,038	0,050	0,038	0,273	0,241	0,346	0,256
shewa1294	10,0	3219S	SO_2009	Q8EFG4	0,173	0,154	0,158	0,107	0,119	0,101	0,314	0,277	0,315	0,276

shewa1569	10,0	-	SO_3526	Q8EBJ0	0,226	0,169	0,215	0,117	0,046	0,098	0,332	0,211	0,277	0,277
shewa1582	10,0	1115E	SO_3541	Q8EBH5	0,163	0,202	0,129	0,086	0,034	0,062	0,321	0,240	0,310	0,267
shewa1414	10,0	0810M	SO_1828	Q8EFY6	0,116	0,090	0,107	0,036	0,050	0,015	0,418	0,318	0,368	0,373
shewa1390	10,0	4171V	SO_1803	Q8EG11	0,087	0,095	0,089	0,034	0,023	0,049	0,259	0,216	0,253	0,225
shewa282	10,0	3978S	SO_4346	Q8E9D8	0,288	0,198	0,296	0,150	0,024	0,103	0,268	0,265	0,337	0,358
shewa1310	10,0	0483G	SO_2260	Q8EEV3	0,111	0,131	0,100	0,028	0,052	0,052	0,406	0,388	0,458	0,455
shewa205	10,0	-	SO_4467	Q8E927	0,219	0,231	0,226	0,081	0,031	0,156	0,411	0,254	0,336	0,305
shewa1773	10,0	1376S	SO_3748	Q8EAY9	0,169	0,187	0,282	0,050	0,014	0,103	0,281	0,288	0,350	0,368
shewa1717	10,0	0403E 1003E	SO_0781	Q8EIQ6	0,089	0,078	0,091	0,040	0,049	0,039	0,402	0,411	0,420	0,424

Tab. T2. In questa tabella sono riportati i valori relativi ai geni ortologhi delle Shewanellaceae selezionati dal *software* SAM, i primi 34 geni corrispondono all'FDR 5%, i rimanenti al 10%. Nelle varie colonne si trovano rispettivamente: (1) codice ID del database MySQL, (2) valore di *q* ottenuto dall'analisi con SAM, (3) codice della categoria di COG, (4) Locus tag dell'NCBI, (5) codide del databse TrEMBL, (6)-(11) valori di ω relativi ai diversi confronti; (12)-(15) valori di CAI calcolati per ogni organismo. I nomi delle specie sono stati abbreviati per questioni di spazio: *S. baltica (S.balt), S. oneidensis (S.onei), S. benthica (S.bent), S. frigidimarina (S.frigi).*

ID Geni Ortologhi	***q*-val (%)**	**COG**	**Locus Tag**	**TrEMBL**	*V.vul P.pro*	*V.par P.pro*	*V.fis P.pro*	*V.par V.vul*	*V.fis V.vul*	*V.fis V.par*	CAI *P.pro*	CAI *V.vul*	CAI *V.par*	CAI *V.fis*
vibrio121	0,0	2089M	PBPRA2707	Q6LNP1	0,228	0,193	0,193	0,032	0,035	0,030	0,504	0,230	0,304	0,264
vibrio1363	0,0	**0659P**	PBPRA2914	Q6LN69	0,316	0,331	0,366	0,065	0,045	0,038	0,494	0,167	0,215	0,201
vibrio1627	0,0	**0053P**	PBPRB1810	Q6LGB2	0,275	0,359	0,282	0,062	0,067	0,089	0,568	0,180	0,220	0,215
vibrio2028	0,0	0547E	PBPRB0206	Q6LKM9	0,403	0,322	0,291	0,105	0,127	0,122	0,527	0,230	0,296	0,277
vibrio1679	0,0	3161H	PBPRA0162	Q6LVS4	2,039	1,777	1,141	0,076	0,127	0,135	0,543	0,164	0,230	0,184
vibrio1763	0,0	3312C	PBPRA3611	Q6LLG1	0,380	0,387	0,441	0,232	0,247	0,227	0,461	0,202	0,244	0,198
vibrio1730	0,0	**0041F**	PBPRA3574	Q6LLJ8	0,204	0,223	0,179	0,066	0,076	0,068	0,656	0,229	0,393	0,365
vibrio1849	0,0	0697GER	PBPRB0012	Q6LKT6	0,303	0,287	0,287	0,096	0,126	0,131	0,494	0,156	0,255	0,236
vibrio1873	0,0	**4454P**	PBPRA1002	Q6LTG3	0,142	0,142	0,150	0,048	0,059	0,070	0,550	0,269	0,395	0,271
vibrio1925	0,0	-	PBPRB1915	Q6LG17	0,472	0,409	0,625	0,047	0,094	0,118	0,517	0,383	0,474	0,472
vibrio2074	0,0	0625O	PBPRB0681	Q6LJH6	0,172	0,165	0,161	0,071	0,053	0,045	0,604	0,264	0,354	0,239
vibrio2083	0,0	2084I	PBPRA2273	Q6LPW7	0,356	0,398	0,316	0,068	0,056	0,055	0,547	0,263	0,305	0,356
vibrio1298	0,0	1066O	PBPRA0637	Q6LUG7	0,259	0,242	0,174	0,032	0,018	0,027	0,537	0,190	0,268	0,293
vibrio1519	0,0	1210M	PBPRA2642	Q6LNV6	0,175	0,167	0,169	0,029	0,014	0,039	0,636	0,325	0,413	0,456
vibrio1254	0,0	2991S	PBPRA0830	Q6LTY2	0,287	0,253	0,193	0,023	0,050	0,039	0,581	0,258	0,420	0,455
vibrio907	0,0	0764I	PBPRA1773	Q6LR96	0,129	0,127	0,135	0,023	0,046	0,043	0,769	0,355	0,429	0,535
vibrio1265	0,0	-	PBPRA0815	Q6LTZ7	0,459	0,495	0,430	0,072	0,076	0,057	0,468	0,205	0,236	0,259
vibrio1646	1,9	0531E	PBPRA1577	Q6LRT8	0,431	0,355	0,594	0,078	0,062	0,096	0,524	0,279	0,452	0,344
vibrio2104	1,9	-	PBPRB0975	Q6LIN3	0,246	0,242	0,295	0,127	0,142	0,140	0,538	0,201	0,262	0,265
vibrio2115	1,9	0454KR	PBPRB0529	Q6LJX8	0,139	0,142	0,140	0,061	0,076	0,080	0,445	0,217	0,264	0,228
vibrio1769	1,9	-	PBPRB0792	Q6LJ66	0,260	0,209	0,242	0,103	0,130	0,107	0,542	0,236	0,246	0,261
vibrio1245	3,3	2978H	PBPRA0142	Q6LVU4	0,081	0,084	0,085	0,028	0,034	0,021	0,636	0,267	0,367	0,383
vibrio1831	3,3	**0607P**	PBPRA0382	Q6LV61	0,254	0,275	0,342	0,145	0,134	0,152	0,466	0,244	0,263	0,206
vibrio1476	3,3	3113R	PBPRA3244	Q6LMC7	0,201	0,203	0,203	0,151	0,145	0,155	0,398	0,194	0,234	0,245
vibrio1953	3,3	1979C	PBPRA2759	Q6LNJ5	0,274	0,298	0,199	0,063	0,095	0,093	0,696	0,278	0,416	0,358
vibrio808	3,3	2885M	PBPRA0057	Q6LW13	0,216	0,373	0,355	0,076	0,057	0,084	0,619	0,316	0,455	0,528
vibrio1584	3,3	1012C	PBPRA0291	Q6LVE5	0,118	0,112	0,097	0,052	0,043	0,040	0,517	0,218	0,350	0,212
vibrio870	3,3	1940KG	PBPRA1747	Q6LRC2	0,133	0,117	0,098	0,034	0,041	0,033	0,624	0,337	0,427	0,311
vibrio1931	3,3	4964U	PBPRA2496	Q6LP98	0,306	0,316	0,234	0,105	0,120	0,152	0,487	0,238	0,273	0,247
vibrio2006	3,3	3851T	PBPRA2346	Q6LPP5	0,201	0,162	0,177	0,067	0,098	0,080	0,433	0,193	0,266	0,203
vibrio2086	3,3	1762GT 1925G	PBPRA2718	Q6LNN0	0,656	0,439	0,486	0,052	0,212	0,195	0,463	0,186	0,244	0,225
vibrio199	3,3	-	PBPRA0394	Q6LV49	0,184	0,188	0,220	0,033	0,089	0,085	0,714	0,700	0,674	0,636
vibrio1822	4,3	0110R	PBPRB0926	Q6LIT2	0,109	0,116	0,120	0,057	0,059	0,069	0,510	0,197	0,266	0,238
vibrio1783	4,3	0350L	PBPRB0210	Q6LLC3	0,148	0,167	0,142	0,059	0,087	0,077	0,458	0,198	0,226	0,282
vibrio2101	4,3	**0581P**	PBPRA1392	Q6LSC3	0,122	0,108	0,104	0,015	0,027	0,053	0,610	0,183	0,343	0,276
vibrio718	4,3	2230M	PBPRB1573	Q6LGZ5	0,195	0,204	0,208	0,032	0,098	0,102	0,483	0,185	0,248	0,203
vibrio617	4,3	0841V	PBPRA2721	Q6LNM7	0,108	0,097	0,087	0,035	0,039	0,042	0,503	0,182	0,274	0,213

vibrio191	4,3	**1178P**	PBPRA0412	Q6LV31	0,185	0,198	0,190	0,113	0,107	0,065	0,444	0,142	0,188	0,160
vibrio1094	4,4	**0605P**	PBPRA2578	Q6LP19	0,103	0,153	0,117	0,037	0,037	0,045	0,846	0,607	0,714	0,607
vibrio940	4,4	-	PBPRB1501	Q6LH63	0,133	0,118	0,138	0,074	0,080	0,072	0,779	0,375	0,463	0,464
vibrio503	4,4	4786N	PBPRA0906	Q6LTQ8	0,080	0,075	0,078	0,037	0,041	0,036	0,574	0,250	0,335	0,451
vibrio1493	4,4	0564J	PBPRA0408	Q6LV35	0,096	0,092	0,091	0,026	0,049	0,039	0,523	0,197	0,330	0,236
vibrio698	4,4	-	PBPRA2547	Q6LP50	0,088	0,075	0,076	0,020	0,036	0,024	0,619	0,294	0,376	0,348
vibrio470	4,4	0621J	PBPRA2878	Q6LNA5	0,091	0,072	0,082	0,015	0,031	0,023	0,656	0,399	0,484	0,464
vibrio1572	4,4	2319R	PBPRA0307	Q6LVC9	0,135	0,141	0,129	0,089	0,067	0,073	0,470	0,181	0,258	0,248
vibrio1864	4,4	0810M	PBPRA2103	Q6LQC5	0,346	0,331	0,318	0,243	0,192	0,141	0,569	0,201	0,275	0,359
vibrio1633	4,4	2197TK	PBPRA3395	Q6LLZ9	0,101	0,092	0,115	0,049	0,041	0,045	0,496	0,192	0,224	0,200
vibrio400	4,4	1574R	PBPRA1472	Q6LS43	0,540	0,417	0,302	0,081	0,163	0,133	0,521	0,199	0,293	0,256
vibrio858	4,5	3307M	PBPRA0357	Q6LV79	0,329	0,357	0,315	0,107	0,169	0,236	0,458	0,164	0,222	0,215
vibrio1240	4,5	0564J	PBPRA2973	Q6LN19	0,102	0,130	0,133	0,061	0,059	0,053	0,554	0,196	0,272	0,244
vibrio733	4,5	0140E	PBPRA1085	P62349	0,125	0,116	0,136	0,021	0,043	0,072	0,587	0,217	0,276	0,249
vibrio1573	4,5	2900S	PBPRA0306	Q6LVD0	0,122	0,114	0,102	0,036	0,061	0,058	0,657	0,246	0,328	0,293
vibrio1838	4,5	0560E	PBPRB0401	Q6LKA6	0,180	0,236	0,165	0,064	0,095	0,090	0,556	0,234	0,276	0,237
vibrio1386	4,5	**4536P**	PBPRA3044	Q6LMV4	0,112	0,093	0,082	0,043	0,032	0,034	0,477	0,193	0,292	0,229
vibrio768	4,5	0558I	PBPRA2117	Q6LQB1	0,175	0,199	0,124	0,076	0,050	0,056	0,518	0,178	0,201	0,180
vibrio209	4,5	1076O	PBPRA0406	Q6LV37	0,206	0,177	0,176	0,044	0,121	0,067	0,554	0,266	0,286	0,315
vibrio2073	4,5	0673R	PBPRA1901	Q6LQX1	0,111	0,103	0,101	0,063	0,065	0,068	0,525	0,227	0,275	0,207
vibrio991	4,5	0558I	PBPRA2236	Q6LQ04	0,151	0,154	0,103	0,031	0,063	0,046	0,612	0,207	0,297	0,248
vibrio31	4,5	3245C	PBPRA0092	Q6LVY8	0,182	0,209	0,172	0,118	0,101	0,120	0,564	0,305	0,417	0,365
vibrio219	4,5	2202T	PBPRB0486	Q6LK21	0,267	0,249	0,215	0,167	0,154	0,155	0,538	0,201	0,232	0,244
vibrio1533	4,8	1406N	PBPRA3327	Q6LM52	0,077	0,069	0,084	0,015	0,038	0,024	0,586	0,295	0,330	0,359
vibrio118	4,8	0463M	PBPRA0213	Q6LVM3	0,329	0,194	0,359	0,126	0,103	0,074	0,435	0,223	0,277	0,192
vibrio51	4,8	1826U	PBPRA0119	Q6LVW7	0,137	0,116	0,122	0,084	0,063	0,064	0,665	0,361	0,483	0,401
vibrio1099	4,8	0136E	PBPRA2579	Q6LP18	0,060	0,081	0,074	0,024	0,025	0,022	0,664	0,360	0,399	0,283
vibrio693	4,8	0848U	PBPRA2552	Q6LP45	0,143	0,237	0,164	0,074	0,056	0,061	0,541	0,228	0,269	0,280
vibrio595	4,8	-	PBPRA2020	Q6LQK3	0,107	0,099	0,097	0,054	0,068	0,057	0,521	0,220	0,228	0,200
vibrio1907	4,8	0282C	PBPRB0074	Q6LLD4	0,094	0,096	0,103	0,024	0,040	0,055	0,699	0,354	0,478	0,460
vibrio372	4,8	0342U	PBPRA0745	Q6LU67	0,092	0,091	0,082	0,047	0,051	0,054	0,614	0,274	0,386	0,365
vibrio1264	4,8	0652O	PBPRA0817	Q6LTZ5	0,111	0,124	0,114	0,074	0,064	0,079	0,616	0,250	0,353	0,354
vibrio2144	8,3	1999R	PBPRB0745	Q6LJB3	0,218	0,142	0,150	0,077	0,109	0,093	0,564	0,192	0,311	0,223
vibrio2131	8,3	0778C	PBPRA2417	Q6LPH5	0,100	0,104	0,096	0,053	0,062	0,068	0,585	0,195	0,263	0,294
vibrio1280	8,3	2271G	PBPRA0158	Q6LVS8	0,075	0,098	0,083	0,045	0,038	0,040	0,583	0,304	0,311	0,280
vibrio1285	8,3	**3213P**	PBPRB0907	Q6LIV1	0,307	0,325	0,162	0,041	0,104	0,105	0,543	0,158	0,179	0,246
vibrio1295	8,3	**2059P**	PBPRB0963	Q6LIP5	0,299	0,241	0,331	0,143	0,194	0,196	0,507	0,175	0,258	0,265
vibrio2110	8,3	**0813F**	PBPRB0062	Q6LLA7	0,051	0,090	0,079	0,032	0,026	0,033	0,732	0,401	0,692	0,542
vibrio1301	8,3	**0813F**	PBPRA0633	Q6LUH1	0,139	0,066	0,092	0,042	0,029	0,025	0,743	0,361	0,492	0,404
vibrio1482	8,3	3117S	PBPRA3252	Q6LMB9	0,233	0,184	0,283	0,076	0,152	0,161	0,534	0,210	0,253	0,333
vibrio1366	8,3	1309K	PBPRA3181	Q6LMJ0	0,128	0,130	0,126	0,051	0,087	0,094	0,567	0,317	0,396	0,374
vibrio1344	8,3	3525G	PBPRA0518	Q6LUT4	0,095	0,100	0,088	0,046	0,067	0,063	0,537	0,235	0,315	0,287
vibrio1337	8,3	0747E	PBPRA0525	Q6LUS7	0,152	0,156	0,159	0,123	0,071	0,094	0,686	0,531	0,597	0,555
vibrio1432	8,3	-	PBPRA3119	Q6LMP3	0,345	0,262	0,219	0,130	0,116	0,156	0,645	0,234	0,273	0,249
vibrio1316	8,3	1185J	PBPRA0616	Q6LUI8	0,163	0,156	0,139	0,054	0,099	0,104	0,800	0,558	0,633	0,646
vibrio1447	8,3	-	PBPRA3140	Q6LMM2	0,071	0,099	0,059	0,024	0,040	0,014	0,582	0,253	0,280	0,280
vibrio1307	8,3	0456R	PBPRA0624	Q6LUI0	0,132	0,132	0,119	0,071	0,097	0,090	0,483	0,252	0,242	0,265
vibrio1478	8,3	-	PBPRA3247	Q6LMC4	0,109	0,092	0,113	0,045	0,074	0,066	0,459	0,173	0,281	0,153
vibrio1443	8,3	2356L	PBPRA3136	Q6LMM6	0,101	0,091	0,125	0,040	0,069	0,068	0,501	0,272	0,385	0,326
vibrio1545	8,3	0337E	PBPRA0281	Q6LVF5	0,115	0,130	0,088	0,076	0,050	0,051	0,570	0,215	0,240	0,294
vibrio1550	8,3	3166NU	PBPRA0276	Q6LVG0	0,306	0,225	0,311	0,071	0,215	0,139	0,426	0,195	0,264	0,209
vibrio1743	8,3	0348C	PBPRA3588	Q6LLI4	0,111	0,102	0,104	0,055	0,061	0,076	0,485	0,229	0,316	0,244
vibrio2011	8,3	2704R	PBPRA1178	Q6LSY7	0,094	0,100	0,114	0,031	0,078	0,051	0,566	0,206	0,329	0,262
vibrio1779	8,3	0786E	PBPRA2089	Q6LQD5	0,186	0,187	0,122	0,081	0,093	0,101	0,574	0,177	0,236	0,265
vibrio1784	8,3	3274S	PBPRA2021	Q6LQK2	0,267	0,312	0,221	0,187	0,102	0,196	0,464	0,156	0,209	0,207
vibrio1793	8,3	1280E	PBPRA0299	Q6LVD7	0,202	0,184	0,145	0,099	0,103	0,112	0,468	0,154	0,205	0,195
vibrio2010	8,3	-	PBPRA1682	Q6LRI7	0,176	0,138	0,119	0,073	0,088	0,074	0,506	0,214	0,273	0,263
vibrio1881	8,3	-	PBPRA2313	Q6LPS8	0,219	0,283	0,166	0,104	0,061	0,100	0,560	0,176	0,232	0,213
vibrio1882	8,3	**0798P**	PBPRA1540	Q6LRX5	0,100	0,121	0,109	0,074	0,039	0,030	0,538	0,156	0,231	0,307
vibrio1884	8,3	3314S	PBPRA3547	Q6LLL1	0,188	0,180	0,115	0,056	0,090	0,081	0,589	0,274	0,397	0,350
vibrio1892	8,3	0695O	PBPRA1742	Q6LRC7	0,137	0,129	0,151	0,113	0,090	0,090	0,556	0,241	0,309	0,227
vibrio1899	8,3	3313R	PBPRB0524	Q6LJY3	0,197	0,192	0,218	0,157	0,162	0,135	0,515	0,247	0,310	0,296
vibrio2153	8,3	1434S	PBPRB0990	Q6LIL8	0,104	0,147	0,110	0,066	0,050	0,071	0,558	0,191	0,277	0,298
vibrio1983	8,3	-	PBPRB0930	Q6LIS8	0,248	0,250	0,213	0,055	0,155	0,146	0,514	0,282	0,338	0,264
vibrio1734	8,3	1652S	PBPRA3578	Q6LLJ4	0,164	0,176	0,195	0,137	0,135	0,123	0,478	0,190	0,263	0,219

vibrio1713	8,3	**0248FP**	PBPRA3543	Q6LLL4	0,131	0,165	0,122	0,031	0,096	0,054	0,541	0,200	0,281	0,252
vibrio1684	8,3	0552U	PBPRA0153	Q6LVT3	0,141	0,149	0,155	0,026	0,112	0,089	0,601	0,253	0,372	0,327
vibrio1607	8,3	1162R	PBPRA3371	Q6LM22	0,127	0,123	0,087	0,054	0,050	0,057	0,572	0,235	0,368	0,333
vibrio1610	8,3	**0607P**	PBPRA0225	Q6LVL1	0,200	0,160	0,203	0,058	0,118	0,106	0,615	0,338	0,449	0,305
vibrio1626	8,3	0205G	PBPRA0234	Q6LVK2	0,084	0,086	0,093	0,061	0,063	0,059	0,687	0,472	0,563	0,509
vibrio2067	8,3	-	PBPRA1961	Q6LQR1	0,218	0,204	0,290	0,039	0,109	0,159	0,726	0,482	0,663	0,675
vibrio2015	8,3	1251C	PBPRA1428	Q6LS87	0,106	0,105	0,075	0,036	0,036	0,045	0,669	0,333	0,498	0,406
vibrio1993	8,3	-	PBPRB0674	Q6LJI3	0,111	0,114	0,115	0,091	0,082	0,076	0,692	0,430	0,516	0,654
vibrio1636	8,3	0591ER 0642T	PBPRA3399	Q6LLZ5	0,153	0,150	0,112	0,067	0,087	0,077	0,441	0,163	0,198	0,189
vibrio2014	8,3	**2146PR**	PBPRA1427	Q6LS88	0,146	0,243	0,151	0,086	0,088	0,090	0,591	0,313	0,333	0,457
vibrio1649	8,3	**0151F**	PBPRA3420	Q6LLX4	0,094	0,080	0,080	0,033	0,057	0,046	0,641	0,300	0,463	0,438
vibrio1650	8,3	-	PBPRA3421	Q6LLX3	0,407	0,343	0,196	0,149	0,146	0,142	0,567	0,217	0,257	0,269
vibrio1672	8,3	-	PBPRA3466	Q6LLU1	0,225	0,256	0,231	0,102	0,115	0,199	0,600	0,162	0,319	0,227
vibrio1680	8,3	1580N	PBPRA0161	Q6LVS5	0,159	0,158	0,139	0,066	0,105	0,115	0,527	0,204	0,278	0,209
vibrio2151	8,3	0612R	PBPRB0732	Q6LJC6	0,364	0,312	0,166	0,118	0,132	0,129	0,462	0,206	0,280	0,233
vibrio704	8,3	-	PBPRA2471	Q6LPC3	0,228	0,243	0,177	0,079	0,150	0,069	0,615	0,330	0,337	0,373
vibrio489	8,3	0008J	PBPRA0874	Q6LTT8	0,104	0,089	0,068	0,048	0,031	0,049	0,694	0,397	0,559	0,452
vibrio463	8,3	-	PBPRA2888	Q6LN95	0,165	0,183	0,156	0,028	0,121	0,084	0,804	0,350	0,414	0,386
vibrio449	8,3	**2011P**	PBPRA2941	Q6LN51	0,188	0,190	0,114	0,060	0,060	0,071	0,538	0,200	0,282	0,336
vibrio428	8,3	-	PBPRA0800	Q6LU12	0,062	0,063	0,050	0,025	0,035	0,023	0,652	0,348	0,448	0,383
vibrio406	8,3	0576O	PBPRA0696	Q6LUA8	0,129	0,183	0,197	0,064	0,103	0,106	0,720	0,341	0,491	0,481
vibrio373	8,3	0341U	PBPRA0746	Q6LU66	0,093	0,076	0,084	0,044	0,045	0,045	0,616	0,267	0,384	0,320
vibrio331	8,3	1047O	PBPRA0593	Q6LUK9	0,086	0,084	0,089	0,031	0,049	0,049	0,488	0,241	0,292	0,290
vibrio330	8,3	0597MU	PBPRA0592	Q6LUL0	0,141	0,132	0,099	0,052	0,056	0,086	0,582	0,216	0,258	0,277
vibrio325	8,3	**1283P**	PBPRA0582	Q6LUM0	0,102	0,103	0,093	0,050	0,074	0,070	0,591	0,291	0,382	0,353
vibrio493	8,3	3418NUO	PBPRA0896	Q6LTR8	0,155	0,228	0,206	0,103	0,137	0,132	0,520	0,251	0,244	0,275
vibrio498	8,3	1815N	PBPRA0901	Q6LTR3	0,147	0,153	0,142	0,028	0,079	0,109	0,555	0,251	0,304	0,291
vibrio672	8,3	1609K	PBPRA2109	Q6LQB9	0,092	0,091	0,090	0,057	0,053	0,070	0,588	0,301	0,424	0,332
vibrio660	8,3	1187J	PBPRA1148	Q6LT17	0,094	0,105	0,143	0,052	0,061	0,062	0,508	0,211	0,279	0,296
vibrio658	8,3	-	PBPRA1010	Q6LTF5	0,167	0,169	0,157	0,058	0,121	0,122	0,517	0,228	0,253	0,277
vibrio655	8,3	4215E	PBPRA2740	Q6LNL4	0,126	0,097	0,109	0,069	0,064	0,037	0,505	0,177	0,214	0,233
vibrio616	8,3	0845M	PBPRA2722	Q6LNM6	0,258	0,225	0,107	0,107	0,078	0,051	0,571	0,187	0,283	0,289
vibrio530	8,3	0367E	PBPRA1027	Q6LTD8	0,075	0,075	0,038	0,011	0,025	0,017	0,664	0,348	0,458	0,360
vibrio519	8,3	1028IQR	PBPRA2818	Q6LND6	0,128	0,118	0,106	0,052	0,079	0,070	0,436	0,190	0,249	0,234
vibrio516	8,3	1585OU	PBPRB1019	Q6LII9	0,273	0,312	0,170	0,152	0,112	0,121	0,428	0,204	0,207	0,212
vibrio502	8,3	4787N	PBPRA0905	Q6LTQ9	0,136	0,153	0,086	0,044	0,043	0,060	0,566	0,221	0,320	0,337
vibrio316	8,3	2204T 1221KT	PBPRA0572	Q6LUN0	0,109	0,105	0,096	0,046	0,069	0,070	0,467	0,216	0,293	0,295
vibrio307	8,3	0513LKJ	PBPRA0562	Q6LUP0	0,105	0,160	0,110	0,045	0,056	0,079	0,654	0,305	0,451	0,398
vibrio95	8,3	**2233F**	PBPRA0186	Q6LVQ0	0,102	0,111	0,101	0,056	0,081	0,065	0,567	0,198	0,257	0,302
vibrio71	8,3	-	PBPRA3484	Q6LLS3	0,149	0,126	0,115	0,094	0,058	0,070	0,678	0,302	0,462	0,491
vibrio59	8,3	-	PBPRA3498	Q6LLQ9	0,148	0,156	0,142	0,043	0,083	0,117	0,559	0,300	0,256	0,302
vibrio50	8,3	0661R	PBPRA0118	Q6LVW8	0,072	0,093	0,091	0,054	0,039	0,034	0,533	0,216	0,255	0,241
vibrio47	8,3	0697GER	PBPRA0114	Q6LVX2	0,105	0,137	0,107	0,079	0,072	0,056	0,431	0,179	0,215	0,241
vibrio43	8,3	0589T	PBPRA0125	Q6LVW1	0,149	0,186	0,196	0,036	0,114	0,062	0,550	0,223	0,301	0,264
vibrio38	8,3	1249C	PBPRA3546	Q6LLL2	0,070	0,081	0,081	0,048	0,045	0,046	0,690	0,336	0,526	0,468
vibrio16	8,3	-	PBPRA0059	Q6LW11	0,187	0,315	0,242	0,032	0,132	0,127	0,433	0,185	0,211	0,188
vibrio4	8,3	0594J	PBPRA0004	Q6LW54	0,090	0,086	0,067	0,015	0,040	0,047	0,498	0,283	0,298	0,316
vibrio120	8,3	-	PBPRA0210	Q6LVM6	0,131	0,125	0,129	0,093	0,103	0,088	0,506	0,207	0,243	0,212
vibrio135	8,3	1575H	PBPRA0252	Q6LVI4	0,226	0,234	0,157	0,102	0,090	0,163	0,477	0,176	0,275	0,229
vibrio299	8,3	0460E 0527E	PBPRA0552	Q6LUQ0	0,099	0,096	0,069	0,052	0,030	0,024	0,588	0,238	0,341	0,313
vibrio298	8,3	-	PBPRA0549	Q6LUQ3	0,233	0,168	0,145	0,036	0,081	0,096	0,508	0,181	0,244	0,199
vibrio283	8,3	0494LR	PBPRA3207	Q6LMG4	0,132	0,132	0,124	0,045	0,075	0,097	0,514	0,261	0,301	0,245
vibrio272	8,3	0770M	PBPRA3219	Q6LMF2	0,175	0,169	0,181	0,127	0,133	0,077	0,529	0,190	0,274	0,258
vibrio270	8,3	0768M	PBPRA3221	Q6LMF0	0,183	0,189	0,144	0,069	0,106	0,101	0,488	0,204	0,272	0,274
vibrio269	8,3	3116D	PBPRA3222	Q6LME9	0,227	0,181	0,156	0,095	0,128	0,132	0,384	0,219	0,298	0,283
vibrio213	8,3	0473CE	PBPRA0418	Q6LV25	0,072	0,075	0,049	0,018	0,033	0,027	0,684	0,331	0,418	0,415
vibrio204	8,3	0639T	PBPRA0400	Q6LV43	0,226	0,232	0,204	0,144	0,196	0,132	0,490	0,267	0,306	0,254
vibrio141	8,3	0254J	PBPRA0258	Q6LVH8	0,237	0,335	0,295	0,202	0,178	0,228	0,810	0,725	0,783	0,731
vibrio2	8,3	0486R	PBPRA0002	Q6LW56	0,106	0,095	0,070	0,031	0,037	0,050	0,565	0,237	0,341	0,341
vibrio716	8,3	0132H	PBPRA2326	Q6LPR5	0,205	0,170	0,156	0,074	0,127	0,085	0,609	0,254	0,309	0,304
vibrio1248	8,3	-	PBPRA2947	Q6LN45	0,130	0,094	0,081	0,031	0,039	0,031	0,611	0,247	0,380	0,324
vibrio970	8,3	1448E	PBPRA2341	Q6LPQ0	0,117	0,107	0,074	0,039	0,041	0,042	0,701	0,418	0,491	0,527
vibrio1000	8,3	0613R	PBPRA2484	Q6LPB0	0,164	0,160	0,140	0,087	0,117	0,065	0,515	0,198	0,279	0,222

vibrio1235	8,3	-	PBPRA2967	Q6LN25	0,168	0,154	0,143	0,125	0,120	0,101	0,758	0,766	0,759	0,770
vibrio1063	8,3	3318R	PBPRA2611	Q6LNY6	0,218	0,150	0,182	0,077	0,122	0,104	0,664	0,220	0,323	0,347
vibrio1065	8,3	-	PBPRA2797	Q6LNF7	0,178	0,168	0,134	0,100	0,039	0,084	0,533	0,198	0,286	0,227
vibrio1068	8,3	4608E	PBPRA1135	Q6LT30	0,095	0,080	0,064	0,031	0,028	0,033	0,642	0,316	0,355	0,304
vibrio1084	8,3	4658C	PBPRA2563	Q6LP34	0,172	0,184	0,131	0,081	0,125	0,080	0,496	0,195	0,232	0,244
vibrio1137	8,3	-	PBPRA2694	Q6LNQ4	0,125	0,110	0,119	0,045	0,083	0,069	0,525	0,266	0,443	0,300
vibrio1145	8,3	0183I	PBPRA0961	Q6LTK4	0,086	0,081	0,066	0,034	0,040	0,053	0,567	0,227	0,344	0,215
vibrio1177	8,3	1317NU	PBPRA0925	Q6LTN9	0,132	0,154	0,135	0,084	0,104	0,102	0,550	0,297	0,327	0,326
vibrio1185	8,3	-	PBPRA0916	Q6LTP8	0,281	0,300	0,259	0,148	0,239	0,211	0,464	0,241	0,229	0,219
vibrio1205	8,3	**0150F**	PBPRA2910	Q6LN73	0,107	0,108	0,076	0,040	0,038	0,055	0,629	0,356	0,534	0,464
vibrio1216	8,3	-	PBPRA2922	Q6LN61	0,273	0,271	0,178	0,112	0,135	0,104	0,549	0,163	0,242	0,252
vibrio1233	8,3	-	PBPRA2965	Q6LN27	0,107	0,090	0,097	0,064	0,068	0,067	0,713	0,512	0,706	0,491
vibrio1236	8,3	-	PBPRA2968	Q6LN24	0,149	0,132	0,136	0,048	0,097	0,096	0,830	0,687	0,754	0,708
vibrio1239	8,3	-	PBPRA2971	Q6LN21	0,078	0,073	0,075	0,043	0,050	0,050	0,565	0,248	0,367	0,309
vibrio1241	8,3	-	PBPRA2974	Q6LN18	0,420	0,359	0,318	0,317	0,243	0,246	0,539	0,186	0,292	0,162
vibrio938	8,3	1538MU	PBPRB0583	Q6LJS4	0,101	0,101	0,101	0,042	0,062	0,068	0,579	0,249	0,284	0,276
vibrio1060	8,3	2186K	PBPRA2608	Q6LNY9	0,099	0,095	0,095	0,014	0,047	0,049	0,643	0,250	0,346	0,287
vibrio924	8,3	2996S	PBPRA2015	Q6LQK8	0,222	0,255	0,205	0,047	0,149	0,132	0,630	0,277	0,364	0,298
vibrio729	8,3	0241E	PBPRA1089	P62455	0,068	0,068	0,069	0,018	0,039	0,041	0,587	0,340	0,322	0,314
vibrio760	8,3	0841V	PBPRA2406	Q6LPI6	0,099	0,106	0,096	0,046	0,065	0,068	0,527	0,195	0,250	0,265
vibrio769	8,3	4136R	PBPRA2116	Q6LQB2	0,103	0,098	0,099	0,060	0,068	0,043	0,429	0,187	0,260	0,248
vibrio770	8,3	4135R	PBPRA2115	Q6LQB3	0,341	0,184	0,234	0,080	0,183	0,094	0,475	0,173	0,230	0,203
vibrio775	8,3	0745TK	PBPRB0844	Q6LJ14	0,039	0,052	0,048	0,014	0,021	0,021	0,534	0,201	0,297	0,257
vibrio823	8,3	-	PBPRA1684	Q6LRI5	0,105	0,162	0,160	0,049	0,074	0,079	0,580	0,217	0,259	0,252
vibrio847	8,3	-	PBPRA3002	Q6LMZ0	0,092	0,077	0,088	0,043	0,057	0,061	0,498	0,206	0,266	0,261
vibrio851	8,3	5018R	PBPRB0011	Q6LKT7	0,079	0,083	0,084	0,029	0,052	0,052	0,525	0,255	0,302	0,293
vibrio856	8,3	3206M	PBPRA0354	Q6LV82	0,185	0,274	0,166	0,084	0,095	0,097	0,493	0,179	0,250	0,225
vibrio915	8,3	0308E	PBPRA1764	Q6LRA5	0,082	0,078	0,081	0,058	0,050	0,051	0,650	0,284	0,368	0,302
vibrio857	8,3	0438M	PBPRA0356	Q6LV80	0,189	0,163	0,166	0,109	0,129	0,105	0,472	0,191	0,260	0,251
vibrio910	8,3	1092R 0116L	PBPRA1769	Q6LRA0	0,116	0,109	0,116	0,052	0,049	0,093	0,718	0,292	0,356	0,394
vibrio1742	9,9	-	PBPRA3587	Q6LLI5	0,196	0,222	0,122	0,041	0,128	0,086	0,644	0,404	0,444	0,454
vibrio934	9,9	-	PBPRA2191	Q6LQ39	0,085	0,085	0,075	0,015	0,056	0,044	0,509	0,306	0,432	0,419
vibrio1559	9,9	2352C	PBPRA0265	Q6LVH1	0,122	0,090	0,099	0,053	0,067	0,070	0,571	0,273	0,382	0,376
vibrio2124	9,9	1387ER	PBPRB2022	Q6LFR3	0,098	0,099	0,097	0,040	0,071	0,069	0,492	0,190	0,252	0,274
vibrio739	9,9	**1435F**	PBPRA1084	Q6LT81	0,049	0,052	0,049	0,009	0,028	0,025	0,537	0,257	0,337	0,267
vibrio737	9,9	0652O	PBPRA1097	Q6LT68	0,105	0,104	0,151	0,027	0,078	0,069	0,800	0,525	0,618	0,629
vibrio1551	9,9	4972NU	PBPRA0275	Q6LVG1	0,282	0,252	0,199	0,176	0,178	0,175	0,423	0,189	0,277	0,226
vibrio1021	9,9	0845M	PBPRB0372	Q6LKD5	0,180	0,178	0,134	0,044	0,098	0,125	0,578	0,194	0,284	0,215
vibrio1057	9,9	**0572F**	PBPRA1170	Q6LSZ5	0,079	0,081	0,055	0,030	0,037	0,041	0,577	0,324	0,378	0,342
vibrio2005	9,9	2197TK	PBPRA2347	Q6LPP4	0,069	0,065	0,063	0,044	0,037	0,045	0,450	0,206	0,280	0,245
vibrio1334	9,9	0628R	PBPRA0528	Q6LUS4	0,219	0,215	0,167	0,065	0,129	0,157	0,552	0,201	0,257	0,269
vibrio523	9,9	-	PBPRA1021	Q6LTE4	0,219	0,158	0,196	0,059	0,136	0,136	0,545	0,206	0,308	0,253
vibrio128	9,9	1086MG	PBPRA2698	Q6LNQ0	0,096	0,106	0,110	0,083	0,073	0,070	0,473	0,216	0,250	0,219
vibrio1165	9,9	0455D	PBPRA0938	Q6LTM7	0,092	0,075	0,079	0,030	0,061	0,043	0,538	0,204	0,318	0,236
vibrio1457	9,9	-	PBPRA3155	Q6LMK7	0,064	0,048	0,059	0,027	0,034	0,026	0,697	0,362	0,470	0,473
vibrio174	9,9	**0737F**	PBPRA3313	Q6LM66	0,104	0,122	0,114	0,077	0,071	0,091	0,669	0,309	0,334	0,419
vibrio1417	9,9	1159R	PBPRA3088	Q6LMS3	0,138	0,107	0,101	0,024	0,074	0,073	0,650	0,285	0,384	0,389
vibrio1923	9,9	-	PBPRB1801	Q6LGC1	0,438	0,314	0,207	0,122	0,202	0,149	0,611	0,282	0,385	0,341

Tab. T3. In questa tabella sono riportati i valori relativi ai geni ortologhi delle Vibrionaceae selezionati dal *software* SAM, i primi 34 geni corrispondono all'FDR 5%, i rimanenti al 10%. Nelle varie colonne si trovano rispettivamente: (1) codice ID del database MySQL, (2) valore di *q* ottenuto dall'analisi con SAM, (3) codice della categoria di COG, (4) Locus tag dell'NCBI, (5) codide del databse TrEMBL, (6)-(11) valori di ω relativi ai diversi confronti; (12)-(15) valori di CAI calcolati per ogni organismo. I nomi delle specie sono stati abbreviati per questioni di spazio: V. *fischeri (V.fis), V. vulnificus (V.vul), P. profundum (P.pro), V. parahaemolyticus (V.par).*