

# UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia “Galileo Galilei”

Corso di Laurea in Fisica

Tesi di Laurea

Un modello per la dinamica evolutiva nell’ambito della  
meccanica statistica di non-equilibrio

Relatore

Prof. Sandro Azaele

Laureanda

Alice Maschera

Anno Accademico 2022/2023

# Indice

<b>I</b>	<b>Introduzione</b>	<b>1</b>
<b>II</b>	<b>Il modello di dinamica RVS</b>	<b>3</b>
1	Definizioni	3
2	Master equation	4
3	Forze evolutive	5
4	Condizioni per dinamica conservativa	6
5	$\Phi_n$ e robustezza	6
6	Regime di piccole probabilità di variazione	8
7	Relazione di fluttuazione	10
8	Modello a due tipi	11
<b>III</b>	<b>Algoritmi evolutivi</b>	<b>14</b>
9	Modello a griglia	15
10	Distribuzioni di abbondanza delle specie	17
11	Modello asimmetrico e survival of the flattest	24
<b>IV</b>	<b>Conclusioni</b>	<b>26</b>
<b>A</b>	<b>Appendice</b>	<b>27</b>
A.1	Processo di Markov . . . . .	27
A.2	Equazione di Chapman-Kolmogorov . . . . .	27
A.3	Determinazione dei contributi alla direzionalità evolutiva . . . . .	27
A.4	Calcoli per il regime di piccole probabilità di variazione . . . . .	28
A.5	Il modello a quasispecie . . . . .	29

## Parte I

# Introduzione

Lo studio dell'**evoluzione di una popolazione** di organismi attraverso il meccanismo di selezione naturale è un problema complesso al quale contribuiscono numerosi fattori. In natura l'evoluzione di una popolazione, che si può definire come il cambiamento dei genotipi osservati nella popolazione nel corso del tempo, è un processo che avviene su **scale temporali** molto lunghe, da decine di migliaia a milioni d'anni, ed è perciò impossibile da osservare direttamente. Per questo i **modelli matematici evolutivi** e le **simulazioni numeriche** sono strumenti fondamentali per indagare in che modo sia regolata l'evoluzione delle specie e quali siano i principali attori coinvolti nel fenomeno.

Ogni modello teorico che provi a descrivere l'evoluzione di una popolazione dovrà necessariamente semplificare il processo: la **genetica delle popolazioni**, ad esempio, si concentra principalmente sulle variazioni nelle frequenze degli alleli presenti negli organismi, mentre il **modello a quasispecie**, introdotto da Manfred Eigen nel 1971, dà una descrizione deterministica dell'evoluzione che trascura gli effetti stocastici. Inoltre è importante valutare gli effetti delle **approssimazioni** che si introducono nei modelli: cosa accade se ipotizziamo che le generazioni di organismi non si sovrappongano tra loro? Quanto siamo giustificati ad estendere le conclusioni ottenute considerando l'ambiente costante, a situazioni reali in cui i fattori ambientali variano continuamente?

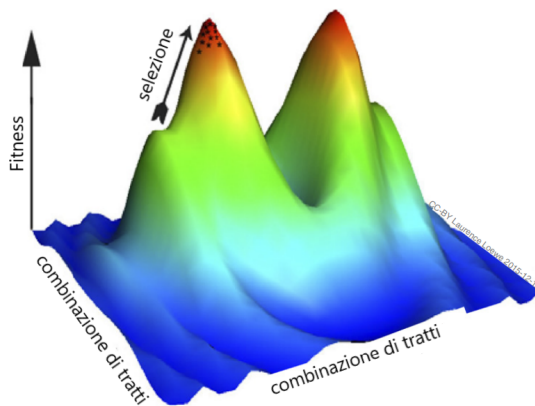


Figura 1: Esempio di fitness landscape [5]

abiotici (la presenza di altre specie o la composizione inorganica di un certo habitat), i cui cambiamenti interessano scale spaziotemporali molto diverse tra loro. Matematicamente un determinato tipo di organismi è favorito o meno dalla selezione in base al suo valore di **fitness**, che ne definisce il successo riproduttivo.

La combinazione dei tre fattori appena citati dà origine alle cosiddette **"forze evolutive"** che regolano la dinamica evolutiva dei sistemi. Tra queste annoveriamo la forza dovuta alla selezione, la forza entropica, la forza che tiene conto degli effetti di variazione e deriva genetica, e la forza associata agli effetti combinati di selezione e variazioni. Ogni forza contribuisce a direzionare l'evoluzione della popolazione verso determinate configurazioni.

In questa tesi ci concentreremo su un modello dinamico che prevede tre fasi: riproduzione, variazione e selezione, e che prende il nome di **modello RVS**. L'obiettivo di questo modello è ricercare il maggior grado di generalità e semplicità possibile e riuscire a determinare in che modo agiscono le forze

evolutive. Per farlo considera la dinamica evolutiva come un **processo termodinamico stocastico**, trovando analogie tra i due ambiti.

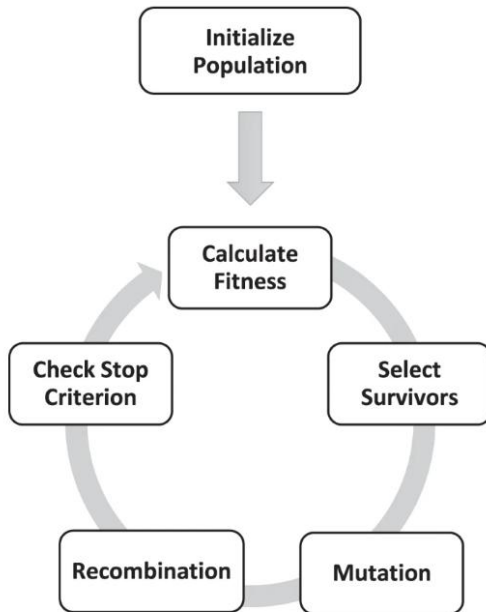


Figura 2: Esempio di algoritmo evolutivo per ottimizzazione. (Lerch, F., Ultsch, A. Lötsch, J. *Distribution Optimization: An evolutionary algorithm to separate Gaussian mixtures*. Sci Rep 10, 648 (2020))

regimi di selezione più o meno forti e diversi tassi di mutazione.

Infine ci concentreremo sullo studio della **distribuzione dell'abbondanza delle specie (SAD)** nelle popolazioni, evidenziando similarità e differenze tra gli andamenti di questi grafici in popolazioni sottoposte a diversi regimi, e verificheremo la comparsa di un fenomeno caratteristico chiamato **"survival of the flattest"**.

Per quanto riguarda gli **algoritmi evolutivi** e la loro applicazione in simulazioni numeriche, storicamente la loro nascita risale agli anni Cinquanta del '900 con Nils Barricelli che scrisse il suo primo lavoro sull'argomento *Esempi numerici di processi di evoluzione* [Barricelli, 1954], e Alexander Fraser seguì poco dopo con l'articolo *Simulation of genetic systems by automatic digital computers* [Fraser, 1957]. Negli anni '60 molti altri biologi, matematici e fisici seguirono i loro passi e usarono i computer per simulare l'evoluzione dei sistemi biologici. Da queste prime esperienze si originò un ampio campo di ricerca incentrato sugli **algoritmi genetici di ottimizzazione**: l'obiettivo era, ed è tutt'oggi, studiare dei problemi reali di ottimizzazione simulando gli effetti della selezione naturale.

In questo elaborato vedremo delle simulazioni numeriche che emulano l'evoluzione di una popolazione e noteremo come questo processo non possa essere ricondotto esclusivamente ad un meccanismo di ottimizzazione, ma risenta molto degli effetti delle mutazioni e delle fluttuazioni nel processo di selezione. In particolare useremo un **modello a griglia 2D** e studieremo popolazioni caratterizzate da

## Parte II

# Il modello di dinamica RVS

Il modello che stiamo per analizzare incorpora i tre elementi essenziali per descrivere la dinamica evolutiva (riproduzione, variazione, selezione) ed è per questo che prende il nome di modello RVS; come conseguenza la dinamica evolutiva da esso descritta viene chiamata dinamica RVS. Il modello ricerca semplicità e generalità, perciò non specifica nè la natura delle variabili ereditarie, nè quella delle variazioni: possono essere genetiche, epigenetiche o fenotipiche. Inoltre assume che le popolazioni siano di grandi dimensioni e che l'ambiente rimanga costante.

Per studiare la dinamica nella sua generalità vengono utilizzati degli strumenti ispirati alla meccanica statistica di non equilibrio. In questo modo è possibile definire la nozione di forze evolutive e spiegare la loro relazione con riproduzione, variazioni e selezione. Grazie alla semplicità del modello siamo in grado di ottenere espressioni analitiche per i termini legati alle forze e analizzare esplicitamente i loro ruoli.

In particolare questo modello ci permetterà di interpretare la *robustezza* alle variazioni di alcuni tipi come una conseguenza diretta della dinamica evolutiva, per la presenza di un termine che "genera robustezza" nella forza evolutiva. Questa forza esprime la tendenza delle varianti selezionate ad essere circondate (nello spazio dei tipi) da altre varianti con valori di fitness simili.

## 1 Definizioni

Definiamo le variabili utilizzate nel modello RVS per la descrizione dell'evoluzione dinamica di una popolazione:

- $\gamma$ : variabili ereditarie, determinano il successo riproduttivo di ciascun organismo; matematicamente sono delle variabili multidimensionali continue o discrete. Un organismo caratterizzato dalle variabili  $\gamma$  sarà detto un tipo- $\gamma$ .
- $f_\gamma$ : tasso di riproduzione (o fitness); legato al numero tipico di discendenti che ciascun tipo- $\gamma$  genera nel corso di una generazione.
- $n = (n_\gamma)$ : vettore che denota la composizione della popolazione; ogni entrata rappresenta il numero di tipi- $\gamma$  presenti nella popolazione.
- $N(n) := \sum_\gamma n_\gamma$  è la dimensione della popolazione totale in funzione di  $n$ .
- $\pi_{\gamma \leftarrow \gamma'}$ : probabilità che un tipo- $\gamma'$  vari in un tipo- $\gamma$  ( $\gamma \leftarrow \gamma'$ ) nel corso di una sola generazione ( $\pi_{\gamma \leftarrow \gamma'} \geq 0$ ,  $\sum_\gamma \pi_{\gamma \leftarrow \gamma'} = 1$ ). Noi considereremo solo generazioni che non si sovrappongono e variazioni potenzialmente reversibili:  $\pi_{\gamma \leftarrow \gamma'} \geq 0 \leftrightarrow \pi_{\gamma' \leftarrow \gamma} \geq 0$ .  $\pi_{\gamma \leftarrow \gamma'}$  è funzione di  $\gamma'$ , è soggetto ad evoluzione e può dipendere da  $n$ .
- $\mu_\gamma := \sum_{\gamma' \neq \gamma} \pi_{\gamma' \leftarrow \gamma} = 1 - \pi_{\gamma \leftarrow \gamma}$  definisce la probabilità complessiva che un organismo di tipo  $\gamma$  vari nel corso di una generazione.
- $W_{n \leftarrow n'}$ : *kernel di transizione*, determina la probabilità che la popolazione transiti da  $n'$  a  $n$  in una generazione. La sua forma precisa dipende dal modello considerato e può cambiare in accordo con i mutamenti dell'ambiente.
- $\mathcal{F}_{n \leftarrow n'} := \ln \frac{W_{n \leftarrow n'}}{W_{n' \leftarrow n}}$  prende il nome di *direzionalità evolutiva* e determina verso quale configurazione è più probabile che il sistema evolva. Se  $W_{n \leftarrow n'} \simeq W_{n' \leftarrow n}$  allora  $\mathcal{F}_{n \leftarrow n'} \simeq 0$  e nessuna direzione è probabilisticamente favorita; viceversa se  $W_{n \leftarrow n'} \gg W_{n' \leftarrow n}$  allora  $\mathcal{F}_{n \leftarrow n'} > 0$  ed è più probabile che la popolazione evolva nella direzione  $n \leftarrow n'$  (un caso analogo si ha per  $W_{n \leftarrow n'} \ll W_{n' \leftarrow n}$ ).

## 2 Master equation

Partiamo da un sistema generico con popolazione piccola e fluttuante. La selezione delle varianti che si riproducono è casuale ma dipende dalla fitness di ogni variante. Inizialmente non trascuriamo le interazioni tra organismi, perciò sia  $\pi_{\gamma \leftarrow \gamma'}$  che  $f_\gamma$  potranno dipendere dalla composizione della popolazione  $n$ . Inoltre consideriamo il caso in cui la dimensione della popolazione totale  $\mathcal{N}(n)$  cambia a causa delle fluttuazioni dovute a riproduzione e selezione, ma l'ambiente mantiene la popolazione media costante, pari a  $N$ .

Matematicamente la dinamica evolutiva su molte generazioni può essere descritta da un *processo di Markov* (A.1), stocastico, con il tempo preso come discreto. L'equazione che descrive il processo è un'equazione di *Chapman-Kolmogorov* (A.2) della forma seguente:

$$p_n(\tau + 1) = \sum_{n'} W_{n \leftarrow n'} p_{n'}(\tau) \quad (1)$$

dove  $p_n(\tau + 1)$  è la probabilità di osservare una certa composizione  $n$  alla generazione  $(\tau + 1)$ -esima, data la probabilità  $p_{n'}(\tau)$  di osservare una composizione  $n'$  alla generazione  $\tau$ -esima. Nel modello RVS il kernel di transizione è espresso come un prodotto di processi di Poisson:

$$W_{n \leftarrow n'} = e^{-N} \prod_{\gamma} \frac{1}{n_{\gamma}!} [\Lambda_{\gamma}(n')]^{n_{\gamma}} \quad (2)$$

dove

$$\Lambda_{\gamma}(n) := \frac{N f_{\gamma}(n) \sum_{\gamma'} \pi_{\gamma \leftarrow \gamma'}(n) n_{\gamma'}}{\sum_{\gamma''} f_{\gamma''}(n) \pi_{\gamma'' \leftarrow \gamma'}(n) n_{\gamma''}} \quad (3)$$

descrive il tasso di crescita medio del tipo  $\gamma$  in una generazione, con  $f_\gamma$  e  $\pi_{\gamma \leftarrow \gamma'}$  che possono essere funzioni della composizione della popolazione  $n$ .

Nell'equazione si possono riconoscere gli eventi stocastici che caratterizzano ogni generazione:  $(\pi n)_{\gamma} := \sum_{\gamma'} \pi_{\gamma'' \leftarrow \gamma'}(n) n_{\gamma''}$  è il valore atteso del numero di organismi di tipo  $\gamma$  a seguito delle variazioni e  $f_\gamma$  tiene conto del successo riproduttivo delle varianti. Nell'equazione (3) il prodotto  $f_\gamma(n) \sum_{\gamma'} \pi_{\gamma \leftarrow \gamma'}(n) n_{\gamma'}$  (valore atteso del numero di individui di tipo  $\gamma$  generati) è riscalato in modo da mantenere la popolazione media costante:

$$\sum_n \mathcal{N}(n) p_n(\tau + 1) = \sum_{nn'} \mathcal{N}(n) W_{n \leftarrow n'} p_{n'}(\tau) = \sum_{\gamma n'} \Lambda_{\gamma}(n') p_{n'}(\tau) = N, \quad \forall \tau \quad (4)$$

dove sono state usate le proprietà delle distribuzioni di Poisson. Analogamente la varianza della dimensione della popolazione rimane costante e scala come  $N$ :

$$\sigma_N^2 = \sum_n (\mathcal{N}(n) - N)^2 p_n(\tau + 1) = \sum_{nn'} (\mathcal{N}(n) - N)^2 W_{n \leftarrow n'} p_{n'}(\tau) = \sum_{\gamma n'} \Lambda_{\gamma}(n') p_{n'}(\tau) = N, \quad \forall \tau \quad (5)$$

Per  $N$  grande,  $N \gg 1$ , si trova che  $\sigma_N/N = 1/\sqrt{N} \ll 1$ ; quindi le fluttuazioni della dimensione della popolazione possono essere trascurate,  $\mathcal{N}(n) \simeq N$ , e il kernel di transizione può essere riscritto come una distribuzione multinomiale:

$$W_{n \leftarrow n'} = N^N e^{-N} \prod_{\gamma} \frac{1}{n_{\gamma}!} \left[ \frac{\Lambda_{\gamma}(n')}{N} \right]^{n_{\gamma}} \simeq N! \prod_{\gamma} \frac{1}{n_{\gamma}!} \left[ \frac{f_{\gamma}(n') \sum_{\gamma'} \pi_{\gamma \leftarrow \gamma'}(n') n'_{\gamma'}}{\sum_{\gamma''} f_{\gamma''}(n') \pi_{\gamma'' \leftarrow \gamma'}(n') n'_{\gamma''}} \right]^{n_{\gamma}} \quad (6)$$

dove, nella seconda uguaglianza, è stata usata l'approssimazione di Stirling.

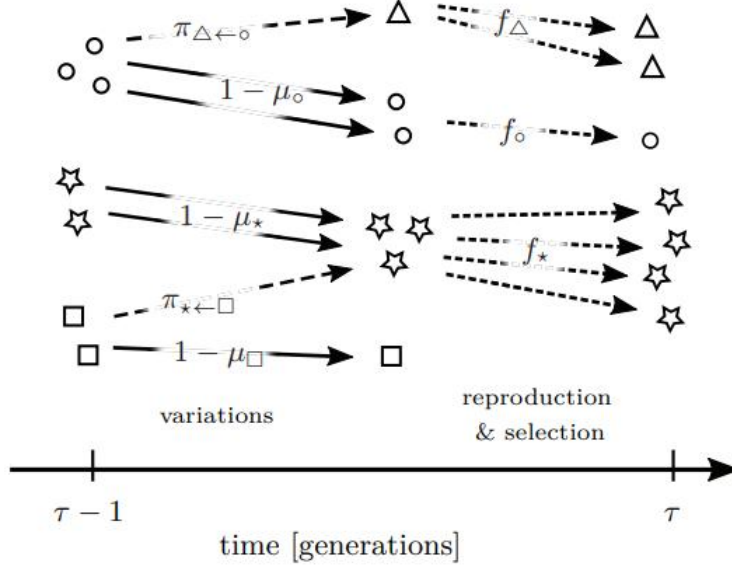


Figura 3: Rappresentazione schematica della dinamica evolutiva come descritta in eq.(6). La produzione di varianti è rappresentata dalle linee tratteggiate più a sinistra, mentre la selezione dei tipi che si riproducono, la cui prole costituirà la generazione successiva, è simboleggiata dalle linee tratteggiate più a destra. La probabilità che un organismo di tipo  $\gamma$  sia scelto è determinata dal tasso di riproduzione  $f_\gamma$ . [1]

### 3 Forze evolutive

La direzionalità evolutiva può essere espressa genericamente come la somma di due contributi, detti forze evolutive:

$$\mathcal{F}_{n \leftarrow n'} := \ln \frac{W_{n \leftarrow n'}}{W_{n' \leftarrow n}} = (\Psi_n - \Psi_{n'}) + \zeta_{nn'} \quad (7)$$

il primo contributo può essere visto come una differenza di potenziale, perciò la forza corrispondente prende il nome di *forza evolutiva conservativa*, mentre il secondo contributo è detto *forza evolutiva non conservativa*. Se  $\zeta_{nn'} > 0$  e  $\Psi_n > \Psi_{n'}$  allora i due termini contribuiscono positivamente a  $\mathcal{F}_{n \leftarrow n'}$  e favoriscono l'evoluzione nella direzione  $n \leftarrow n'$ .

A partire dalla forma per il kernel di transizione trovata in (6), seguiamo un procedimento puramente algebrico (A.3) per separare i contributi che possono essere scritti come differenze di potenziale da quelli che non possono; otteniamo:

$$\Phi_n = N \ln \sum_{\gamma'} f_\gamma(n) \pi_{\gamma \leftarrow \gamma'}(n) n_{\gamma'} \quad (8)$$

$$\Psi_n = \Phi_n - \sum_{\gamma} \ln n_\gamma! \quad (9)$$

$$\zeta_{nn'} = \ln \prod_{\gamma} \frac{[\sum_{\gamma'} f_\gamma(n') \pi_{\gamma \leftarrow \gamma'}(n') (n'_{\gamma'})]^{n_\gamma}}{[\sum_{\gamma'} f_\gamma(n) \pi_{\gamma \leftarrow \gamma'}(n) (n_{\gamma'})]^{n'_\gamma}} \quad (10)$$

Quando trascuriamo la possibilità che le interazioni abbiano effetto sui tassi di variazione e sulla riproduzione ( $\pi_{\gamma \leftarrow \gamma'}(n) = \pi_{\gamma \leftarrow \gamma'}$ ,  $f_\gamma(n) = f_\gamma$ ) per tutti gli  $n$ , allora il contributo della selezione ( $f_\gamma$ ) alla forza non conservativa  $\zeta_{nn'}$  diventa conservativo e può essere incluso nel termine  $\Psi_n$ :

$$\Psi_n = \sum_{\gamma} n_\gamma \ln f_\gamma - \sum_{\gamma} \ln n_\gamma! + \Phi_n \quad (11)$$

$$\Phi_n = N \ln \sum_{\gamma\gamma'} f_\gamma \pi_{\gamma \leftarrow \gamma'} n_{\gamma'} \quad (12)$$

$$\zeta_{nn'} = \ln \prod_{\gamma} \frac{[\sum_{\gamma'} \pi_{\gamma \leftarrow \gamma'} n_{\gamma'}]^{n_\gamma}}{[\sum_{\gamma'} \pi_{\gamma \leftarrow \gamma'} n_{\gamma'}]^{n'_\gamma}} \quad (13)$$

$\zeta_{nn'}$  tiene conto degli effetti di variazione e di deriva genetica; ciò è evidenziato dalla presenza delle probabilità di variazione  $\pi_{\gamma \leftarrow \gamma'}$  nell'equazione (13).

$\Psi_n$  è il potenziale associato alla forza evolutiva conservativa ed è composto da tre termini.

Il primo contributo è legato alla *selezione*, e la corrispondente forza evolutiva  $\sum_{\gamma} (n_\gamma - n_{\gamma'}) \ln f_\gamma$  porta la popolazione verso configurazioni con un più alto tasso di riproduzione logaritmico.

Il secondo è il *contributo entropico* che deriva dall'aver assunto come indistinguibili gli organismi dello stesso tipo. La forza corrispondente  $\sum_{\gamma} \ln n'_\gamma! - \sum_{\gamma} \ln n_\gamma!$  favorisce popolazioni con composizione eterogenea.

Il terzo termine è detto *potenziale di crescita atteso* per la popolazione con composizione  $n$ ,  $\Phi_n$ ; è associato agli effetti combinati di selezione e variazioni ed è più grande per popolazioni con alta probabilità di generare varianti con un'alta fitness. Questo termine genera robustezza.

## 4 Condizioni per dinamica conservativa

Le condizioni per una dinamica conservativa si hanno quando il contributo non conservativo può essere visto come un termine aggiuntivo di  $\Psi_n$ , quindi quando  $\zeta_{nn'}$  può essere scritto come una differenza di potenziale. Perchè ciò sia possibile il termine in parentesi quadre nell'equazione (13) non deve dipendere dalla composizione della popolazione, perciò le probabilità di variazione devono dipendere solo dal "tipo-target" ( $\pi_{\gamma \leftarrow \gamma'} = \pi_\gamma \forall \gamma, \gamma'$ ). Otteniamo:

$$\ln \prod_{\gamma} \frac{(N\pi_\gamma)^{n_\gamma}}{(N\pi_\gamma)^{n'_\gamma}} = \ln \frac{N^N}{N^N} + \ln \prod_{\gamma} (\pi_\gamma)^{n_\gamma} - \ln \prod_{\gamma} (\pi_\gamma)^{n'_\gamma}$$

$$\Psi_n = \sum_{\gamma} n_\gamma \ln f_\gamma + \sum_{\gamma} n_\gamma \ln \pi_\gamma - \sum_{\gamma} \ln n_\gamma! + N \ln N \sum_{\gamma} f_\gamma \pi_\gamma \quad (14)$$

dove è stato usato che  $\sum_{\gamma'} \pi_{\gamma \leftarrow \gamma'} = N\pi_\gamma$ . In questa forma il termine  $\ln \pi_\gamma$  rappresenta quanto un certo tipo è favorito dalle variazioni.

## 5 $\Phi_n$ e robustezza

La robustezza misura il grado di insensibilità di un organismo, o di una sua caratteristica fenotipica, alle variazioni intrinseche (es. mutazioni) ed estrinseche (es. cambiamenti dell'ambiente). Ad esempio la *robustezza mutazionale* misura quanto una caratteristica fenotipica cambia a seguito di una mutazione genetica.

Definiamo la sensibilità (inverso della robustezza) dei tassi di riproduzione alle variazioni:

$$\omega_\gamma := \sum_{\gamma'} \left( \frac{f_\gamma - f_{\gamma'}}{f_\gamma} \right) \frac{\pi_{\gamma' \leftarrow \gamma}}{\mu_\gamma} < 1 \quad (15)$$

dove  $\mu_\gamma = \sum_{\gamma' \neq \gamma} \pi_{\gamma' \leftarrow \gamma}$  è la probabilità di variazione totale per un individuo di tipo  $\gamma$ , ed è stata omessa la dipendenza da  $\pi$  e  $n$  per semplicità di notazione.



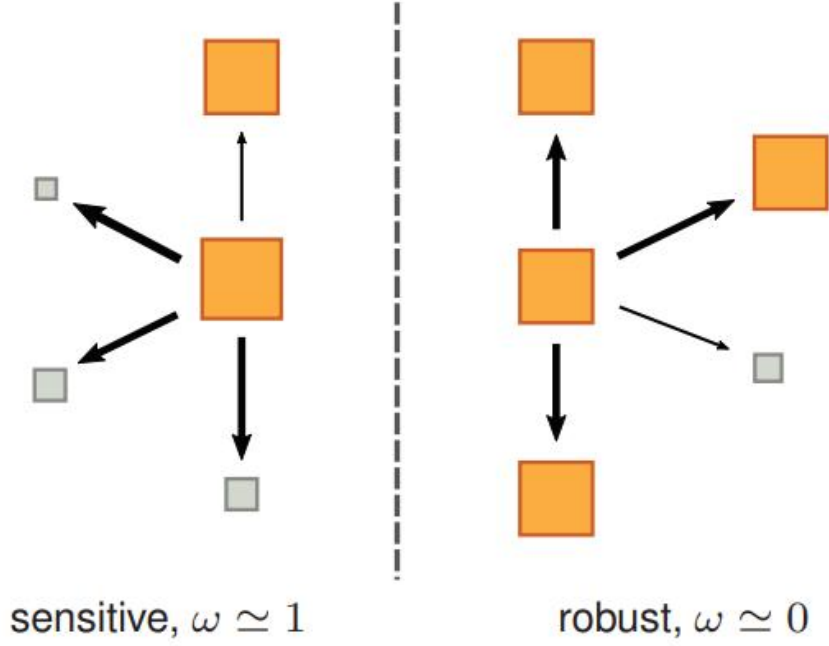


Figura 4: Ogni quadrato rappresenta un diverso tipo: i tipi che si riproducono più velocemente sono rappresentati da quadrati grandi arancioni, mentre quelli con basso tasso di riproduzione sono rappresentati da quadrati piccoli grigi. Le frecce identificano le possibili variazioni e il loro spessore riflette la probabilità di variazione. (Sinistra) Un tipo con alta sensibilità è probabile vari in un tipo con più basso tasso di riproduzione. (Destra) Un tipo robusto varia con alta probabilità in tipi con tassi di riproduzione ugualmente alti. [1]

Grandi valori di  $\omega_\gamma$  ( $\omega_\gamma \lesssim 1$ ) implicano che la decrescita relativa del tasso di riproduzione  $(f_\gamma - f_{\gamma'})/f_\gamma$  mediata su tutte le possibili variazioni  $\pi_{\gamma' \leftarrow \gamma}/\mu_\gamma$  è grande. Quindi tipi con  $\omega_\gamma$  grande hanno grande sensibilità (equivalentemente bassa robustezza) alle variazioni; viceversa tipi con  $\omega_\gamma \simeq 0$  hanno tassi di riproduzione insensibili alle variazioni, e sono quindi robusti.

Il potenziale di crescita atteso può ora essere espresso come funzione decrescente di  $\omega_\gamma$  (sfruttando il fatto che  $\sum_{\gamma'} f_{\gamma'} \pi_{\gamma' \leftarrow \gamma} = f_\gamma (1 - \mu_\gamma \omega_\gamma)$ ):

$$\Phi_n = N \ln \left( \sum_{\gamma} f_\gamma(n) n_\gamma - \sum_{\gamma} \mu_\gamma(n) \omega_\gamma(n) f_\gamma(n) n_\gamma \right) \quad (16)$$

Come esempio consideriamo il caso in cui la composizione della popolazione  $\bar{n}$  è tale che il tasso di riproduzione dei tipi in  $\bar{n}$  è alto rispetto alle loro varianti ed è approssimativamente uguale per tutti:  $f_\gamma(\bar{n}) \simeq f_{high}$  per ogni  $\gamma$  tale che  $\bar{n}_\gamma > 0$ ; cioè stiamo considerando un caso in cui solo i tipi che si riproducono più velocemente possono sopravvivere.

$$\Phi_{\bar{n}} = N \ln f_{high} + N \ln \left( N - \sum_{\gamma} \mu_\gamma \omega_\gamma \bar{n}_\gamma \right) \simeq N \ln(f_{high} N) - \sum_{\gamma} \mu_\gamma \omega_\gamma \bar{n}_\gamma = - \sum_{\gamma} \mu_\gamma \omega_\gamma \bar{n}_\gamma + cost \quad (17)$$

dove è stato usato lo sviluppo in serie di Taylor del logaritmo (approssimazione giustificata dal fatto che il prodotto  $\mu_\gamma \omega_\gamma$  è assunto piccolo).

La forza che si origina da  $\Phi_{\bar{n}} - \Phi_{\bar{n}'}$  favorisce popolazioni con organismi che esibiscono o bassa probabilità di variazione ( $\mu_\gamma$ ) o alta robustezza (bassa  $\omega_\gamma$ ).

## 6 Regime di piccole probabilità di variazione

Studiamo ora cosa accade nel limite di probabilità di variazione piccole e identiche,  $\mu_\gamma = \mu \ll 1$ , e assumiamo che non ci siano interazioni tra gli organismi della popolazione. In questo regime il kernel di transizione può essere approssimato come (per i calcoli si veda A.4):

$$W_{n \leftarrow n'} \simeq \mu^{\sum_{\gamma: n'_\gamma=0} n_\gamma} N! \left\{ \prod_{\gamma: n'_\gamma \neq 0} \frac{1}{n_\gamma!} \left[ \frac{f_\gamma n'_\gamma}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}} \right]^{n_\gamma} \right\} \left\{ \prod_{\gamma: n'_\gamma=0} \frac{1}{n_\gamma!} \left[ \frac{f_\gamma \sum_{\gamma' (\neq \gamma)} \bar{\pi}_{\gamma \leftarrow \gamma'} n'_{\gamma'}}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}} \right]^{n_\gamma} \right\} \quad (18)$$

dove abbiamo usato le probabilità condizionate di transizione, definite per  $\gamma \neq \gamma'$  come:

$$\bar{\pi}_{\gamma \leftarrow \gamma'} = \frac{\pi_{\gamma \leftarrow \gamma'}}{\mu} \quad (19)$$

Ne consegue che il kernel di transizione scala come una potenza di  $\mu$ , il cui esponente è il numero di organismi di nuovi tipi che compaiono nella popolazione. Il primo termine tra parentesi graffe è riferito alla selezione di tipi che esistono già nella popolazione  $n'$ , e seleziona quelli che crescono più velocemente (alto  $f_\gamma$ ) o sono più abbondanti (alto  $n'_\gamma$ ). Il secondo termine tra parentesi graffe seleziona i tipi che compaiono a seguito delle variazioni e dipende fortemente dalle probabilità di variazione.

Per  $\pi_{\gamma \leftarrow \gamma'}$  sufficientemente piccole,  $\mu \ll 1/N$ , è improbabile che più di una variazione abbia luogo in una generazione, perciò le transizioni tipiche  $n \leftarrow n'$  coinvolgono nessuna variazione o solo una.

Come caso esemplificativo possiamo considerare una popolazione composta da un solo tipo  $\gamma$  che ha una transizione verso una composizione in cui compare un nuovo tipo  $\bar{\gamma}$ ; il kernel di transizione risulta allora:

$$W_{n \leftarrow n'} = \frac{\mu N f_{\bar{\gamma}} \bar{\pi}_{\bar{\gamma} \leftarrow \gamma}}{f_\gamma}$$

Inoltre, nel limite di  $N$  grande possiamo esprimere anche la probabilità della transizione inversa in termini semplici:

$$W_{n' \leftarrow n} = \left[ \frac{(N-1)f_\gamma}{f_{\bar{\gamma}} + (N-1)f_\gamma} \right]^N \simeq \exp\left(-\frac{f_{\bar{\gamma}}}{f_\gamma}\right)$$

dove si evidenzia la dipendenza esponenziale di  $W_{n' \leftarrow n}$  dal rapporto dei tassi di riproduzione.

Ora studiamo la forma dei contributi alla direzionalità evolutiva nel regime di piccole probabilità di variazione. Espandendo  $\Phi_n$  in serie di Taylor in termini di  $\mu$  otteniamo:

$$\Phi_n \simeq N \ln \sum_{\gamma} f_\gamma n_\gamma - \mu N \frac{\sum_{\gamma} f_\gamma n_\gamma - \sum_{\gamma, \gamma' (\neq \gamma)} f_{\gamma'} \pi_{\gamma' \leftarrow \gamma} n_\gamma / \mu}{\sum_{\gamma} f_\gamma n_\gamma} \equiv N \ln F_n - \mu N \frac{F_n - F_{\pi n}}{F_n} \quad (20)$$

dove  $F_n := \sum_{\gamma} f_\gamma n_\gamma$  è la *crescita cumulativa* della popolazione e  $F_{\pi n} := \sum_{\gamma, \gamma' (\neq \gamma)} f_{\gamma'} (\pi_{\gamma' \leftarrow \gamma} / \mu) n_\gamma$  è la *crescita cumulativa a seguito delle variazioni*.

Il potenziale associato alla forza conservativa risulta dunque:

$$\Psi_n \simeq \sum_{\gamma} n_\gamma \ln f_\gamma - \sum_{\gamma} \ln n_\gamma! + N \ln \sum_{\gamma} n_\gamma f_\gamma - \mu N \frac{F_n - F_{\pi n}}{F_n} \quad (21)$$

I termini di ordine 0 in  $\mu$  tengono conto della selezione e dei contributi entropici; il contributo di ordine 1 favorisce popolazioni robuste (come vedremo nella sezione 7).

Approssimando i termini logaritmici del contributo non conservativo, possiamo riscrivere la forza non conservativa come:

$$\zeta_{nn'} \simeq -l_{nn'} \ln \mu + \sum_{\gamma: n_\gamma \neq 0, n'_\gamma \neq 0} [n_\gamma \ln n'_\gamma - n'_\gamma \ln n_\gamma] + \sum_{\gamma: n'_\gamma = 0} n_\gamma \ln \sum_{\gamma'} \bar{\pi}_{\gamma \leftarrow \gamma'} n'_\gamma - \sum_{\gamma: n_\gamma = 0} n'_\gamma \ln \sum_{\gamma'} \bar{\pi}_{\gamma \leftarrow \gamma'} n_\gamma \quad (22)$$

dove  $l_{nn'} := \sum_{\gamma: n_\gamma = 0} n'_\gamma - \sum_{\gamma: n'_\gamma = 0} n_\gamma$  può essere interpretato come il numero netto di tipi persi nella transizione  $n \leftarrow n'$ .

Per  $l_{nn'} \neq 0$ ,  $\zeta_{nn'}$  diverge logaritmicamente nel limite  $\mu \rightarrow 0$  e il segno di  $l_{nn'}$  determina il segno della forza:  $\zeta_{nn'}$  è positiva quando la perdita di tipi è più grande nella direzione  $n \leftarrow n'$ . Questo risultato è in accordo con l'osservazione che la selezione delle varianti che si riproducono sfavorisce la diversità perchè privilegia la selezione dei tipi più abbondanti. Questo effetto è parzialmente bilanciato dal termine entropico in equazione (11), che invece favorisce composizioni eterogenee. Usando l'approssimazione di Stirling possiamo scrivere la somma di forza non conservativa e forza entropica come:

$$\zeta_{nn'} - \sum_{\gamma} \ln \frac{n_\gamma!}{n'_\gamma!} \simeq \sum_{\gamma: n'_\gamma = 0, n_\gamma \neq 0} n_\gamma \ln \frac{\mu N}{n_\gamma} - \sum_{\gamma: n'_\gamma \neq 0, n_\gamma = 0} n'_\gamma \ln \frac{\mu N}{n'_\gamma} - \sum_{\gamma: n'_\gamma \neq 0, n_\gamma \neq 0} \left[ n_\gamma \ln \frac{n_\gamma}{n'_\gamma} - n'_\gamma \ln \frac{n'_\gamma}{n_\gamma} \right] \quad (23)$$

dove sono stati trascurati gli ultimi due termini in equazione (45). Il contributo maggiore in questa equazione deriva nuovamente dai primi due termini; infatti dal momento che le variazioni sono rare, le transizioni tipiche non prevedono la comparsa di molti individui dello stesso tipo: se  $n'_\gamma = 0$  e  $n_\gamma \neq 0$  allora  $n_\gamma \simeq 1$ , analogamente se  $n'_\gamma \neq 0$  e  $n_\gamma = 0$  allora  $n'_\gamma \simeq 1$ . Inoltre le transizioni tipiche mantengono il numero dei tipi più abbondanti pressochè costante,  $n_\gamma \sim n'_\gamma$ , quindi l'ultimo termine nell'equazione precedente è trascurabile. Quindi:

$$\zeta_{nn'} - \sum_{\gamma} \ln \frac{n_\gamma!}{n'_\gamma!} \simeq -l_{nn'} \ln \mu N \quad (24)$$

La diversità all'interno della popolazione è preservata o sfavorita in base al fatto che  $\mu N$  sia maggiore o minore di 1; in particolare per  $l_{nn'} > 0$ , se  $\mu N \ll 1$  la forza non conservativa prevale su quella entropica e l'eterogeneità è sfavorita.

Come esempio consideriamo nuovamente il caso di una transizione in cui un tipo  $\bar{\gamma}$  compare a partire da una popolazione monomorfa di tipo  $\gamma$ , e  $\mu N \ll 1$ . Allora:

$$\Psi_n - \Psi_{n'} \simeq \ln \frac{f_{\bar{\gamma}}}{f_\gamma} + \ln N + \frac{f_{\bar{\gamma}}}{f_\gamma} \quad (25)$$

$$\zeta_{nn'} \simeq \ln \mu + \ln \bar{\pi}_{\bar{\gamma} \leftarrow \gamma} \quad (26)$$

Per variazioni tra tipi con uguale tasso di crescita,  $f_{\bar{\gamma}}/f_\gamma \simeq 1$ , la comparsa di un nuovo tipo è favorita dal termine  $\ln N > 0$  proveniente dal contributo entropico, e ostacolata dal termine  $\ln \mu$  proveniente dal contributo non conservativo.

## 7 Relazione di fluttuazione

Come prima cosa definiamo cosa si intende con *regime di forte selezione*: in questo regime solo i tipi con un'alta fitness sopravvivono; per formalizzare questo concetto possiamo riscrivere i tassi di riproduzione come:

$$f_\gamma = \exp\{\beta\lambda_\gamma\} \quad (27)$$

L'esponente può essere interpretato come il tasso di crescita logaritmico parametrizzato da  $\beta$ . Per grandi valori di  $\beta$ , piccole differenze di  $\lambda_\gamma$  danno origine a grandi differenze di  $f_\gamma$ ; viceversa per piccoli valori di  $\beta$ , anche grandi differenze di  $\lambda_\gamma$  originano piccole differenze in  $f_\gamma$ . Ne consegue che  $\beta$  racchiude l'informazione su quanto è forte la selezione:  $\beta \geq 1$  implica un regime di forte selezione.

Con queste premesse possiamo discutere un caso particolare di rilevanza biologica per il quale abbiamo una soluzione analitica approssimata. Trascuriamo la forza non conservativa  $\zeta_{nn'}$ ; questa operazione è giustificata se le probabilità di transizione  $\mu_\gamma$  non sono troppo piccole e la selezione è forte, in modo che  $\zeta_{nn'}$  rimanga finita e non risulti rilevante rispetto alla forza conservativa  $(\Psi_n - \Psi_{n'})$ , perchè non dipende da quanto è forte la selezione.

Sotto queste assunzioni la dinamica è approssimativamente conservativa e la distribuzione di probabilità che compare in equazione (1) rilassa ad una distribuzione di Boltzmann, perciò per tempi lunghi le popolazioni con  $\Psi_n$  più alto sono esponenzialmente più probabili:

$$p_n(\infty) \simeq \frac{\exp \Psi_n}{Z} \quad Z := \sum_n \exp \Psi_n \quad (28)$$

dove, sfruttando la forma di  $\Phi_n$  in equazione (20),  $\Psi_n$  è della forma :

$$\Psi_n \simeq \sum_\gamma n_\gamma \ln f_\gamma - \sum_\gamma \ln n_\gamma! + N \ln F_n - \mu N \Omega_n \quad (29)$$

con  $\Omega_n := \frac{F_n - F_{\pi n}}{F_n}$  la sensibilità della crescita della popolazione alle variazioni e  $\mu_\gamma = \mu \forall \gamma$ .

La relazione di fluttuazione segue direttamente dai calcoli:

$$\begin{aligned} \frac{1}{N} \frac{d \ln Z}{d\mu} &\simeq \langle \Omega_n \rangle \\ -\frac{1}{N} \frac{d \langle \Omega_n \rangle}{d\mu} &\simeq \frac{1}{N^2} \frac{d^2 \ln Z}{d\mu^2} \simeq \langle \Omega_n^2 \rangle - \langle \Omega_n \rangle^2 \end{aligned} \quad (30)$$

con  $\langle \Omega_n \rangle = \sum_n p_n(\infty) \Omega_n$  la sensibilità media all'equilibrio.

Questa relazione lega le fluttuazioni della sensibilità alle variazioni della sua media, e l'equazione (30) mostra come le fluttuazioni di  $\Omega_n$  siano più alte quando  $\langle \Omega_n \rangle$  è più suscettibile alle variazioni di  $\mu$ .

Ci sono tre importanti implicazioni della relazione di fluttuazione.

La prima è che la sensibilità media della popolazione  $\langle \Omega_n \rangle$  decresce come funzione della probabilità di variazione  $\frac{d \langle \Omega_n \rangle}{d\mu} \lesssim 0$ . Dal momento che l'inverso della sensibilità misura la robustezza della popolazione alle variazioni, questo risultato mostra che la robustezza aumenta con  $\mu$ .

La seconda è che le fluttuazioni della sensibilità tendono a zero per  $N \rightarrow \infty$ .

La terza è che, assumendo che  $\Omega_n$  saturi ad alti valori di  $\mu$  raggiungendo un plateau, ci aspettiamo che le fluttuazioni  $\langle \Omega_n^2 \rangle - \langle \Omega_n \rangle^2$  decrescano all'aumentare di  $\mu$ .

Possiamo notare come, nel regime di equilibrio asintotico (28), la popolazione si comporti in modo simile ad un sistema termodinamico in equilibrio con l'ambiente, con  $\Omega_n$  che ha il ruolo dell'energia e  $\mu N$  dell'inverso della temperatura.

La relazione di fluttuazione (30) è il risultato principale ottenuto a partire dal modello RVS al quale vengano imposte ulteriori condizioni. Nonostante ciò e nonostante il fatto che la relazione valga solo in modo approssimato, essa sembra consistente con i risultati di esperimenti eseguiti su popolazioni virali e con le simulazioni numeriche.

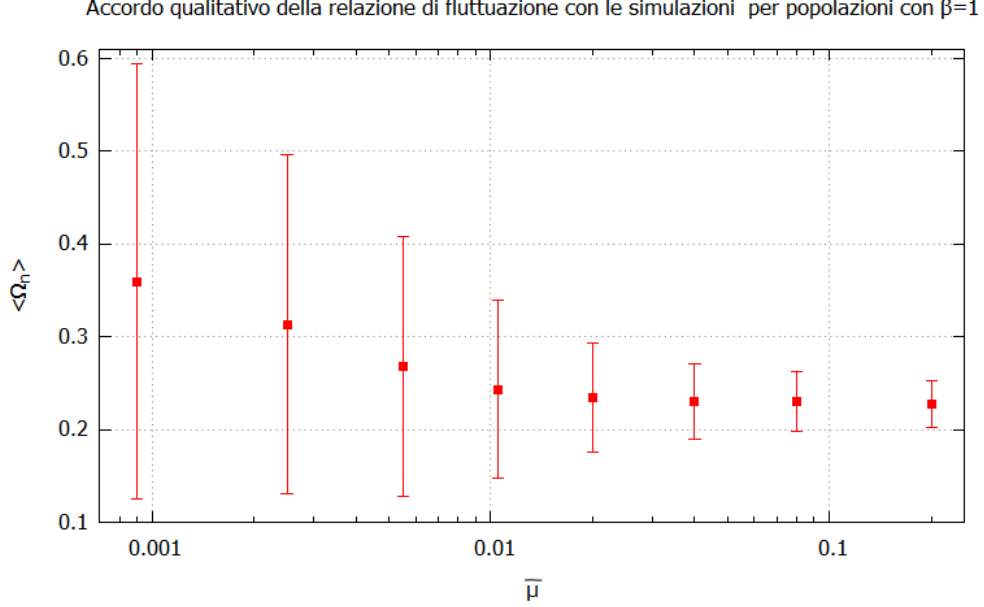


Figura 5: Il grafico rappresenta l'andamento della sensibilità media all'equilibrio in funzione del tasso di variazione  $\bar{\mu}$  per un campione di popolazioni aventi tutte  $\beta = 1$ . I risultati sono ottenuti a partire da simulazioni numeriche, come descritto nella sezione (9) con  $\bar{\mu} = \mu/4$ . Notiamo che, in accordo con la relazione di fluttuazione, la sensibilità media  $\langle \Omega_n \rangle$  diminuisce con l'aumentare di  $\mu$  fino a raggiungere un plateau per alti valori di  $\mu$ ; allo stesso modo le fluttuazioni della sensibilità media tendono ad annullarsi per alti valori di  $\mu$ .

## 8 Modello a due tipi

Consideriamo un esempio di dinamica RVS con 4 organismi,  $N=4$ , e due tipi,  $\gamma_1$  e  $\gamma_2$ . Entrambi i tipi hanno la stessa probabilità di variazione,  $\pi_{\gamma_1 \leftarrow \gamma_2} = \pi_{\gamma_2 \leftarrow \gamma_1} = \mu < 1/2$  e scegliamo  $f_{\gamma_1} > f_{\gamma_2}$ .

La forza legata alla selezione può essere espressa come:

$$\sum_{\gamma} (n_{\gamma} - n'_{\gamma}) \ln f_{\gamma} = (n_1 - n'_1) \ln \frac{f_{\gamma_1}}{f_{\gamma_2}}$$

e favorisce la popolazione con la più alta proporzione di tipi che si riproducono velocemente:  $(n_1, n_2) = (4, 0)$ .

Vale:

$4 \ln \frac{f_{\gamma_1}}{f_{\gamma_2}}$  per la transizione  $(4, 0) \leftarrow (0, 4)$

$3 \ln \frac{f_{\gamma_1}}{f_{\gamma_2}}$  per le transizioni  $(4, 0) \leftarrow (1, 3)$ ,  $(3, 1) \leftarrow (0, 4)$

$2 \ln \frac{f_{\gamma_1}}{f_{\gamma_2}}$  per le transizioni  $(4, 0) \leftarrow (2, 2)$ ,  $(3, 1) \leftarrow (1, 3)$ ,  $(2, 2) \leftarrow (0, 4)$

$1 \ln \frac{f_{\gamma_1}}{f_{\gamma_2}}$  per le transizioni  $(4, 0) \leftarrow (3, 1)$ ,  $(3, 1) \leftarrow (2, 2)$ ,  $(2, 2) \leftarrow (1, 3)$ ,  $(1, 3) \leftarrow (0, 4)$

La forza entropica è diversa da zero solo quando la composizione della popolazione cambia il suo grado di omogeneità, ed è pari a  $\sum_{\gamma} (\ln n'_{\gamma}! - \ln n_{\gamma}!)$ . Favorisce popolazioni con composizione disomogenea.

Vale:

0 per le transizioni  $(4, 0) \leftarrow (0, 4)$ ,  $(1, 3) \leftarrow (3, 1)$   
 $\ln 6$  per le transizioni  $(2, 2) \leftarrow (0, 4)$ ,  $(2, 2) \leftarrow (4, 0)$   
 $\ln(3/2)$  per le transizioni  $(2, 2) \leftarrow (1, 3)$ ,  $(2, 2) \leftarrow (3, 1)$   
 $\ln 4$  per le transizioni  $(3, 1) \leftarrow (0, 4)$ ,  $(1, 3) \leftarrow (4, 0)$ ,  $(3, 1) \leftarrow (4, 0)$ ,  $(1, 3) \leftarrow (0, 4)$

La forza non conservativa può essere espressa come:

$$\zeta_{(n_1, n_2) \leftarrow (n'_1, n'_2)} = \ln \frac{[(1 - \mu)n'_1 + \mu n'_2]^{n_1}}{[(1 - \mu)n_1 + \mu n_2]^{n'_1}} + \ln \frac{[(1 - \mu)n'_2 + \mu n'_1]^{n_2}}{[(1 - \mu)n_2 + \mu n_1]^{n'_2}} \quad (31)$$

La differenza con le forze viste in precedenza sta nel fatto che questa (così come i rapporti tra i diversi termini che vi compaiono) dipende esplicitamente da  $\mu$ . Per  $\mu N \ll 1$  questa forza favorisce gli stati omogenei.

Vale:

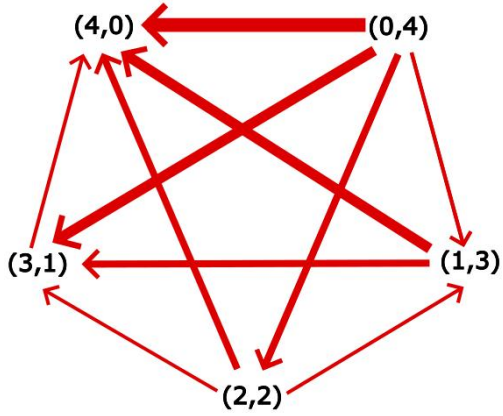
0 per le transizioni  $(4, 0) \leftarrow (0, 4)$ ,  $(1, 3) \leftarrow (3, 1)$   
 $\ln \frac{1}{16\mu^2(1-\mu)^2}$  per le transizioni  $(0, 4) \leftarrow (2, 2)$ ,  $(4, 0) \leftarrow (2, 2)$   
 $\ln \frac{16}{(3+4\mu-4\mu^2)^2}$  per le transizioni  $(1, 3) \leftarrow (2, 2)$ ,  $(3, 1) \leftarrow (2, 2)$   
 $\ln \frac{(3-2\mu)^4}{256(1-\mu)^3\mu}$  per le transizioni  $(0, 4) \leftarrow (1, 3)$ ,  $(4, 0) \leftarrow (3, 1)$   
 $\ln \frac{(1+2\mu)^4}{256(1-\mu)\mu^3}$  per le transizioni  $(0, 4) \leftarrow (3, 1)$ ,  $(4, 0) \leftarrow (1, 3)$

Infine calcoliamo il potenziale di crescita atteso:

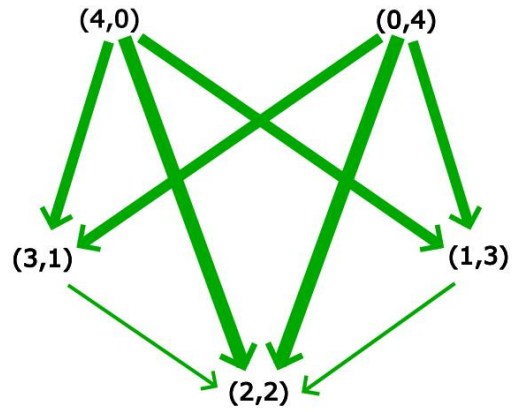
$$\Phi_{(n_1, n_2)} = 4 \ln[f_{\gamma_1}((1 - \mu)n_1 + \mu n_2) + f_{\gamma_2}(\mu n_1 + (1 - \mu)n_2)] \quad (32)$$

$4 \ln[4\mu f_{\gamma_1} + 4(1 - \mu)f_{\gamma_2}]$  per la configurazione  $(0, 4)$   
 $4 \ln[(1 + 2\mu)f_{\gamma_1} + (3 - 2\mu)f_{\gamma_2}]$  per la configurazione  $(1, 3)$   
 $4 \ln[2f_{\gamma_1} + 2f_{\gamma_2}]$  per la configurazione  $(2, 2)$   
 $4 \ln[(3 - 2\mu)f_{\gamma_1} + (1 + 2\mu)f_{\gamma_2}]$  per la configurazione  $(3, 1)$   
 $4 \ln[4(1 - \mu)f_{\gamma_1} + 4\mu f_{\gamma_2}]$  per la configurazione  $(4, 0)$

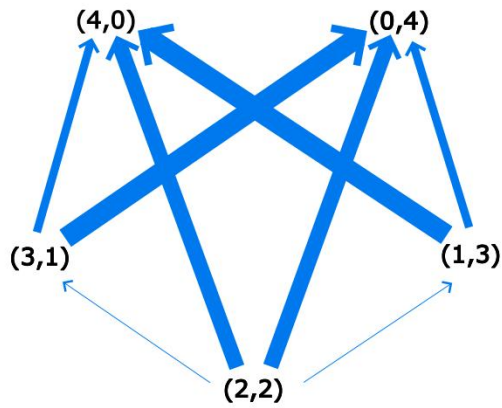
Vediamo che la forza associata  $\Phi_n - \Phi_{n'}$  favorisce semplicemente la popolazione con la crescita attesa più alta alla generazione successiva, quindi porta la popolazione dalla configurazione  $(0, 4)$  verso la configurazione  $(4, 0)$  (a causa della semplicità della popolazione scelta non è possibile evidenziare l'aspetto di questa forza legato alla robustezza). Tuttavia l'intensità relativa di questo tipo di forze dipende dalla particolare scelta dei parametri  $\mu$ ,  $f_{\gamma_1}$  e  $f_{\gamma_2}$ , perciò nella pagina seguente non riporteremo la rappresentazione grafica di questa forza.



((a)) *Contributo dovuto alla riproduzione*



((b)) *Contributo entropico*



((c)) *Contributo dovuto alla forza non conservativa*

Figura 6: Rappresentazione grafica dell'azione della forza legata alla riproduzione, della forza entropica e della forza non conservativa (quest'ultima calcolata per  $\mu = 0.0025$ ). Lo spessore delle frecce indica l'intensità della forza rapportata alle altre dello stesso schema. Si può notare come per  $\mu \ll 1$  la forza entropica tenda a controbilanciare quella non conservativa.

## Parte III

# Algoritmi evolutivi

Gli algoritmi evolutivi indicano sequenze di istruzioni da eseguire in modo ciclico per simulare l'evoluzione di una popolazione, generando una nuova popolazione a partire da quella precedente ad ogni iterazione. Se la popolazione è aploide (gli organismi possiedono un solo set cromosomico) l'algoritmo dovrà tenere conto dei processi di variazione, selezione e riproduzione; se la popolazione è diploide (ogni cromosoma è presente in due copie) si dovrà simulare anche il processo di crossing-over da cui risulta la ricombinazione di geni materni e paterni nella progenie.

Per un sistema aploide, si inizia con una popolazione di  $N$  organismi, ognuno dei quali è caratterizzato da alcune variabili ereditarie  $\gamma$ . Ogni iterazione consiste in 2 passi:

1. Le variabili ereditarie di ogni organismo sono fatte variare casualmente in accordo con le probabilità di variazione  $\pi_{\gamma' \leftarrow \gamma}$ . Il risultato di questo primo passo è una popolazione di varianti caratterizzate ognuna da un certo tasso di riproduzione  $f_{\gamma}$ .
2.  $N$  nuovi organismi vengono estratti e vanno a rimpiazzare i precedenti, la selezione degli organismi scelti dipende dai loro tassi di riproduzione (fitness)  $f_{\gamma}$ . La popolazione estratta procede all'iterazione successiva.

La scelta degli organismi che si riproducono avviene sulla base del loro valore di fitness. I tipi con fitness più alta avranno maggiore probabilità di sopravvivere e riprodursi. Un modo comune di selezione è chiamato selezione proporzionale alla fitness, il cui funzionamento può essere immaginato come una roulette con settori proporzionali ai valori di fitness. A scopo illustrativo supponiamo di avere 4 individui nella popolazione e supponiamo che la loro fitness individuale sia valutata nel modo seguente:

Individuo 1: Fitness = 10

Individuo 2: Fitness = 20

Individuo 3: Fitness = 30

Individuo 4: Fitness = 40

Creiamo una ruota in cui ciascun settore corrisponde alla fitness di un individuo; lanciamo virtualmente una pallina, il settore nel quale la pallina si ferma indica che individuo si riprodurrà. Ripetiamo l'operazione 4 volte per ottenere la composizione della popolazione alla generazione successiva.

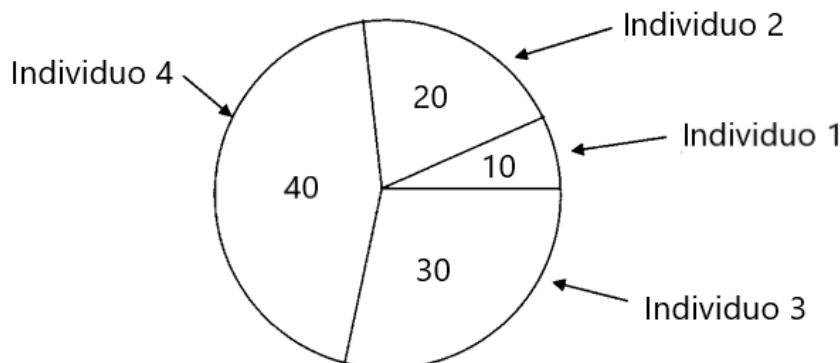


Figura 7: Questo grafico a torta illustra la selezione degli individui che si riproducono per una popolazione di 4 organismi. A ciascun individuo è assegnato un settore proporzionale al suo valore di fitness. La selezione di ogni individuo ai fini della riproduzione è proporzionale all'ampiezza del settore.



## 9 Modello a griglia

Utilizziamo un modello a griglia per simulare l'evoluzione di una popolazione. Ad ogni casella della griglia 2D associamo un determinato tipo caratterizzato dalle variabili ereditarie  $\gamma$ , che nel nostro caso sono rappresentate dalle coordinate  $(x,y)$  della casella all'interno della griglia. Per ogni popolazione consideriamo  $N=144$  individui che inizialmente distribuiamo in modo casuale tra le caselle della griglia. Le variazioni da un tipo ad un altro corrispondono a transizioni tra primi vicini e sono regolate dal coefficiente di variazione, che poniamo pari ad un valore  $\bar{\mu}$  costante. Di conseguenza la probabilità di variazione totale per un tipo- $\gamma$  sarà pari a  $4\bar{\mu}$  per i tipi interni alla griglia,  $3\bar{\mu}$  per i tipi sui lati, e  $2\bar{\mu}$  per i tipi sugli angoli.

I tassi di riproduzione sono parametrizzati come in equazione (27), e i valori di  $\lambda_\gamma$  sono mostrati in figura (8). Isole di tipi che si riproducono velocemente (quadrati grandi arancioni,  $\lambda_\gamma = \lambda_* = 5$ ) sono circondate da un mare di tipi con tassi di riproduzione più bassi (quadrati piccoli grigi,  $\lambda_\gamma = \lambda_- = 1$ ). Ogni isola di tipi che si riproducono velocemente si differenzia dalle altre per la topologia, cioè per il numero di tipi adiacenti con basso tasso di riproduzione.

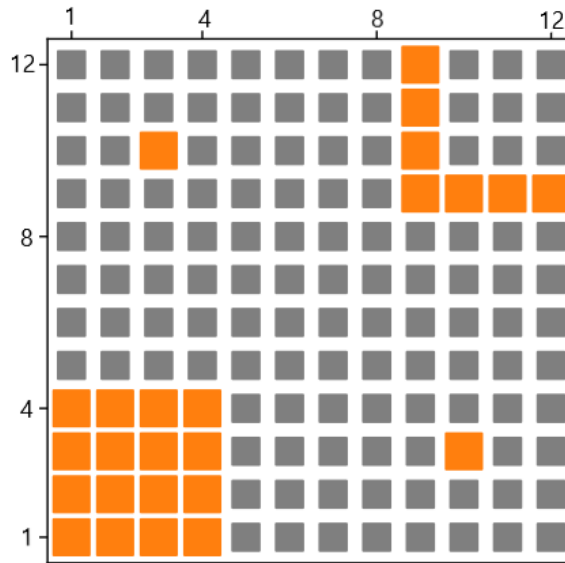


Figura 8: Rappresentazione del modello a griglia, dimensioni 12x12. Ogni casella rappresenta un diverso tipo. I quadrati arancioni rappresentano i tipi con  $\lambda_\gamma = \lambda_* = 5$ , mentre quelli grigi hanno  $\lambda_\gamma = \lambda_- = 1$ .

La figura (10) mostra la distribuzione dei tipi per diversi valori della forza di selezione  $\beta$  e del coefficiente di variazione  $\bar{\mu}$ . Per ognuna delle nove griglie sono state considerate  $N_{pop} = 2000$  popolazioni; ogni popolazione è stata lasciata evolvere fino all'equilibrio e poi sono state estratte casualmente alcune configurazioni all'interno di 1000 generazioni successive. Il risultato finale mostrato in figura è dato dalla media di queste configurazioni. Il raggiungimento dell'equilibrio si può verificare graficando il valore medio del numero di individui contenuti in ogni casella non vuota in funzione del numero di generazioni (figura 9): quando il grafico diventa pressochè costante - a meno di fluttuazioni - la popolazione ha raggiunto l'equilibrio.

Notiamo che i tipi che appartengono all'isola 2D in basso a sinistra hanno una più alta probabilità di essere presenti nella popolazione rispetto a quelli nell'isola uno-dimensionale e a quelli isolati. Anche all'interno dell'isola 2D i tipi interni, che sono completamente circondati da tipi con alto tasso di riproduzione, sono favoriti rispetto ai tipi esterni.

Inoltre, ad eccezione di forza di selezione estremamente elevata e coefficiente di variazione molto basso (griglia nell'angolo in alto a sinistra della figura), si osservano sistematicamente degli individui appartenenti a tipi con basso tasso di riproduzione vicini alle isole ad alto tasso di riproduzione. Entrambi questi effetti diventano più importanti con l'aumentare di  $\bar{\mu}$  e il diminuire della forza di selezione  $\beta$ .

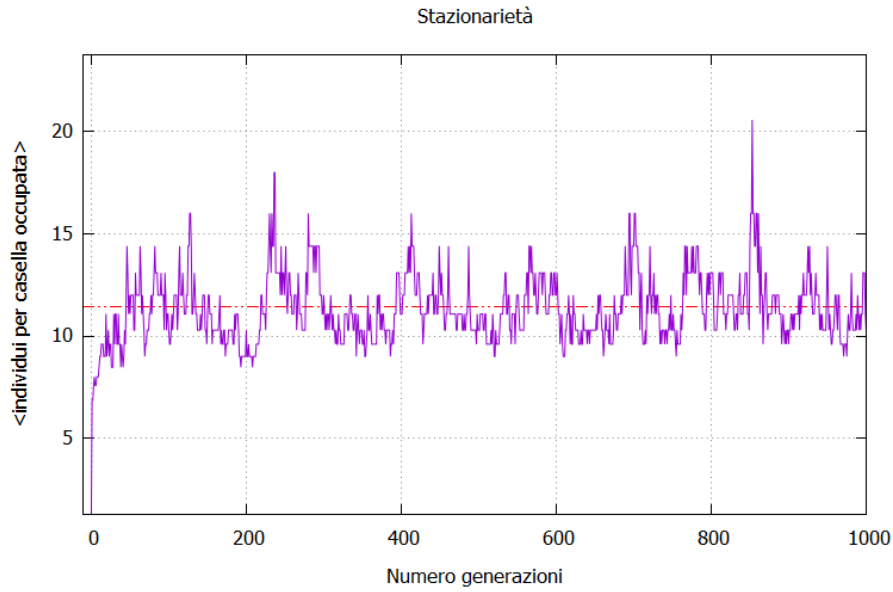


Figura 9: Andamento del numero medio di individui per casella occupata in funzione del numero di generazioni. Si noti come la popolazione considerata raggiunga l'equilibrio velocemente, dopo qualche centinaio di generazioni.

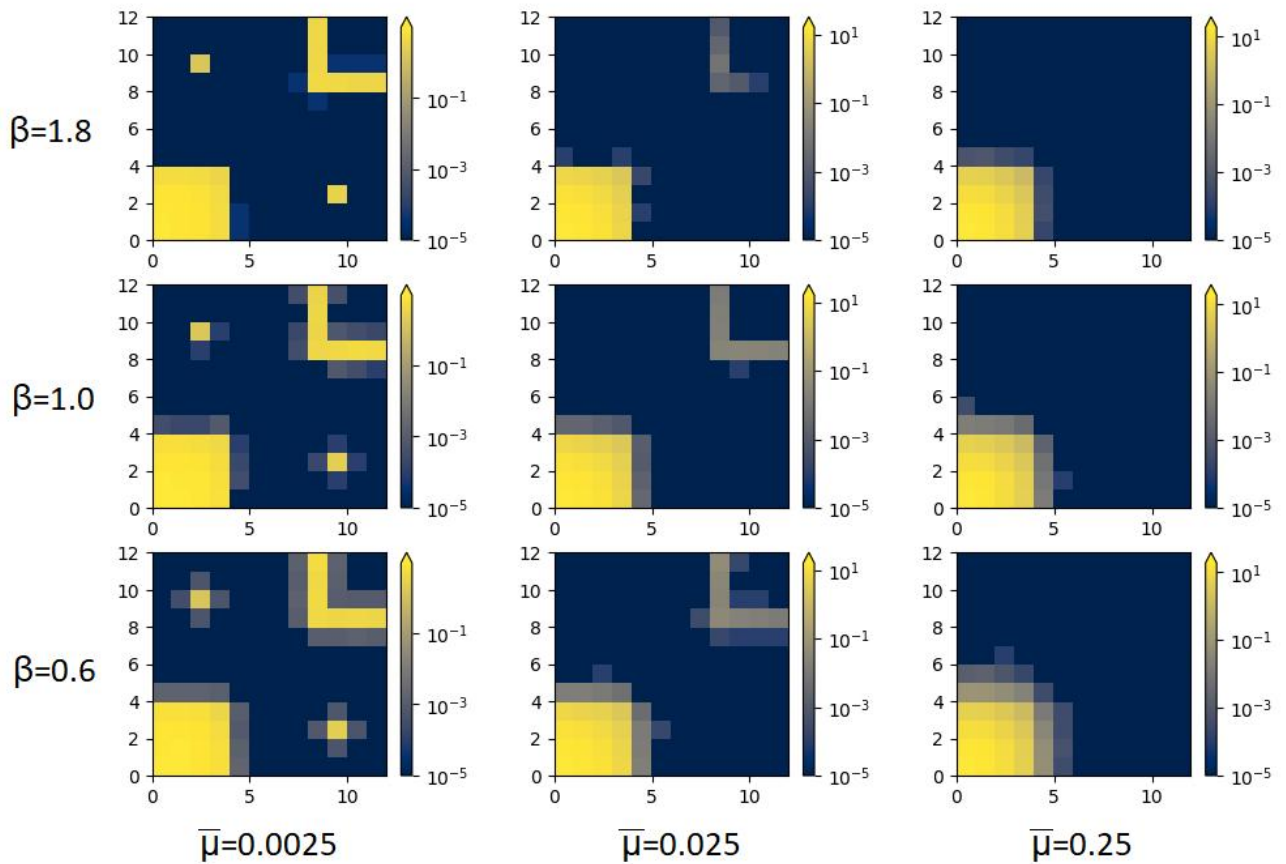


Figura 10: Configurazioni delle popolazioni all'equilibrio per diversi regimi di forza di selezione ( $\beta$ ) e probabilità di variazione ( $\bar{\mu}$ ). I diversi colori indicano quanti individui di un determinato tipo ci si aspetta di trovare per tempi lunghi.

Queste osservazioni sono in accordo con il fenomeno conosciuto come *survival of the flattest* nel modello a quasispecie (Appendice A.5). Questo effetto, in contrasto con il più conosciuto "survival of the fittest", può essere riassunto nel modo seguente: quando i tassi di variazione sono alti, la selezione naturale non favorisce necessariamente i tipi con più alto tasso di riproduzione/con valori di fitness più alti; infatti per alti valori di  $\mu$  i tipi che occupano regioni piatte del fitness landscape hanno una maggiore probabilità di sopravvivere rispetto a quelli che occupano regioni con fitness più alto ma che sono isolate, perchè è più probabile che le mutazioni abbiano effetti nocivi su questi ultimi.

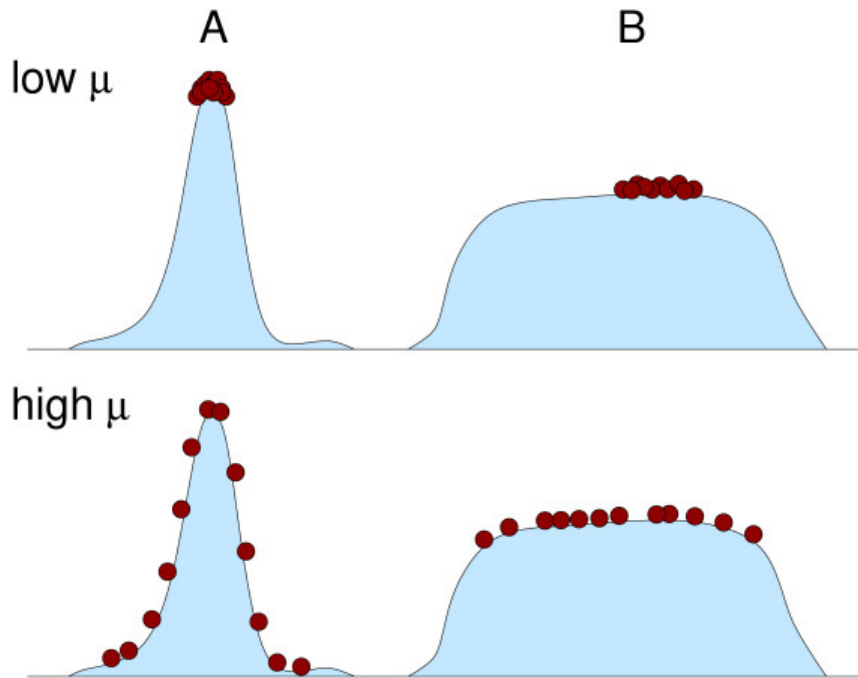


Figura 11: Rappresentazione schematica dell'effetto chiamato "survival of the flattest". Per basse probabilità di variazione gli individui che si accumulano in corrispondenza della cima del picco di fitness (A) sono molti di più rispetto a quelli che si trovano sul tratto piatto (B). Per alte probabilità di variazione invece, la maggior parte degli individui che si trovano sul picco sono in corrispondenza dei suoi pendii, a bassi valori di fitness; mentre gli individui sul tratto piatto continuano ad avere valori di fitness vicini al massimo locale. Come conseguenza la fitness media degli individui in B è più grande della fitness media degli individui in A, perciò B risulterà favorito. [3]

## 10 Distribuzioni di abbondanza delle specie

Possiamo mettere in evidenza l'effetto di survival of the flattest, così come confrontare tra loro diverse popolazioni corrispondenti a diverse scelte dei parametri  $\bar{\mu}$  e  $\beta$ , graficando la distribuzione del numero di tipi con  $n$  individui in funzione di  $n$ . Questo tipo di grafici è largamente utilizzato nello studio delle popolazioni e prende il nome di "species abundance distribution" (SAD). L'istogramma che ne risulta in ecologia ha un andamento pressochè universale con poche specie molto abbondanti e molte specie rare.

La peculiarità della SAD è che trascura l'informazione sulle caratteristiche dei diversi tipi e così facendo permette di mettere a confronto anche popolazioni che non hanno specie in comune, mettendo in evidenza, ad esempio, quali popolazioni siano più ricche di specie rare o quali presentino fenomeni distintivi come la *monodominance* (condizione nella quale più del 60% degli individui è di una stessa specie).

Costruendo i grafici della distribuzione dell'abbondanza delle specie per i due casi limite di figura 10 ( $\beta = 1.8$ ,  $\bar{\mu} = 0.0025$  e  $\beta = 0.6$ ,  $\bar{\mu} = 0.25$ ) vediamo che in entrambi i casi si ha un picco per  $n=0$ , cioè la maggior parte delle caselle non è occupata. Nel regime di forte selezione e bassa probabilità di

variazione ci sono tipi con un numero di individui pari all'intera popolazione, cioè si verificano delle configurazioni in cui tutti gli individui della popolazione sono dello stesso tipo, mentre nel regime con  $\beta = 0.6$  e alta probabilità di variazione si nota come gli individui siano più distribuiti tra i tipi, infatti non ci sono caselle che contengono più di 60 individui, in accordo con il fenomeno di survival of the flattest (gli individui si distribuiscono tra i tipi dell'isola 2D in basso a sinistra nella griglia).

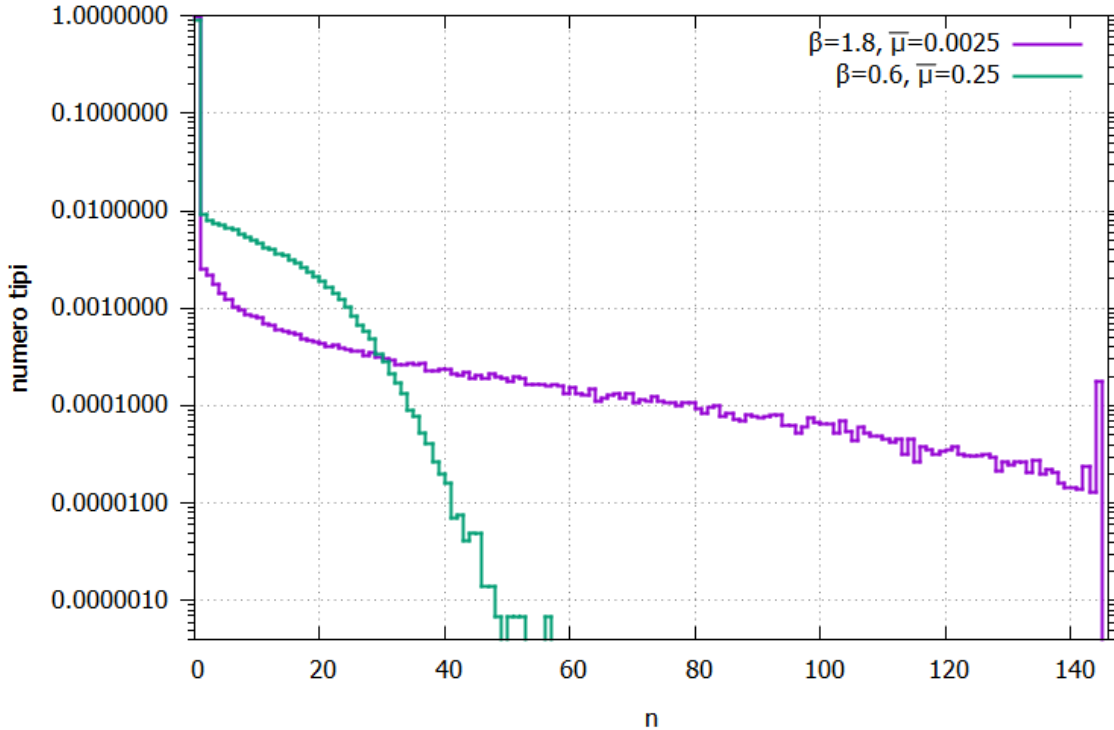


Figura 12: Confronto tra la SAD di un insieme di 9951 configurazioni di popolazioni all'equilibrio con alta forza di selezione e basso tasso di mutazione ( $\beta = 1.8$ ,  $\bar{\mu} = 0.0025$ ), e la SAD di un insieme di 9992 configurazioni di popolazioni all'equilibrio in un regime di bassa selezione e alto tasso di mutazione ( $\beta = 0.6$ ,  $\bar{\mu} = 0.25$ ). Le due distribuzioni sono normalizzate con  $N_{tipi} = 1$ . Si noti il decadimento di tipo esponenziale per  $n > 40$  per il grafico viola.

Vogliamo ora indagare come cambiano le distribuzioni di abbondanza delle specie mantenendo fisso uno dei due parametri che regolano l'evoluzione delle popolazioni ( $\bar{\mu}$ ,  $\beta$ ) e facendo variare l'altro. Notiamo che ponendo  $\beta = 1$  e diminuendo progressivamente il valore di  $\bar{\mu}$  si ottengono delle distribuzioni con una rapida decrescita iniziale, con gli  $n$  più grandi che si popolano al diminuire di  $\bar{\mu}$ . Per  $\bar{\mu} \lesssim 144^{-1}$  si assiste ad una ricrescita finale della distribuzione dovuta ad un effetto di scala, legato al fatto che non stiamo lavorando con un sistema infinito.

Graficando le distribuzioni con gli assi in scala logaritmica possiamo mettere in evidenza il fatto che queste siano delle *power-law* almeno per bassi valori di  $n$ , seguite poi da un *cutoff* esponenziale. In formule ci aspettiamo di poter descrivere le distribuzioni nel modo seguente:

$$T(n) = kn^{-a}e^{-bn} \quad \ln(T(n)) = c - a \ln(n) - bn \quad (33)$$

con  $a, b, k$  opportuni parametri,  $c = \ln(k)$  e  $T(n)$  che rappresenta il numero di tipi con  $n$  individui.

Ripetendo il procedimento mantenendo  $\bar{\mu}$  fissato, pari a 0.00033, e diminuendo progressivamente  $\beta$ , si ottiene un andamento analogo a quello appena descritto per le distribuzioni di abbondanza delle specie.

Inoltre, indagando come cambia la SAD al variare del numero di individui, vediamo che mantenendo  $\bar{\mu}$  e  $\beta$  fissati e variando  $N$  la pendenza della *power-law* rimane la stessa, ciò che cambia è il comportamento per grandi valori di  $n$ .

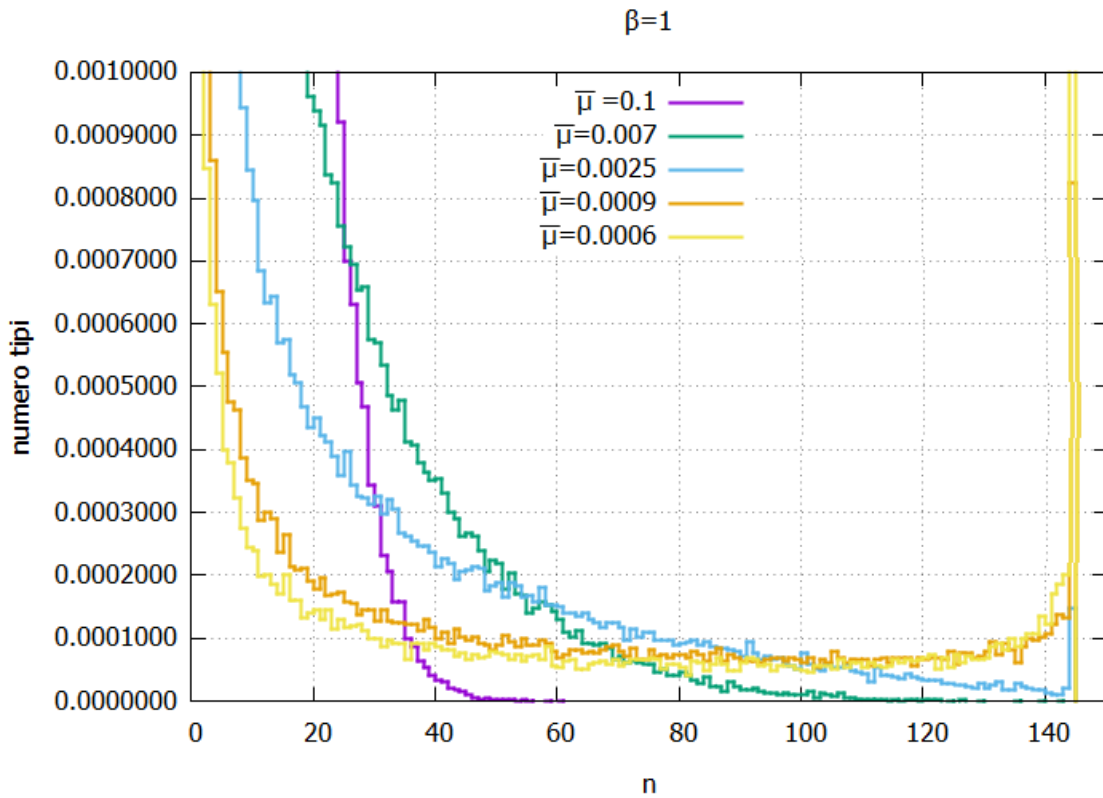


Figura 13: Distribuzione del numero di tipi con  $n$  individui per diverse popolazioni con  $\beta = 1$ .

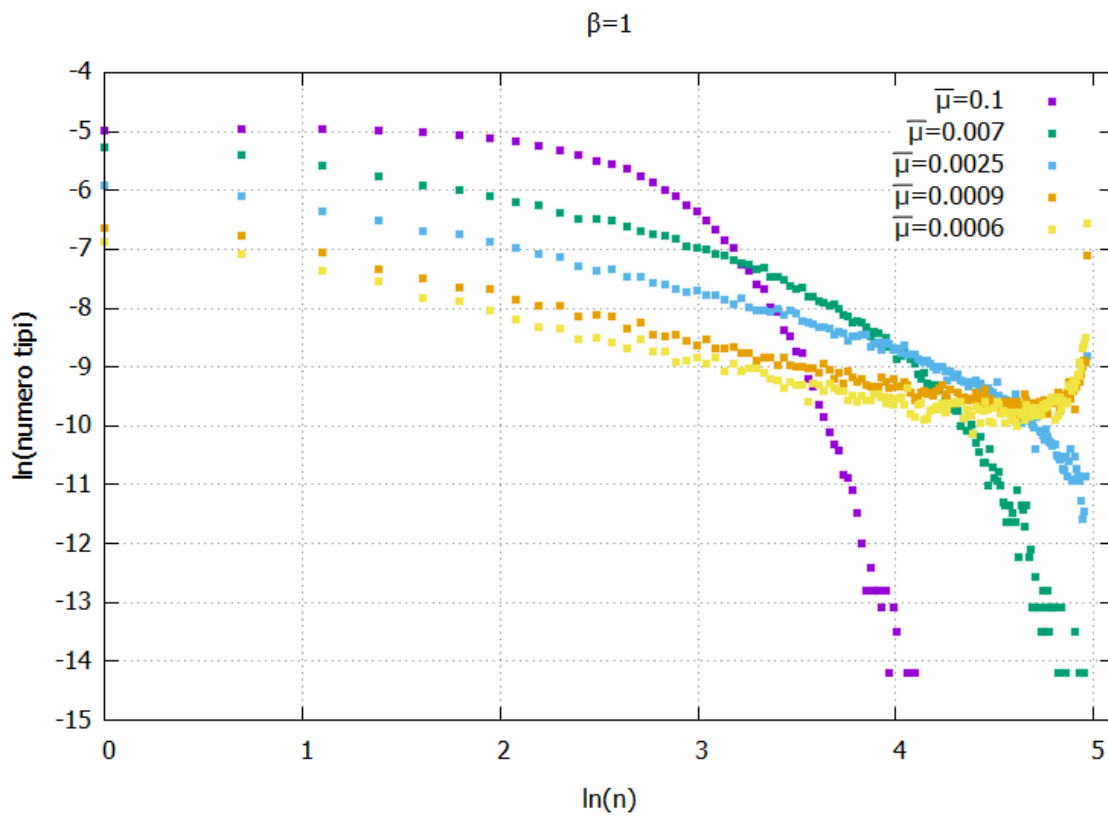


Figura 14: Distribuzione del numero di tipi con  $n$  individui per diverse popolazioni con  $\beta = 1$ .

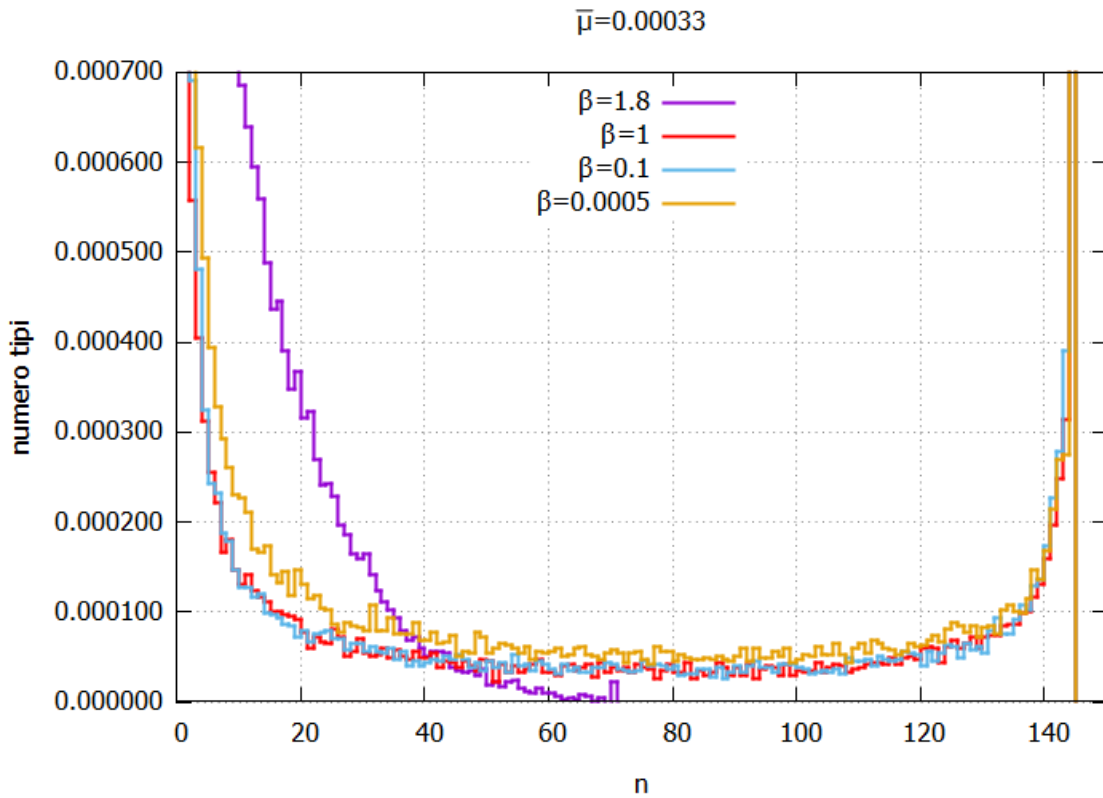


Figura 15: Distribuzione del numero di tipi con  $n$  individui per diverse popolazioni con  $\bar{\mu} = 0.00033$ .

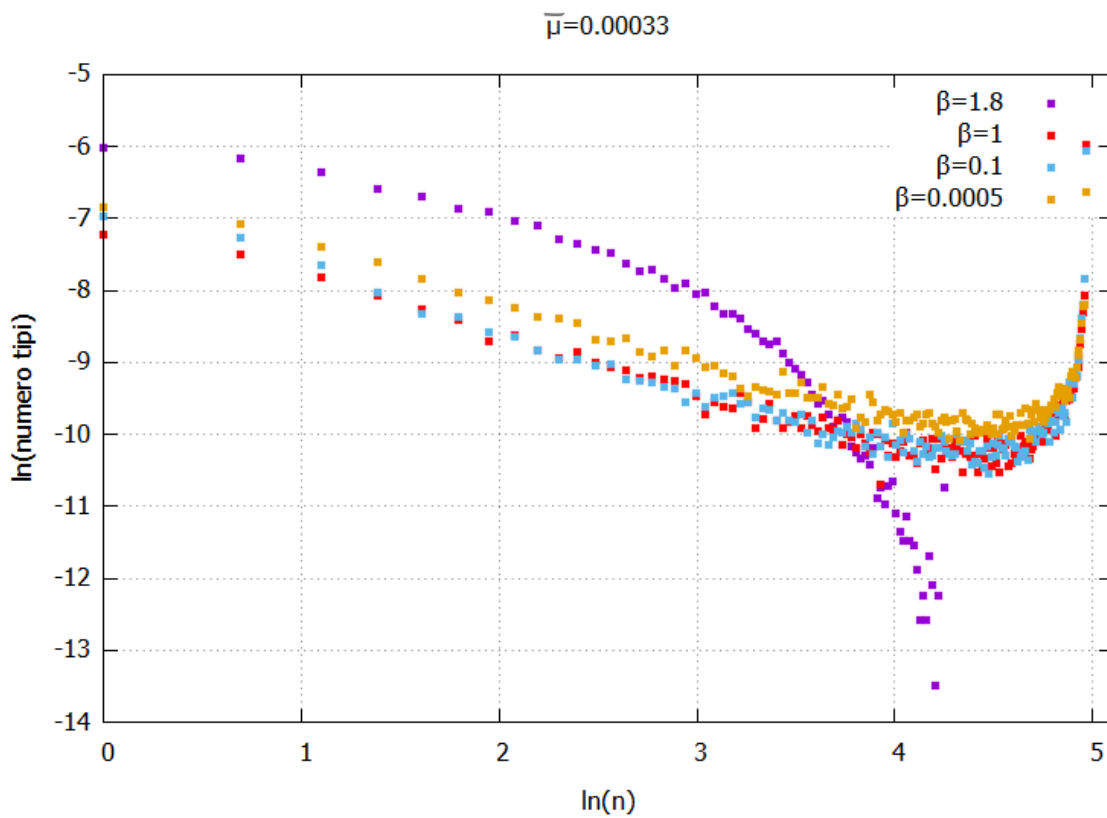


Figura 16: Distribuzione del numero di tipi con  $n$  individui per diverse popolazioni con  $\bar{\mu} = 0.00033$ .

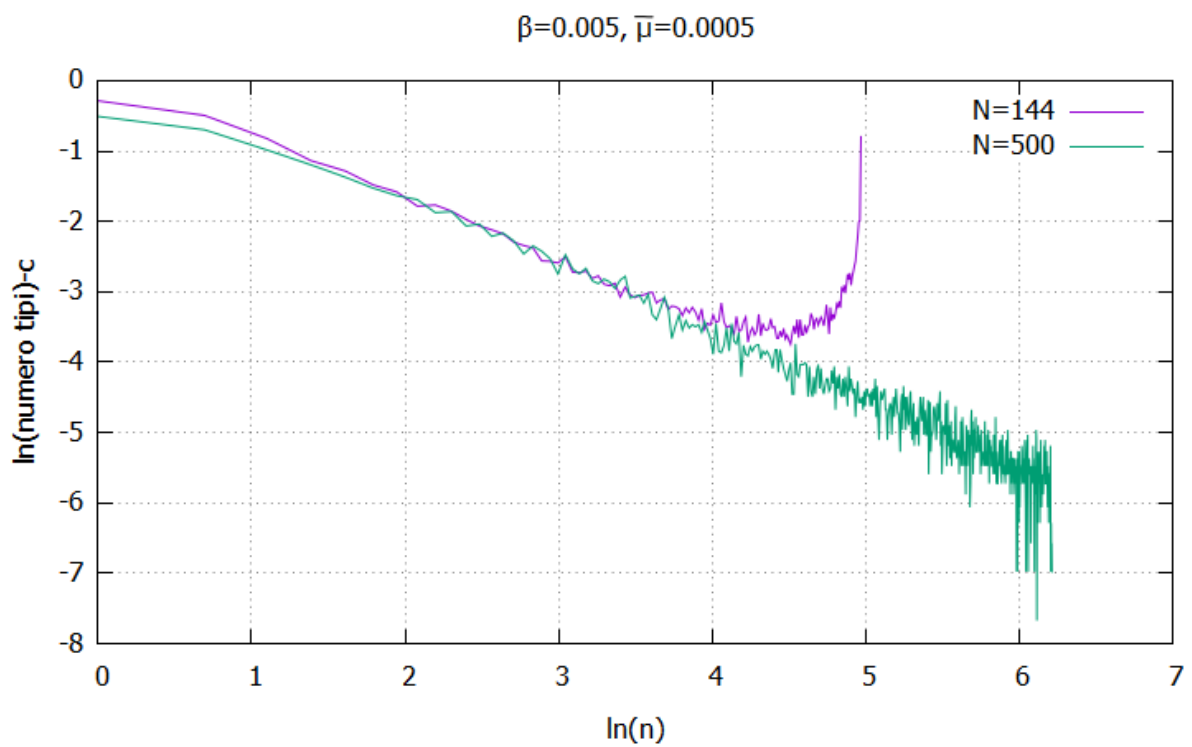
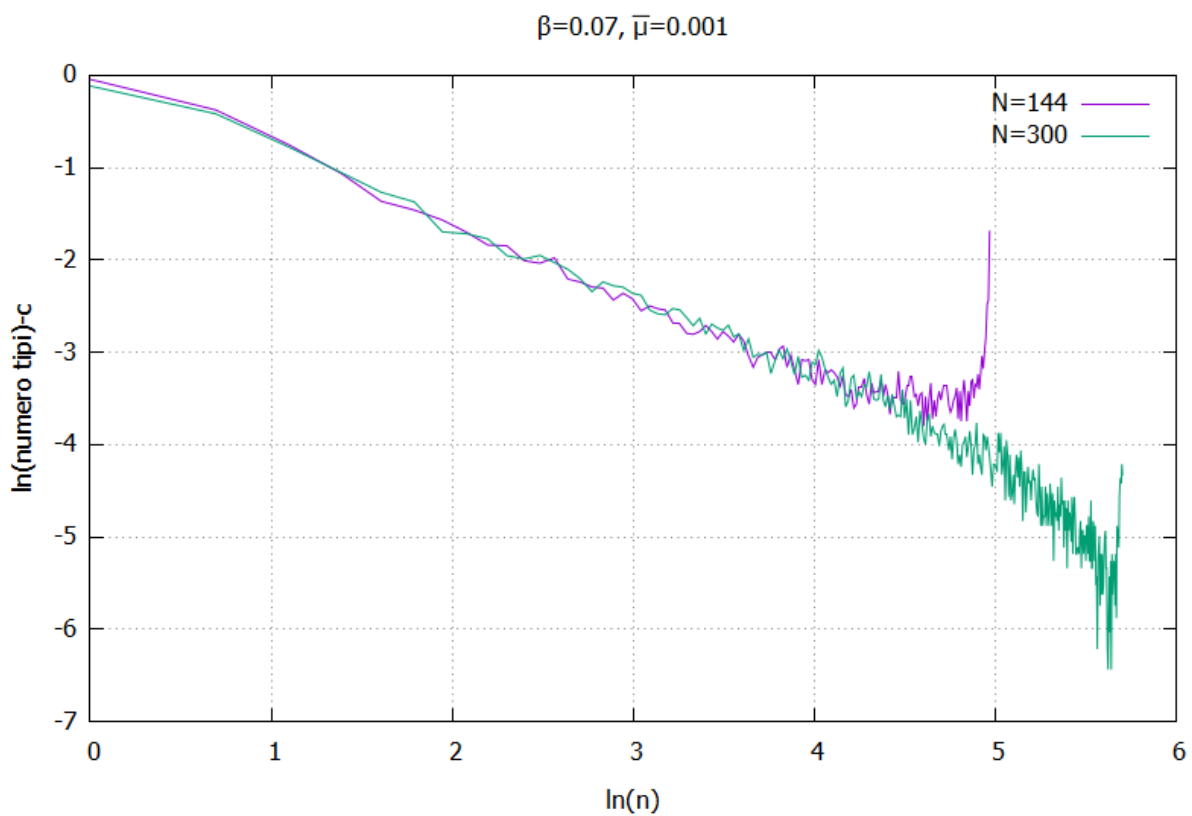


Figura 17: Confronto delle distribuzioni di abbondanza delle specie per due diverse coppie  $(\beta, \bar{\mu})$  e diverso numero di individui all'interno della popolazione,  $N$ .

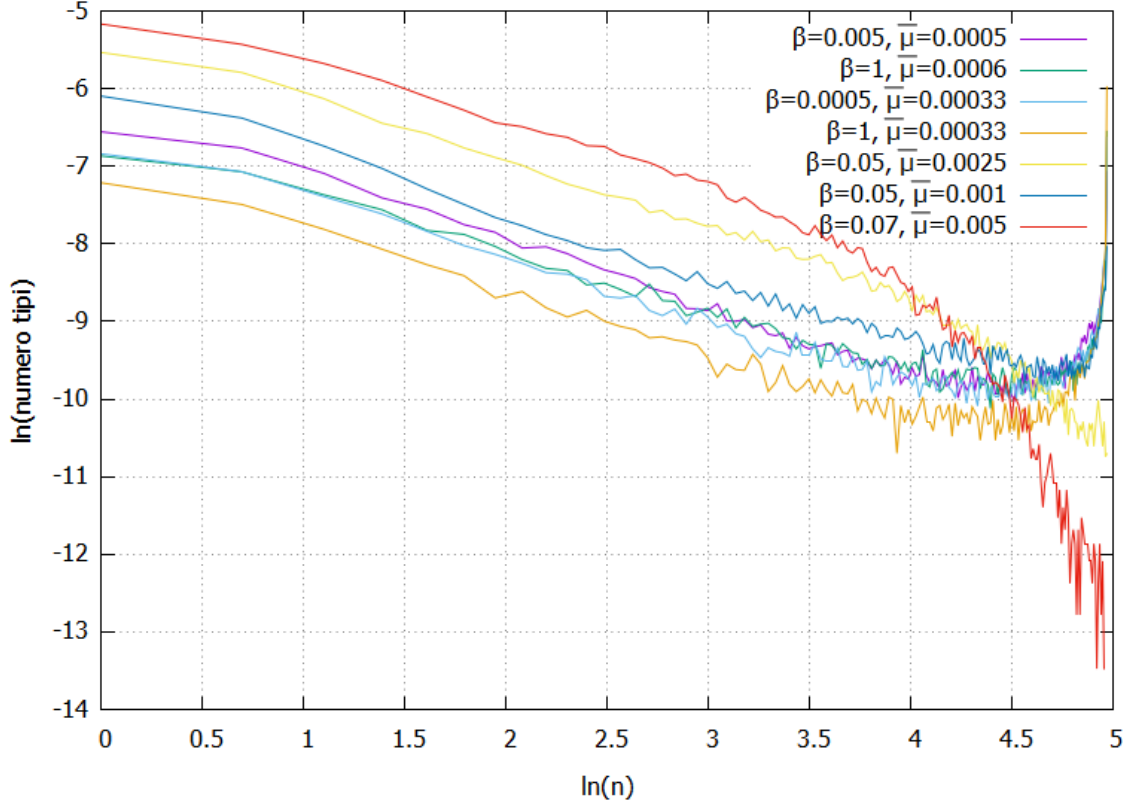


Figura 18: Distribuzioni di abbondanza delle specie per diverse popolazioni ( $N=144$ ), ad evidenziare il fatto che la pendenza delle power law per  $\beta$  e  $\bar{\mu}$  sufficientemente piccoli è la stessa.

Effettuando dei fit lineari sulle distribuzioni in scala logaritmica delle pagine precedenti ci possiamo accorgere del fatto che, per valori sufficientemente piccoli di  $\bar{\mu}$  e  $\beta$ , la pendenza delle distribuzioni è la stessa. Inoltre ci aspettiamo che il parametro  $b$  in eq. (33) possa essere scritto come un'opportuna combinazione di  $\beta$ ,  $\bar{\mu}$  e  $N$  per ogni popolazione; un *educated guess* basato sui valori di  $b$  ottenuti a partire dai fit (solo per le distribuzioni che crollano verso il basso per alti valori di  $n$ ) è che  $b$  possa essere proporzionale al prodotto  $\beta\bar{\mu}/N$ .

Ipotizziamo allora che le diverse distribuzioni di abbondanza delle specie in realtà collassino in un unico grafico se sostituiamo alla variabile  $n$  la variabile  $m = n\beta\bar{\mu}/N$ , indicando che la dinamica delle popolazioni risulta esattamente la stessa.

Ricaviamo la forma della nuova distribuzione in funzione di  $m$ :

$$T(m) = T(n) \frac{dn}{dm} = T(n) \frac{N}{\beta\bar{\mu}} \quad (34)$$

Perciò, facendo riferimento ad eq. (33), possiamo scrivere:

$$T(m) = k' m^{-a} e^{-qm} \frac{N}{\beta\bar{\mu}} \quad \ln(T(m)) = c' - a \ln(m) - qm + \ln \frac{N}{\beta\bar{\mu}} \quad (35)$$

Il risultato di questa operazione, visibile nella figura alla pagina seguente, mostra come le distribuzioni  $T(m)$  effettivamente collassano lungo un'unica curva. Tuttavia le distribuzioni che hanno una repentina crescita per alti valori di  $m$  risultano tutte più basse rispetto alle altre, indicando la possibile presenza di altri effetti di cui non è stato tenuto conto.

Prendendo separatamente ognuna delle distribuzioni della figura (19) ed eseguendo un fit lineare sulla parte della distribuzione che segue una legge di potenza, si trova che la pendenza del best fit è sempre compresa tra -0.76 e -0.88. Inoltre considerando tutte le distribuzioni assieme in un unico set di dati (solo le parti delle distribuzioni che possono essere approssimate con una power-law) troviamo che la pendenza del best fit è pari a  $-0.866 \pm 0.006$ , dove però l'errore risulta sottostimato per le considerazioni appena fatte.



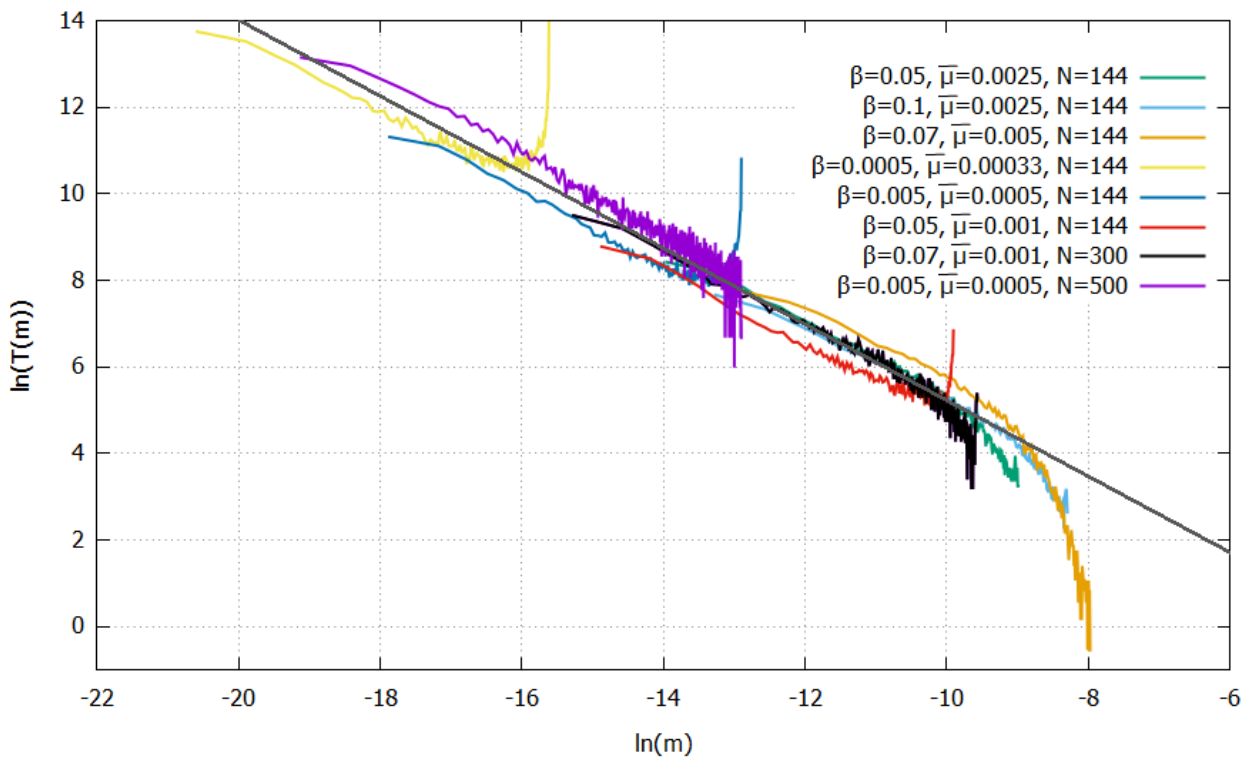


Figura 19: Distribuzione del numero di tipi in funzione della variabile  $m = n\beta\bar{\mu}/N$ . Le distribuzioni collasano tutte lungo un'unica curva ma si osservano degli scostamenti in base all'andamento finale di  $T(m)$ . In grigio scuro il risultato del best fit lineare ottenuto considerando tutte le distribuzioni.

## 11 Modello asimmetrico e survival of the flattest

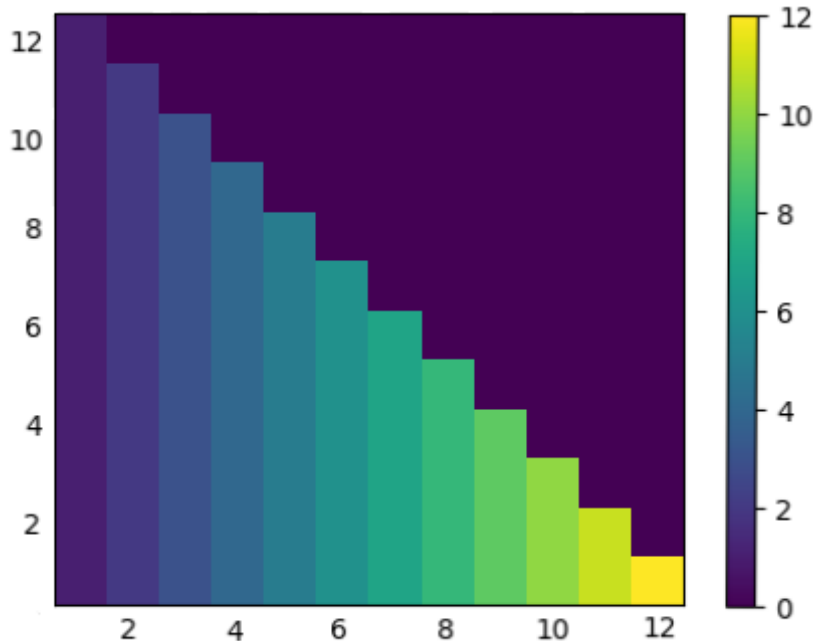


Figura 20: Fitness landscape: ad ogni colore corrisponde un valore di fitness diverso; si hanno così 12 sottoinsiemi che si differenziano per fitness e dimensione. (La simulazione è costruita in modo che gli individui non possano accedere alla parte più scura della griglia.)

Consideriamo ora un diverso modello, asimmetrico, in cui ad ogni tipo/casella sono associati due soli tratti, uno selettivo corrispondente alla coordinata lungo l'asse  $x$ , e uno neutrale corrispondente alla coordinata lungo l'asse  $y$ . Inoltre, detta  $\mu$  la probabilità di variazione complessiva di ogni individuo, si avrà che la probabilità di variazione in un tipo qualunque della stessa riga sarà pari a  $\mu/2$  (e analogamente la probabilità di variazione in un tipo qualunque della stessa colonna sarà pari a  $\mu/2$ ). Stiamo implicitamente assumendo che tutti i tipi con la stessa  $x$  (analogamente con la stessa  $y$ ) si trovino ad una sola variazione di distanza l'uno dall'altro.

Ogni colonna della griglia, oltre ad essere caratterizzata da un diverso valore del tratto selettivo e quindi da un diverso valore di fitness, contiene un diverso numero di tipi: il fitness landscape è tale che la dimensione dei sottoinsiemi neutri varia in modo opposto rispetto alla fitness. Questo modello ci permette di indagare tre diversi regimi, corrispondenti a valori di  $\beta$  e  $\mu$  diversi: il regime *redundancy-dominated* in cui il tratto più ridondante è il più presente nella popolazione, il regime *subottimale* in cui il tratto che si presenta maggiormente si trova ad un valore di fitness intermedio, e infine il regime di *survival of the fittest* che si verifica per alti valori della forza di selezione, quando tutta la popolazione si concentra nei tipi con fitness più elevata.

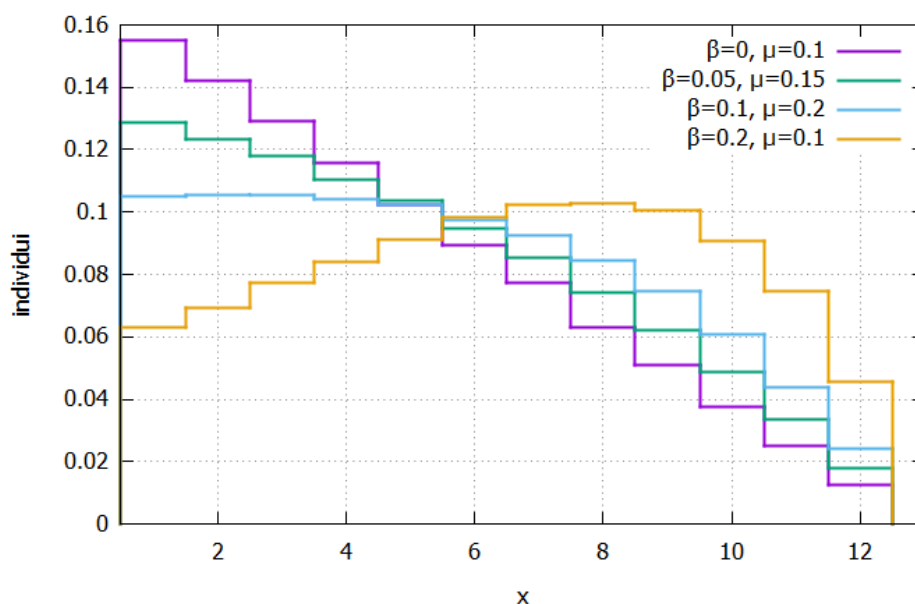


Figura 21: Distribuzione normalizzata del numero di individui in funzione del tratto selettivo  $x$  (legato ai valori di fitness) per diverse popolazioni.

Osservando la figura notiamo che per regimi in cui la forza di selezione è debole, i profili decrescono monotonicamente e si ha il massimo per  $x = 1$  (corrispondente a  $\lambda = 1$  per come è stata costruita la griglia), quindi la distribuzione dei valori di fitness è dominata dal sottoinsieme con dimensione maggiore (redundancy-dominated regime). Aumentando la forza di selezione e diminuendo la probabilità di variazione i profili smettono di essere monotoni a causa dell'emergenza di un picco per valori di  $x$  intermedi: ci troviamo nel regime subottimale, così chiamato perché il massimo corrisponde ad un valore di fitness intermedio e non a quello ottimale. Aumentando il rapporto  $\beta/\mu$  il picco si sposta verso valori di  $x$  sempre più grandi finché, per  $\beta/\mu \rightarrow \infty$  tutta la popolazione è concentrata nell'unico tipo con fitness massima (caso non rappresentato in figura per questioni grafiche). Quest'ultimo caso corrisponde al famoso paradigma di "survival of the fittest".

Riconoscere questa diversità di comportamenti all'interno delle popolazioni, può aiutare a migliorare la comprensione dell'evoluzione delle epidemie e di alcune patologie, perché l'informazione sulla forma della distribuzione dei patogeni in funzione del loro valore di fitness può essere utilizzata per progettare strategie di risposta atte a debellare la malattia.

Ad esempio, in una popolazione virale o batterica, supponiamo che  $x$  rappresenti la resistenza ad una medicina o ad un antibiotico, in modo che alti valori di  $x$  conferiscano una fitness più alta agli individui. Allora ci si potrebbe aspettare che la popolazione sia dominata dagli individui con la più alta resistenza, e questa ipotesi ci porterebbe a sviluppare una terapia che massimizza l'intervento sui tratti associati a valori più alti di resistenza. Tuttavia se il tratto selettivo  $x$  fosse accompagnato da un tratto neutrale  $y$  (che non ha effetto sulla resistenza del patogeno) e ci fosse un fitness landscape come quello da noi appena analizzato, allora la distribuzione potrebbe essere dominata da individui con una resistenza subottimale e la terapia si concentrerebbe sui tratti sbagliati. Il risultato complessivo potrebbe quindi involontariamente favorire gli individui subottimali e le loro mutazioni verso tipi con fitness più alta.

## Parte IV

# Conclusioni

In questa tesi abbiamo visto un possibile modello teorico per la descrizione dell'evoluzione di popolazioni generiche: il modello di dinamica RVS. Trattando il modello nell'ambito della meccanica statistica di non equilibrio abbiamo potuto appurare l'insorgenza di analogie tra i risultati della dinamica evolutiva e i processi termodinamici stocastici, e abbiamo verificato l'accordo qualitativo tra previsioni teoriche e risultati delle simulazioni numeriche attraverso la relazione di fluttuazione.

Nella seconda parte della tesi, di tipo computazionale, abbiamo utilizzato un particolare tipo di algoritmo evolutivo per simulare l'evoluzione di una popolazione in diversi regimi di forza di selezione e probabilità di variazione, andando così ad indagare come cambia la configurazione della popolazione all'equilibrio. I risultati delle simulazioni ci hanno permesso di mettere in evidenza un fenomeno molto importante sia in ecologia che nello studio di popolazioni batteriche o virali, che prende il nome di "survival of the flattest" in contrapposizione al darwiniano "survival of the fittest". Inoltre abbiamo studiato le distribuzioni di abbondanza delle specie (SAD) per diverse popolazioni in stato stazionario, trovando un tipico andamento iniziale a legge di potenza, con un esponente di scala comune alle diverse distribuzioni per forza di selezione e probabilità di variazione sufficientemente basse.

I possibili sviluppi futuri in questi ambiti sono molteplici e vanno dall'applicazione del modello RVS a popolazioni specifiche, all'estensione del modello al caso di un fitness landscape continuo, all'approfondimento dello studio sulla forma delle distribuzioni di abbondanza delle specie.

Infine facciamo notare, come illustrato nell'ultima sezione, che lo studio del fenomeno di "survival of the flattest" può risultare molto utile nel progettare farmaci atti a contrastare i patogeni in modo efficace, rivestendo quindi un ruolo importante anche a livello medico.

## A Appendice

### A.1 Processo di Markov

Un processo di Markov è caratterizzato dal fatto che la probabilità di transizione, che determina il passaggio del sistema ad un determinato stato, dipende solo dallo stato del sistema immediatamente precedente e non da come vi si è giunti. Matematicamente ciò corrisponde a richiedere che la probabilità condizionata sia determinata interamente dalla conoscenza della condizione più recente:

$$p(x_1, t_1; x_2, t_2; \dots | y_1, \tau_1; y_2, \tau_2; \dots) = p(x_1, t_1; x_2, t_2; \dots | y_1, \tau_1) \quad (36)$$

con

$$t_1 \geq t_2 \geq t_3 \geq \dots \geq \tau_1 \geq \tau_2 \geq \dots$$

Da ciò e dalla definizione di distribuzione di probabilità condizionata segue che:

$$p(x_1, t_1; x_2, t_2; x_3, t_3; \dots x_n, t_n) = p(x_1, t_1 | x_2, t_2) p(x_2, t_2 | x_3, t_3) p(x_3, t_3 | x_4, t_4) \dots p(x_{n-1}, t_{n-1} | x_n, t_n) p(x_n, t_n) \quad (37)$$

con

$$t_1 \geq t_2 \geq t_3 \geq \dots \geq t_{n-1} \geq t_n$$

che prende il nome di catena di Markov.

### A.2 Equazione di Chapman-Kolmogorov

In generale per probabilità congiunte di eventi vale la relazione:

$$\sum_B P(A \cap B \cap C) = P(A \cap C) \quad (38)$$

quando questa viene applicata a processi stocastici si ottengono le seguenti equazioni:

$$p(x_1, t_1) = \int dx_2 p(x_1, t_1; x_2, t_2) = \int dx_2 p(x_1, t_1 | x_2, t_2) p(x_2, t_2) \quad (39)$$

$$p(x_1, t_1 | x_3, t_3) = \int dx_2 p(x_1, t_1; x_2, t_2 | x_3, t_3) = \int dx_2 p(x_1, t_1 | x_2, t_2; x_3, t_3) p(x_2, t_2 | x_3, t_3) \quad (40)$$

Introducendo ora l'assunzione di Markov e ponendo  $t_1 \geq t_2 \geq t_3$  si ottiene l'equazione di Chapman-Kolmogorov:

$$p(x_1, t_1 | x_3, t_3) = \int dx_2 p(x_1, t_1 | x_2, t_2) p(x_2, t_2 | x_3, t_3) \quad (41)$$

Nel caso di variabili discrete si sostituisce la sommatoria con l'integrale e si ottiene:

$$P(n_1, t_1 | n_3, t_3) = \sum_{n_2} P(n_1, t_1 | n_2, t_2) P(n_2, t_2 | n_3, t_3) \quad (42)$$

### A.3 Determinazione dei contributi alla direzionalità evolutiva

Procedimento algebrico per separare i contributi alla direzionalità evolutiva che possono essere scritti come differenze di potenziale da quelli che non possono:

$$\begin{aligned} \mathcal{F}_{n \leftarrow n'} := \ln \frac{W_{n \leftarrow n'}}{W_{n' \leftarrow n}} &= \ln \left( \frac{\prod_{\gamma} \frac{1}{n_{\gamma}!} [\Lambda_{\gamma}(n')]^{n_{\gamma}}}{\prod_{\gamma} \frac{1}{n_{\gamma}'!} [\Lambda_{\gamma}(n)]^{n_{\gamma}'}} \right) = \ln \prod_{\gamma} \left( \frac{n_{\gamma}'!}{n_{\gamma}!} \right) + \ln \prod_{\gamma} \frac{[\sum_{\gamma'} f_{\gamma}(n') \pi_{\gamma \leftarrow \gamma'}(n') (n_{\gamma}')^{n_{\gamma}}]}{[\sum_{\gamma'} f_{\gamma}(n) \pi_{\gamma \leftarrow \gamma'}(n) (n_{\gamma}')^{n_{\gamma}}]} + \\ &+ \ln \prod_{\gamma} \left( \frac{N^{n_{\gamma}}}{N^{n_{\gamma}'}} \frac{[\sum_{\gamma'} f_{\gamma''}(n) \pi_{\gamma'' \leftarrow \gamma'}(n) (n_{\gamma}')^{n_{\gamma}'}]}{[\sum_{\gamma'} f_{\gamma''}(n') \pi_{\gamma'' \leftarrow \gamma'}(n') (n_{\gamma}')^{n_{\gamma}'}]} \right) = \end{aligned}$$

$$= \sum_{\gamma} \ln(n'_{\gamma}!) - \sum_{\gamma} \ln(n_{\gamma}!) + \zeta_{nn'} + N \ln \sum_{\gamma\gamma'} f_{\gamma}(n) \pi_{\gamma\leftarrow\gamma'}(n) n_{\gamma'} - N \ln \sum_{\gamma\gamma'} f_{\gamma}(n') \pi_{\gamma\leftarrow\gamma'}(n') n'_{\gamma'}$$

dove abbiamo usato la forma di  $W_{n'\leftarrow n}$  di eq. (6), il fatto che  $\ln \prod_{\gamma} (N^{n_{\gamma}}/N^{n'_{\gamma}}) = \ln(N^N/N^N) = 0$  e  $\prod_{\gamma} x^{n_{\gamma}} = x^{\sum_{\gamma} n_{\gamma}} = x^N$ .

#### A.4 Calcoli per il regime di piccole probabilità di variazione

Approssimazione per il kernel di transizione:

$$\begin{aligned} W_{n\leftarrow n'} &= N! \prod_{\gamma} \frac{1}{n_{\gamma}!} \left[ \frac{f_{\gamma}[(1-\mu)n'_{\gamma} + \mu \sum_{\gamma'(\neq\gamma)} \bar{\pi}_{\gamma\leftarrow\gamma'} n'_{\gamma'}]}{\sum_{\gamma''} f_{\gamma''} [(1-\mu)n'_{\gamma''} + \mu \sum_{\gamma'(\neq\gamma'')} \bar{\pi}_{\gamma''\leftarrow\gamma'} n'_{\gamma'}]} \right]^{n_{\gamma}} \simeq \\ &\simeq N! \left\{ \prod_{\gamma: n'_{\gamma} \neq 0} \frac{1}{n_{\gamma}!} \left[ \frac{f_{\gamma} n'_{\gamma}}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}} \right]^{n_{\gamma}} \right\} \left\{ \prod_{\gamma: n'_{\gamma} = 0} \frac{1}{n_{\gamma}!} \left[ \frac{\mu f_{\gamma} \sum_{\gamma'(\neq\gamma)} \bar{\pi}_{\gamma\leftarrow\gamma'} n'_{\gamma'}}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}} \right]^{n_{\gamma}} \right\} = \\ &= \mu^{\sum_{\gamma: n'_{\gamma} = 0} n_{\gamma}} N! \left\{ \prod_{\gamma: n'_{\gamma} \neq 0} \frac{1}{n_{\gamma}!} \left[ \frac{f_{\gamma} n'_{\gamma}}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}} \right]^{n_{\gamma}} \right\} \left\{ \prod_{\gamma: n'_{\gamma} = 0} \frac{1}{n_{\gamma}!} \left[ \frac{f_{\gamma} \sum_{\gamma'(\neq\gamma)} \bar{\pi}_{\gamma\leftarrow\gamma'} n'_{\gamma'}}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}} \right]^{n_{\gamma}} \right\} \end{aligned}$$

dove  $\bar{\pi}_{\gamma\leftarrow\gamma'} = \frac{\pi_{\gamma\leftarrow\gamma'}}{\mu}$  e nel primo passaggio abbiamo sfruttato la relazione seguente:

$$\sum_{\gamma'} \pi_{\gamma\leftarrow\gamma'} n_{\gamma'} = (1-\mu)n_{\gamma} + \mu \sum_{\gamma'(\neq\gamma)} \frac{\pi_{\gamma\leftarrow\gamma'}}{\mu} n_{\gamma'} = (1-\mu)n_{\gamma} + \mu \sum_{\gamma'(\neq\gamma)} \bar{\pi}_{\gamma\leftarrow\gamma'} n_{\gamma'}$$

Per  $\pi_{\gamma\leftarrow\gamma'}$  sufficientemente piccole,  $\mu \ll 1/N$ , è improbabile che più di una variazione abbia luogo in una generazione, perciò le transizioni tipiche  $n \leftarrow n'$  coinvolgono nessuna variazione o solo una:

$$0 \text{ variazioni: } W_{n\leftarrow n'} = N! \prod_{\gamma: n'_{\gamma} \neq 0} \frac{1}{n_{\gamma}!} \left[ \frac{f_{\gamma} n'_{\gamma}}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}} \right]^{n_{\gamma}}, \text{ e } n_{\gamma} = 0 \text{ se } n'_{\gamma} = 0, \forall \gamma$$

$$1 \text{ variazione: } W_{n\leftarrow n'} = \mu N! \left( \prod_{\gamma: n'_{\gamma} \neq 0} \frac{1}{n_{\gamma}!} \left( \frac{f_{\gamma} n'_{\gamma}}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}} \right)^{n_{\gamma}} \right) \frac{f_{\bar{\gamma}} \sum_{\gamma'(\neq\bar{\gamma})} \bar{\pi}_{\bar{\gamma}\leftarrow\gamma'} n'_{\gamma'}}{\sum_{\gamma'} f_{\gamma'} n'_{\gamma'}},$$

dove  $n_{\gamma} = 0$  se  $n'_{\gamma} = 0$ , per tutti i  $\gamma$  ad eccezione del tipo che si origina dalla variazione,  $\bar{\gamma}$ , per il quale  $n_{\bar{\gamma}=1}$  e  $n'_{\bar{\gamma}=0}$ .

Per caratterizzare il contributo non conservativo, riscriviamolo inizialmente come:

$$\zeta_{nn'} := \sum_{\gamma} \ln \frac{[(1-\mu)n'_{\gamma} + \mu \sum_{\gamma'(\neq\gamma)} \bar{\pi}_{\gamma\leftarrow\gamma'} n'_{\gamma'}]^{n_{\gamma}}}{[(1-\mu)n_{\gamma} + \mu \sum_{\gamma'(\neq\gamma)} \bar{\pi}_{\gamma\leftarrow\gamma'} n_{\gamma'}]^{n_{\gamma}}} \quad (43)$$

ora approssimiamo ogni termine logaritmico nel modo seguente:

$$\ln \frac{[(1-\mu)n'_{\gamma} + \mu \sum_{\gamma'(\neq\gamma)} \bar{\pi}_{\gamma\leftarrow\gamma'} n'_{\gamma'}]^{n_{\gamma}}}{[(1-\mu)n_{\gamma} + \mu \sum_{\gamma'(\neq\gamma)} \bar{\pi}_{\gamma\leftarrow\gamma'} n_{\gamma'}]^{n_{\gamma}}} \simeq \begin{cases} n_{\gamma} \ln n'_{\gamma} - n'_{\gamma} \ln n_{\gamma} & \text{se } n_{\gamma} \neq 0, n'_{\gamma} \neq 0 \\ n_{\gamma} \ln \mu + n_{\gamma} \ln \sum_{\gamma'} \bar{\pi}_{\gamma\leftarrow\gamma'} n'_{\gamma'} & \text{se } n_{\gamma} \neq 0, n'_{\gamma} = 0 \\ -n'_{\gamma} \ln \mu - n'_{\gamma} \ln \sum_{\gamma'} \bar{\pi}_{\gamma\leftarrow\gamma'} n_{\gamma'} & \text{se } n_{\gamma} = 0, n'_{\gamma} \neq 0 \end{cases} \quad (44)$$

Allora la forza non conservativa diventa:

$$\zeta_{nn'} \simeq -l_{nn'} \ln \mu + \sum_{\gamma: n_{\gamma} \neq 0, n'_{\gamma} \neq 0} [n_{\gamma} \ln n'_{\gamma} - n'_{\gamma} \ln n_{\gamma}] + \sum_{\gamma: n'_{\gamma} = 0} n_{\gamma} \ln \sum_{\gamma'} \bar{\pi}_{\gamma\leftarrow\gamma'} n'_{\gamma'} - \sum_{\gamma: n_{\gamma} = 0} n'_{\gamma} \ln \sum_{\gamma'} \bar{\pi}_{\gamma\leftarrow\gamma'} n_{\gamma'} \quad (45)$$

dove  $l_{nn'} := \sum_{\gamma: n_{\gamma} = 0} n'_{\gamma} - \sum_{\gamma: n'_{\gamma} = 0} n_{\gamma}$ .

## A.5 Il modello a quasispecie

Il modello a quasispecie è stato introdotto da Manfred Eigen nel 1971 e descrive popolazioni di individui asessuati tramite equazioni deterministiche. Il modello si concentra sulla competizione tra selezione naturale e mutazioni casuali e viene spesso utilizzato per descrivere l'evoluzione di virus a RNA. L'equazione fondamentale di questa teoria è:

$$\dot{x}_i(t) = [W_{ii} - E(t)]x_i(t) + \sum_{j \neq i} W_{ij}x_j(t) \quad (46)$$

dove  $x_i(t)$  è la concentrazione della sequenza  $i$  all'interno della popolazione,  $W_{ij} = f_j \pi_{ij}$  è il prodotto del tasso di riproduzione di  $j$  e della probabilità di mutazione  $\pi_{ij}$  da  $j$  a  $i$  e  $E(t)$  esprime la produzione totale di nuove sequenze:

$$E(t) = \sum_k \sum_j W_{kj}x_j(t) \quad (47)$$

L'equazione (46) può essere mappata in un sistema lineare e risolta diagonalizzando la matrice  $W$ , purchè i suoi elementi vengano considerati costanti.

## Riferimenti bibliografici

- [1] R. Rao, S. Leibler, Evolutionary dynamics, evolutionary forces, and robustness: A nonequilibrium statistical mechanics perspective, *Proc. Natl. Acad. Sci. USA* 119, 2022.
- [2] L. Miele, R.M.L. Evans, S. Azaele, Redundancy-selection trade-off in phenotype-structured populations, *Journal of Theoretical Biology* 531, 2021.
- [3] C. O Wilke, Quasispecies theory in the context of population genetics, *BMC Evolutionary Biology* 5:44, 2005.
- [4] B. J McGill et al., Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework, *Ecology letters* 10:995-1015, 2007.
- [5] L. Loewe, *Systems in Evolutionary Systems Biology*, Encyclopedia of Evolutionary Biology, vol 4, pp 297–318, Academic Press, Oxford, UK, 2016.
- [6] D. Simon, *Evolutionary Optimization Algorithms*, John Wiley & Sons, Hoboken, New Jersey, 2013.
- [7] C.W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Science*, Springer-Verlag, Berlin, 1985.