UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Leggi statistiche generali in sistemi con molti componenti

Relatore

Prof. Sandro Azaele

Correlatore

Dr. Emanuele Pigani

Laureando

Simone Toso

Anno Accademico 2021/2022

Indice

Capitolo 1

Introduzione

I sistemi complessi sono sistemi costituiti da un grande numero di componenti microscopiche che interagiscono localmente tra loro. La caratteristica che contraddistingue tali sistemi è la possibilità di originare spontaneamente fenomeni di auto-organizzazione globale collettiva a livello macroscopico, che difficilmente possono essere spiegati e predetti sulla base della sola conoscenza dei loro costituenti microscopici. Lo studio dei sistemi complessi si focalizza dunque su come pattern su grande scala possano emergere da interazioni su piccola scala.

Un'importante classe di sistemi complessi è data dai *sistemi modulari*: insiemi composti da un numero elevato di elementi microscopici preesistenti che, combinandosi fra loro in svariate proporzioni, danno origine a diverse *realizzazioni*. I set di LEGO, i capitoli di un libro, le famiglie di proteine presenti in un genoma e le molecole di mRNA nei trascrittomi di cellule appartenenti a tessuti diversi sono esempi di sistemi modulari.

Una caratteristica peculiare di questi sistemi è la presenza di leggi empiriche che emergono esaminando la loro composizione e che sono comuni a sistemi anche molto diversi tra loro, suggerendo quindi l'esistenza di leggi universali che non risentono dei dettagli particolari di ciascun sistema. La **legge di Zipf**, ad esempio, fa riferimento a una distribuzione a legge di potenza delle frequenze con cui compaiono gli elementi microscopici che compongono un sistema modulare in funzione della posizione che tali elementi occupano in una classifica stilata in ordine di frequenza decrescente. Tale legge deve il nome al linguista americano George Kingsley Zipf [1], che l'ha osservata studiando la distribuzione delle parole più frequenti in un *corpus* di testi in lingua inglese, ed è stata riscontrata, ad esempio, anche in sistemi biologici, in campo economico, urbanistico e sociale [2]. Altre leggi empiriche osservate sono la crescita sublineare del numero di elementi diversi in funzione della dimensione del sistema (**legge di Heaps**) e una dipendenza a legge di potenza della varianza in funzione della media delle frequenze con cui appare un dato elemento (**legge di Taylor**). Un'ulteriore caratteristica di questi sistemi è il fatto che la distribuzione delle *componenti comuni* risulti seguire una "distribuzione a U": compaiono quindi molti elementi comuni, molti elementi rari e pochi elementi intermedi.

Sono stati proposti numerosi modelli generativi per la realizzazione simulata di sistemi modulari: molti di questi modelli si basano sul principo dell'*attaccamento preferenziale* [3], andando gradualmente a ridurre il vocabolario possibile per spiegare l'emergenza della legge di Zipf nella distribuzione delle frequenze. In ogni caso, indipendentemente dalla natura del modello utilizzato, un meccanismo che voglia descrivere statisticamente la composizione dei sistemi modulari deve riproporre le leggi empiriche finora descritte.

Una possibilità a tal proposito è data dai *null models*: modelli generativi di natura statistica che non sono soggetti a vincoli legati alla specificità del sistema simulato e che definiscono la probabilità di estrazione di ogni elemento in base alla frequenza empiricamente osservata. Studiando la differenza tra la composizione sistemi reali e dei sistemi simulati con un *null model*

si può capire come e in che misura tali leggi siano frutto di effetti stocastici e quanto invece siano date da specifici vincoli strutturali. In questo modo è possibile mettere in evidenza meccanismi che soggiaciono alla realizzazione dei sistemi esaminati.

In questa tesi descriveremo da un punto di vista matematico il comportamento di un *null model* a estrazione casuale, soffermandoci in particolare su come tale modello preveda elementi più o meno comuni e derivando alcune espressioni analitiche che ne descrivono l'azione. Applicheremo poi tale modello allo studio di due dataset: il database `SUPERFAMILY`, che riporta la composizione dei proteomi di organismi procarioti e che è stato già analizzato in [4], e alcuni dati meta-omici raccolti dalla spedizione TARA [5] su una classe di organismi unicellulari eucarioti autotrofi, le diatomee. L'analisi del database `SUPERFAMILY` permette di ricondurre alcuni aspetti della distribuzione delle occorrenze a fenomeni di natura stocastica che emergono anche da un processo di generazione casuale; d'altro canto, l'analisi svolta ha messo in evidenza un risultato già riportato da Mazzolini et. al [4], ovvero l'esistenza, nei proteomi dei batteri, di componenti che compaiono esattamente una volta in ogni individuo e che non possono essere ricondotte a fenomeni statistici. Nel corso dell'analisi dei dati di `TARA`, invece, si è proposto un modello generativo diverso e si è studiato sotto quali ipotesi tale modello possa spiegare l'andamento generale delle occorrenze.

Capitolo 2

Null model a estrazione casuale

Una vasta gamma di sistemi complessi possono essere visti come *collezioni* di elementi fondamentali, estratti in quantità diverse da un *vocabolario*: definiamo *realizzazione* un sistema di questo tipo. Siccome elementi diversi compariranno in quantità diverse, possiamo dire che alcuni elementi sono più *frequenti* di altri (*definiremo in modo più rigoroso questo concetto nelle prossime righe*). Possiamo immaginare queste diverse realizzazioni come il risultato di un processo di estrazione casuale da un'unica distribuzione "vera", a noi in genere non nota. Questo modello generativo è un *null model*, in quanto non fa alcuna ipotesi sull'esistenza di vincoli strutturali legati alla natura del sistema realizzato.

Conoscendo il comportamento di un simile modello, possiamo verificare se la composizione di veri sistemi complessi modulari possa essere ricondotta a un processo di estrazione casuale, permettendo quindi di capire se un determinato pattern sia sintomatico di vincoli strutturali o se possa semplicemente emergere per effetti statistici.

Nelle prossime pagine studieremo analiticamente il comportamento di un tale modello generativo.

2.1 Definizioni

2.1.1 Notazione e rappresentazione matriciale

Si consideri un insieme di R realizzazioni aventi dimensione $M_1, M_2, \ldots, M_R$ e si supponga che le loro componenti provengano da un vocabolario di N elementi distinti. Possiamo rappresentare questo dataset come una matrice A di dimensione $N \times R$: ogni colonna rappresenta una realizzazione e ogni riga uno specifico elemento. In questo modo l'entrata A_{ij} corrisponde al numero di volte in cui la componente i-esima compare nella $j-$esima realizzazione: definiamo quindi A_{ij} come l'*abbondanza* dell'elemento i nella realizzazione j.

Focalizziamoci ora su un'unica realizzazione. Se denotiamo con a_i l'abbondanza dell'elemento i in tale realizzazione, segue naturalmente la definizione di *frequenza* dell'elemento i nella realizzazione j: $f_i = a_i / \sum_{k=1}^{N} a_k$. Si noti che per definizione si ha $\sum_{i=1}^{N} f_i = 1$ e $f_i \geq 0$, $\forall i = 1, \ldots, N$. Il caso $f_i = 0$ corrisponde all'eventualità in cui l'elemento i non compare mai nella realizzazione considerata.

Siamo inoltre interessati alla frazione di realizzazioni in cui compare un dato elemento, indipendentemente dalla sua abbondanza: definiamo quindi l'*occorrenza* $o_i := \frac{1}{R} \sum_{j=1}^{R} (1 - \delta_{A_{ij},0})$, dove $\delta_{\alpha,\beta}$ è la funzione delta di Kronecker.

2.1.2 Null-model a estrazione casuale con sostituzione

Definiamo ora il concetto di *null-model a estrazione casuale con sostituzione*, supponendo di partire da un set empirico di realizzazioni. Prendiamo in considerazione R_{obs} realizzazioni aventi dimensione $M_1, M_2, \ldots, M_{R_{obs}}$: ognuno degli N elementi del vocabolario comparirà con una frequenza diversa nelle varie realizzazioni.

Supponiamo ora di voler generare una realizzazione di dimensione M: abbiamo bisogno di una distribuzione che descriva la probabilità di estrazione dei vari componenti. Ciò può essere fonte di ambiguità: vorremmo infatti giungere alla stima di una *frequenza teorica* a partire dai dati empirici.

La f dei vari elementi è soggetta a una variabilità intrinseca legata alla diversa natura di ogni realizzazione. In particolare, definendo $\vec{f} = \begin{bmatrix} f_1 & \ldots & f_N \end{bmatrix}$ il vettore composto dalle N frequenze, supponiamo che tale variabilità intrinseca si manifesti come una distribuzione di probabilità $\rho(\vec{f})$. Chiamando $\vec{a} = [a_1, ..., a_N]$ l'array che contiene le abbondanze dei vari elementi in questa realizzazione, avremo

$$P(\vec{a}|M) = \int_0^1 d\vec{f}\rho\left(\vec{f}\right) \frac{M!}{\prod_i a_i!} \prod_i (f_i)^{a_i}. \tag{2.1}$$

Come si evince dall'equazione 2.1, la variabilità nell'abbondanza dell'elemento i è data non solo dalla presenza della $\rho(\vec{f})$, ma anche da un effetto di campionamento, rappresentato dal termine $\frac{M!}{\prod_i a_i!} \prod_i (f_i)^{a_i}$. La conoscenza della $\rho(\vec{f})$ è spesso impossibile; se, tuttavia, supponiamo che il suo effetto sulla variabilità delle abbondanze sia trascurabile rispetto al rumore di campionamento, possiamo approssimare la statistica del null model con un processo di tipo poissoniano. Il valore "vero" della frequenza f_i potrà infine essere stimato dal valor medio delle frequenze con cui l'elemento i si presenta nelle varie realizzazioni. Stiamo dunque supponendo che per la frequenza f_i la densità di probabilità sia data da $\rho(f_i) = \delta(f_i - \overline{f_i})$, ove $\overline{f_i}$ è la frequenza media data da

$$\overline{f_i} = \frac{1}{R_{obs}} \sum_j \frac{a_{ij}}{M_j}. \tag{2.2}$$

Il sistema così ottenuto può essere dunque descritto da un vettore di abbondanze $\vec{a}$; alla luce delle considerazione appena esposte, gli array seguiranno una distribuzione multinomiale:

$$P(\vec{a}|\vec{f}) = \frac{M!}{\prod_{j=1}^N a_j!} \prod_{j=1}^N f_j^{a_j}. \tag{2.3}$$

Osserviamo che nell'equazione 2.3 stiamo trascurando le correlazioni, ipotizzando quindi che l'estrazione di un elemento i non influenzi la probabilità di una successiva estrazione di un altro elemento j.

La probabilità che il singolo elemento i compaia a_i volte sarà data invece da una distribuzione binomiale, ottenuta marginalizzando la 2.3:

$$P(a_i|f_i) = \binom{M}{a_i} (f_i)^{a_i} (1 - f_i)^{M - a_i} \approx e^{-Mf_i} \frac{(Mf_i)^{a_i}}{a_i!}. \tag{2.4}$$

Il limite poissoniano è giustificato nel caso di basse frequenze e alte dimensioni ($Mf_i = \text{cost.}, f_i \to 0^+, M \to +\infty$); la validità di tale approssimazione dipenderà dalle dimensioni dei dataset considerati.

Il *null model* consiste dunque in questo: è un'estrazione di elementi dal vocabolario, adottando come distribuzione di probabilità l'array delle frequenze medie $\vec{f}$. Tale modello è talvolta detto "con sostituzione" poiché la probabilità di estrazione dei vari elementi non cambia man mano che vengono estratti.

Nota importante: Siccome, visto quanto detto finora, lavoreremo sempre supponendo che la $\rho(f)$ sia una delta piccata sul valor medio, nell'analisi dei dataset adotteremo la notazione $f_i := \overline{f_i}$.

2.1.3 RAD e SAD

Da un punto di vista sperimentale, l'informazione che abbiamo su un insieme di realizzazioni è l'insieme di elementi diversi campionati e la loro abbondanza.
È usanza comune lavorare quindi con *rank-abundance distributions* (RAD): gli elementi vengono etichettati con un indice $i = 1, 2, ..., N_{obs}$, andando in ordine decrescente dall'elemento con f maggiore ($i = 1$) a quello con f minore ($i = N_{obs}$).

Definiamo invece *species abundance distribution* (SAD) la densità di probabilità per la frequenza f; in altre parole essa descrive l'istogramma in cui contiamo quanti elementi hanno una data frequenza (detta anche "abbondanza relativa" in alcuni articoli). Talvolta la SAD viene indicata come RSA (*relative species abundance*). Se abbiamo un'espressione analitica $q(f)$ per la SAD, il rank di un elemento può essere visto come una variabile continua $i(f)$ che sottostà alla relazione differenziale $di = -N_{obs}q(f)df$; un'espressione analitica $f(i)$ per la RAD può essere ricavata integrando tale espressione e invertendola.

Termine	Definizione
Frequenza	Frazione di elementi di una realizzazione appartenenti al tipo i. Nell'analisi di dataset reali stimiamo il suo valore "vero" con la frequenza media tra tutte le realizzazioni
Occorrenza	Frazione di realizzazioni in un dataset in cui compare almeno una volta l'elemento in questione
RAD	Funzione $f(i)$, ove i è la posizione in classifica della frequenza f ($i = 1$ per l'elemento a frequenza maggiore, $i = N_{obs}$ per l'elemento a frequenza minore)
SAD	Densità di probabilità per la frequenza f.

Tabella 2.1: Principali termini introdotti nella sezione precedente e relative definizioni.

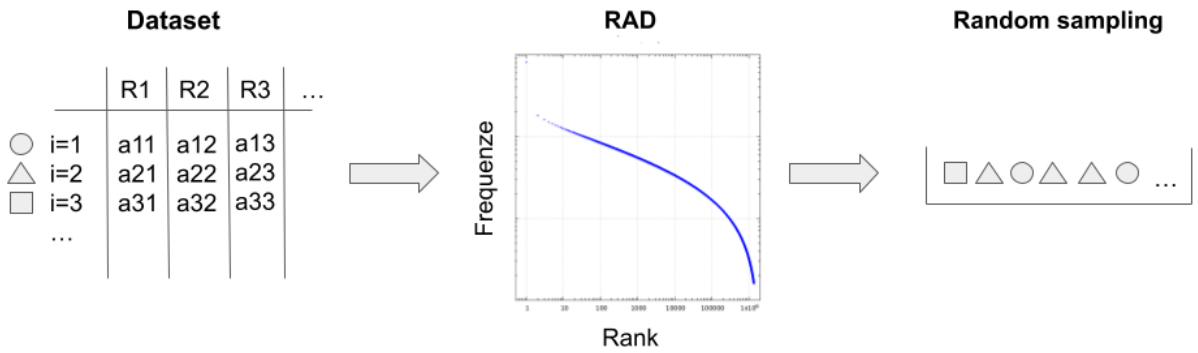


Figura 2.1: Schema riassuntivo della procedura di random sampling. La RAD è ottenuta mettendo in ordine decrescente le frequenze medie dei vari elementi; tali valori costituiscono le probabilità di estrazione, attraverso le quali viene fatto il random sampling.

2.2 Distribuzione di Mandelbrot e legge di Zipf

Supponiamo di avere un vocabolario composto da N elementi e di indicizzarli in base alla loro posizione i in una classifica in ordine di frequenza decrescente, andando dunque a definirne la RAD. La **distribuzione di Mandelbrot**, motivata dall'osservazione empirica delle frequenze in tali sistemi, è data da

$$f_i = \frac{1}{\alpha}(c+i)^{-\gamma} \tag{2.5}$$

con $c, \gamma > 0$ parametri reali e $\alpha = \sum_{i=1}^{N}(c+i)^{-\gamma}$ fattore di normalizzazione. Nel caso $c = 0$ ricadiamo nella distribuzione di Zipf-Pareto, la cui rilevanza in natura è nota come **legge** empirica **di Zipf**. Questo andamento delle frequenze è di particolare rilevanza: la legge di Zipf, infatti, nasce dall'osservazione empirica che in numerosi sistemi modulari la RAD segue un andamento a legge di potenza [1]. Essa compare, ad esempio, nel caso delle famiglie di proteine, oggetto d'analisi del prossimo capitolo.

Per questo motivo, è utile calcolare analiticamente il comportamento atteso di un *null model* che segua la legge di Zipf (il caso più generale definito da una distribuzione di Mandelbrot prevede un calcolo del tutto analogo e conduce alla stessa espressione analitica). In particolare, supponiamo di comporre R realizzazioni di abbondanza totale M, estraendo gli elementi da un vocabolario avente una distribuzione del tipo:

$$f_i = \frac{1}{\alpha} i^{-\gamma} \quad \text{con} \quad \alpha = \sum_{1}^{N} i^{-\gamma} \tag{2.6}$$

In generale, l'elemento i avrà occorrenza $\mathscr{O}_i$, che nell'ipotesi di null model è una variabile aleatoria distribuita in modo gaussiano. Definendo $\vec{M} = \begin{bmatrix} M_1 & \dots & M_R \end{bmatrix}$ l'array contenente le abbondanze totali delle R realizzazioni, il suo valore atteso o_i è dato da

$$o_i(f_i|R, \vec{M}) = 1 - \frac{1}{R}\sum_{j=1}^{R} (1-f_i)^{M_j}. \tag{2.7}$$

Possiamo spiegare intuitivamente il significato della 2.7: la sommatoria è la somma su ogni realizzazione della probabilità che l'elemento *non* compaia; l'idea dunque è che l'occorrenza attesa sia ottenuta sottraendo a 1 la frazione attesa di realizzazioni in cui l'elemento non compare.

Se supponiamo che ogni realizzazione abbia la stessa abbondanza totale M, la 2.7 si riduce a

$$o_i(f_i|R, M) = 1 - (1-f_i)^M. \tag{2.8}$$

Sempre supponendo che ogni realizzazione abbia la medesima abbondanza totale, il valore di aspettazione del *vocabolario osservato*, i.e., il numero totale di elementi diversi campionati, sarà

$$N_{obs}(N|MR) = N - \sum_{i=1}^{N}(1-f_i)^{MR}. \tag{2.9}$$

Statistica delle componenti condivise nel caso di una power-law distribution

Componendo realizzazioni simulate secondo una distribuzione di Zipf, possiamo immaginare di riempire un istogramma con le occorrenze attese o_i. Qual è l'espressione analitica $p(o)$ che descrive l'istogramma così ottenuto?
Osserviamo innanzitutto che la posizione in classifica i di un dato componente è data dal numero di elementi aventi occorrenza attesa o maggiore; è bene ricordare, per evitare di cadere in

errori grossolani, che questa osservazione è vera solo se si parla di valori attesi, come stiamo facendo adesso. Ricordando che $N_{obs} = N_{obs}(N|MR)$ è il valore atteso del vocabolario osservato, abbiamo:

$$i(o) = \sum_{o'=o}^{o_1} N_{obs} p(o') \approx N_{obs} \int_o^{o_1} p(o')do' \tag{2.10}$$

ove $o_1 := o(f_1)$ è il valore di aspettazione dell'occorrenza per l'elemento con frequenza maggiore. Il passaggio dal discreto al continuo è giustificato dal fatto che, per $N_{obs} \gg 1$, gli incrementi possono essere considerati infinitesimi, vedendo quindi il rank i come funzione continua di o. Usando queste due formule si può ricavare analiticamente l'espressione di

$$p : [o_{N_{obs}}, o_1] \to \mathbb{R}^+.$$

Per semplicità, nei prossimi calcoli supponiamo che ogni realizzazione abbia abbondanza totale M e facciamo uso dell'approssimazione $N_{obs}(N|MR) \approx N$: tale limite è giustificato se $MR \gg 1$, come si evince dalla 2.9. Abbiamo, rifacendoci alla 2.7,

$$\begin{aligned} o_i &= 1 - \frac{1}{R}\sum_{j=1}^{R}(1 - \frac{1}{\alpha}i^{-\gamma})^{M_j} = 1 - (1 - \frac{1}{\alpha}i^{-\gamma})^M \\ &= 1 - \left(1 - \frac{1}{\alpha}\left(N\int_o^{o_1} p(o')do'\right)^{-\gamma}\right)^M \end{aligned} \tag{2.11}$$

Da cui, derivando rispetto a o, sostituendo con le espressioni note e sfruttando la relazione 2.11, troviamo:

$$1 = -M\left(1 - \frac{1}{\alpha}\left(N\int_o^{o_1} p(o')do'\right)^{-\gamma}\right)^{M-1} \cdot \frac{\gamma}{\alpha}\left(N\int_o^{o_1} p(o')do'\right)^{-\gamma-1} \cdot (-Np(o))$$

$$= M(1-o)^{\frac{M-1}{M}}(\frac{\gamma}{\alpha})Np(o)\left[\alpha(1-(1-o)^{\frac{1}{M}})\right]^{1+\frac{1}{\gamma}}$$

$$\implies p(o) = \frac{\alpha}{\gamma MN}\frac{(1-o)^{\frac{1}{M}-1}}{\alpha^{1+\frac{1}{\gamma}}[1-(1-o)^{\frac{1}{M}}]^{1+\frac{1}{\gamma}}}$$

da cui troviamo la forma analitica:

$$p(o) = \frac{(1-o)^{\frac{1}{M}-1}}{\gamma MN\alpha^{\frac{1}{\gamma}}[1-(1-o)^{\frac{1}{M}}]^{1+\frac{1}{\gamma}}} \tag{2.12}$$

È facile verificare che questa espressione, in approssimazione $N \gg 1$, soddisfi la condizione di normalizzazione $\int_{o_{min}}^{o_1} p(o)do = 1$. Infatti la sostituzione $1-(1-o)^{\frac{1}{M}} = f$ porge

$$\int_{o_{N_{obs}}}^{o_1} p(o)do = \frac{1}{\gamma N\alpha^{\frac{1}{\gamma}}}\int_{f_{min}}^{f_{max}} df\, f^{-\frac{1}{\gamma}-1} = \frac{N-1}{N} \approx 1,$$

ove si è fatto uso della sostituzione $(\alpha f)^{-\frac{1}{\gamma}} = i$. Il termine $\frac{N-1}{N}$ deriva dal fatto che stiamo approssimando il "conteggio" degli elementi $\sum_{i=1}^{N} 1$ con un integrale $\int_1^N di$.

Nel limite di componenti rare $o \ll 1$ la 2.12 tende a

$$p(o) \approx \frac{M^{\frac{1}{\gamma}}}{\alpha^{\frac{1}{\gamma}}\gamma N} o^{-\frac{1}{\gamma}-1} \tag{2.13}$$

La distribuzione $p(o)$ ha dunque, nel limite di componenti rare $o \to 0^+$, una dipendenza a legge di potenza con esponente $\beta := \frac{1}{\gamma} + 1$.
È infine facile vedere, come anticipato prima, che l'espressione 2.12 si estende anche al caso di una più generica distribuzione di Mandelbrot $f_i = a(c+i)^{-\gamma}$: la derivazione della formula procede allo stesso identico modo e conduce allo stesso risultato, osservando che $\frac{d}{do}(c+i(o)) = \frac{di(o)}{do} = -Np(o)$; si è quindi scelto di ricavare la $p(o)$ per una legge di Zipf senza perdita di generalità.

Riportiamo in figura 3.3 il confronto tra la $p(o)$ appena ricavata e l'istogramma delle occorrenze per un sistema simulato; la simulazione è stata effettuata estraendo elementi da una RAD a power-law con probabilità data dalla loro frequenza.

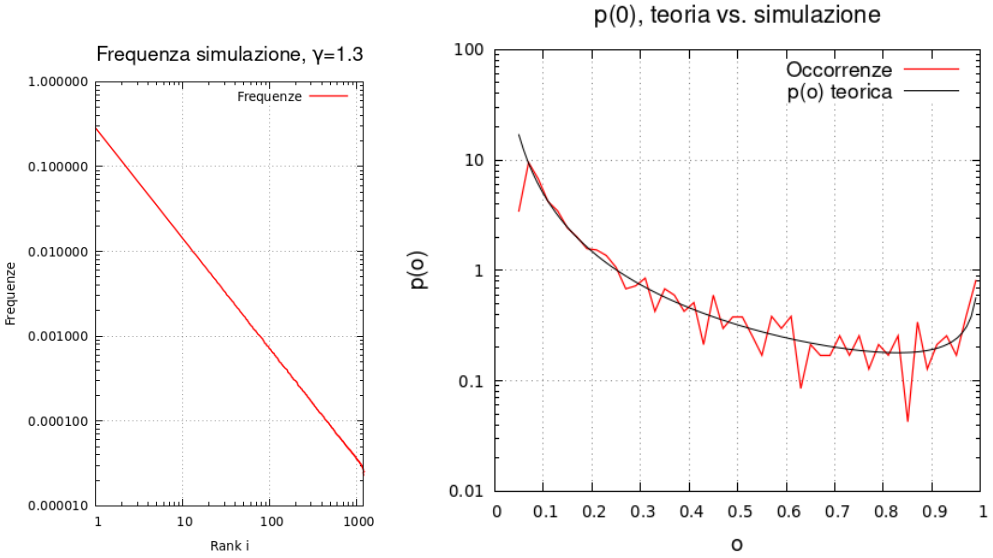


Figura 2.2: *Sinistra*: RAD a power-law utilizzata per la simulazione di un insieme di realizzazioni; la simulazione segue la procedura di estrazione casuale prevista dal *null model*. *Destra*: Confronto tra distribuzione delle occorrenze teorica e simulata. Parametri impostati per il random sampling: set di $R = 2000$ realizzazioni con $M = 2000$ elementi estratti da un vocabolario di $N = 1200$ parole.)

2.2.1 Dimensione del *core*

Definiamo *core size* del vocabolario la frazione di componenti che compaiono con un'occorrenza $o > \theta_c$, con $\theta_c \in [0, 1]$ parametro arbitrario che indica la soglia di occorrenza oltre la quale una componente può essere considerata "comune". È di interesse definire una formula analitica per la dimensione del *core* per confrontarla in un secondo momento con i dati empirici.
Supponiamo una distribuzione delle frequenze a *power-law*, avendo quindi $p(o)$ data da 2.12. La *core size* c sarà quindi data da:

$$\begin{aligned} c &= \int_{\theta_c}^{o_{max}} p(o)do = \int_{\theta_c}^{o_{max}} do \frac{(1-o)^{\frac{1}{M}-1}}{\gamma M N \alpha^{\frac{1}{\gamma}} [1-(1-o)^{\frac{1}{M}}]^{\frac{1}{\gamma}+1}} \\ &= -\frac{1}{N\alpha^{\frac{1}{\gamma}}} \left[(1-(1-o)^{\frac{1}{M}})^{-\frac{1}{\gamma}} \right]_{\theta_c}^{o_{max}} \end{aligned}$$

e infine, osservando che, per $M \gg 1$ e o fissato,

$$(1-(1-o)^{\frac{1}{M}})^{-\frac{1}{\gamma}} = \left(1-\exp\left(\frac{1}{M}\ln(1-o)\right)\right)^{-\frac{1}{\gamma}} \approx \left(-\frac{1}{M}\ln\left(1-o\right)\right)^{-\frac{1}{\gamma}} \tag{2.14}$$

arriviamo a

$$\begin{cases} c \approx k(-\ln(1-\theta_c))^{-\frac{1}{\gamma}},\ k := \frac{M^{-\frac{1}{\gamma}}}{N\alpha^{\frac{1}{\gamma}}},\ \theta_c > o_{min} \\ c = 1,\ \theta_c < o_{min} \end{cases} \tag{2.15}$$

Abbiamo quindi che, fissati i parametri θ_c e γ, la dimensione del *core* cresce linearmente con k.

Legge di Heaps come conseguenza della legge di Zipf

Come visto in precedenza, il vocabolario misurato cresce in funzione di M e R. In particolare la legge di Heaps osserva che tale crescita è sublineare [6].
Sempre supponendo una distribuzione delle frequenze come da equazione 2.6, possiamo studiare analiticamente la dipendenza del vocabolario N di una realizzazione dal numero di elementi m (il caso di R set di dimensione M è analogo al caso di un unico set di dimensione $m = MR$ vista la natura del *null model*).
Abbiamo che il valore atteso N è dato da

$$N_{obs}(m) = N - \sum_{i=1}^{N}\left(1-\frac{i^{-\gamma}}{\alpha}\right)^{m} = N - \sum_{i=1}^{N} e^{m\ln(1-\frac{i^{-\gamma}}{\alpha})}$$

Da qui, approssimando la sommatoria con un integrale e supponendo $m \gg 1$, troviamo

$$N - \frac{1}{\gamma}\int_1^N di\ e^{(-m\frac{i^{-\gamma}}{\alpha})} \overset{\substack{di=(\frac{m}{\alpha})^{\frac{1}{\gamma}}(-\frac{1}{\gamma})z^{-\frac{1}{\gamma}-1}dz \\ z=m\frac{i^{-\gamma}}{\alpha}}}{=} N - \int_{\frac{m}{\alpha N^{\gamma}}}^{\frac{m}{\alpha}} dz\left(\frac{m}{\alpha}\right)^{\frac{1}{\gamma}} z^{-\frac{1}{\gamma}-1}e^{-z}$$
$$= N - \left(\frac{m}{\alpha}\right)^{\frac{1}{\gamma}}\frac{1}{\gamma}\left[\Gamma\left(-\frac{1}{\gamma},x\right)\right]\Big|_{x=\frac{m}{\alpha}}^{x=\frac{m}{\alpha N^{\gamma}}}$$

ove si è fatto uso della funzione gamma incompleta $\Gamma(s,t) := \int_t^{+\infty} z^{s-1}e^{-z}dz$.

L'integrale ottenuto andrebbe calcolato come differenza tra due termini, proporzionali rispettivamente a $\Gamma(-\frac{1}{\gamma},\frac{m}{\alpha})$ e $\Gamma(-\frac{1}{\gamma},\frac{m}{\alpha N_{obs}^{\gamma}})$; il primo termine può tuttavia essere trascurato nell'approssimazione $\frac{m}{\alpha} \to +\infty$. Abbiamo quindi trovato l'espressione

$$N_{obs}(m) \approx N - \left(\frac{m}{\alpha}\right)^{\frac{1}{\gamma}}\frac{1}{\gamma}\Gamma\left(-\frac{1}{\gamma},\frac{m}{\alpha N^{\gamma}}\right). \tag{2.16}$$

Per studiare il comportamento di questa funzione per grandi realizzazioni possiamo usare l'espansione asintotica in un intorno di $+\infty$ data da $\Gamma(s,z) \approx z^{s-1}e^{-z}\sum_{k=0}^{\infty}\frac{\Gamma(s)}{\Gamma(s-k)}z^{-k}$ trovando

$$N_{obs}(m) = N - \left(\frac{m}{\alpha}\right)^{\frac{1}{\gamma}}\frac{1}{\gamma}\left[(\frac{m}{\alpha N^{\gamma}})^{-\frac{1}{\gamma}-1}e^{-\frac{m}{\alpha N^{\gamma}}}\right] = N\left[1-\frac{\alpha N^{\gamma}}{m}\frac{1}{\gamma}e^{-\frac{m}{\alpha N^{\gamma}}}\right] \tag{2.17}$$

Osserviamo quindi che il vocabolario raggiunge la saturazione per dimensioni dell'ordine di αN^{γ}. Ciò implica che la velocità con cui il sistema "scopre" tutti gli elementi possibili, raggiungendo quindi il regime di saturazione $N_{obs}(m) \approx N$, dipende dall'esponente della power-law distribution.

Si può inoltre dimostrare analiticamente [6] che prima del regime di saturazione, ovvero quando $N \gg m$, la distribuzione di Mandelbrot (e quindi anche la legge di Zipf) implica che la

crescita del vocabolario sia ben descritta da una crescita a legge di potenza $N(m) \propto m^{\zeta}$, con $\zeta = \frac{1}{\gamma}$[1].

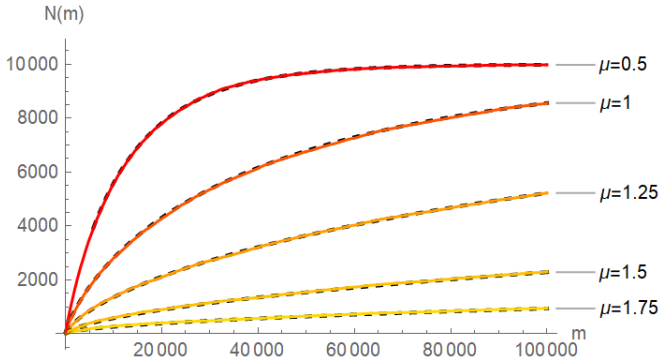


Figura 2.3: Andamento del vocabolario osservato N_{obs} al variare dell'abbondanza totale m: confronto tra teoria (linee tratteggiate) e simulazione (linee colorate). Il null model a estrazione casuale con frequenze a power-law distribution rispetta l'espressione analitica.

2.3 Statistica delle componenti comuni per una distribuzione esponenziale

Nella sezione 2.2 avevamo osservato che, nel caso di una distribuzione di Zipf-Pareto, le occorrenze sono distribuite secondo una distribuzione di probabilità $p(o)$ che, per componenti rare (ovvero $o \ll 1$), segue un andamento a legge di potenza con esponente $-\frac{1}{\gamma} - 1$. Estendendo la trattazione al caso di frequenza distribuite secondo un'esponenziale decrescente possiamo osservare che l'andamento a legge di potenza delle occorrenze emerge nuovamente. Ricaviamo dunque la $p(o)$ per questa nuova distribuzione delle frequenze.
Data

$$f(i) = \frac{1}{\alpha} e^{-\lambda i}, \qquad \alpha = \sum_{i=1}^{\tilde{N}} e^{-\lambda i} \tag{2.18}$$

varrà ancora la relazione data da 2.7:

$$o = 1 - (1 - f)^M = 1 - (1 - \frac{1}{\alpha} e^{-\lambda i})^M.$$

Derivando rispettoa o si trova

$$1 = -M(1 - \frac{e^{-\lambda i}}{\alpha})^{M-1} (\frac{\lambda}{\alpha} e^{-\lambda i})(-Np(o))$$

che, applicando sostituzioni analoghe a quelle svolte nella derivazione di 2.12, conduce a

$$p(o) = \frac{(1-o)^{\frac{M-1}{M}}}{\lambda M N [1 - (1-o)^{\frac{1}{M}}]} \tag{2.19}$$

Prendendo il limite di grandi campioni $M \gg 1$ si trova quindi

$$M[1 - (1-o)^{\frac{1}{M}}] = M[1 - e^{\frac{1}{M}\ln(1-o)}] \approx (-\ln(1-o)) \implies \lim_{M \to \infty} p(o) = \frac{(1-o)^{-1}}{\lambda N ln[(1-o)^{-1}]}$$

[1]In particolare la legge trovata da van Leijenhorst e van der Weide è $N(m) = \Gamma(1 - \frac{1}{\mu}) \alpha_{\infty}^{\frac{1}{\mu}} m^{\frac{1}{\gamma}}$, ove α_{∞} è il limite per $N \to \infty$ del fattore di normalizzazione α.

Nel limite di componenti rare $o \ll 1$ possiamo fare uno sviluppo in serie, ottenendo

$$p(o) \approx -\frac{1+o}{\lambda N(-o)} = \frac{1+o}{N\lambda o} \approx \frac{1}{N\lambda} o^{-1}$$

Si ha quindi che anche per questa distribuzione si ottiene una decrescita a legge di potenza per la $p(o)$ nel limite di componenti rare, in questo caso con un esponente -1 indipendente dalla λ che caratterizza l'esponenziale.

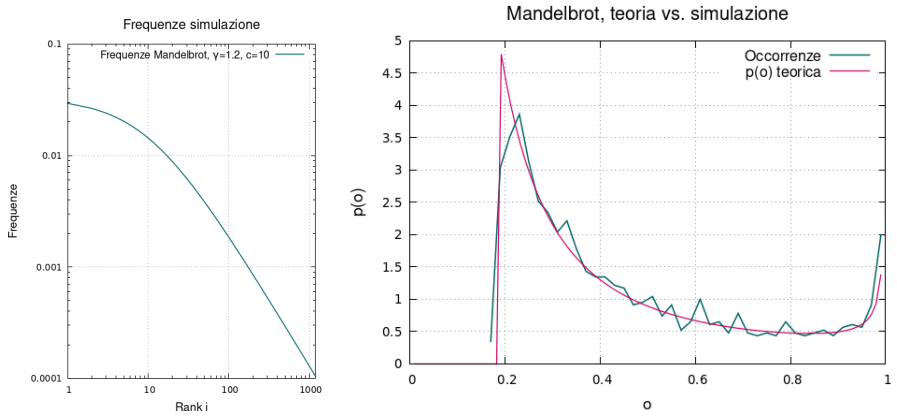


Figura 2.4: *Sinistra*: RAD, distribuzione di Mandelbrot. *Destra*: Confronto della distribuzione delle occorrenze teorica con quella simulata con un processo di random sampling (set di $R = 2000$ realizzazioni con $M = 2000$ elementi da un vocabolario di $N = 1200$ parole.)

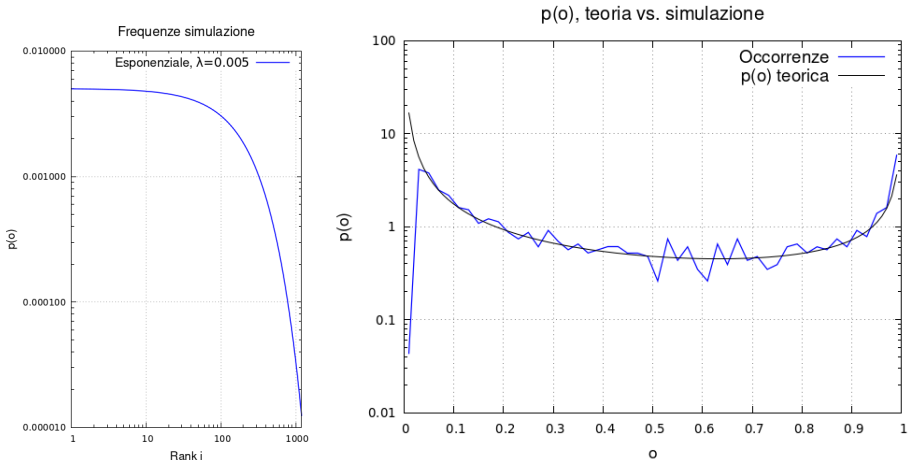


Figura 2.5: *Sinistra*: RAD, distribuzione a esponenziale decrescente. Destra: Confronto della distribuzione delle occorrenze teorica con quella simulata tramite un processo di random sampling (numero di realizzazioni $R = 2000$, $M = 2000$ elementi per ogni realizzazione, estratti da un vocabolario di $N = 1200$ parole.)

2.4 Distribuzione delle occorrenze nel caso di una RAD a power-law con cutoff esponenziale

Abbiamo ricavato un'espressione analitica per l'istogramma normalizzato delle occorrenze, prima ipotizzando una RAD a legge di potenza, poi nel caso di una distribuzione a esponenziale

decrescente. Abbiamo ottenuto delle espressioni già note [7].
Una RAD che più volte compare nello studio di dataset empirici è data da una legge di potenza, che viene modulata da una coda esponenziale, effetto della taglia finita del sistema. Proviamo quindi a ricavare un'espressione analitica per la $p(o)$ ottenuta da tale distribuzione.

La RAD è data da:

$$f(i) = \frac{1}{\alpha} i^{-\lambda} e^{-\mu i} \tag{2.20}$$

Proviamo a ottenere una espressione per la $p(o)$ descritta da questa distribuzione.
Supponiamo, come fatto per la distribuzione di Zipf (2.2)e per l'esponenziale (2.3), di estrarre da un vocabolario di N parole, componendo R realizzazioni di abbondanza totale M. Ricordiamo inoltre che si suppone sempre che il vocabolario osservato $N_{obs}(N|MR)$ possa essere approssimato da quello totale N: $N_{obs}(N|MR) \approx N$.

Avremo

$$\begin{aligned} o = 1 - (1-f)^M \implies 1 &= M(1-f)^{M-1} \frac{df}{di}(-Np(o)) \\ &= MN(1-o)^{1-\frac{1}{M}} p(o) \left[1 - (1-o)^{\frac{1}{M}}\right] \left[\frac{\lambda}{i} + \mu\right] \end{aligned}$$

Abbiamo $i = \frac{\lambda}{\mu} W\left(\frac{\mu(\alpha f)^{\frac{-1}{\lambda}}}{\lambda}\right)$, ove W è la funzione di Lambert t.c. $W(z) = w \iff we^w = z$.
Arriviamo dunque a

$$p(o) = \frac{1}{\mu MN} \frac{(1-o)^{\frac{1}{M}-1}}{1-(1-o)^{\frac{1}{M}}} \cdot \left[1 + \frac{1}{W\left(\frac{\mu}{\lambda} \alpha^{-\frac{1}{\lambda}} (1-(1-o)^{\frac{1}{M}})^{-\frac{1}{\lambda}}\right)}\right]^{-1} \tag{2.21}$$

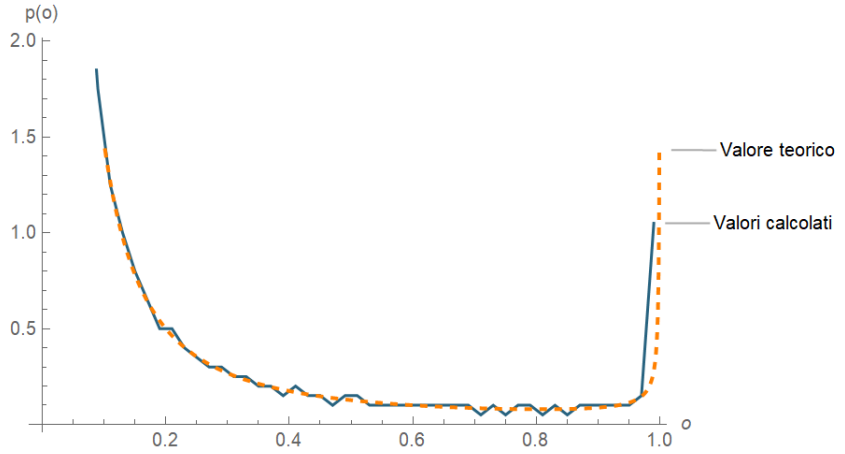


Figura 2.6: Confronto, nel caso di RAD a power-law modulata da un esponenziale decrescente, tra realizzazioni ottenute tramite simulazione e $p(o)$ ricavata analiticamente.

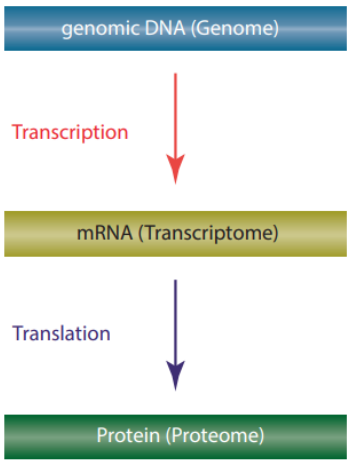
Capitolo 3

Un'applicazione pratica: i domini strutturali

Nella sezione precedente abbiamo esaminato alcune leggi statistiche che, sotto l'ipotesi di un' estrazione casuale, predicono la dimensione del vocabolario e la distribuzione delle occorrenze; è interessante adesso confrontare tali risultati con un dataset reale.
Inizieremo dal database `SUPERFAMILY`: le *realizzazioni* saranno i proteomi di organismi procarioti, mentre le *componenti* saranno domini strutturali, sezioni autonome della catena di aminoacidi di una proteina.

Genoma, trascrittoma e proteoma

Acidi nucleici (DNA e RNA) e proteine sono macromolecole che svolgono un ruolo fondamentale in ogni forma di vita: il DNA contiene l'informazione genetica che codifica la composizione delle proteine, mentre l'RNA è coinvolto nella loro biosintesi. I costituenti di DNA e RNA sono i nucleotidi, mentre le proteine sono composte da aminoacidi [8].



I quattro nucleotidi (A, T, C, G) che compongono il DNA codificano la composizione in aminoacidi delle proteine sintetizzate dall'organismo. Negli organismi viventi vengono sintetizzati 20 aminoacidi diversi, ognuno dei quali è codificato nel DNA da una terna di basi; queste terne prendono il nome di *codoni*[1].
Nella codifica di una proteina l'informazione viene trasferita dal DNA ai ribosomi tramite l'*RNA messaggero* (mRNA), il quale trascrive la sequenza di nucleotidi che codifica la proteina: nelle cellule eucariote l'mRNA deve essere trasportato dal nucleo al citoplasma, mentre in una cellula procariota (quindi priva di nucleo) il processo di sintesi per azione dei ribosomi può avere inizio subito dopo la trascrizione. L'intero *pool* di mRNA presente in un organismo è detto *trascrittoma*; similmente il termine *proteoma* indica l'insieme delle proteine sintetizzate da un organismo.

[1] A dire il vero, codoni diversi possono codificare lo stesso aminoacido: abbiamo 4^3 possibili terne di nucleotidi contro un totale di 20 aminoacidi diversi. Per questo motivo il codice genetico è detto *degenere*.

Proteine e domini strutturali

La sequenza di aminoacidi di una proteina è detta *struttura primaria*; la conoscenza della sola struttura primaria di una proteina non permette di predirne univocamente la funzione: quest'ultima dipende infatti dalla configurazione spaziale degli aminoacidi [4],[8],[9].

Questa struttura, detta *struttura terziaria*, è estremamente complessa; tuttavia, è possibile distinguere sezioni della proteina la cui disposizione spaziale è determinata indipendentemente dal resto della catena. In altre parole, la sequenza di aminoacidi può essere divisa in sezioni che si dispongono indipendentemente l'una dall'altra: queste sezioni prendono il nome di *domini strutturali*. Un singolo dominio ha una lunghezza compresa tra circa 50 e 250 aminoacidi; svariate proteine sono costituite da più di un dominio. Domini diversi possono essersi evoluti da un medesimo dominio e possono dunque essere classificati in una *superfamiglia*.

È possibile classificare le superfamiglie in base al ruolo che esse tipicamente ricoprono. La **SCOP annotation** (Structural Classification of Proteins) [10] mappa le superfamiglie in sette categorie:

1. **Informazione**: funzioni legate alla conservazione del codice genetico, alla replicazione del DNA e a processi di trascrizione.
2. **Regolazione**: regolazione dell'espressione genica[2] e risposta agli input ambientali.
3. **Metabolismo**: processi necessari alla respirazione cellulare
4. **Processi intracellulari**: mobilità della cellula, trasporto intracellulare, secrezione
5. **Processi extracellulari**: adesione cellulare e altre funzioni inter- e extracellulari.
6. **Funzioni generali**: domini strutturali con molteplici funzioni
7. **Altro**: funzioni non ancora note.

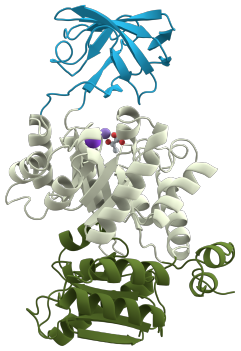


Figura 3.1: Proteina (*Pyruvate kinase*) in cui si distinguono tre domini strutturali

Il database `SUPERFAMILY` [9] contiene la classificazione in superfamiglie delle sequenze di aminoacidi estratte dal genoma di svariati organismi procarioti.
Possiamo considerare i proteomi come sistemi complessi modulari le cui componenti sono le superfamiglie di proteine: analizzando la composizione in superfamiglie di diversi genomi possiamo provare ad applicare le considerazioni teoriche svolte finora a un contesto reale. In particolare noi prenderemo in analisi un set di $R = 1082$ proteomi con un vocabolario totale di $N = 1530$ superfamiglie. In un proteoma del dataset analizzato compaiono in media $\overline{M} = 3388$ superfamiglie diverse.

3.1 Distribuzione delle frequenze e delle occorrenze

3.1.1 Rad e distribuzione delle occorrenze

In figura 3.2 riportiamo le frequenze con cui le superfamiglie compaiono nei genomi analizzati. Nei sistemi reali non è raro osservare una distribuzione delle frequenze che presenta diversi

[2]Per *espressione genica* si intende il processo in un cui l'informazione di un gene risulta nella produzione della proteina da esso codificata.

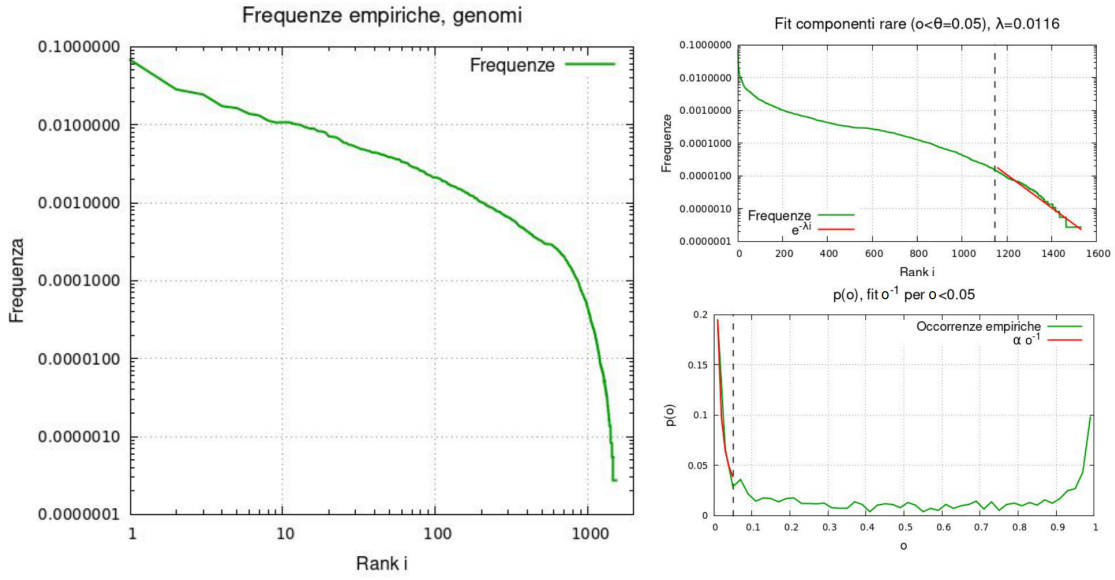


Figura 3.2: *Sinistra:* frequenze delle superfamiglie nel dataset utilizzato. *Destra, in alto:* fit delle componenti rare (ovvero con $o(i) < 0.05$) tramite una funzione esponenziale. *Destra, in basso:* fit per $o < 0.05$ con una funzione $\propto o^{-1}$ in virtù di quanto previsto per dalla teoria per un random sampling.

regimi; in questo caso, ad esempio, la distribuzione presenta un andamento a power-law per le componenti ad alte frequenze e una decrescita esponenziale nelle componenti più rare. Possiamo tuttavia estendere la trattazione teorica svolta finora restringendo l'analisi a un sottoinsieme delle frequenze possibili.
Nel primo capitolo avevamo osservato che da una distribuzione a esponenziale decrescente segue un andamento $\propto o^{-1}$ per le componenti rare. Possiamo verificare la persistenza di tale andamento effettuando un fit esponenziale dei punti (i, f) con $o(i) < \theta_r$, ove θ_r è un parametro arbitrario al di sotto del quale le componenti sono considerate "rare". Ricordiamo che l'aspettativa teorica delle occorrenze è data da 2.7. Il modello teorico prevede quindi una dipendenza monotona crescente dell'occorrenza in funzione della frequenza; gli elementi con $o(i) < \theta_r$ saranno dunque tutti e soli gli elementi di posizione in classifica $i > i(\theta_r)$.
Scegliamo di usare $\theta_r = 0.05$. In figura 3.2 riportiamo il risultato di tale fit (che fornisce un esponente $\lambda = 0.0116 \pm 0.0001$) e la distribuzione delle occorrenze $p(o)$ avendo fittato i punti $o < \theta_r$ con una funzione Ao^{-1}. Non si osservano divergenze apprezzabili dalla predizione teorica: l'elevata frazione di componenti rare è una conseguenza dell'eterogeneità delle frequenze ed è ben descritta da un modello teorico basato sul random sampling. Essa non è quindi indice di particolari vincoli strutturali legati alla natura dei genomi.

3.1.2 Occorrenze: confronto tra dati e simulazione

Esaminiamo ora la distribuzione delle occorrenze, confrontandola con quella descritta dal *null model*.
Riportiamo in figura 3.3 l'istogramma delle occorrenze empiriche e simulate. La simulazione è stata effettuata estraendo elementi con probabilità data dalla RAD empirica; per ogni realizzazione j del dataset, data la sua abbondanza totale M_j, sono stati estratti M_j elementi, preservando

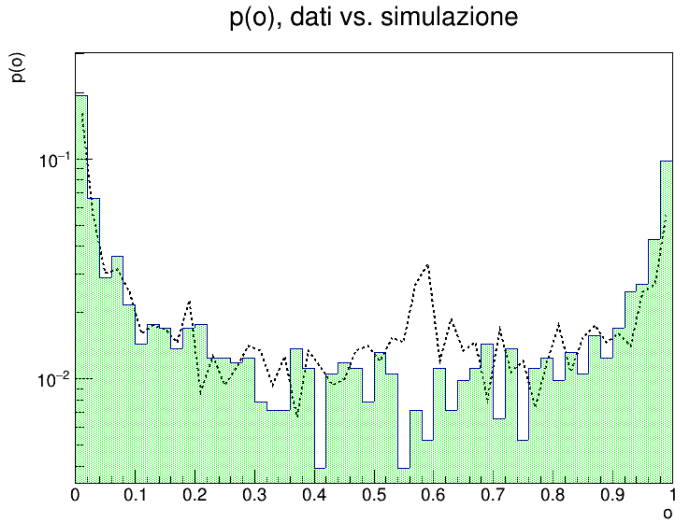


Figura 3.3: Confronto della distribuzione delle occorrenze: dati empirici (istogramma) e simulazione (linea tratteggiata). Si osserva una discrepanza tra *null model* e dati intorno a $o \approx 0.6$.

.

quindi l'abbondanza totale empirica.
Il dataset presenta l'andamento a U tipico dei sistemi modulari; come descritto nel paragrafo precedente, la presenza di numerose componenti rare è spiegata dalla distribuzione molto eterogenea delle frequenze. Osserviamo invece un discostamento notevole della simulazione dai dati empirici attorno a $o \approx 0.6$.

Possiamo esaminare meglio tale discrepanza studiando la distribuzione delle occorrenze al variare della classificazione delle superfamiglie prevista dalla *SCOP annotation* [10]. Riportiamo in figura 3.4, a titolo di esempio, uno scatter-plot dei punti empirici (f, o) per le superfamiglie appartenenti rispettivamente alla classi funzionali *Informazione* e *Metabolismo*. Il valore atteso delle occorrenze nel caso di un null model, dato dall'equazione 2.7, descrive solo parzialmente i dati empirici; in particolare osserviamo che, per f fissato, le occorrenze si distribuiscono su un range ampio attorno al valore atteso.

Osserviamo inoltre che il valore di o è superiormente limitato: un elemento di abbondanza a_i, infatti, avrà occorrenza massima data da $o_{max} = \min(a_i/R, 1)$; il caso $o = a_i/R$ corrisponde alla situazione in cui ognuna delle $a_i < R$ copie dell'elemento si trova in una realizzazione diversa.

Mentre le superfamiglie della classe funzionale *Metabolismo* risultano omogeneamente distribuite attorno al valore teorico, le superfamiglie del gruppo *Informazione* sono concentrate sul punto di incontro tra la curva delle massime occorrenze e la retta $o = 1$. Ciò suggerisce la presenza di superfamiglie della classe *Informazione* che compaiono una sola volta in ogni genoma. Osserviamo che una superfamiglia che compaia esattamente una volta in ogni realizzazione avrebbe un'occorrenza attesa, data dalla formula 2.7, pari a $o \approx 0.59$, coerentemente con la discrepanza tra *null model* e dataset osservata in figura 3.3.
La sovrabbondanza di famiglie del gruppo *Informazione* che compaiono una sola volta in ogni realizzazione appare evidente anche osservando l'istogramma 2D dei punti (f, o) (figura 3.5).

Il grafico delle rank-abundances (figura 3.6) mostra effettivamente un *plateau* nella zona di frequenze che vengono mappate dal random sampling nel range di occorrenze del picco.
Osserviamo inoltre che escludendo le famiglie del gruppo *Informazione* dal grafico delle *rank-abundance* il plateau non è più apprezzabile. Escludendo, infatti, le famiglie del gruppo *In-*

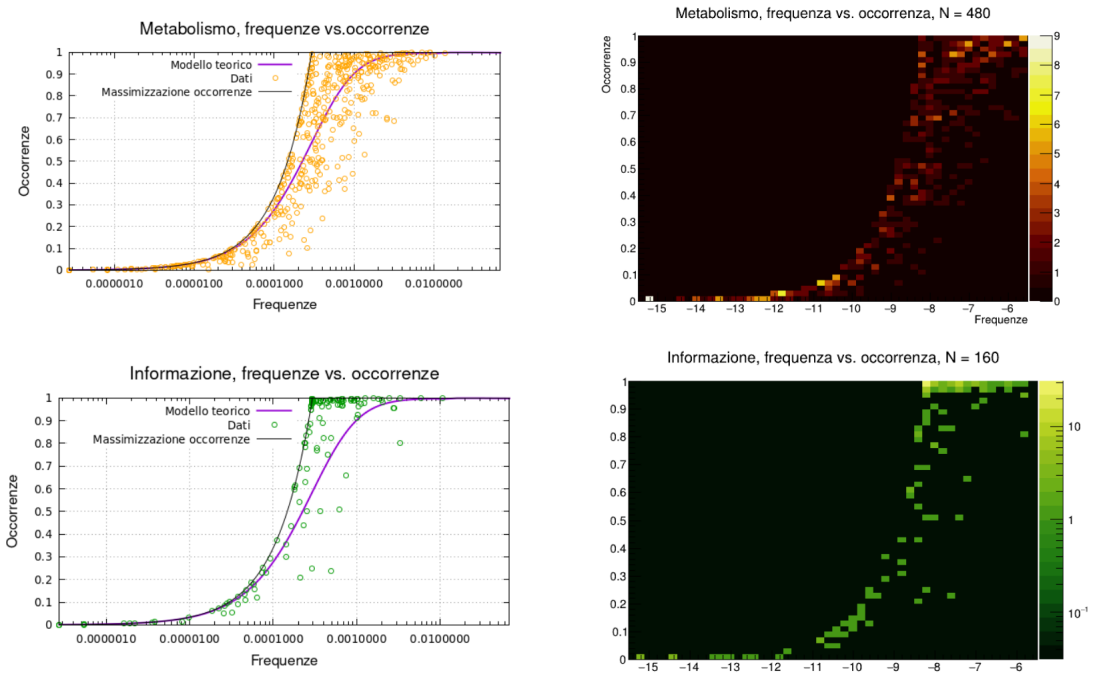


Figura 3.4: Scatterplot dei punti (f, o) per superfamiglie appartenenti alle categorie *Informazione* e *Metabolismo*. Le occorrenze delle superfamiglie appartenti alla classe *Informazione* si concentrano all'intersezione tra la curva delle massime occorrenze e $o = 1$.

.

formazione abbiamo una diminuzione significativa del picco di occorrenze a 0.6 (figura 3.7)

Comportamento di una distribuzione teorica con plateau

Nota importante: nel resto della tesi, per semplicità, si è sempre definito il valore di aspettazione dell'occorrenza simulata come E[o](f):=o(f). Siccome, in questa sezione, dovremo prendere in considerazione la variabilità del null model, torneremo momentaneamente alla notazione originale: il valore di aspettazione dell'occorrenza sarà dunque definito come E[o](f).

Verifichiamo che la presenza, in una RAD, di un plateau alla frequenza $\tilde{f}$, si manifesta nella distribuzione delle occorrenze come un picco centrato attorno al valore di aspettazione $E[o](\tilde{f})$.

Osserviamo innanzitutto che, ipotizzando una distribuzione uniforme delle frequenze ($f(i) \equiv \frac{1}{N}$, con N numero di parole nel vocabolario), otteniamo una $p(o)$ pressochè gaussiana, centrata attorno a $E[o](\tilde{f})$ (*fig.* 3.9). Ciò è spiegabile osservando che x, il numero di realizzazioni in cui compare un elemento di frequenza f, segue un andamento poissoniano con valore di aspettazione $\lambda := E[o](\tilde{f}) \cdot R$. Questa pdf, per $R \gg 1$, tende a una gaussiana centrata su λ e con deviazione standard $\sqrt{\lambda}$. La distribuzione delle occorrenze è ricavabile riscalando l'asse delle ascisse tramite il cambio di variabile $o = \frac{x}{R}$. In questo modo il valor medio è $E[o](\tilde{f})$ e la deviazione standard della gaussiana diventa $\sigma = \frac{\sqrt{E[o](\tilde{f}) \cdot R}}{R} = \sqrt{\frac{E[o](\tilde{f})}{R}}$.
Possiamo studiare il caso di una distribuzione teorica $f(i)$ a cui agguiungiamo un plateau a una data frequenza $\tilde{f}$; ci aspettiamo di veder comparire un picco gaussiano attorno a $E[o](\tilde{f})$ in maniera analoga a quanto osservato sperimentalmente con i genomi.

Possiamo modellizzare il risultato atteso vedendo l'istogramma come somma di un "fondo" a *power-law* e un "segnale" gaussiano.

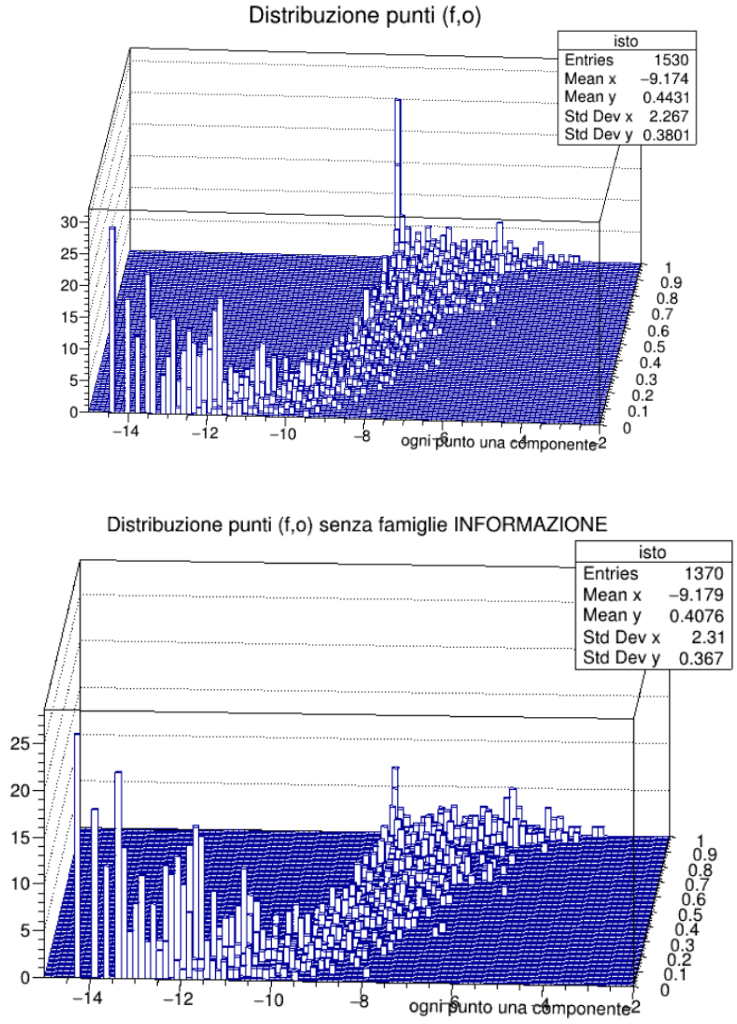


Figura 3.5: Istogramma dei punti empirici (f, o) per tutti i gruppi funzionali (*sopra*) ed escludendo le famiglie della famiglia *Informazione* (*sotto*); si nota la scomparsa di un picco posto all'intersezione tra la curva di massima occorrenza

Denoteremo N_{pow} il numero di elementi distribuiti secondo la power-law e N_{plat} il numero di elementi contenuti nel plateau.
L'espressione della $p(o)$ è quindi:

$$p(o) = \frac{N_{pow}}{N_{pow} + N_{plat}} \cdot p_{pl}(o) + \frac{N_{plat}}{N_{pow} + N_{plat}} \cdot \mathcal{G}(o)$$

con $p_{pl}(o) := \frac{(1-o)^{\frac{1}{M}-1}}{\gamma M N \alpha^{\frac{1}{\gamma}} [1-(1-o)^{\frac{1}{M}}]^{1+\frac{1}{\gamma}}}$ e $\mathcal{G}(o) := \frac{1}{\sqrt{2\pi E[o](\tilde{f})/R}} \exp\left[-\frac{1}{2}\frac{(o-E[o](\tilde{f}))^2}{E[o](\tilde{f})/R}\right]$. Tale espressione è riportata in figura 3.8, sovrapposta all'istogramma delle occorrenze simulate.

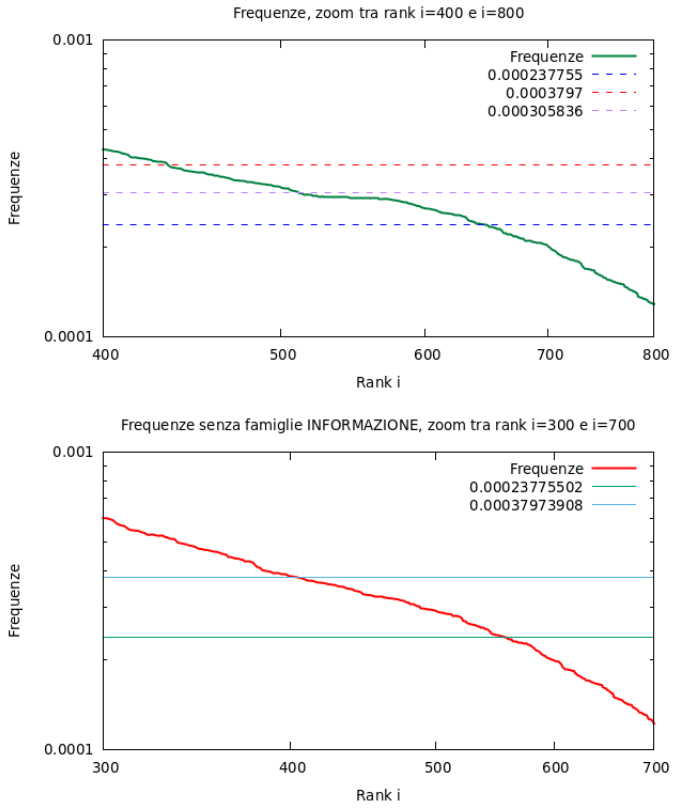


Figura 3.6: `Rad`, zoom sulle frequenze mappate nel picco a $o = 0.6$, con (*sopra*) e senza (*sotto*) famiglie dell'informazione: dopo la rimozione, si nota la scomparsa del plateau. Esso era infatti per lo più da elementi mappati nella funzione *Informazione*.

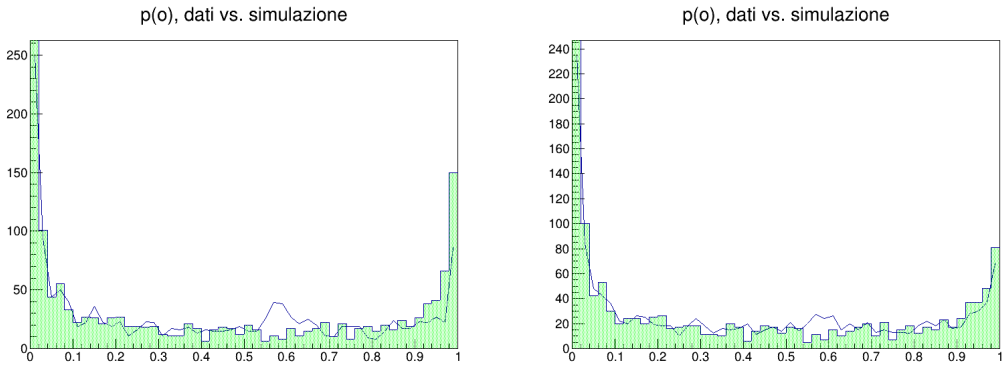


Figura 3.7: Distribuzione delle occorrenze con e senza le famiglie del gruppo *Informazione*: il picco è meno notevole dopo la rimozione.

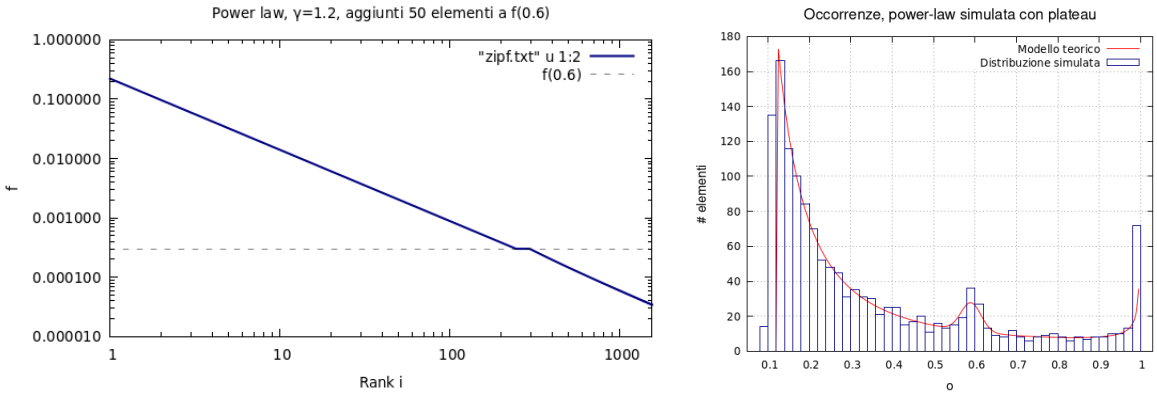


Figura 3.8: Distribuzione teorica con plateau e relativa distribuzione delle occorrenze: il plateau alla frequenza $\tilde{f}$ si manifesta sotto forma di un picco gaussiano centrato a $o(\tilde{f})$.

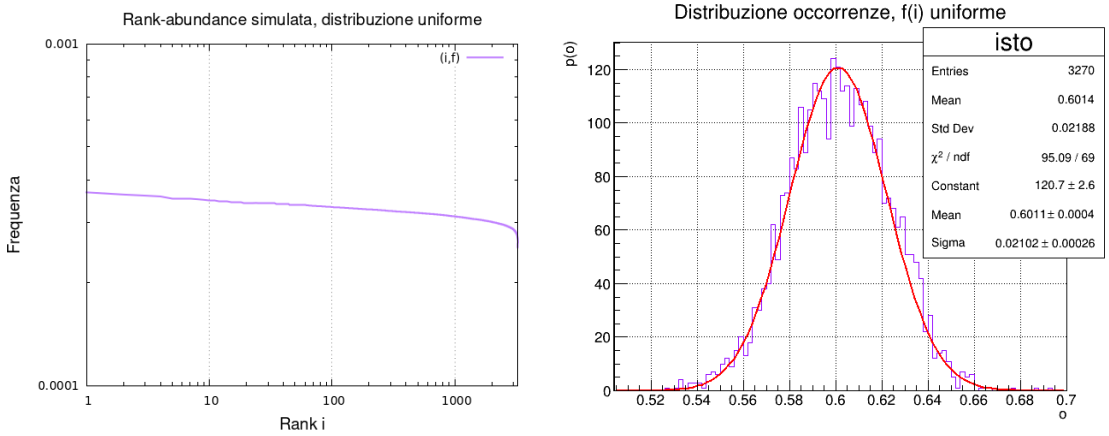


Figura 3.9: Distribuzione delle occorrenze partendo da una RAD costante (la leggera decrescita è dovuta a un effetto di campionamento). L'istogramma delle occorrenze è ben descritto da una gaussiana centrata sul valore atteso $o(f)$.

Capitolo 4

La spedizione TARA

In questa sezione studiamo un altro sistema reale: un dataset di metagenomica e metatrascrittomica raccolto nel corso della spedizione TARA. Come vedremo nelle prossime pagine, i dati raccolti sono per loro natura molto diversi da quelli del database `SUPERFAMILY`. Ciò rende poco sensato l'utilizzo di un null-model identico a quello usato in quella circostanza; proviamo dunque a proporre un null-model diverso, verificando sotto quali ipotesi esso possa ben descrivere la distribuzione delle occorrenze osservata empiricamente. Nello studio dei proteomi avevamo visto, infatti, che l'andamento complessivo della $p(o)$ empirica (forma a U, andamento a legge di potenza) era spiegabile nell'ottica di un random sampling; ci chiediamo quindi se un modello leggermente modificato possa ancora descrivere l'andamento generale delle occorrenze.

4.1 Obiettivi della spedizione TARA

La spedizione *TARA Oceans* ha avuto luogo tra il 2009 e il 2013: essa è consistita nell'estrazione di campioni di plankton in 210 punti (*stazioni*) dell'oceano. Dal 2009 al 2013, la spedizione ha percorso 140000 km, raccogliendo campioni di plankton a diverse profondità. Il plankton raccolto è stato separato in gruppi a seconda della dimensione, estraendone successivamente campioni di acidi nucleici per generare dati di *metabarcoding*, *metagenomica* e *metatrascrittomica*.
I dati utilizzati nel resto dell'analisi sono relativi alle *diatomee*, organismi unicellulari di dimensioni che vanno da 1 μm a pochi mm. Le diatomee sono presenti in tutto il mondo: la loro presenza è maggiore in acque artiche (più torbide e ricche di nutrienti), ma sono presenti anche in altre zone dell'oceano, dove si possono trovare in stati dormienti in condizioni di particolari stress di nutrienti. Questi organismi svolgono un ruolo chiave nella catena alimentare marina e nell'estrazione di CO_2 dall'atmosfera, in quantità paragonabili a quelle di tutte le foreste pluviali messe assieme [5].

4.2 *Meta-omics* e struttura del dataset

I termini *metagenomica*, *metatrascrittomica* e *metabarcoding*, noti collettivamente sotto il nome di *meta-omics*, indicano lo studio di materiale genetico estratto da un campione che contiene contributi provenienti da molte (o tutte) le specie presenti in un ecosistema. Il termine metabarcoding fa riferimento all'identificazione di diverse specie (*OTU, Operational Taxonomic Units*) da un campione di questo tipo; l'insieme dei genomi e dei trascrittomi presenti nel campione costituiscono rispettivamente il metagenoma e il metatrascrittoma.

I dati utilizzati sono divisi in tre dataset: `metaG`, `metaT` e `metaB`, corrispondenti rispettivamente ai dati meta-genomica, -trascrittomica e -barcoding. Essi riportano, per 77 stazioni della

spedizione TARA, le abbondanze con cui compaiono gli unigenes (nel caso di `metaG` e `metaT`) e le OTU (nei `metaB`). Nell'ottica di un sistema complesso modulare, possiamo dire che in ciascuno di questi tre dataset ogni stazione costituisce una realizzazione i cui elementi sono gli unigenes o le OTUs. Naturalmente, l'abbondanza di un unigene è influenzata dall'abbondanza delle diatomee che lo codificano: se, in una certa stazione, una diatomea compare con abbondanza molto elevata, gli unigenes tipicamente espressi da tale diatomea compariranno con alta abbondanza. Le distribuzioni del `metaG` e `metaT` in ciascuna stazione risentiranno dunque della distribuzione delle diatomee ivi presenti, ovvero dal `metaB`. In questo senso, tale sistema è molto diverso da quello del database `SUPERFAMILY`, in cui ogni realizzazione è costituita dal proteoma di un batterio e in cui la popolarità dei batteri non influenza l'abbondanza dei domini strutturali.

4.3 RAD empiriche nei dati di TARA

Prima di procedere con la formulazione di un null model che tenga in conto la natura del dataset considerato, riportiamo in figura 4.1 le RAD empiriche relative ai tre dataset: `metaB`, `metaG` e `metaT`. I dati di `metaG` e `metaT` seguono un andamento approssimativamente a power law

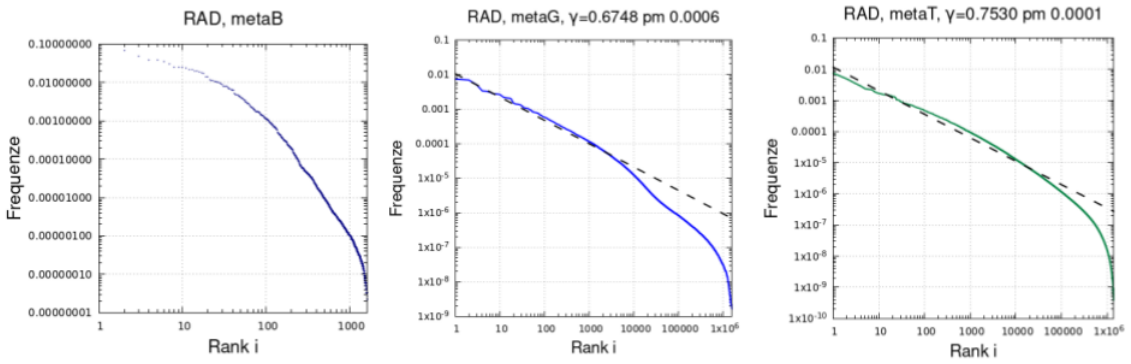


Figura 4.1: RAD empiriche dai dati di TARA. Nei dati di `metaG` e `metaT` si osserva il classico andamento a power-law con modulazione esponenziale.

per i primi quattro ordini di grandezza, con esponenti rispettivamente $\gamma_G = 0.67474 \pm 0.0003$ e $\gamma_T = 0.7272 \pm 0.0006$. La coda esponenziale che compare per elementi rari è frutto della dimensione finita del dataset, similmente a quanto visto anche nel database `SUPERFAMILY`.

Nei dati di `metaB` l'andamento è meno chiaramente a legge di potenza; un eventuale fit potrebbe essere ottenuto ipotizzando una power-law con offset e decadimento esponenziale.

4.4 Analisi delle occorrenze e null-model a estrazione pesata

Il problema del sottocampionamento nei dati di TARA rende particolarmente utile lo studio delle *occorrenze*: infatti l'ipotesi di conoscere, in ogni stazione, quali elementi compaiono e quali non compaiono, è meno stringente rispetto all'ipotesi di conoscerne esattamente le abbondanze.
Avevamo visto, nel caso dei proteomi, che l'andamento complessivo della $p(o)$ empirica (forma a U, a meno del picco gaussiano dovuto al plateau) era spiegato da fenomeni statistici. Dedichiamo quindi le prossime pagine a uno studio delle occorrenze nei dati di `metaT` e alla proposta di un null model che possa a grandi linee rendere conto di questa distribuzione. In particolare, ricaveremo una ipotesi di RAD che possa riproporre l'andamento osservato.

Andamento delle occorrenze

L'istogramma delle occorrenze nei `metaT` presenta un andamento particolare: esso è ben descritto da un esponenziale decrescente (4.2).

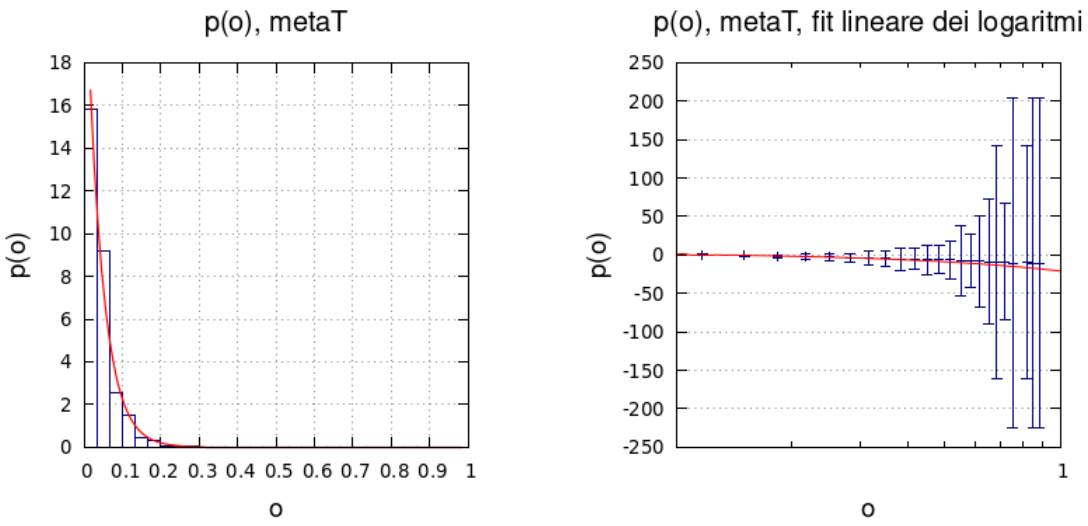


Figura 4.2: Istogramma normalizzato delle occorrenze nei dati di `metaT` e relativo fit esponenziale. Ai conteggi è stata associata un'incertezza poissoniana.

Proviamo a vedere se un modello a estrazione casuale, con gli opportuni accorgimenti, possa riproporre questo andamento. Ricordiamo che ogni realizzazione è una stazione della spedizione TARA e che ogni elemento è un unigene.

Ipotesi del modello

Il modello casuale partirà dalle seguenti ipotesi:

1. I trascrittomi sono ottenuti con un processo di estrazione casuale da una RAD, analogamente a quanto fatto con i proteomi. Questa RAD sarà diversa da quella osservata empiricamente, siccome quest'ultima è frutto del conteggio degli unigenes ed è quindi influenzata dalla diversa popolarità delle varie diatomee.

2. In una stazione, le diatomee della stessa specie compaiono con lo stesso trascrittoma.

3. L'espressione media di una diatomea, ovvero il numero di unigenes che compaiono in un trascrittoma, contati con la loro abbondanza, è la stessa per tutte le diatomee di tutte le stazioni campionate. Indicheremo tale quantità con la lettera a.

4. La presenza o assenza di una specie in una stazione è fornita dai dati di `metaB`.

La seconda ipotesi viene fatta per semplicità e può risultare stringente, in quanto stiamo supponendo che, in una stazione, tutte le diatomee di una certa specie stiano esprimendo gli stessi geni, ovvero stiano "facendo la stessa cosa" (fotosintesi, riproduzione, etc.). Allo stesso modo, anche la terza ipotesi può risultare semplicistica, dal momento che l'attività di una diatomea può variare a seconda di molti fattori, quali le condizioni ambientali o le diverse funzioni che una

diatomea può svolgere. Dunque, un modello più dettagliato potrebbe supporre una variabilità nell'espressione all'interno di una stessa specie.

Consideriamo per semplicità le ipotesi appena introdotte. L'idea è la seguente: avvalersi dei dati di `metaB` per sapere quali diatomee compaiono in una stazione e generarne il trascrittoma con un random sampling. Un'analisi matematica del modello può portare a una stima di quale RAD meglio si addica allo scopo.

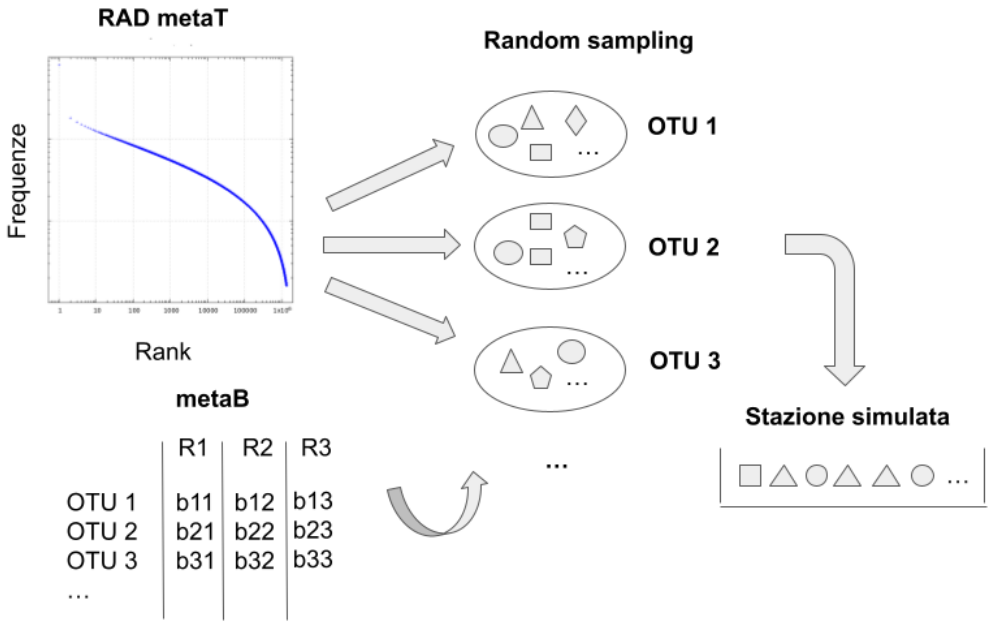


Figura 4.3: Schema riassuntivo del modello utilizzato. I dati di `metaB` vengono utilizzati per sapere quali diatomee sono presenti in una data stazione; i trascrittomi delle singole OTU vengono realizzati da una procedura di random sampling analoga a quella vista finora.

Descrizione matematica del modello e stima della RAD

Sia S_j il numero di specie diverse (*ricchezza*) all'interno della $j-$esima stazione e sia f_i la frequenza dell'$i-$esimo elemento. Ricordiamo che, in questo contesto, le realizzazioni sono le stazioni e gli elementi sono gli unigenes.
Partendo dalle ipotesi appena esposte, avremo che il valore atteso dell'occorrenza è dato da

$$o_i = 1 - \frac{1}{N_{stazioni}} \sum_{j \in stazioni} (1 - f_i)^{aS_j} \tag{4.1}$$

ove ricordiamo che a è l'abbondanza totale media di un trascrittoma.
Osservando l'istogramma delle ricchezze empiriche si nota che l'andamento è ben descritto da un'esponenziale decrescente (4.4).

Abbiamo quindi un'espressione analitica per la distribuzione delle ricchezze:

$$\begin{aligned} p(S) &= \frac{1}{c} e^{-\mu S} \\ c &= \int_{N_{min}}^{N_{max}} e^{-\mu S} dS \end{aligned} \tag{4.2}$$

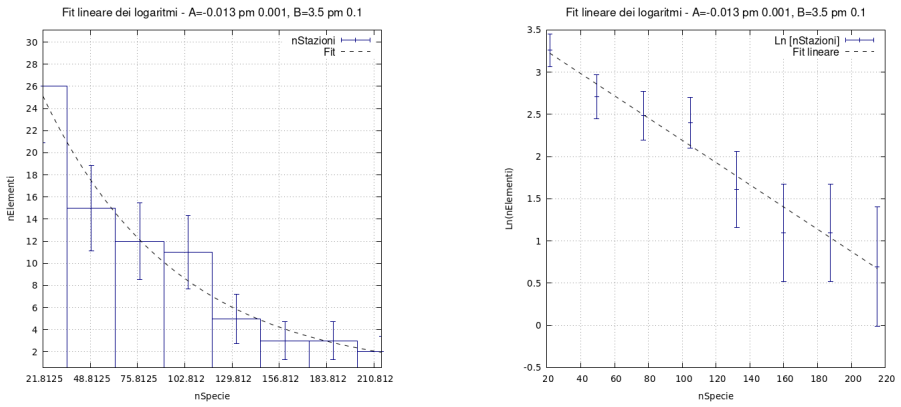


Figura 4.4: Distribuzione della *ricchezza* dei `metaB`: essa è ben descritta da un esponenziale.

con N_{min} e N_{max} minima e massima ricchezza possibile. Per agevolare i calcoli supponiamo $N_{min} = 0$ e $N_{max} = +\infty$, ottenendo $c = \frac{1}{\mu}$.
Approssimando la sommatoria di 4.1 con un integrale (il che è giustificato dal grande numero di unigenes rilevati sperimentalmente) avremo

$$o(f) = 1 - \int_0^{+\infty} (1-f)^{aS} \mu e^{-\mu S} dS$$

che porge

$$o(f) = 1 - \frac{1}{1 - \frac{a}{\mu} \ln[1-f]} \tag{4.3}$$

Invertendo la relazione troviamo un'espressione per $f(o)$:

$$f(o) = 1 - \exp\left[\frac{\mu}{a}(1 - \frac{1}{1-o})\right] \tag{4.4}$$

Ricordiamo che questi passaggi da frequenze a occorrenze attese si basano sulla corrispondenza monotona crescente $f \to o$ del modello casuale e sono di fatto analoghi a dei cambi di variabile.
Ci chiediamo ora che legame esista tra la distribuzione delle occorrenze attese $p(o)$ e la distribuzione delle frequenze $q(f)$. La $q(f)$ è, in altre parole, la SAD: essa descrive l'istogramma normalizzato delle f_i.
La $q(f)$ è collegata alla distribuzione delle occorrenze $p(o)$ dalla seguente relazione:

$$q(f)df = p(o)do$$
$$q(f) = p(o)\Big|_{o=o(f)} \frac{do}{df}$$

che porge

$$q(f) = \frac{a}{\mu} \cdot \frac{1}{1-f} \cdot \frac{1}{\left[1 - \frac{a}{\mu}\ln(1-f)\right]^2} p(o)\Big|_{o(f)} \tag{4.5}$$

Nei `metaT` abbiamo osservato empiricamente un andamento esponenziale delle occorrenze. Ponendo dunque $p(o) = A\exp[-Bo]$, $p : [o_{min}, o_{max}] \to \mathbb{R}^+$, troviamo:

$$q(f) = A\exp\left[\frac{Ba}{\mu}\frac{\ln[1-f]}{1 - \frac{a}{\mu}\ln[1-f]}\right] \cdot \frac{a}{\mu}\frac{1}{1-f}\frac{1}{\left[1 - \frac{a}{\mu}\ln[1-f]\right]^2} \tag{4.6}$$

con supporto $[f(o_{min}), f(o_{max})]$. La $q(f)$ così ottenuta, inserendo per A, B, μ, a i valori del dataset `metaT`, è rappresentata in figura 4.6. La normalizzazione della $p(o)$ implica la normalizzazione della $q(f)$, come è facilmente verificabile integrando la 4.6.

Proviamo ora a ricavare un'espressione per la RAD. Sia N il numero di elementi diversi (in questo caso il numero di unigenes diversi). Partiamo dalla relazione

$$Nq(f)df = -di$$
$$\frac{1}{N}i(f) = -\int q(f)df + c$$

L'integrale di $q(f)$ si svolge applicando la sostituzione $s = 1 - \frac{a}{\mu}\ln[1-f]$.
Otteniamo

$$\begin{aligned} i(f) &= \frac{NA}{B}\exp\left[B\left(\frac{1}{1-\frac{a}{\mu}\ln[1-f]} - 1\right)\right] + Nc \\ \implies f(i) &= 1 - \exp\left[\frac{\mu}{a}\left(1 - \frac{1}{1+\frac{1}{B}\ln[\frac{B}{NA}(i-Nc)]}\right)\right] \end{aligned} \tag{4.7}$$

con c costante di integrazione; il suo valore viene fissato imponendo che $i(f_{max}) \stackrel{!}{=} 1$, ottenendo

$$c = \frac{1}{N} - \frac{A}{B}\exp\left[B\left(\frac{1}{1-\frac{a}{\mu}\ln[1-f_{max}]} - 1\right)\right] \tag{4.8}$$

La *rank-abundance* così ricavata dipende dai parametri μ, A, B, N e a. Mentre il valore dei primi quattro è noto dai dati sperimentali (μ, A e B sono note dai fit esponenziali e N è il vocabolario totale dei `metaT`), l'abbondanza totale a di un trascrittoma è più ambigua: non solo l'abbondanza media per diatomea varia notevolmente di stazione in stazione, ma i dati empirici sono anche affetti da fenomeni di campionamento che non permettono di conoscere l'effettiva abbondanza dei singoli trascrittomi. D'altra parte, anche se ciò fosse possibile, l'espressione media per OTU è estremamente variabile e sensibile alle condizioni ambientali, come mostrato in figura 4.5.

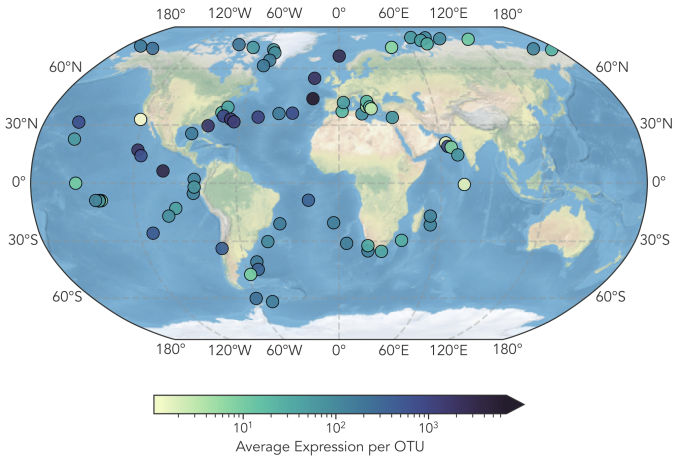


Figura 4.5: Espressione media per OTU nelle stazioni della spedizione TARA. I valori variano su due ordini di grandezza.

Possiamo però stimare il parametro a ricordando la condizione $\sum_i f_i = 1 \implies \int_{f_{min}}^{f_{max}} f q(f)df = 1/N$. Risolvendo con metodi numerici rispetto ad a e imponendo la validità della condizione

si giunge alla stima $a \approx 1073$. Come ordine di grandezza tale valore è in linea con quanto atteso: precedenti studi di dataset di trascrittomica (applicati però a specie diverse dal plankton) avevano fornito per l'attività media stime dell'ordine di ≈ 1200 elementi per cellula [7]. Possiamo

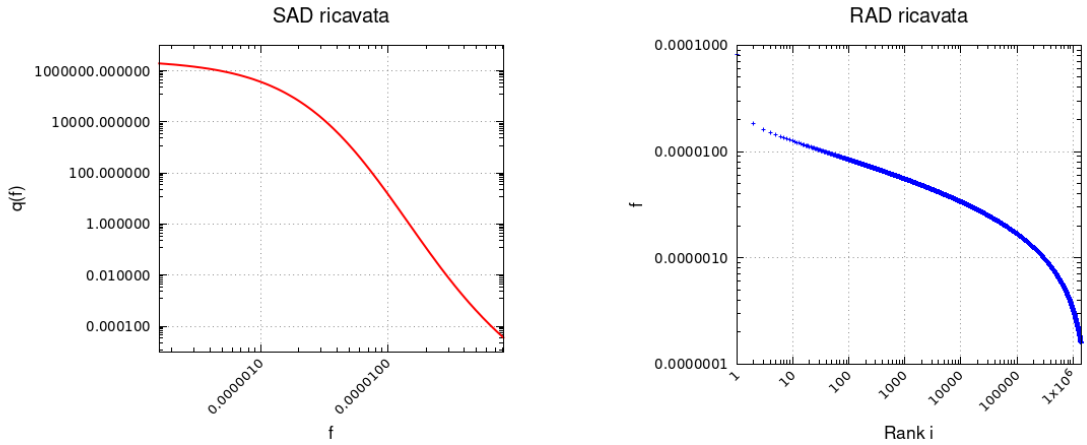


Figura 4.6: *Sinistra:* SAD (sinistra) e RAD (destra) ricavate dall'istogramma delle occorrenze dei `metaT`.

verificare la correttezza dei calcoli generando delle "stazioni simulate" con la procedura descritta e raccogliendone l'istogramma delle occorrenze: i risultati sono riportati in figura 4.7. Il buon

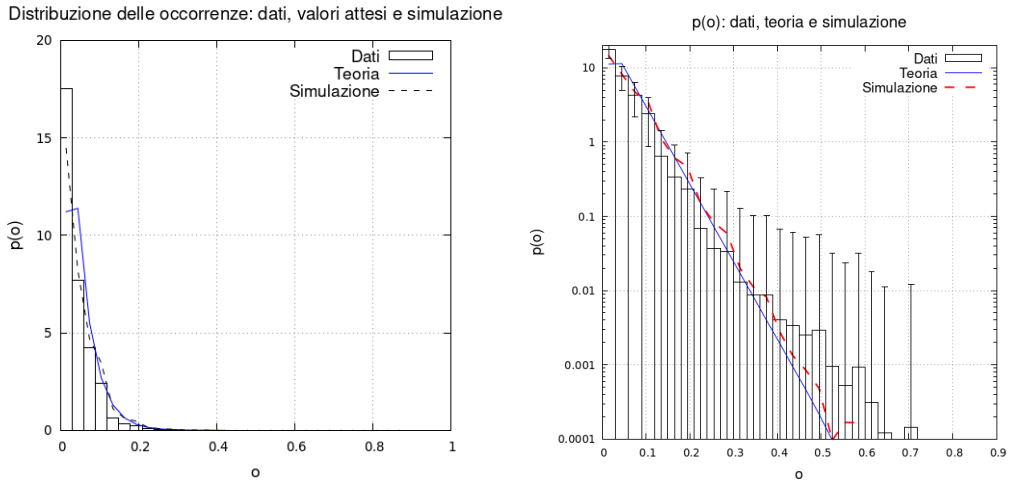


Figura 4.7: *Sinistra:* confronto tra l'istogramma delle occorrenze empiriche, la predizione teorica e i risultati della simulazione. *Destra*: idem, scala logaritmica.

accordo tra dati, simulazioni e teoria implica solo la consistenza dei calcoli effettuati e non è necessariamente prova di un significato fisico del modello utilizzato. Lavori futuri potrebbero testare la validità del modello confrontandolo con aspetti statistici diversi dalla distribuzione delle occorrenze.

Capitolo 5

Conclusioni

In questo lavoro di tesi abbiamo innanzitutto presentato alcuni aspetti matematici dello studio dei sistemi modulari, soffermandoci in particolar modo sul legame tra frequenze e occorrenze; sono state derivate espressioni analitiche per la distribuzione delle occorrenze definita da una RAD a legge di potenza, a esponenziale decrescente e, infine, a legge di potenza con modulazione esponenziale.

Tali concetti sono stati messi in pratica studiando il database `SUPERFAMILY`, che riporta la composizione in domini strutturali di svariati proteomi procarioti. L'analisi ha messo in evidenza alcuni aspetti già noti dalla letteratura, come la presenza di domini che compaiono una volta sola in ogni proteoma, discostandosi quindi dallla statistica di un'estrazione casuale. Altri aspetti, come la forma a U dell'istogramma delle occorrenze, sono riproposti dal null model e sono dunque riconducibili a fenomeni statistici [4].

Ci si è infine dedicati all'analisi di alcuni dataset della spedizione `TARA`, i quali consistono in dati di metagenomica, metatrascrittomica e metabarcoding relativi a campioni di plankton estratti dalle acque oceaniche. Questi dati presentano aspetti molto diversi rispetto al dataset precedentemente studiato; per questo motivo, si è apportata una leggera modifica al modello generativo finora utilizzato e si è ricavata una RAD da cui esso potesse riproporre la statistica delle occorrenze empiricamente osservata. Studi futuri potrebbero provare ad applicare un modello simile allo studio di aspetti statistici diversi dalla distribuzione delle occorrenze.

Appendice

Dimensione del vocabolario e legge di Heaps

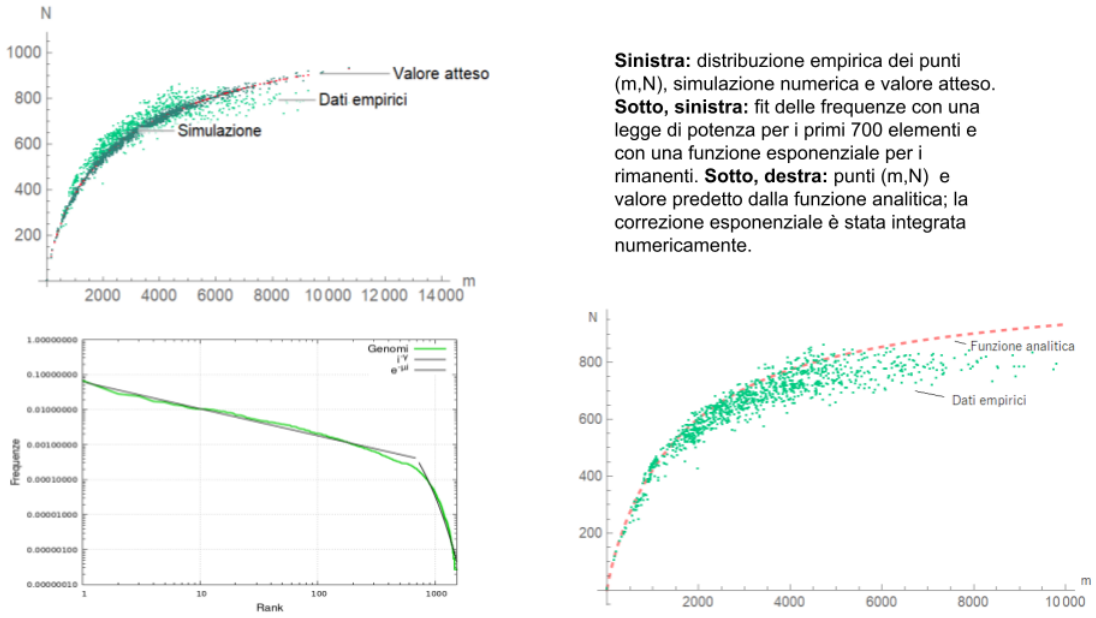


Figura 5.1: Andamento del vocabolario campionato in funzione dell'abbondanza totale.

Riportiamo in figura 5.1 le coppie (m, N_{obs}) del dataset, ove m è la dimensione del genoma e N_{obs} è il vocabolario misurato in quella specifica realizzazione. Osserviamo che sia la simulazione che i dati empirici sono ben descritti dal modello teorico (la curva *Valore atteso*); la differenza principale tra simulazione e dataset è data dalla diversa distribuzione dei punti (m, N_{obs}) rispetto al valore atteso: la simulazione è infatti distribuita in modo più compatto attorno alla curva teorica.
Visto l'andamento non interamente a power-law, la dimensione del vocabolario effettivo in funzione della dimensione del genoma non potrà essere descritta dalla formula analitica 2.16 poiché essa non tiene conto della coda esponenziale. Possiamo giungere a un'espressione *ad hoc* tenendo conto dei due diversi regimi (power-law ed esponenziale) e avvalendoci del fit che ben approssima la distribuzione delle frequenze. Tali fit è riportato in figura 5.1: esso tiene conto di un decadimento a legge di potenza e di una coda esponenziale.

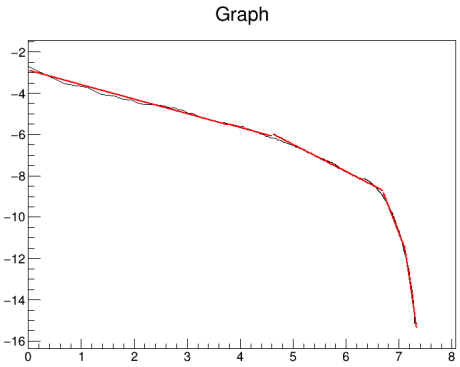


Figura 5.2: Un fit più accurato potrebbe tenere conto di due diversi regimi a legge di potenza; riportiamo un esempio con un regime a power-law per i rank da 1 a 100, una seconda power-law da 101 a 700 e una coda esponenziale da 701 a 1530.

Abbiamo, ponendo $N_1 = 700$, $N_2 = 1530$:

$$\begin{aligned}\langle \tilde{N}(m) \rangle &= N - \sum_{i=1}^{N_1}(1 - Ai^{-\mu})^m - \sum_{i=N_1}^{N_2}(1 - Be^{-\mu i}) = N - \sum_{i=1}^{N} e^{m \ln(1-\frac{i^{-\mu}}{\alpha})} - \sum_{i=N_1}^{N2} e^{-mBe^{-\mu i}} \\ &= N - \frac{(mA)^{\frac{1}{\gamma}}}{\gamma}\Gamma(\frac{1}{\gamma}, \frac{mA}{N_1^{\gamma}}) - \int_{N_1}^{N_2} e^{-mBe^{-\mu i}} di \\ &= N - \frac{(mA)^{\frac{1}{\gamma}}}{\gamma}\Gamma(\frac{1}{\gamma}, \frac{mA}{N_1^{\gamma}}) - \frac{1}{\mu}\int_{mBe^{-N_2\mu}}^{mBe^{-N_1\mu}} \frac{e^{-z}}{z} dz\end{aligned}$$

ove l'ultimo passaggio è stato ottenuto applicando la sostituzione $z = mBe^{-\mu i}$. Abbiamo in sostanza apportato una correzione per tenere conto della coda esponenziale: la correzione può essere valutata usando metodi numerici (e può essere riscritta in termini della funzione integrale esponenziale $Ei(x) := -\int_{-x}^{+\infty} \frac{e^{-t}}{t} dt$).
In figura 5.1 riportiamo le coppie (m, N) empiriche, confrontandole con i risultati ottenuti dalla simulazione e con l'espressione numerica appena ricavata. Osserviamo che il modello teorico tende a sovrastimare leggermente l'andamento effettivo dei dati; ciò è dovuto al fatto che il fit effettuato tende a sovrastimare le frequenze per i rank $i \in [200, 800]$, nella regione quindi in cui c'è una transizione tra power law e decadimento esponenziale. Ciononostante il buon accordo fra dati, simulazione e teoria implica che la legge di Heaps nel caso dei genomi può essere spiegata nell'ottica di un *random sampling* e non è dunque influenzata in modo apprezzabile da vincoli strutturali. [4]

Bibliografia

[1]G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology* (Ravenio Books, 2016).

[2]M. Newman, "Power laws, Pareto distributions and Zipf's law", Contemporary Physics **46**, 323–351 (2005).

[3]A.-L. Barabási e R. Albert, "Emergence of scaling in random networks", science **286**, 509–512 (1999).

[4]A. Mazzolini, M. Gherardi, M. Caselle, M. Cosentino Lagomarsino e M. Osella, "Statistics of Shared Components in Complex Component Systems", Phys. Rev. X **8**, 021023 (2018).

[5]S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, F. Lombard et al., "Tara Oceans: towards global ocean ecosystems biology", Nature Reviews Microbiology **18**, 428–445 (2020).

[6]D. van Leijenhorst e T. van der Weide, "A formal derivation of Heaps' Law", Information Sciences **170**, 263–272 (2005).

[7]S. Lazzardi, F. Valle, A. Mazzolini, A. Scialdone, M. Caselle e M. Osella, "Emergent Statistical Laws in Single-Cell Transcriptomic Data", bioRxiv, `10.1101/2021.06.16.448706` (2022).

[8]P. Selzer, R. Marhöfer e A. Rohwer, *Applied Bioinformatics: An Introduction* (Springer Berlin Heidelberg, 2008).

[9]*SUPERFAMILY database*, `https://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/index.html`.

[10]*SCOP annotation*, `https://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/function.html`.