SPAZIO DEI GENI, SPAZIO DEI GENOMI E DETERMINAZIONE DI UN GENOMA MINIMALE

Alessandra Carbone

Génomique Analytique, Université Pierre et Marie Curie, INSERM U511, 91, Bd de l'Hôpital, 75013 Paris, France; e-mail: carbone@ihes.fr

Abstract

La nuova corsa verso la sintesi di un genoma batterico che sopravviva in laboratorio e che realizzi cicli metabolici desiderati, lanciata 3 anni fa da Venter, Smith e Hutchison, richiede di chiarire i meccanismi biologici di base delle cellule viventi. Il problema si riconduce alla ricerca di quale sia l'insieme minimale di geni che siano essenziali alla vita dell'organismo. Esperienze di mutagenesi aleatoria su batteri specifici hanno permesso di proporre un insieme di tali geni. Presenteremo qualche semplice idea matematica che combina statistica e algoritmica, e che permette di ritrovare una gran parte dei geni individuati sperimentalmente, e di proporne altri di cui in biologia non si conoscono ancora le funzioni.

L'approccio presentato si propone di derivare informazione di interesse biologico da un'analisi puramente statistica dei genomi e da una concezione appropriata di algoritmi.

MOTIVAZIONI TEORICHE E SPERIMENTALI

Vi parlerò di come una visione matematica del genoma, e la definizione di misure appropriate per analizzarlo può permetterci di dedurre dell'informazione biologica interessante. Vorrei iniziare questa lezione col presentarvi un problema che interessa la ricerca in genomica in questi ultimi anni. Abbiamo vissuto dal 1996 gli anni del sequenzaggio, la disponibilitá di numerosi genomi completamente sequenziati, un'esplosione delle banche di dati di sequenze di DNA e di sequenze proteiche. Siamo ora nel periodo della post-genomica dove questi dati devono venire compresi e messi insieme per dedurre informazioni biologiche, per rispondere a domande ancora senza risposta in biologia e per capire cos'altro possa esserne derivato. La storia che vi racconteró all'inizio della lezione riguarda un nuovo progetto, proposto da uno degli attori maggiori del sequenzaggio americano, Craig Venter.

Nel novembre del 2002, Venter annuncia un progetto di costruzione di un cromosoma batterico: il cromosoma, una sequenza di DNA sufficientemente lunga per codificare i geni necessari alla vita del battere, sará sintetizzata e inserita in una cellula vivente il cui materiale genetico é stato rimosso, per verificare se puó dirigere le normali funzioni di un organismo. Il progetto di Venter si cala in un momento in cui altri esperimenti contribuiscono indirettamente all'idea. Clyde Hutchison [Hutchison et al. 1999] ha preso il battere con il piú piccolo genoma conosciuto, quello di *Mycoplasma genitalium*, e ha bloccato l'attività dei geni (seguendo un processo di mutagenesi transposomale) per verificare quali fossero i geni necessari alla sopravvivenza del battere. Sui 517 geni di questo genoma particolarmente compatto (costituito da

Esperienze di Mutagenesis

Organismo	geni essenziali	geni testati	referenza
Bacillus subtilis	300	≈ 4000	[Itaya 1995]
Mycoplasma genitalium	265	517	[Hutchison et al. 1999]
Haemophilus influenzae	670	1272	[Akerley et al. 2002]
Escherichia coli	620	3746	[Gerdes et al. 2003]
Saccharomices cerevisiae	1105	5916	[Giaever et al 2002]
Caenorhabditis elengans	1722	19427	[Kamath et al. 2003]

Genomica Comparativa

Numero di genomi	geni omnipresenti	referenza
2	256	[Mushegian e Koonin 1996]
34	80	[Harris et al. 2003]
100	60	[Koonin 2003]
147	35	[Charlebois e Doolittle 2004]

Tavola 1. Lista di esperimenti realizzati *in silico* e *in vitro*, e numero di geni essenziali determinati.

580.000 basi nucleotidiche), é stato stimato che circa 300 geni sono necessari per la sopravvivenza dell'organismo. Qualche anno dopo, Eckard Wimmer ha sintetizzato il primo genoma virale che iniettato in una cellula la infetta [Cello et al.2002]. Si tratta del genoma del poliovirus, costituito da 7500 basi nucleotidiche di RNA. Poiché la sintesi chimica del RNA é molto piú difficile di quella del DNA, il genome completo del poliovirus é stato assemblato come DNA e poi trasformato in RNA. Test di viabilitá del poliovirus sono stati fatti *in vitro* e *in vivo*, e in ambedue i casi, le proteine di replicazione della cellula riproducono copie del poliovirus. Inoltre, la neutralizzazione del virus tramite immuno-specifici sera é stata favorevolmente testata, e se ne é concluso che il materiale genetico di un poliovirus sintetico é capace di riprodurre virus attivi nella stessa maniera di un poliovirus naturale.

Il progetto di Venter riguarda, come il progetto di Hutchison, il genoma minimale, ma l'approccio è differente. Venter vuole costruire il genoma dal niente, come Wimmer. Creare un genoma batterico è però molto piu complicato che per un virus visto che si tratta di un assemblaggio, realizzato senza errori, di qualche centinaia di migliaia di basi nucleotidiche. Una volta sintetizzata, la sequenza di DNA verrá inserita in una cellula per vederla funzionare.

Ricercare quali siano i geni necessari e dunque definire un genoma minimale non é semplice. Infatti, dietro a questa idea di trovare un genoma minimale e sintetizzarlo, c'é l'ambizione di aggiungergli poi altri geni e trasformare il "mycoplasma minimale" in un battere utile. I geni aggiunti dovrebbero servire, ad esempio, come rimedio contro l'inquinamento ambientale e la distruzione di tossine, o per la produzione di nuove sostanze chimiche industriali (come il carburante), l'insulina etc.

Metodi per determinare l'insieme minimale. Fino ad oggi, per determinare l'insieme minimale di geni essenziali ad un organismo sono state seguite due strade. L'una sperimentale che ha messo in evidenza che i geni essenziali sono sempre di piú, man mano che consideriamo organismi piú complessi e in particolare se li vogliamo vedere vivere in condizioni di vita particolari, in un ambiente particolare. L'altra strada d'analisi é la genomica comparativa, che cerca i

geni condivisi da genomi di speci diverse e trova sempre meno geni con l'aumentare del numero delle speci. La Tavola 1 da un'idea dei dati ottenuti con i due approcci.

Il problema con questi due approcci risiede nella dipendenza dalle condizioni ambientali per le esperienze, e dagli strumenti di deduzione delle omologie per la genomica comparativa. In particolare, tra i geni dedotti come omnipresenti con la genomica comparativa, non troviamo geni di funzione non conosciuta (che invece appartengono agli insiemi determinati sperimentalmente; ad esempio, 111 dei geni ottenuti in [Hutchison et al. 1999] hanno funzione non conosciuta) e neppure geni dipendenti dalle condizioni ambientali che sono specifici di un organismo dato. Nel seguito di questa lezione vi proporró un nuovo metodo [Carbone 2005] che ci permette di riottenere le liste di geni omnipresenti determinati dalla genomica comparativa ma che contiene anche i geni desiderati sopraindicati.

Qualche richiamo di biologia molecolare. Prima di continuare mi é necessario ricordare succintamente qualche nozione di base in biologia. Un genoma é costituito da due filamenti orientati in senso opposto, dove i geni sono letti dalla polimerase rispettando l'orientamento, per poi venir "copiati" in un filamento, della lunghezza del gene, chiamato RNA del gene. Sará l'RNA che verrá poi letto dal ribosoma, il complesso che traduce un gene in proteina. Il linguaggio del DNA é composto da 4 lettere AGCT, corrispondenti alle 4 basi nucleotidiche della sequenza di DNA, mentre il linguaggio delle proteine é composto da ben 20 lettere, corrispondenti ai 20 amino acidi che formano i blocchi di base delle proteine. (Per una veloce introduzione, dedicata ai matematici, delle basi della biologia molecolare e di modelli potenzialmente interessanti da studiare, vedi [Carbone e Gromov 2001].)

Un *codone* e' una sequenza costituita da tre nucleotidi. Gli amino-acidi sono codificati sul DNA da codoni. Codoni diversi possono codificare lo stesso amino-acido, e in questo caso si parla di *codoni sinonimi*. Abbiamo esempi di amino-acidi, come la leucina, che sono codificati da ben 6 codoni. Le ragioni per una tale ridondanza non si conoscono, anche se varie teorie hanno cercato di spiegare il fenomeno senza completamente convincere. Puó essere osservato peró che codoni che codano per uno stesso amino acido non sono usati in maniera uniforme nel genoma. Un tale *bias* sui codoni sinonimi puó essere *globale* e coinvolgere il genoma completo, o *locale* e riguardare parti diverse dei cromosomi. La distribuzione dell'uso dei codoni varia da organismo a organismo. In particolare, organismi che si dividono rapidamente sfruttano un uso privilegiato di certi codoni per codificare proteine essenziali per la riproduzione.

L'uso di codoni privilegiati puó risultare da una diversitá di fattori: la preferenza nell'uso di G e C (bias GC), la preferenza nell'uso di G e C alla terza posizione nucleotidica dei codoni (bias GC3) [Lafay et al. 1999], un filamento di DNA piú ricco in G+T che il filamento opposto [Lafay et al. 1999], un trasferimento orizzontale di geni che induce segmenti di cromosomi ad avere inusuali composizioni [Moszer et al. 1999], e in particolare, il bias dovuto alla traduzione che é stato spesso notato in organismi procarioti e eucarioti che si dividono rapidamente [Sharp e Li 1987, Sharp et al. 1986, Médigue et al. 1991, Shields e Sharp 1987, Sharp et al. 1988, Stenico et al. 1994]. Tre fatti principali appoggiano l'idea di "translational impact": i geni fortemente espressi tendono ad usare solo un numero limitato di codoni e presentano un forte bias verso i codoni usati [Grantham et al. 1980, Sharp e Li 1987], codoni preferiti e numero di RNA di trasferimento nella cellula sono in forte correlazione positiva [Ikemura 1985, Bennetzen e Hall 1982, Bulmer 1987, Gouy e Gautier 1982], e i pool di tRNA di trasferimento disponibili nella cellula influenzano la velocitá di elongazione polipeptidica [Varenne et al. 1984, Buckingham e Grosjean 1986].

Nell'analisi dei genomi presentata in questa lezione, eviteró di fare ipotesi biologiche e rimarró in seno ad una pura analisi statistica delle sequenze. Per essere precisi, i soli segnali biologici indirettamente utilizzati, sono quelli relativi all'inizio e alla fine delle regioni codanti. Un genoma, per noi, sará inteso come l'insieme delle sue regioni codanti. Il background biologico resterá sempre presente nella nostra analisi e ci aiuterá a verificare i risultati intermedi ottenuti.

I geni sono visti come insiemi di codoni, dove l'ordine di apparizione del codone nella sequenza é perso. Li rappresentiamo come vettori normalizzati, dove la normalizzazione é usata per poter considerare codoni particolarmente usati e codoni rari nella stessa maniera. Piú formalmente, una sequenza codante è rappresentata da un vettore a 64 dimensioni, le cui coordinate corrispondono alle 64 frequenze relative dei codoni nella sequenza. Ricordiamo che la *frequenza* di un codone i nella sequenza g è il numero di occorrenze di i in g (dove g é un gene inteso come diviso in triplette nucleotidiche consecutive e non sovrapposte che corrispondono alla decomposizione in amino acidi), e che la *frequenza relativa* di i in g é la frequenza di i in g divisa per il numero di codoni in g. Per ogni vettore rappresentante una sequenza codante, la somma delle coordinate deve essere uguale a 1. Si ha quindi che una sequenza codante é un punto in uno spazio a 64 dimensioni $[0 \ldots 1]^{64}$, dove nessuna ipotesi é stata fatta sulla geometria dello spazio né sul sistema di coordinate scelto.

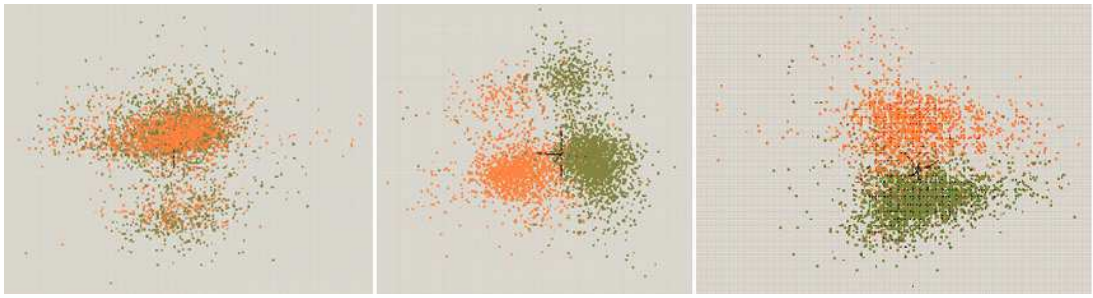

Figura 1. Tre viste diverse dello spazio di proiezione (ottenuto con l'ACP) dei genomi di *Haemophilus influenzae* (arancione) e di *Staphylococcus aureus* (verde).

Visualizzazione dei genomi e loro comparazione. Un genoma completo puó essere guardato come un insieme di punti in uno spazio a 64 dimensioni. Nella Figura 1 vediamo due genomi batterici in una proiezione fatta con l'Analisi delle Componenti Principali (ACP; le coordinate principali sono scelte in modo tale che la dispersione dei punti proiettati sia massima; la direzione del vettore di varianza massimale é la direzione dell'eigenvector associato al piú grande eigenvalue della matrice di varianza; e cosí di seguito, per le altre direzioni di varianza massimale). I punti sono proiettati sulle tre componenti principali, in modo tale che la perdita di informazione dovuta alla proiezione sia minima. Osserviamo che i due batteri sembrano avere una geometria di punti molto simile e sembrano essenzialmente traslati nello spazio. Per i due organismi, la nuvola di punti é organizzata in due cluster principali, e il piu grande dei cluster presenta due estensioni laterali meno dense, due specie di "orecchie di coniglio" [Médigue et al. 1991].

La forma geometrica delle nuvole di punti associate ai due genomi che percepiamo visivamente come simili, suggerisce la possibilitá che le relazioni tra geni siano preservate da genoma

a genoma, che esista una sorta di ”traduzione” di un genoma in un altro, e che la composizione nucleotidica dei geni preservi la mappa di traduzione da un genoma all'altro.

La forma geometrica dei due genomi illustrati nella Figura 1, si ritrova in molti altri batteri. La loro localizzazione nello spazio peró puó variare sensibilmente. La traslazione spaziale delle due nuvole di punti, osservata nella Figura 1, é indotta da frequenze diverse per codoni codanti uno stesso amino-acido. Accanto alla traslazione, nella comparazione con altri genomi troviamo esempi di rotazione (non illustrati) di nuvole di punti. Le combinazioni possibili di traslazione, rotazione e di variazione nella distanza tra nuvole di punti, creano un panorama di relazioni tra organismi che sarebbe interessante studiare in un quadro comune e uniforme. Dati due genomi, un tale quadro dovrebbe saper dire a cosa sia dovuta la loro differenza. Vedremo piú avanti una proposta formale di un tale spazio.

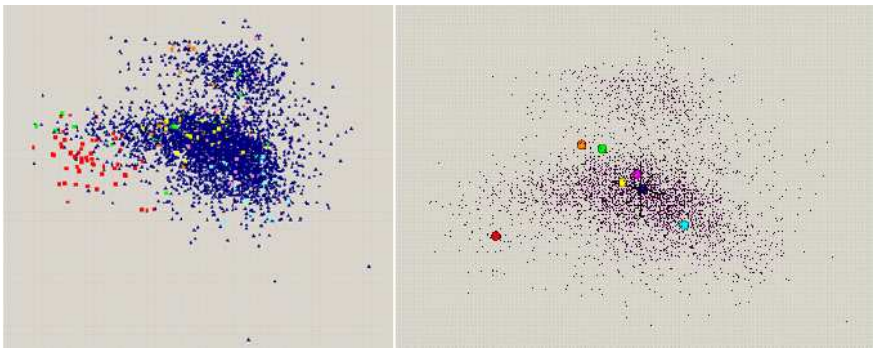

Figura 2. A sinistra, proiezione del genoma di *E. coli* dove gruppi di proteine annotate sono differenziati da colori diversi: proteine ribosomali (rosso), ATP binding proteins (giallo), IS proteins (turchese), proteine NADH (arancione), Proteine della biosintesi flagellare (rosa), lipoproteine, proteine membranali e proteine di trasporto (verde). A destra, la stessa proiezione rappresentata a sinistra ma dove i centroidi delle proteine annotate sono messi in evidenza con un punto espanso colorato (i colori sono gli stessi che a sinistra).

Interpretazione biologica delle rappresentazioni. Se guardiamo da vicino dove le famiglie di proteine annotate sono localizzate nello spazio dei codoni, scopriamo che formano gruppi fisicamente prossimi. In Figura 2 é illustrata la posizione di gruppi di geni annotati per il battere *E. coli* (sinistra) e la posizione dei centroidi dei gruppi caratterizzati funzionalmente (destra). La disposizione spaziale relativa dei centroidi rimane la stessa da genoma a genoma? Una risposta a questa domanda puó essere data dopo aver definito formalmente una nozione appropriata di ”rete” di centroidi e di ”deformazione accettabile” della rete. La questione costituisce un problema aperto.

Gli organismi che si riproducono rapidamente presentano una localizzazione molto particolare delle proteine ribosomali, che si trovano nella grande maggioranza localizzate in una delle ”orecchie di coniglio” introdotte sopra (notare la posizione delle proteine ribosomali colorate di rosso in Figura 2, a sinistra). In altre parole, le proteine ribosomali sono codificate con un insieme di codoni caratteristico. In generale, possiamo dire che tali organismi hanno un insieme di geni, costituente l'1% dei geni del genoma, che [Grantham et al. 1980, Sharp e Li 1987]:

- sono costituiti da codoni preferenziali

- sono necessari alla traduzione, oppure sono geni che devono essere tradotti rapidamente ed essere presenti nella cellula in grandi quantita' ad un momento dato.

Per studiare gli effetti del bias sulle proteine ribosomali, chiamato *bias traduzionale*, sull'espressione dei geni, Sharp & Li hanno suggerito di calcolare dei "pesi" rappresentanti la preferenza di certi codoni rispetto ad altri e di estrapolare il comportamento degli altri geni del genoma a partire da questa informazione. L'idea é che un tale processo permette di individuare quali siano i geni piú espressi e quelli meno espressi. E' importante tener presente che questa proposta, di natura computazionale, era stata fatta nel 1987, quando non esistevano ancora tecniche di analisi sperimentale come i microarrays per studiare il livello di espressione dei geni.

Piú precisamente, Sharp & Li [Sharp e Li 1987] hanno proposto di associare ad ogni gene di un genoma un valore numerico, chiamato *Indice di Adattamento dei Codoni* (CAI), che esprime il bias dei codoni sinonimi tra loro (una definizione formale é data nel seguito). Ad ogni codone verrá associato un *peso* (rappresentante il suo adattamento relativo) calcolato a partire dalla sua frequenza in un pool relativamente piccolo di geni altamente espressi S. Tali pesi saranno poi combinati per definire il peso di un gene g, indicato con $CAI(g)$. Per Sharp *et al.*, l'ipotesi che giustifica la scelta di S é che, per certi organismi, i geni piú frequentemente espressi in una cellula hanno un bias di codoni piú forte, e che questi geni, costituiti principalmente da codoni preferiti, siano rappresentativi del bias. Basato su un tale razionale, si sceglierá un pool di proteine ribosomali, fattori di elongazione, proteine partecipanti alla glicolisi, proteine degli istoni (in eucarioti) e di membrane esterne (in procarioti) o eventuali altre selezioni fatte tra geni conosciuti per essere altamente espressi, e costituirá l'insieme rappresentativo S. In tale maniera, verranno calcolati i valori CAI per ogni gene del genoma. I geni con i valori di CAI piú elevati saranno validati contro geni, non usati nell'insieme S, ma che si sanno essere molto espressi nell'organismo in esame. Parallelamente, un test su geni con livelli di espressione bassa dovrá venir fatto. Una validazione positiva sui due gruppi di geni permetterá, con un livello di confidenza relativamente alto, di estrapolare il comportamento dei livelli di espressione dei geni restanti, ed anche per quelli le cui funzioni non sono ancora state sperimentalmente determinate. Anche se concettualmente chiaro, questo schema di ragionamento ha incontrato un gran numero di usi incorretti nella letteratura e risultati biologici erronei sono stati piú volte determinati per organismi che non soddisfano bias traduzionale, come discusso ad esempio in [Grocock e Sharp 2002]. Questa confusione motiva la ricerca di una metodologia basata su un precisa formulazione matematica del problema per determinare l'esistenza del bias traduzionale. É, a questo proposito, importante precisare, che nemmeno da un punto di vista biologico, per una gran parte di genomi, é evidente come definire l'insieme di riferimento S. In particolare, la gran quantitá di organismi sequenziati che presentano informazioni biologiche poco note, aumenta l'interesse nell'analisi. Vedremo nel resto della lezione come, la "vecchia" nozione del bias dei codoni, puó ancora presentare delle sorprese.

Definizione formale del bias dei codoni. Per ogni genoma G e un insieme di sequenze codanti S in G, il *bias dei codoni* é misurato rispetto all'uso dei codoni che ne sono suoi sinonimi, cioé quei codoni che codano per lo stesso amino acido. Dato un amino acido j, i suoi codoni sinonimi possono avere frequenze diverse in S; se $x_{i,j}$ é il numero di volte che il codone i per un amino acido j occurre in S, allora associamo a i un *peso* $w_{i,j}$ relativo al suo sinonimo di frequenza massima y_j in S

$$w_{i,j} = \frac{x_{i,j}}{y_j}.$$

Un codone con frequenza massimale in S é chiamato *preferito* tra i suoi codoni sinonimi. L'*Indice di Adattamento dei Codoni* (CAI) associato da Sharp & Li [Sharp e Li 1987] al gene

g in G, é un valore in $[0, 1]$, definito come una media geometrica dei pesi dei codoni

$$CAI(g) = (\prod_{k=1}^{L} w_k)^{1/L}$$

dove L é il numero di codoni nel gene, e w_k é il peso del k-esimo codone nella sequenza del gene. Geni con valore di CAI vicino ad 1 sono costituiti da codoni particolarmente frequenti.

Una determinazione automatica del bias dei codoni. Abbiamo proposto un algoritmo per determinare il bias dei codoni dominante in un genoma [Carbone et al. 2003]. Una ricerca esaustiva di un tale insieme di geni é irrealizzabile per via dei tempi esponenziali di ricerca. L'algoritmo proposto é basato su una formulazione matematica precisa del problema che conduce all'uso di CAI come misura *universale* del bias dei codoni, cioé una misura per bias di origine differente (e non solo giustificata dalla traduzione, come era stata spiegata all'origine). A partire dal solo insieme di sequenze codanti come pool d'informazione di origine biologica, l'algoritmo produce un insieme di riferimento S costituito dai geni che sono i piu rappresentativi del bias dei codoni nel genoma. L'idea dell'algoritmo é semplice. Si tratta di un algoritmo iterativo:

1. calcolare il peso dei codoni sull'intero genoma e calcolare in seguito i valori di CAI per tutti i geni
2. selezionare il 50% dei geni con il piú alto valore di CAI
3. ripetere i passi 1 e 2 ma selezionare il 25% dei geni con il piú alto valore di CAI
4. e cosí via: ripetere i passi 1 e 2 ma selezionando il 12% dei geni e poi il 6%... fino ad arrivare all'1% dei geni.
5. ripetere i passi 1 e 2 sull'1% dei geni con il piú alto valore CAI fino ad ottenere la convergenza dell'insieme S.

L'insieme S é usato per calcolare l'Indice di Adattamento dei Codoni di geni di organismi procarioti e eucarioti, inclusi quelli la cui annotazione funzionale non é ancora disponibile. Un'applicazione importante dell'algoritmo riguarda la determinazione del bias traduzionale che correla i livelli d'espressione in molti batteri e eucarioti inferiori [Carbone et al. 2003, Carbone et al. 2004]; determina anche il bias della posizione dei geni sui filamenti d'ADN, il bias di contenuto GC, il bias GC3, e il trasferimento orizzontale dei geni. In generale, l'algoritmo diventa uno strumento di misura per predire i livelli d'espressione, per guidare la ricostruzione di circuiti regolatori di geni, e per comparare speci diverse. L'approccio é stato validato su 96 batteri e archee di divisione lenta, rapida, e sugli eucarioti *Saccharomyces cerevisiae*, *Plasmodium falciparum*, *Caenorhabditis elegans* e *Drosophila melanogaster*.

Versione randomizzata dell'algoritmo. Esiste una versione randomizzata dell'algoritmo che permette di mettere in luce in maniera ancora piú forte il fatto che una "traccia" del bias dei codoni é presente in tutti i geni del genoma e non solo in alcuni. L'idea é quella di partire da un insieme scelto aleatoriamente di geni e costituito dall'1% dei geni del genoma G, piuttosto che da tutto il genoma G. Si calcolano poi i pesi dei codoni e i valori CAI per tutti i geni del genoma, si seleziona l'1% dei geni con il piú forte valore di CAI e si reitera il processo fino a convergenza su un insieme che contiene l'1% dei geni del genoma. Sui genomi considerati (descritti sopra) abbiamo notato che questa versione aleatoria dell'algoritmo converge piú rapidamente della version originale, ed in particolare converge allo stesso insieme. Alcuni casi di

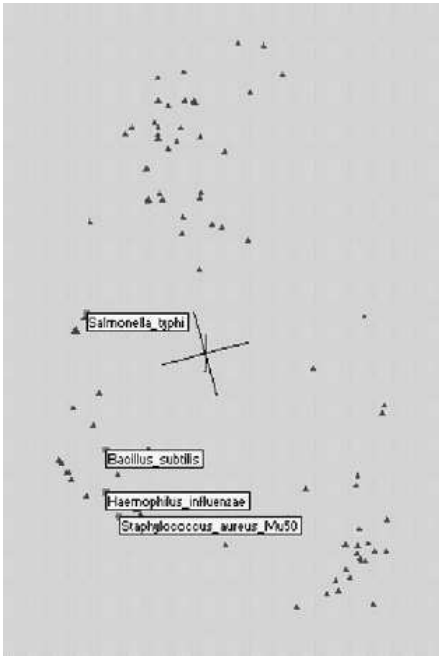

Figura 2. Vista della proiezione tridimensionale (realizzata con ACP) dello spazio a 64 dimensioni degli organismi. Rappresenta la posizione relativa di 96 batteri e archee. Notare la posizione prossima di *H. influenzae* e *S. aureus* le cui rappresentazioni tridimensionali sono illustrate in Figura 1.

non convergenza, ma giustificabili con considerazioni di forte interesse biologico, sono discussi in [Carbone et al. 2003].

”IMPRONTE” GENOMICHE E SPAZIO DI GENOMI ADATTO ALLA LORO COMPARAZIONE

Basati sull’analisi precedente, abbiamo proposto un nuova struttura formale per interpretare le relazioni tra organismi derivate dell’analisi di genomi completi piuttosto che da *loci* individuali (come, ad esempio, é fatto in filogenia considerando le proteine ribosomali). In un tale quadro, un organismo viene rappresentato dai pesi caratteristici dei suoi codoni, diventando cosi un punto in uno spazio a 64 dimensioni (vedi Figura 2). Quando questo spazio in 64 dimensioni di genomi é proiettato in 3 dimensioni dall’ACP, la componente principale é interpretabile con il contenuto GC et la seconda componente principale con la temperatura di crescita ottimale [Carbone et al. 2003]. É interessante notare come la prima componente principale abbia una ”natura” statistica e la seconda una ”natura” biologica.

Questo spazio permette l’analisi di insiemi di organismi legati tra loro da un pool di caratteristiche comuni riguardanti il bias del loro genoma. A volte, piú di un tipo di bias influisce sulla sequenza genomica di un organismo e l’insieme di bias sovrapposti definisce l’*impronta* di un genoma [Carbone et al. 2004]. Abbiamo fornito un insieme di criteri numerici per inferire il bias avente origine nel contenuto G e C dei filamenti di DNA, di origine nella traduzione e quello dipendente dalla posizione del gene sui filamenti. Abbiamo mostrato, in un contesto formale uniforme, che i genomi appartenenti a famiglie filogenetiche differenti possono condividere un bias di codoni simile; altri genomi, d’abitudine raggruppati insieme da vari metodi

filogenetici, appaiono come suddivisi, risultano divisi in sottogruppi con simili bias di codoni; archee e batteri sono caratterizzati dagli stessi codoni preferiti quando il loro bias dominante é AT3 o GC3; i genomi delle archee che sono di bias traduzionale sono caratterizzati da codoni preferiti nettamente distinti dal pool di codoni preferiti di genomi batterici aventi lo stesso bias. La nostra analisi, basata su 96 genomi batterici e archee, apre le porte all'idea che questo spazio possa riflettere la geometria di uno "spazio fisiologico" procariota. Se fosse vero, la combinazione di nuovi genomi sequenzializzati e la determinazione delle impronte di organismi diventerebbe un metodo utile per inferire informazione biologica sulla fisiologia, l'ecologia e possibilmente sulle condizioni ambientali sotto le quali organismi batterici e archee evolvono. Per molti organismi una tale informazione sarebbe impossibile da determinare con approcci sperimentali.

STUDIO DELLE RETI METABOLICHE ATTRAVERSO L'ANALISI DELLE SEQUENZE E DI DATI TRASCRITTOMICI E DETERMINAZIONE DELL'INSIEME DI GENI MINIMALE

Rifacciamo il punto sulla "logica" seguita fino a questo momento in questa presentazione. La nostra analisi dei genomi é partita da segnali di selezione basati sulla traduzione, abbiamo poi usato questi segnali per parlare in maniera formale e universale di bias dei codoni, in seguito abbiamo usato questa nozione matematica per associare ad un genoma una rappresentazione formale che permetta di vederlo come un punto in uno spazio di organismi a 64 dimensioni, numerosi test di validazione di questo spazio sono stati sviluppati e realizzati. Questa "logica" é validata dalla coerenza dei risultati una volta interpretati nel contesto biologico. Ci domandiamo allora, se tali segnali possono essere usati per catturare altra informazione di carattere biologico. In particolare, ci siamo chiesti se potessimo determinare quali fossero le piú importanti reti metaboliche di un organismo? In breve, la risposta é si. I primi indizi incoraggianti, che hanno condotto alla verifica di questa ipotesi, sono stati suggeriti dai geni caratterizzati da un forte bias che occorrono nei genomi di organismi a divisione rapida: in *Synechocystis* troviamo geni dedicati al cammino metabolico della fotosintesi; in *Methanosarcina acetivorans* troviamo quelli coinvolti nel metabolismo del metano; in *Pyrococcus abyssi* quelli del metabolismo della ferrodoxina; in *Streptococcus mutans* troviamo geni coinvolti nel metabolismo dei carboidrati. Tali geni riflettono le condizioni metaboliche conosciute di queste speci. Da questi esempi, ne deriviamo che geni con un elevato bias dei codoni descrivono in maniera significativa le caratteristiche fisiologiche di un organismo e sono rappresentativi di specifici usi metabolici [Carbone e Madden 2005].

Vorremmo dimostrare ora che, a parte l'alta espressivitá di certe proteine durante la fase di crescita rapida dell'organismo o le attivitá metaboliche di glicolisi che sono state notate spesse volte nella litteratura, la necessitá di sopravvivere sotto condizioni biologiche specifiche hanno lasciato una traccia nel codice genetico [Carbone e Madden 2005]. Questa osservazione apre la possibilitá alla predizione, a partire dall'analisi statistica del genoma, di attivitá metaboliche rare ma necessarie alla sopravvivenza dell'organismo.

Per rispondere alla questione, proponiamo di considerare gli insiemi di geni coinvolti in una rete metabolica data e di calcolare una "media" dei valori di CAI per i geni che ne fanno parte [Carbone e Madden 2005]. Si scopre in questa maniera che parecchi cammini metabolici dell'energia sono correlati con un *elevato* bias dei codoni in organismi con fisiologie differenti, organismi che non sono necessariamente caratterizzati da una divisione rapida e i cui genomi sono considerati essere omogenei. Piú in generale, deriviamo una classificazione di cammini

metabolici indotta dall'analisi dei codoni, che dimostra come il codice genetico di organismi differenti sia modulato a partire da cammini metabolici specifici e come questa osservazione abbia un carattere universale.

La composizione in codoni di enzimi che partecipano alla glicolisi ad esempio, che spesso richiedono di essere tradotti rapidamente, presenta un forte bias determinato dal bias dei codoni dominante (in altre parole questi enzimi hanno un valore elevato di CAI), e questa proprieta sembra verificata indipendentemente da quale sia l'organismo considerato. In organismi che si riproducono rapidamente, l'evidenza numerica di questa proprietá é notevolmente piú forte che per altri organismi (cioé, la differenza assoluta tra il valore di CAI di questi enzimi e la media dei valori CAI per i geni del genoma e' "grande"), ma anche per *Helicobacter pylori*, un genoma la cui composizione in codoni risulta piuttosto omogenea, gli enzimi che partecipano al cammino metabolico glicolitico sono "biased" al di sopra della media. Nello stesso modo, uno determina il ruolo cruciale dei cammini della fotosintesi per *Synechocystis* o del metabolismo del metano per *Methanobacterium*. Un esempio importante é dato dall'analisi dei cammini metabolici del *Mycobacterium tuberculosis* che ha permesso di identificare i cammini metabolici *essenziali* per questo battere. Tali cammini sono stati tutti validati precedentemente in laboratorio e corrispondono in maniera precisa a quelli determinati dalla nostra analisi statistica del genoma dell'organismo.

I livelli trascrizionali di mRNA collezionati durante il ciclo cellulare sotto condizioni di "diauxic shift" per *Saccharomices cerevisiae* [deRisi et al. 1997] (dove il glucosio decresce nel medium durante il ciclo cellulare, e il lievito passa dalla fermentazione alla respirazione aerobica), sono stati poi da noi usati per analizzare le reti metaboliche del lievito in modo simile a ció che abbiamo fatto per il bias dei codoni. Abbiamo proposto una classificazione dei cammini metabolici basati sui dati trascrittomici, e abbiamo dimostrato che la classificazione metabolica ottenuta attraverso l'analisi dei codoni essenzialmente "coincide" con quella basata sul pool di dati trascrittomici a disposizione (c'é da notare che un tale pool di dati é grande e differenziato, e dunque affidabile). Un tale risultato apre la via ad una spiegazione dei fenomeni di pressione evolutiva e di selezione naturale che affetta organismi che crescono nel loro ambiente naturale, aiuta a spiegare il metabolismo di batteri che crescono lentamente, e, potenzialmente, a suggerire le migliori condizioni di crescita in laboratorio.

Verso un genoma minimale. Gli insiemi dei geni codificati con codoni preferenziali in un organismo sono proposti come geni appartenenti al genoma minimale. Si tratta di geni con funzioni non identificate, geni dipendenti da specifiche condizioni ambientali e geni molto espressi (e appartenenti a tutte le speci). Il numero di geni facenti parte di un insieme minimale che proponiamo va da 200 a 500 geni. Tali geni corrispondono a quelli individuati sperimentalmente cosi come negli studi di genomica comparativa. (Referenze a tali studi sono date nella Tavola 1.) Otteniamo in questa maniera una cartografia di tutti i geni essenziali per un organismo. Essa ricopre tutte le principali reti metaboliche. Lo studio é stato fatto su 27 genomi di organismi a bias traduzionale [Carbone 2005].

APPENDICE: COMMENTO SUI METODI MATEMATICI

Tutti i risultati citati in questo articolo sono ottenuti usando strumenti matematici e algoritmici molto semplici che sono descritti nel dettaglio in [Carbone et al. 2003, Carbone et al. 2004, Carbone e Madden 2005]. L'analisi statistica e i sogli numerici che abbiamo proposto sono realizzati nello spazio dei codoni a 64 dimensioni. I metodi statistici di analisi multivariata sono

stati utilizzati come metodi di visualizzazione, ma nessuno dei risultati formali né le conclusioni biologiche sono dedotte dalla proiezione in 3 dimensioni. Lo spazio dei geni e lo spazio degli organismi sono defniti in 64 dimensioni, e le distanze tra organismi sono definite come distanze ℓ_1. Tutte le viste delle proiezioni ottenute con l'ACP sono state realizzate con il programma VidaExpert, disponibile all'URL www.ihes.fr/˜zinovyev.

REFERENCES

[Ajdić et al. 2002] D. Ajdić, W.M. McShan, R.E. McLaughlin, G. Savic, J. Chang, M.B. Carson, C. Primeaux, R. Tian, S. Kenton, H. Jia *et al.*, Genome sequence of *Streptococcus mutans UA159*, a cariogenic dental pathogen. *Proceedings of the National Academy of Sciences USA*, **99**, 14434–14439 (2002).

[Akerley et al. 2002] B.J. Akerley, E.J. Rubin, V.L. Novick, K. Amaya, N. Judson, J.J. Mekalanos, A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proceedings of the National Academy of Sciences USA*, **99**, 966–971 (2002).

[Bennetzen e Hall 1982] Bennetzen, J.L., Hall, B.D., Codon selection in Yeast. *Journal of Biological Chemistry*, **257**, 3026–3031 (1982).

[Brown et al. 2001] J.R. Brown, C.J. Douady, M.J. Italia, W.E. Marshall, M.J. Stanhope, Universal trees based on large combined protein sequence data sets. *Nature Genetics*, **28**, 281–285 (2001).

[Buckingham e Grosjean 1986] Buckingham, R.H., Grosjean, H., The accuracy of mRNA-tRNA recognition. In *Accuracy in molecular processes: its control and relevance to living systems*, ed. T.B.L. Kirkwood, R. Rosenberger and D.J. Galas, Chapman & Hall Publishers, London, 83-126 (1986).

[Bulmer 1987] Bulmer, M., Coevolution of codon usage and transfer RNA abundance. *Nature*, **325**, 728-730 (1987).

[Carbone e Gromov 2001] Carbone, A., Gromov, M., Mathematical slices of molecular biology. *La Gazette des Mathematiciens*, Numero speciale 88:11-80, Société Mathématique de France (2001).

[Carbone et al. 2004] A. Carbone, F. Képès, A. Zinovyev, Codon bias signatures, organisation of microorganisms in codon space and lifestyle. *Molecular Biology and Evolution*, **22**, 547-561 (2004).

[Carbone et al. 2003] A. Carbone, A. Zinovyev, F. Képès, Codon Adaptation Index as a measure of dominating codon bias. *Bioinformatics*, **19**, 2005–2015 (2003).

[Carbone e Madden 2005] A. Carbone, R. Madden, Insights on the evolution of metabolic networks of unicellular translationally biased organisms from transcriptomic data and sequence analysis. *Journal of Molecular Evolution*, in press, (2005).

[Carbone 2005] A. Carbone, Computational prediction of genomic functional cores specific to different microbes. Sottomesso a rivista (2005).

[Cello et al.2002] J. Cello, A.V. Paul, E. Wimmer, Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science*, **297**, 1016-1018 (2002).

[Charlebois e Doolittle 2004] R.L. Charlebois, W.F. Doolittle, Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Research*, **14**, 2469–2477 (2004).

[Cohen et al. 2003] G.N. Cohen, V. Barbe, D. Flament, M. Galperin, R. Heilig, O. Lecompte, O. Poch, D. Prieur, J. Querellou, R. Ripp *et al.*, An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. *Molecular Microbiology*, **47**, 1495–1512 (2003).

[deRisi et al. 1997] DeRisi, J.L., Iyer, V.R., Brown, P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686 (1997).

[Gerdes et al. 2003] S.Y. Gerdes, M.D. Scholle, J.W. Campbell, G. Balazsi, E. Ravasz, M.D. Daugherty, A.L. Somera, N.C. Kyrpides, I. Anderson, M.S. Gelfand, A. Bhattacharya *et al.*, Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of Bacteriology*, **185**, 5673–5684 (2003).

[Giaever et al 2002] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre *et al.*, Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391 (2002).

[Gouy e Gautier 1982] Gouy, M. and Gautier, Ch., Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, **10**, 7055-7070 (1982).

[Grantham et al. 1980] Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A., Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, **8**, r49-r62 (1980).

[Grocock e Sharp 2002] Grocock, R.J., Sharp, P.M., Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene*, **289**, 131–139 (2002).

[Harris et al. 2003] J.K. Harris, S.T. Kelley, G.B. Spiegelman, N.R. Pace, The genetic core of the universal ancestor. *Genome Research*, **13**, 407–412 (2003).

[Hutchison et al. 1999] C.A. Hutchison, S.N. Peterson, S.R. Gill, R.T. Cline, O. White, C.M. Fraser, H.O. Smith, J.C. Venter, Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, **286**, 2165–2169 (1999).

[Ikemura 1985] Ikemura, T., Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13-34 (1985).

[Itaya 1995] M. Itaya, An estimation of the minimal genome size required for life. *FEBS Lett.*, **362**, 257–260 (1995).

[Kamath et al. 2003] R.S. Kamath, A.G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann *et al.*, Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237 (2003).

[Koonin et al. 1997] E.V. Koonin, A.R. Mushegian, M.Y. Galperin, D.R. Walker, Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin of the archaea. *Mol. Microbiology*, **25**, 619–637 (1997).

[Koonin 2000] E.V. Koonin, How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum.Genet.*, **1**, 99–116 (2000).

[Koonin 2003] E.V. Koonin, Comparative genomics, minimal gene sets and the last common ancestor. *Nature Reviews Microbiology*, **1**, 127–136 (2003).

[Lafay et al. 1999] Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M. and Wolfe, K.H., Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Research*, **27**, 1642–1649 (1999).

[Makarova et al. 1999] K.S. Makarova, L. Aravind, M.Y. Galperin, N.V. Grishin, R.L. Tatusov, Y.I. Wolf, E.V. Koonin, Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Research*, **9**, 608–628 (2003).

[Médigue et al. 1991] Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A., Evidence for horizontal gene transfer in *Escherichia coli* speciation. *Journal of Molecular Biology*, **222**, 851-856 (1991).

[Moszer et al. 1999] Moszer, I., Rocha, E.P.C., Danchin, A., Codon usage and lateral gene transfer in *Bacillus Subtilis*. *Current Opinion in Microbiology*, **2**, 524–528 (1999).

[Mushegian e Koonin 1996] A.R. Mushegian, E.V. Koonin, A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Science USA*, *93*, 10268–10273 (1996).

[Nesbø et al. 2001] C.L. Nesbø, Y. Boucher, W.F. Doolittle, Defining the core of non-transferable prokaryotic genes: the euryarchaeal core. *Journal of Molecular Evolution*, **53**, 340–350 (2001).

[Sharp e Li 1987] P.M. Sharp, W-H. Li, The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acid Research*, **15**, 1281–1295 (1987).

[Sharp et al. 1986] Sharp, P.M., Tuohy, T.M.F., Mosurski, K.R., Codon usage in yeast: cluster analysis clearly differentiate highly and lowly expressed genes. *Nucleic Acids Research*, **14**, 8207-8211 (1986).

[Sharp et al. 1988] Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomices pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Research*, **16**, 8207-8211 (1988).

[Shields e Sharp 1987] Shields, D.C. and Sharp, P.M. Synonymous codon usage in *Bacillus subtilis* reflects both traditional selection and mutational biases. *Nucleic Acids Research*, **15**, 8023-8040 (1987).

[Smith et al. 2003] H.O. Smith, C.A. Hutchison III, C. Pfannkoch, C. Venter, Generating a synthetic genome by whole genome assembly: $\phi X174$ bacteriophage from synthetic oligonucleotides. *Proceedings of the National Academy of Sciences USA*, **100**, 15440–15445 (2003).

[Stenico et al. 1994] Stenico, M., Loyd, A.T., Sharp, P.M. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acid Research*, **22**, 2437–2446 (1994).

[Varenne et al. 1984] Varenne, S., Buc, J., Lloubès, R. and Lazdunski, C. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of Molecular Biology*, **180**, 549-576 (1984).

[Venter et al. 2003] J.C. Venter, S. Levy, T. Stockwell, K. Remington, A. Halpern, A massive parallelism, randomness and genomic advances. *Nature Genetics*, **33**, 219–227 (2003).

[Zimmer 2003] C. Zimmer, Genomics. Tinker, tailor: can Venter stitch together a genome from scratch? *Science*, **299**, 1006–1007 (2003).