UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea Triennale in Fisica

Tesi di Laurea

Uno studio statistico delle mutazioni di sequenze di RNA virale

Relatore

Prof. Marco Favretti

Correlatore

Prof. Samir Suweis

Laureanda

Laura Dalla Pozza

Anno Accademico 2022/2023

Indice

Abstract

EN

This thesis consists of a statistical analysis (both theoretical and numerical) of mutations in viral RNA sequences, using the Sars-CoV-2 database. The RNA sequences are compared by using the typical tools of information theory: entropy, relative entropy, and reciprocal information. Eventually, a simple model of the mutation mechanism is developed by using the Markov chains model and the Ehrenfest urn model.

IT

La tesi consiste in un' analisi statistica (teorica e numerica) di mutazioni di sequenze di RNA virale, usando il database del Sars-CoV-2. Si confrontano le sequenze di RNA usando strumenti tipici della teoria dell'informazione, quali entropia, entropia relativa e mutua informazione. Infine, si vuole costruire un modello semplice del meccanismo delle mutazioni attraverso il modello delle catene di Markov e il modello delle urne di Ehrenfest.

Introduzione

In biologia le informazioni per la genesi di qualsiasi organismo sono contenute negli acidi nucleici: il DNA e l'RNA, a loro volta composti da una sequenza ordinata di nucleotidi. Esistono cinque diversi tipi di nucleotidi, che differiscono solo per la base azotata: adenina, guanina, citosina, timina (nel DNA) o uracile (nell'RNA). Pertanto, per quanto riguarda le informazioni in esso contenute, l'acido nucleico può essere considerato come una stringa in cui in ogni posizione c'è una scelta di $m = 4$ possibili caratteri: A, C, G, T nel DNA e A, C, G, U nell'RNA. Quando gli organismi si riproducono le informazioni genetiche vengono trasmesse alle generazioni successive, ma la copia del contenuto genetico non è perfetta. Ciò può portare a delle mutazioni, che possono essere viste come la sostituzione di una base con un'altra. Le mutazioni avvengono con piccole variazioni nella sequenza, le quali, con il passare delle generazioni, si accumulano e portano ad una popolazione di organismi diversa e sempre in evoluzione. Le mutazioni che si susseguono nelle stringhe di nucleotidi avvengono in modo stocastico e sono quindi descritte da una distribuzione di probabilità.

In questa tesi considereremo sequenze virali di RNA del Sars-CoV-2 di lunghezza N=29903 nucleotidi, dove $x = (x_1, x_2...x_N)$ è la sequenza di riferimento di RNA del virus e $y = (y_1, y_2...y_N)$ è una mutazione di x, tali sequenze sono state prese dal database in [1]. Siccome in tale database la sequenza virale viene indicata con le basi A,C,G,T, sebbene si tratti di RNA, anche noi adotteremo tale convenzione, quindi $x_i, y_i \in \{A, C, G, T\}$.

Oggetto di questa tesi è l'analisi delle mutazioni rispetto ad una sequenza iniziale attraverso strumenti di meccanica statistica, con l'intenzione di trovare un modello che descriva la natura delle mutazioni virali nelle sequenze di RNA. Per fare ciò si è partiti considerando solamente la frequenza di ogni base, trascurandone dunque la disposizione nelle sequenze. Dopo tale semplificazione è stato osservato che l'entropia in funzione dell'entropia relativa delle sequenze non solo tende a calare, ma in certi casi è addirittura compatibile al minimo valore concesso a meno di piccoli intervalli. Con l'intento di realizzare un modello che descrivesse tale fenomeno, è stato creato per il sistema delle mutazioni un modello Markoviano a tempo discreto, dal quale è stato possibile, tramite l'approssimazione di mean field, ricavare un'ODE che descrivesse l'evoluzione delle frequenze delle basi associate alle sequenze mutate in modo deterministico. Una volta ottenuto il modello è stato possibile, attraverso il database [1], ricavare le probabilità di transizione tra le basi ed il vettore di frequenze q relativo alla sequenza di riferimento, che inseriti nell'ODE hanno portato ad una soluzione deterministica che descrive l'andamento delle frequenze associate alle sequenze in evoluzione. Infine, è stato effettuato un confronto tra l'entropia ricavata dalle soluzioni dell'ODE e il minimo di entropia concesso alle sequenze del database, entrambe in funzione dell'entropia relativa, verificandone quindi una corrispondenza nei rispettivi andamenti.

Entropia di sequenze virali

Questa tesi si interessa di studiare la sequenza di RNA virale del Sars-CoV-2 ed il suo modo di mutare. Nel database del Sars-CoV-2 [1] si trova una sequenza iniziale $x = (x_1, ..x_N)$ fissata, in quanto rappresenta il patrimonio genetico del virus di partenza, e circa 5000 mutazioni, che denoteremo come $y = (y_1, ..y_N)$. Esse sono derivanti da piccole variazioni che si sono susseguite a partire dalla sequenza iniziale, come detto in precedenza. Al fine di ridurre il gran numero di gradi di libertà del problema sarà considerato solo il vettore delle frequenze relativo ad una specifica sequenza di RNA: $p = (p_A, p_C, p_G, p_T)$, le cui componenti sono date da:

$$p_i = \frac{n_i(y)}{N} \qquad i \in \{A, C, G, T\} \tag{1.1}$$

in cui si definisce $n_i(y) := \#(y_\alpha = i)$ il numero di volte in cui viene contata ciascuna base nella sequenza di lunghezza N. Si può osservare che, dalla conoscenza della sequenza iniziale x, le frequenze associate ad essa, che denoteremo con il vettore $q := q(x)$, sono note. Infatti, attraverso il database [1] è possibile ricavare $q = (0.299, 0.184, 0.196, 0.321)$. Le frequenze relative alle mutazioni y invece saranno denotate dal vettore $p := p(y)$, e sono ovviamente variabili a seconda del modo di mutare della sequenza iniziale. Si noti che, quando si passa alle frequenze di una specifica sequenza, l'informazione relativa all'ordine delle basi viene persa. Infatti, due sequenze che si ottengono attraverso una permutazione delle basi l'una dell'altra hanno lo stesso vettore di frequenze. A questo punto è utile definire l'entropia di Shannon e l'entropia relativa, due quantità che svolgono un ruolo chiave nella meccanica statistica e che saranno alla base dell'analisi effettuata in questo capitolo:

Entropia di Shannon

Si può dimostrare che esiste una funzione unica $h(p)$, teorizzata da Claude Shannon, la quale misura l'incertezza di una distribuzione discreta $p = (p_1, ..p_n)$ di probabilità:

$$h(p_1, ..p_n) = -\sum_{i=1}^{n} p_i \ln p_i \tag{1.2}$$

L'entropia di Shannon ha le seguenti proprietà:
i) Può essere estesa con continuità in $0 \ln 0 = 0$, quindi essere definita in $\mathbb{R}^n_+$
ii) È C^∞ nel suo dominio
iii) La sua matrice hessiana è sempre negativa: $Hess(h(p)) = -Diag[\frac{1}{p_i}]$, quindi $h(p)$ è una funzione concava che ha il suo massimo in $p_i = \frac{1}{n}$.

Entropia relativa

L'entropia relativa quantifica le differenze tra due distribuzioni di probabilità. In particolare sarà utile per quantificare le variazioni tra le frequenze della sequenza iniziale $q(x)$ e le frequenze di una qualsiasi mutazione $p(y)$:

$$D(p|q) = \sum_i p_i \ln \frac{p_i}{q_i}, \qquad i \in \{A, C, G, T\} \tag{1.3}$$

L'entropia relativa ha le seguenti proprietà:
i) $D(p|q) \geq 0 \quad \forall p, q$
ii) $D(p|q) = 0 \leftrightarrow p \equiv q$
iii) $D(p|q)$ è una funzione convessa quando q è fissato.

1.1 L'entropia delle sequenze decresce

Si vuole studiare la modalità di mutare del virus in esame utilizzando l'entropia e l'entropia relativa appena definite, con lo scopo di capire l'andamento dell'entropia delle sequenze. E' stato quindi tracciato un grafico delle sequenze contenute nel database del Sars-CoV-2 che mostra il valore dell'entropia al variare dell'entropia relativa di tali sequenze mutate rispetto a q, utilizzando le formule (1.2) e (1.3).

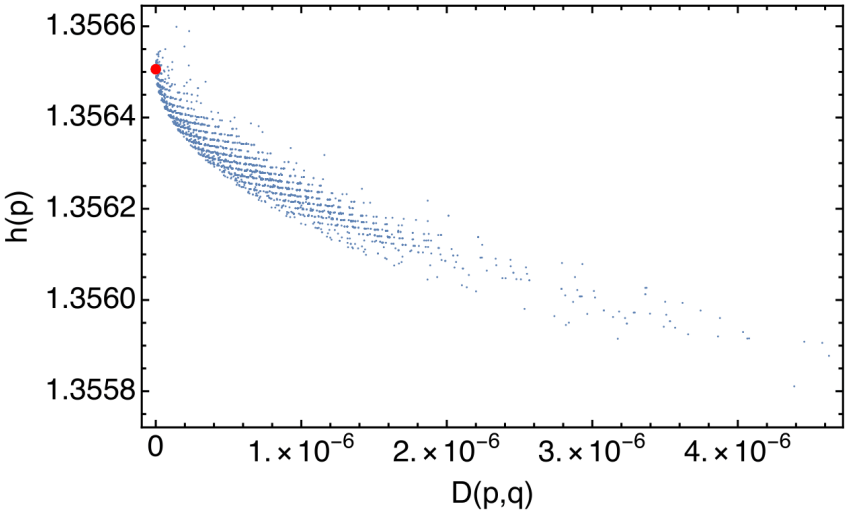

Figura 1.1: Entropia associata alle mutazioni

Ciò che si può notare dal grafico è che, mentre le sequenze evolvono e si allontanano gradualmente da quella iniziale x, l'andamento dell'entropia associata ad esse diminuisce, contrariamente a quello che ci si sarebbe potuti aspettare dal principio di massima entropia nel caso in cui le mutazioni fossero state tutte equiprobabili. Ciò significa che il processo di mutazione virale non è casuale. Ora è naturale chiedersi se esista un minimo di $h(p)$ per ogni arbitrario valore dell'entropia relativa $D(p|q) = d$, e se questo minimo venga raggiunto dai punti della figura (1.1).

1.2 Geometria del problema di estremo vincolato

La domanda appena posta sull'esistenza di un minimo per l'entropia è, dal punto di vista matematico, un problema di estremo vincolato. Uno strumento matematico utile che può essere applicato al nostro problema è il metodo dei moltiplicatori di Lagrange, trattato in [2]. Questo metodo consente di ridurre i punti stazionari di una funzione scalare vincolata, nel nostro caso la funzione è l'entropia $h(p)$ ed i vincoli sono $\sum_i p_i = 1$ e $D(p|q) = d$, ai punti stazionari di $G(p, \lambda, \mu)$ non vincolata. Quest'ultima è chiamata funzione di Lagrange. In particolare $G(p, \lambda, \mu)$ è della forma:

$$G(p, \lambda, \mu) = h(p) - \lambda(D(p|q) - d) - \mu(\sum_i p_i - 1) \tag{1.4}$$

dove λ e μ sono i moltiplicatori di Lagrange.

Teorema 1. *Siano:* $h(p) : \mathbb{R}^4 \to \mathbb{R}$ *una funzione scalare, nel nostro caso l'entropia*

$g(p) : \mathbb{R}^4 \to \mathbb{R}^2$ *la funzione che rappresenta i vincoli* $\sum_i p_i - 1 = 0$, $D(p|q) - d = 0$.

Sia inoltre $rk(dg) = 2$, ovvero i vincoli del problema sono linearmente indipendenti. Si deve minimizzare la funzione $h(p)$ soggetta al vincolo $g(p)$: sia $\hat{p}$ tale che $h(\hat{p}) = 0$ e $g \in C^2$, $h \in C^2$ attorno a $\hat{p}$. Data $G(p, \lambda, \mu) : \mathbb{R}^{4+2} \to \mathbb{R}$ la lagrangiana definita come (1.8), supponiamo che esistano $\hat{\lambda}$, $\hat{\mu}$ tali che

$$\nabla_p G(p, \lambda, \mu)_{|p=\hat{p}, \lambda=\hat{\lambda}, \mu=\hat{\mu}} = 0$$

Allora se

$$z \cdot \nabla^2_{pp} G(\hat{p}, \hat{\lambda}\hat{\mu}) z > 0 \quad \forall z \neq 0, \quad z \in \mathbb{R}^4, \quad z \in ker(dh(\hat{p}))$$

$\hat{p}$ è un minimo per $h(p)$.

Si osservino inoltre le seguenti quantità della meccanica statistica, le quali torneranno utili in seguito:

$$E_i = \ln \frac{1}{q_i} \tag{1.5}$$

$$\mathbb{E}_p(E) = \sum_i p_i E_i \tag{1.6}$$

$$\sum_i q_i^\beta =: Z(\beta) \tag{1.7}$$

Dove la (1.7) è detta funzione di partizione. E' possibile riscrivere l'entropia (1.2) in funzione di (1.3) ed (1.6) nel seguente modo:

$$h(p) = -D(p|q) + \mathbb{E}_p(E) \tag{1.8}$$

Si utilizzi ora il Teorema 1 per trovare i punti stazionari dell'entropia e la loro natura per arbitrari valori di d, con lo scopo di confrontarli con il grafico ottenuto dai punti del database del Sars-CoV-2 in figura (1.1).

Condizione necessaria per massimi o minimi vincolati

La condizione necessaria del teorema dei moltiplicatori di Lagrange è

$$\nabla_p G(p, \lambda, \mu)_{|p=\hat{p}, \mu=\hat{\mu}, \lambda=\hat{\lambda}} = 0 \tag{1.9}$$

dove $\hat{p}$ è il punto stazionario dell'entropia. Calcolando la prima derivata si ottiene:

$$\frac{\partial G}{\partial p_i}(p, \lambda, \mu) = -(\lambda - 1)(\ln p_i + 1) - \lambda E_i - \mu = 0 \ \ \forall i \tag{1.10}$$

Risolvendo queste equazioni e utilizzando le sostituzioni

$$\beta = \frac{\lambda}{\lambda + 1}, \qquad c = (\frac{\mu}{\lambda + 1} + 1) \tag{1.11}$$

si possono facilmente ricavare i punti stazionari $\hat{p}$ per l'entropia $h(p)$:

$$\hat{p_i} = e^{-\beta E_i} e^{-c} \tag{1.12}$$

In questo calcolo β e c dipendono dai vincoli, per tale motivo è possibile ricavare questi valori da tali vincoli.

Il valore di e^{-c} si ricava facilmente imponendo il vincolo $\sum_i p_i = 1$:

$$\begin{aligned} 1 &= e^{-c} \sum_i e^{-\beta E_i} \\ 1 &= e^{-c} Z(\beta) \\ e^{-c} &= \frac{1}{Z(\beta)} \end{aligned} \tag{1.13}$$

Sostituendo questo risultato nell' (1.12) si ottiene infine il punto stazionario $\hat{p}$ in funzione del parametro β:

$$\hat{p_i} = \frac{e^{-\beta E_i}}{Z(\beta)} = \frac{q_i^\beta}{\sum_i q_i^\beta} \tag{1.14}$$

Il valore di β può essere ricavato dal vincolo $D(\hat{p}|q) - d = 0$ attraverso $\lambda(d)$, calcolo che approfondiremo in seguito. Siamo ora interessati a capire quando si tratta di un massimo e quando di un minimo. A tal fine è utile studiare la condizione sufficiente del teorema dei moltiplicatori.

Condizione sufficiente per massimi o minimi vincolati

La condizione sufficiente per un massimo o un minimo è:

$$z \cdot H_G(\hat{p}, \hat{\lambda}, \hat{\mu}) z \begin{cases} > 0 & \forall z : dh(\hat{p})z = 0 \to \hat{p} \;\; e' \;\; min \\ < 0 & \forall z : dh(\hat{p})z = 0 \to \hat{p} \;\; e' \;\; max \end{cases} \tag{1.15}$$

Dove $H_G(\hat{p}, \hat{\lambda}, \hat{\mu})$ è la matrice hessiana della funzione lagrangiana calcolata nel suo punto stazionario. Essa è ottenibile calcolando le derivate seconde miste:

$$\frac{\partial}{\partial p_j}(\frac{\partial G}{\partial p_i}(p, \lambda, \mu)) = -(\lambda + 1)\frac{\partial}{\partial p_j}(\ln p_i + 1) = -(\lambda + 1)\frac{1}{p_i}\delta_{ij} \tag{1.16}$$

La matrice hessiana della funzione lagrangiana pertanto sarà:

$$H_G(p, \lambda, \mu) = \nabla^2 G(p, \lambda, \mu) = -(\lambda + 1) \begin{pmatrix} \frac{1}{p_1} & & \\ & \ddots & \\ & & \frac{1}{p_k} \end{pmatrix} \tag{1.17}$$

Si può osservare facilmente che la matrice $Diag[\frac{1}{p_1}, .. \frac{1}{p_k}]$ è sempre definita positiva, in quanto è una matrice diagonale i cui autovalori sono tutti positivi. Pertanto, l'unico termine rilevante per lo studio del segno della matrice hessiana calcolata in un determinato punto sarà $-(\lambda + 1)$. Con il fine di studiare la condizione necessaria di estremo vincolato bisogna osservare la matrice H_G calcolata nel punto stazionario $H_G(\hat{p}, \hat{\lambda}, \hat{\mu})$, la quale sarà definita positiva(negativa) se $(\hat{\lambda}+1) < (>)0$, dove $\hat{\lambda} = \lambda(d)$ dipende dal valore del vincolo. Quindi lo studio del segno di $H_G(p_i = \hat{p_i}, \lambda = \hat{\lambda}, \mu = \hat{\mu})$ dipende solamente dal valore di $\hat{\lambda} = \lambda(d)$. Utilizzando ora la definizione di β in (1.11) e calcolando il suddetto parametro nel punto stazionario $\hat{\lambda} = \lambda(d)$ si ottiene:

$$\beta(\lambda(d)) = \beta(\hat{\lambda}) = \frac{\hat{\lambda}}{\hat{\lambda} + 1} \tag{1.18}$$

In questo modo è stato ottenuto il valore di $\beta(d)$ in funzione del vincolo. Come conseguenza dello studio (1.17) e del risultato (1.18) che descrive $\beta(\hat{\lambda})$, è possibile calcolare lo studio del segno per la matrice hessiana di $G(\hat{p}, \hat{\lambda})$ in funzione di d attraverso $\beta(\hat{\lambda}) = \beta(d)$:

$$\nabla^2 G(p, \lambda, \mu)_{|p=\hat{p}, \mu=\hat{\mu}, \lambda=\hat{\lambda}} \begin{cases} > 0 & se \quad \beta(d) > 1 \to min \\ < 0 & se \quad \beta(d) < 1 \to max \end{cases} \tag{1.19}$$

Riprendendo il calcolo che era stato lasciato sospeso in (1.14), è possibile ricavare il valore di $\beta(d)$ imponendo il vincolo $D(\hat{p}|q) = d$ e cercando le soluzioni della seguente equazione:

$$\begin{aligned} d &= -\beta \sum_i \hat{p}_i E_i - \ln Z(\beta) + \mathbb{E}_{\hat{p}(E)} \\ &= (1-\beta) E_{\hat{p}}(E) - \ln Z(\beta) =: f(\beta) \end{aligned} \tag{1.20}$$

Risolvendo quindi l'equazione $f(\beta) = d$ si possono trovare i valori del parametro β per un'entropia relativa d fissata. Una volta trovati questi valori sarà possibile sostituirli nella (1.14) così da trovare definitivamente i punti stazionario $\hat{p}$ e conoscerne la natura grazie allo studio (1.19). A questo scopo studieremo la funzione $f(\beta)$ facendo particolare attenzione alla sua invertibilità.

Studio della funzione $f(\beta)$

$$f(\beta) = (1-\beta) \frac{\sum_i e^{-\beta_i E_i} E_i}{\sum_i e^{-\beta E_i}} - \ln \sum_i e^{-\beta E_i} \tag{1.21}$$

Utilizzando la relazione (1.5) è possibile riformulare $f(\beta)$ nel seguente modo:

$$f(\beta) = (1-\beta) \frac{\sum_i (\ln \frac{1}{q_i}) q_i^\beta}{\sum_k q_k^\beta} - \ln \sum_i q_i^\beta \tag{1.22}$$

In questa forma si può facilmente verificare che $f(1) = 0$, ovvero in $\beta = 1$ la funzione raggiunge il minimo possibile per l'entropia relativa, che ricordiamo essere non negativa. Questo significa che quando $\beta = 1$ si ha $\hat{p} = q$. La funzione è globalmente invertibile se $f'(\beta) \neq 0 \quad \forall \beta$. Calcolando la derivata prima si ottiene:

$$f'(\beta) = (1-\beta) \left[\left(\frac{\sum_i (\ln \frac{1}{q_i}) q_i^\beta}{\sum_k q_k^\beta} \right)^2 - \left(\frac{\sum_i (\ln \frac{1}{q_i})^2 q_i^\beta}{\sum_k q_k^\beta} \right) \right] \tag{1.23}$$

e tornando alla scrittura iniziale:

$$\begin{aligned} f'(\beta) &= (1-\beta) \left[\sum_i p_i \left(E_i^2 \right) - \left(\sum_i p_i E_i \right)^2 \right] \\ &= (\beta - 1) \left[\mathbb{E}[E^2] - (\mathbb{E}[E])^2 \right] \\ &= (\beta - 1)[var[E]] \end{aligned} \tag{1.24}$$

La funzione non è invertibile siccome non è iniettiva, infatti la sua derivata prima si annulla in $\beta = 1$. Inoltre, poiché la varianza è sempre una quantità positiva, si può dedurre che $\beta = 1$ è anche l'*unico* punto in cui la derivata prima assume un valore nullo. È quindi ragionevole assumere che negli intervalli che precedono e che seguono $\beta = 1$, la funzione $f(\beta)$ sia monotona, dato che è continua.

In seguito a questo studio analitico della funzione è stato tracciato il grafico di $f(\beta)$ con il valore $q = (0.299, 0.184, 0.196, 0.321)$, ovvero il vettore delle frequenze associato alla sequenza iniziale x nell'ordine (A,C,G,T), ottenuto dal database [1]. Il primo grafico mostra la funzione vicino al valore $\beta = 1$, il secondo, invece, in $(-50, 50)$, con lo scopo di osservarne l'andamento al variare di β.

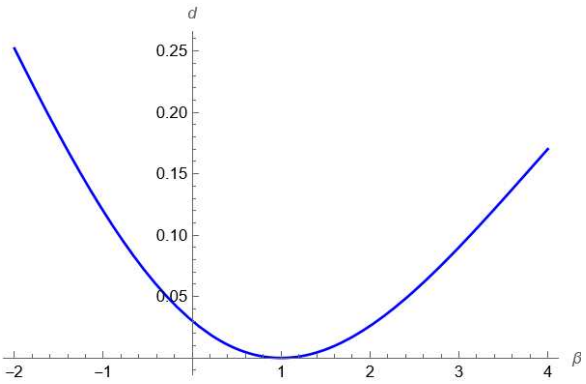

Figura 1.2: $f(\beta)$ vicino $\beta = 1$

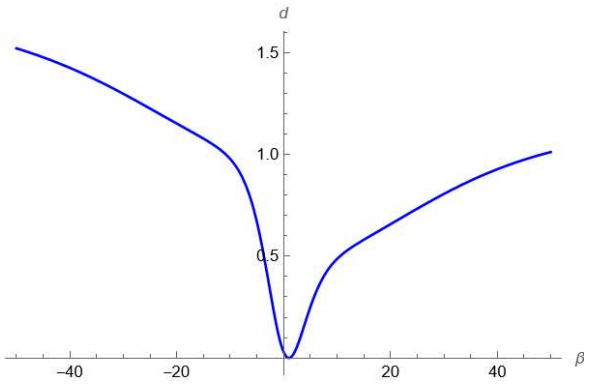

Figura 1.3: $f(\beta)$ in $(50, -50)$

Dai grafici risulta evidente che per ogni valore fissato dell'entropia relativa $d \neq 0$ ci sono due possibili soluzioni per $\beta(d)$. Definiamo questi valori:

$$\beta^+(d) := \beta(d) > 1, \qquad \beta^-(d) := \beta(d) < 1 \tag{1.25}$$

Sostituendo questi valori di $\beta(d)$ nella (1.14) è possibile trovare le frequenze che stazionarizzano l'entropia, ed in seguito capirne la natura attraverso lo studio del segno dell'hessiana in (1.19). Si ottiene:

$$\hat{p}_{imin}(d) = \frac{q_i^{\beta^+(d)}}{\sum_i q_i^{\beta^+(d)}}, \qquad \hat{p}_{imax}(d) = \frac{q_i^{\beta^-(d)}}{\sum_i q_i^{\beta^-(d)}} \tag{1.26}$$

Infine, per mezzo di questi valori, si può calcolare il massimo e il minimo valore di entropia concessi alle sequenze che possiedono un arbitrario valore di entropia relativa d.

$$\begin{aligned} h(\hat{p}(d)) &= -\sum_i \hat{p}_i(d) \ln \hat{p}_i(d) \\ &= \beta(d) \sum_i \hat{p}_i E_i + \ln Z(\beta(d)) \\ &= \beta(d) E_{\hat{p}}(E) + \ln z(\beta(d)) \end{aligned} \tag{1.27}$$

Utilizzando la formula (1.27) calcolata nelle soluzioni in (1.26) si ottengono i valori di massimo e minimo della funzione entropia per arbitrari valori di d fissati:

$$h_{min}(d) = h(\hat{p}_{min}) = h(\beta^+(d)), \qquad h_{max}(d) = h(\hat{p}_{max}) = h(\beta^-(d)) \tag{1.28}$$

E' stato in seguito tracciato il grafico di $h(\beta^+(d))$ ed $h(\beta^-(d))$ al variare di d. Ne riportiamo la sua versione in $d \in [0, 0.5]$ con lo scopo di osservarne l'andamento complessivo, ed in $d \in [0, 5 \cdot 10^{-6}]$ per poter apprezzare la zona con il medesimo ordine di grandezza di entropia relativa dei punti in figura (1.1), quindi la zona a cui il database fa riferimento. Si noti inoltre che $h(0) = 1.356$ è il valore di entropia associato alla frequenza iniziale q, coerentemente con quanto si osserva in figura (1.1)

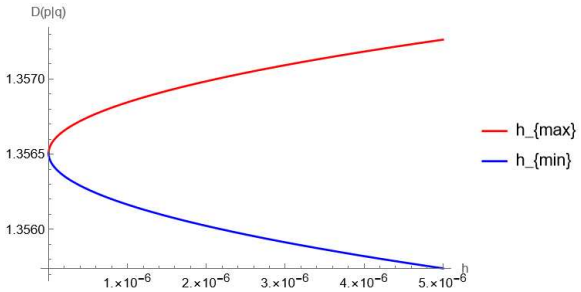

Figura 1.4: h_{max} e h_{min} scala del database

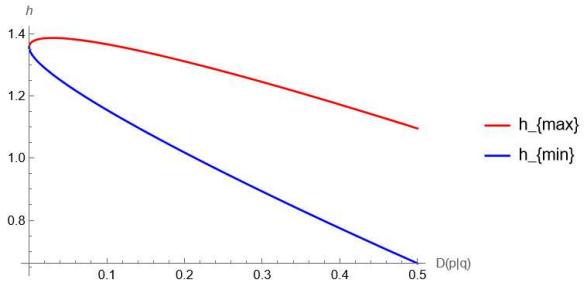

Figura 1.5: andamento di h_{max} e h_{min}

E' possibile osservare che ogni mutazione della sequenza iniziale avrà un certo valore di entropia relativa d calcolabile a partire dalla frequenza $p(y)$ associata alla mutazione, ed un valore di entropia di Shannon che deve trovarsi tra il massimo ed il minimo valore di entropia calcolati in (1.28) per un fissato valore di entropia relativa d della mutazione. Pertanto, le mutazioni osservate nel database del Sars-CoV-2 devono necessariamente trovarsi nella regione di grafico compresa tra $h_{min}(d)$ ed $h_{max}(d)$. Nel prossimo paragrafo si vuole svolgere un confronto più accurato tra i punti rappresentati in figura (1.1) e le soluzioni di entropia massima e minima in funzione di d ottenute in (1.28).

1.3 Confronto tra dati e teoria

Ora che sono stati ricavati i valori di massimo e di minimo concessi all'entropia al variare dell'entropia relativa, è utile confrontarli con i punti in figura (1.1). Sovrapponendo quindi tale figura ai grafici di $h_{max}(d)$ e $h_{min}(d)$ in (1.4) si osserva che i punti rappresentanti le mutazioni giacciono nella regione prevista e, in particolare, si avvicinano notevolmente al grafico di h_{min}. Questo avvenimento inaspettato ha una notevole implicazione, ovvero che le mutazioni non sono casuali: devono infatti necessariamente avvenire secondo delle regole e delle modalità piuttosto efficienti.

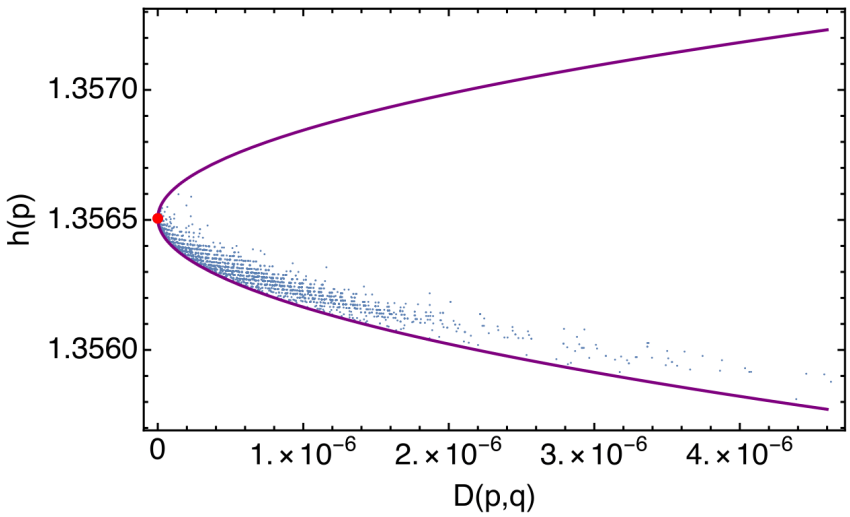

Figura 1.6: Confronto tra i dati in fig. (1.1) e curva teorica in fig.(1.5)

Osservando che, per d fissata, l'entropia di alcune sequenze assume valori compatibili con $h_{min}(d)$ a meno di piccoli intervalli, si ipotizza che queste possano essere descritte da $h_{min}(d) = h(\beta^+(d))$. Ci si domanda quindi come evolva il rapporto tra due frequenze mutate p_i e p_j utilizzando appunto il valore $\beta^+(d)$ al variare dell'entropia relativa. Attraverso le formule (1.5) e (1.26) il calcolo tra i rapporti delle frequenze di sequenze mutate in funzione di d è dato da:

$$\frac{p_i}{p_j} = \frac{e^{-\beta E_i}}{Z(\beta(d))} \frac{Z(\beta(d))}{e^{-\beta E_j}} = (\frac{q_i}{q_j})^{\beta(d)} \tag{1.29}$$

Quindi, avendo le frequenze iniziali delle quattro basi $q = (0.299, 0.184, 0.196, 0.321)$ relative alla sequenza iniziale x nell'ordine A,C,G,T, e conoscendo il valore di $\beta(d)$ è possibile calcolare il valore del rapporto tra le frequenze p_i delle mutazioni in funzione di d via via che queste evolvono. Come detto precedentemente, si utilizza il valore di $\beta^+(d)$ in quanto modellizza bene le mutazioni che sono vicine al minimo dell'entropia. Si può osservare che quando due frequenze di partenza q_i e q_j non sono uguali, il loro rapporto $(\frac{q_i}{q_j})^{\beta^+(d)}$ con $\beta^+ > 1$ tende ad allontanarsi da 1 all'aumentare di d. Mano a mano che il valore dell'entropia relativa aumenta, quindi con l'evoluzione della sequenza iniziale, le frequenze delle basi tendono sempre meno alla distribuzione uniforme. In particolare un valore iniziale di $\frac{q_i}{q_j} < 1$ tenderà a diminuire, mentre un valore iniziale di $\frac{q_i}{q_j} > 1$ tenderà ad aumentare ulteriormente, ed in entrambi i casi il divario tra le cardinalità di due diverse basi aumenta anziché mitigarsi. Osserviamo per curiosità che nel caso in cui fosse stato utilizzato $\beta^- < 1$, in accordo con il principio di massima entropia, i valori iniziali di $(\frac{q_i}{q_j} \neq 1)^{\beta^-(d)}$ con l'evolversi delle mutazioni si sarebbero avvicinati sempre più a $\frac{p_i}{p_j} = 1$, quindi alla distribuzione uniforme.

Si vuole a questo punto osservare l'andamento del rapporto tra $\frac{p_i}{p_j} = (\frac{q_i}{q_j})^{\beta(d)}$ utilizzando il valore che descrive le mutazioni $\beta(d)^+$ ed il vettore q ricavato dal database. A tal proposito è stato eseguito un grafico che mostra l'andamento del rapporto tra le frequenze p_A, p_C, p_G, p_T prese a due a due al variare del valore dell'entropia relativa d, in modo che le frequenze iniziali soddisfino $\frac{q_i}{q_j} > 1$ per poter apprezzare meglio come l'andamento si discosti da 1.

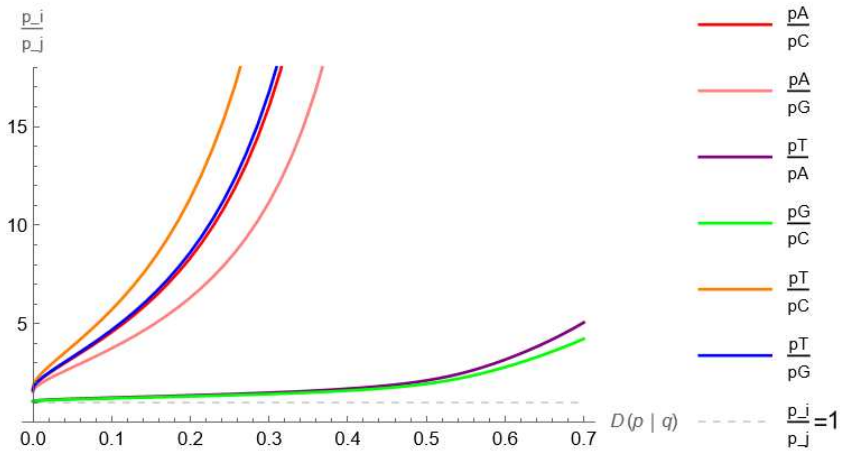

Figura 1.7: $\frac{p_i}{p_j}(d)$ in funzione dell'entropia relativa

Come previsto, all'aumentare dell'entropia relativa il rapporto tra le $\frac{p_i}{p_j}$ tende ad allontanasi da 1 $\forall ij \in \{A, C, G, T\}$, ovvero dal valore atteso nel caso di distribuzione uniforme.

Dinamica Markoviana delle mutazioni

Questo capitolo mira a comprendere il processo di evoluzione delle mutazioni y contenute nel database [1] a partire dalla sequenza iniziale x attraverso la costruzione di un modello. Il modello che sarà utilizzato per descrivere il processo $x \to y$ sarà quello delle catene di Markov, a sua volta ispirato a quello delle urne di Ehrenfest. In seguito descriviamo questi due modelli per poi applicarli al problema in esame.

2.1 Il modello delle urne di Ehrenfest

Questo modello è stato proposto per descrivere, attraverso la meccanica statistica, lo scambio di calore di due sistemi a contatto aventi due diverse temperature [3]. Inizialmente si considerino N particelle complessivamente contenute in due urne e si suppone che al tempo $n \geq 0$ ci siano $X_n = i$ particelle nella prima urna. Ad ogni istante di tempo discreto una particella viene scelta a caso con probabilità $\frac{1}{N}$ e spostata dall'urna in cui essa si trova all'altra urna, così da rendere il sistema all'istante successivo $X_{n+1} = i - 1$ con probabilità $\frac{i}{N}$, oppure $X_{n+1} = i + 1$ con probabilità $\frac{N-i}{N}$. E' facile dimostrare che, attraverso questo processo di scelta delle particelle da spostare da un'urna all'altra, il sistema tende con l'avanzare del tempo alla distribuzione uniforme.

Questo modello è utile in quanto il processo delle mutazioni può essere pensato in modo analogo alle urne di Ehrenfest ponendo il numero di palline N corrispondente alla lunghezza della sequenza, e dividendole in 4 urne, che corrispondono alle basi A,C,G,T. Lo spostamento di una pallina da un'urna ad un'altra sarà quindi analogo alla sostituzione di una base con un'altra nel processo non esatto di trascrizione che origina le mutazioni. E' importante notare che, a differenza del modello di Ehrenfest, il processo di scelta casuale di una base che viene sostituita con un'altra base, non implica per forza che il sistema tenda ad una distribuzione uniforme con l'avanzare del tempo. Infatti, nel caso con k=4 urne si deve tenere presente anche della scelta tra le tre diverse possibili urne in cui la pallina scelta casualmente può essere spostata. Pertanto, sfruttando l'analogia appena descritta, lo studio delle probabilità di transizione tra le urne in un modello a 4 urne e la conoscenza di una certa distribuzione iniziale di palline nelle urne possono descrivere l'evoluzione delle frequenze relative alle sequenze di basi che compongono l'RNA del virus.

Il modello delle urne di Ehrenfest è un esempio di catena di Markov, il quale è un modello ideale per descrivere il problema a 4 urne. Introduciamo in seguito il modello delle catene di Markov nella sua generalità e vediamone poi l'applicazione al processo delle mutazioni.

2.2 Il modello delle catene di Markov

Le sequenze di variabili aleatorie indipendenti e identicamente distribuite sono particolari processi stocastici. Definiamo *processo stocastico a tempo discreto* una sequenza $\{X_n\}, n \geq 0$ di variabili aleatorie con valori in un insieme E detto *spazio degli stati*, dove n indica l'istante di tempo. Lo spazio degli stati è numerabile ed include tutti gli stati possibili in cui la catena può trovarsi, in particolare, se $X_n = i$, si dice che il processo si trova nello stato i all'instante di tempo $t = n$.

Proprietà di Markov: Sia $\{X_n\}_{n\geq 0}$ una sequenza di variabili casuali a valori in E definita come sopra. Se $\forall$ intero $n \geq 0$ e $\forall$ stato $i_0, i_1 ... i_{n-1}, i, j \in E$,

$$P(X_{n+1} = j|X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = 0) = P(X_{n+1} = j|X_n = i) \tag{2.30}$$

questo processo stocastico è detto *catena di Markov*. Si noti che entrambi i membri dell'equazione sono ben definiti. In particolare, se la parte destra di (2.30) non dipende dall'istante di tempo n allora è detta *catena di Markov omogenea*. La proprietà (2.30) è detta *Proprietà di Markov* e stabilisce che il futuro del sistema dipende solamente dal presente, non dal passato.

La matrice

$$P = \{p_{ij}\}_{i,j\in E}, \qquad p_{ij} = P(X_{n+1} = j|X_n = i) \tag{2.31}$$

è la *matrice di transizione* della catena di Markov omogenea. Siccome le sue entrate sono delle probabilità segue che:

$$p_{ij} \geq 0, \qquad \sum_j p_{ij} = 1 \tag{2.32}$$

Teorema 2. *La distribuzione di una catena di Markov omogenea è determinata dalla sua distribuzione iniziale e dalla sua matrice di transizione.*

La distribuzione ad un istante di tempo $n = t$ della catena è il vettore $\pi(n)$:

$$\pi_i(n) = P(X_n = i) \tag{2.33}$$

Dalla regola di Bayes si ottiene che $\pi_j(n+1) = \sum_{i\in E} \pi_i(n)p_{ij}$ che espressa in forma matriciale diventa:

$$\pi(n + 1) = P^T\pi(n) \tag{2.34}$$

A questo punto è possibile iterare questa equazione per ottenere in modo ricorsivo la dipendenza di $\pi(n)$ da π_0 ovvero dal vettore di partenza:

$$\pi(n) = (P^T)^n\pi_0 \tag{2.35}$$

Ora che il modello delle catene di Markov è stato introdotto nella sua generalità, si svilupperà nel prossimo paragrafo un modello composto da 4 urne ed N palline che si spostano tra le urne in modo casuale, influenzato però da delle specifiche probabilità di transizione tra le urne. Le probabilità di transizione, come detto in precedenza, non portano necessariamente alla distribuzione uniforme. Quello che ci si aspetta infatti da questo modello, in cui saranno introdotte le probabilità di transizione ricavate dal database [1], è un andamento asintotico ad una distribuzione non uniforme tra le frequenze p_i dove $i \in \{A, C, G, T\}$, coerentemente con quanto trovato nel capitolo 1.

2.3 Modello intuitivo per le mutazioni

In questa sezione sarà introdotto un modello semplice di catene di Markov per una descrizione euristica del problema. I risultati ottenuti in questo paragrafo saranno formalizzati in seguito. Si considerino k urne ed N palline. Il numero di palline contenute nell' i-esima urna è n_i e vale $\sum_i n_i = N$. Siano inoltre $\pi_i = \frac{n_i}{N}$ le frequenze relative all'i-esima urna. Siamo interessati allo studio della dinamica del sistema, quindi ad osservare come evolve nel tempo il numero di palline in ogni urna, sapendo che la somma complessiva delle palline contenute nelle urne rimane costante. Questo processo è identico a quello delle mutazioni in una sequenza di RNA, quindi allo scambio di una data base con un'altra nella sequenza. Un evento consiste in una pallina del sistema che viene scelta a caso e spostata dall'urna in cui si trova ad un'altra, che può essere diversa o la stessa in cui era. In questo modello gli eventi accadono in istanti discreti ed ordinati cronologicamente, che numeriamo con $t \in \mathbb{N}$.

Sia $\Delta n_i(t)$ la variabile aleatoria che descrive la variazione di palline nell'urna i-esima tra due istanti di tempo consecutivi, nell'ipotesi in cui solo una pallina per volta, scelta casualmente con probabilità $\frac{1}{N}$, venga spostata:

$$\Delta n_i(t) = n_i(t+1) - n_i(t), \qquad i \in \{A, C, G, T\} \tag{2.36}$$

Si noti che nell'ipotesi fatta deve essere $\Delta n_i(t) \in \{1, 0, -1\}$, infatti ad ogni time-step può spostarsi una sola pallina del sistema. Calcolando il valore medio della variazione nella i-esima urna si ricava:

$$\begin{aligned}\langle \Delta n_i(t) \rangle &= 1 \cdot P(\Delta n_i(t) = 1) + 0 \cdot P(\Delta n_i(t) = 0) - 1 \cdot P(\Delta n_i(t) = -1) \\ &= P(\Delta n_i(t) = 1) - P(\Delta n_i(t) = -1)\end{aligned} \tag{2.37}$$

Per trovare il valore medio della variazione di palline nell'urna i-esima, è necessario ricavare le due probabilità in (2.37). Si fa l'ipotesi che il modello sia una catena di Markov omogenea e che quindi la matrice di transizione non dipenda dal tempo. La dipendenza dal tempo della (2.37) è data solamente dallo stato in cui il sistema si trova all'istante t. Siano:

$$P(\Delta n_i = 1)(t) = \sum_{j \neq i} P(j \to i|j)\pi_j(t) = \sum_{j \neq i} P_{ji}\pi_j(t) = (\sum_{j=1}^{k} P_{ji}\pi_j) - P_{ii}\pi_i(t) \tag{2.38}$$

$$P(\Delta n_i = -1)(t) = \sum_{j \neq i} P(i \to j|i)\pi_i(t) = \sum_{j \neq i} P_{ij}\pi_i(t) = (\sum_{j=1}^{k} P_{ij}\pi_i) - P_{ii}\pi_i(t) \tag{2.39}$$

dove $P(j \to i|j)$ è la probabilità condizionata che una pallina salti dall'urna j-esima all'urna i-esima sapendo che si trova nell'urna j-esima, $\forall i, j$. Per ipotesi di catena di Markov omogenea questa probabilità di transizione è costante. In particolare la (2.38) rappresenta la probabilità complessiva delle $k-1$ urne diverse da i che una pallina venga scelta in un'urna j diversa da i con probabilità π_j e spostata nell'urna i-esima, mentre la (2.39) rappresenta la probabilità che una pallina dell'urna i-esima venga scelta con probabilità π_i e spostata in una delle altre $k-1$ urne. Sostituendo (2.38) e (2.39) in (2.37) si ottiene la variazione di palline media nell'urna i-esima tra uno stato t e il suo successivo $t+1$ in funzione delle frequenze dello stato $\pi(t)$:

$$\langle \Delta n_i \rangle(t) = \sum_{j=1}^{4} (P_{ji}\pi_j(t) - P_{ij}\pi_i(t)) = \sum_{j=1}^{4} \mathcal{P}_{ji}(t) - \mathcal{P}_{ij}(t) \tag{2.40}$$

dove $\mathcal{P}_{ji} = P_{ji}\pi_j$ è detta *probabilità congiunta*.

2.3.1 Stati stazionari

Nello studio della dinamica delle soluzioni siamo interessati a capire innanzitutto quali sono gli stati stazionari del sistema. Si può osservare che se vale la condizione di bilancio dettagliato (DBC), la quale prevede che $\mathcal{P}_{ji} = \mathcal{P}_{ij}\ \forall i, j$, allora $\langle \Delta n_i \rangle(t) = 0\ \forall i$. Quindi nel caso di (DBC) lo stato in cui il sistema si trova è stazionario. Si noti che, a causa della sommatoria, la quale non si annulla in modo univoco, il viceversa non è vero. Si cerchino ora le altre soluzioni stazionarie del problema. Dalla proprietà delle catene di Markov vista in (2.32) si ha che $\sum_j P_{ij} = 1\ \forall i$, ed utilizzando questo risultato nella (2.40) si ricava:

$$\begin{aligned}\langle \Delta n_i \rangle &= \sum_j (P_{ji}\pi_j - P_{ij}\pi_i) \\ &= \sum_j P_{ji}\pi_j - \pi_i \\ &= \sum_j (P_{ji}\pi_j - \delta_{ij}\pi_i) \\ &= ((P^T - \mathbb{I})\pi)_i \end{aligned} \tag{2.41}$$

dove nell'ultimo passaggio si è utilizzata la notazione matriciale. Quindi le soluzioni stazionarie sono tutte quelle che annullano la (2.41), ovvero che soddisfano:

$$\langle \Delta n \rangle = 0 \iff (P^T - \mathbb{I})\pi = 0 \iff P^T\pi = \pi \tag{2.42}$$

Quindi le soluzioni stazionarie sono tutti i vettori di frequenze $\hat{\pi}$ che sono autovettori della matrice di transizione P^T di autovalore unitario.

Definizione Se il vettore $\hat{\pi}$ soddisfa $P^T\hat{\pi} = \hat{\pi}$ per una data matrice P, allora $\hat{\pi}$ è detta *distribuzione stazionaria.*

L'applicazione della matrice di transizione della catena di Markov P^T a questi autovettori $\hat{\pi}$ lascia tali vettori invariati. Quando il sistema si trova in uno stato rappresentato da un autovettore di P^T di autovalore 1 questo stato è stazionario.

2.3.2 Comportamento per N elevato

Una volta trovati gli stati stazionari del problema siamo interessati a capire come le mutazioni evolvono a partire da una distribuzione iniziale $\pi_0 := \pi(t_0 = 0)$ e dalle probabilità di transizione della catena di Markov omogenea. Nei sistemi dinamici a tempo continuo il moto, quindi la soluzione del problema, è governato in modo deterministico da un insieme di equazioni differenziali, la cui variabile indipendente è appunto il tempo, e da un certo dato iniziale $x_0 := x(t = 0)$. Nel modello considerato, però, è stata fatta l'ipotesi di eventi che avvengono in istanti di tempo discreti, ciò rende impossibile la descrizione della dinamica per mezzo di un sistema di equazioni differenziali. Un metodo per trattare la dinamica del sistema considerato come fosse generata da un'equazione differenziale che agisce in valori continui del paramtro temporale è l'approssimazione di mean field. L'approssimazione di mean field consiste nel considerare il processo stocastico tale che siano trascurabili le fluttuazioni della variabile n_i, quindi si fa l'ipotesi che la varianza di tale variabile si avvicini sempre più allo zero all'aumentare di N, da cui:

$$\sigma_\pi^2 = \mathbb{E}[(\pi - \mathbb{E}[\pi])^2] \simeq 0 \qquad \rightarrow \qquad \pi \simeq \langle \pi \rangle \tag{2.43}$$

Questa approssimazione non rigorosa è verosimilmente considerata valida nel problema delle mutazioni, in quanto la lunghezza della sequenza N=29903 è sufficientemente elevata. Dividendo ambo i lati dell'equazione (2.41) per N si ottiene la variazione media del vettore di frequenze π tra un istante di tempo ed il suo successivo:

$$\langle \Delta\pi \rangle = \frac{1}{N}(P^T - \mathbb{I})\langle \pi \rangle \in \mathbb{R}^4 \tag{2.44}$$

dove nella parte destra è stato usato il risultato in (2.43) in accordo con l'ipotesi di mean field. Calcolando ora la variazione di $\langle \pi \rangle(t)$ tra un istante t e il suo successivo $t + 1$, e supponendo valida la relazione $\langle \Delta\pi \rangle(t) = \Delta\langle \pi \rangle(t)$ come conseguenza dell'approssimazione di mean field, si ottiene:

$$\langle \pi \rangle(t + 1) = \langle \pi \rangle(t) + \Delta\langle \pi \rangle(t) = \langle \pi \rangle(t) + \langle \Delta\pi \rangle(t) \tag{2.45}$$

dove la misura dell'intervallo ha valore $\Delta t = 1\ \forall t$. Si calcoli in seguito il rapporto incrementale tra una frequenza e la successiva utilizzando l'intervallo generico Δt:

$$\frac{\langle\pi\rangle(t+\Delta t) - \langle\pi\rangle(t)}{\Delta t} = \frac{\Delta\langle\pi\rangle(t)}{\Delta t} = \frac{\langle\Delta\pi\rangle(t)}{\Delta t} = \frac{1}{N\Delta t}(P^T - \mathbb{I})\langle\pi\rangle(t) \tag{2.46}$$

Ponendo $\Delta t = \frac{1}{N}$, dove $N =$ *lunghezza della sequenza* è abbastanza elevato da rendere infinitesimo l'intervallo, si ottiene:

$$\frac{d\langle\pi\rangle(t)}{dt} = \lim_{N\to\infty} \frac{\langle\pi\rangle(t+\frac{1}{N}) - \langle\pi\rangle(t)}{\frac{1}{N}} = \frac{N}{N}(P^T - \mathbb{I})\langle\pi\rangle(t) = (P^T - \mathbb{I})\langle\pi\rangle(t) \tag{2.47}$$

la quale è un'equazione differenziale ordinaria che descrive la dinamica delle mutazioni nell'approssimazione di mean field e nell'approssimazione di una misura infinitesima dt dell'intervallo tra due eventi della catena di Markov. Definendo inoltre la matrice $A := (P^T - \mathbb{I})$ è possibile riscrivere il problema nella seguente formulazione autonoma:

$$\langle\dot{\pi}\rangle = A\langle\pi\rangle, \qquad \pi(0) = \pi_0 \tag{2.48}$$

le cui soluzioni sono date da:

$$\langle\pi\rangle(t) = \pi_0 e^{At} \tag{2.49}$$

e, tramite mean field, $\pi \simeq \langle\pi\rangle$.

2.3.3 L'approssimazione rigorosa di mean field

Questa sezione descrive un risultato contenuto in [4].

Si consideri l'insieme:

$$\hat{\sigma}^N_{\mathbb{Q}} = \{x = (x_1, ..x_4), \quad x_i = \frac{n_i}{N} \geq 0, \quad \sum_i x_i = 1\} \subset \mathbb{Q}^4 \tag{2.50}$$

Sia ora $\hat{\pi}(\tau)$ una catena di Markov a variabili in $\hat{\sigma}^N_{\mathbb{Q}}$ a tempo discreto dove gli eventi nella catena sono descritti agli istanti $\tau \in \mathbb{T} := \{0, \delta, 2\delta..\}$, $\delta = \frac{1}{N}$ e $\tau_n = \frac{n}{N}$. E' possibile costruire una successione di vettori frequenze, ovvero di variabili aleatorie $\hat{\pi}(\tau_n)$, a valori nel simplesso $\hat{\sigma}^N_{\mathbb{Q}}$. Visualizzando le frequenze come punti di $\hat{\sigma}^N_{\mathbb{Q}}$, un vettore di frequenze che subisce una transizione $(i \to j)$ da un'urna ad un'altra si può denotare come:

$$\hat{\pi}(\tau + \Delta\tau) = \hat{\pi}(\tau) + \frac{1}{N}(e_j - e_i) \qquad \forall i, j \tag{2.51}$$

ove e_i con $i \in \{1, .., 4\}$ sono i vettori della base canonica di $\mathbb{R}^4$. Ciò corrisponde, infatti, a sottrarre $\frac{1}{N}$ alla componente $\hat{x}_i$ ed ad aggiungere $\frac{1}{N}$ nella componente $\hat{x}_j$ del vettore $\hat{x}$ associato a $\hat{\pi}(\tau)$. Per ogni coppia di urne (i, j) si assume che esista una funzione $\mathcal{P}_{ij}$ *continua* di $\hat{\pi}$, indipendente da τ ed N quando $\Delta\tau = \frac{1}{N}$, che descrive le probabilità di spostamento di una pallina $(i \to j)$.

$$\begin{aligned} &\mathcal{P}_{ij} : \hat{\sigma}^N_{\mathbb{Q}} \to [0, 1] \\ &\mathcal{P}_{ij}(\hat{\pi}) = Prob[\hat{\pi}(\tau + \Delta\tau) = \hat{\pi} + \frac{1}{N}(e_j - e_i) | \pi(t) = \hat{\pi}(t)] \end{aligned} \tag{2.52}$$

dove $\mathcal{P}_{ij} = 0$ se $x_i = 0$. Si definisce *corrente di probabilità* dell'urna i-esima:

$$F_i(\hat{\pi}) : \hat{\sigma}_{\mathbb{Q}} \to \mathbb{R}$$
$$F_i(\hat{\pi}) = \sum_j \mathcal{P}_{ji}(\hat{\pi}) - \mathcal{P}_{ij}(\hat{\pi}), \qquad i \in \{1, .., 4\} \tag{2.53}$$

Segue dalla definizione (2.53) che $F_i(\hat{\pi})$ è limitata e continua, e che $\sum_i F_i(\hat{\pi}) = 0$ in quanto N è costante. In particolare, si definisce *campo vettoriale indotto* associato alla catena di Markov il campo ottenuto dalle componenti in (2.53):

$$F(\hat{\pi}) = (F_1(\hat{\pi}), .., F_4(\hat{\pi})) \tag{2.54}$$

Al fine di descrivere l'andamento della catena di Markov tramite il flusso di un'equazione differenziale, si noti che $\hat{\sigma}^N_{\mathbb{Q}}$ è *denso* in $\hat{\sigma}_{\mathbb{R}}$ quando si considera il limite di $N \to \infty$. Per tale limite è quindi possibile raggiungere qualsiasi punto del simplesso $\hat{\sigma}_{\mathbb{R}}$ tramite il limite di una successione in $\hat{\sigma}^N_Q$, in quanto $\hat{\sigma}_{\mathbb{R}}$ rappresenta la chiusura di $\hat{\sigma}^N_Q$. Per questo motivo considerando $N \to \infty$ si può estendere il campo continuo $F(\hat{\pi})$ in tutto $\hat{\sigma}_{\mathbb{R}}$ e renderlo un campo vettoriale su un dominio continuo. Si definisca ora lo spazio tangente a $\hat{\sigma}_{\mathbb{R}}$ come l'iperpiano *3*-dimensionale:

$$T\hat{\sigma}_{\mathbb{R}} = \{h \in \mathbb{R}^4 : \sum_i h_i = 0\} \tag{2.55}$$

Nel limite $N \to \infty$ si può considerare il campo $F(\pi)$ come una mappa $F : \hat{\sigma}_{\mathbb{R}} \to T\hat{\sigma}_{\mathbb{R}}$, ove $\pi \in \hat{\sigma}_{\mathbb{R}}$, la quale verifica la condizione $\sum_i F_i(\pi) = 0$. Essendo il campo vettoriale $F(\pi)$ esteso in $\hat{\sigma}_{\mathbb{R}}$ continuo, questo è di conseguenza lipschitziano in $T\hat{\sigma}_{\mathbb{R}}$. Introduciamo un'equazione differenziale ordinaria ed il problema di Cauchy dato da:

$$\dot{\pi} = F(\pi), \qquad \pi(0) = \pi_0 \tag{2.56}$$

la cui soluzione verifica la condizione $\sum_i \pi_i(t) = 1 \ \forall t$. Il sistema autonomo (2.56) definisce una soluzione

$$\begin{aligned} \pi :& \mathbb{R} \times \hat{\sigma}_{\mathbb{R}} \to \hat{\sigma}_{\mathbb{R}} \\ & (t, \pi_0) \to \pi(t, \pi_0) \end{aligned} \tag{2.57}$$

detta *flusso indotto*. Se questa esiste, rappresenta l'approssimazione deterministica del processo stocastico associato alla catena di Markov. Vedremo infatti che il processo stocastico si muoverà con un'alta probabilità vicino alla soluzione identificata da (2.57).

Si vuole ora quantificare l'aderenza della linea spezzata che interpola i singoli punti $\hat{\pi}(\tau)$ della successione all'approssimazione deterministica $\pi(t, \pi_0)$ per intervalli di tempo limitati. Data la catene di Markov a tempo discreto $\hat{\pi}(\tau)$, si consideri il processo stocastico a tempo continuo che interpola la successione individuata da $\hat{\pi}(\tau)$, $\forall t \in [\tau, \tau + \frac{1}{N}] = [\frac{m}{N}, \frac{m+1}{N}]$:

$$\begin{aligned} \pi_c(t) &= \hat{\pi}(\frac{m}{N}) + \frac{t - \frac{m}{N}}{\frac{1}{N}}[\hat{\pi}(\frac{m+1}{N}) - \hat{\pi}(\frac{m}{N})] \\ &= \hat{\pi}(\frac{m}{N}) + (Nt - m)[\hat{\pi}(\frac{m+1}{N}) - \hat{\pi}(\frac{m}{N})] \end{aligned} \tag{2.58}$$

Al fine di quantificare quanto $\pi_c(t)$ possa discostarsi dalla soluzione $\pi(t, \pi_0)$, si definisce la norma L^∞:

$$\begin{aligned} &\| \cdot \|_\infty : \mathbb{R}^k \to \mathbb{R} \\ &\| \pi_c(t) - \pi(t, \pi_0) \|_\infty = max_{i=1,..k} |(\pi_c)_i(t) - (\pi(t, \pi_0))_i| \end{aligned} \tag{2.59}$$

Si consideri ora la variabile aleatoria

$$D_T(\pi_0) = max_{t\in[0,T]} \parallel \pi_c(t) - \pi(t,\pi_0) \parallel_\infty \tag{2.60}$$

che quantifica la deviazione massima durante un intervallo limitato [0,T]. La giustificazione rigorosa dell'approssimazione di mean field è basata sulla seguente proposizione:

Proposizione 1. *[4]* $\exists c > 0$ *tale che* $\forall \epsilon > 0\ \forall T > 0$ *ed N sufficientemente grande:*

$$Prob[D_T(\pi_0) \geq \epsilon \mid \hat{\pi(0)} = \pi_0] \leq 2me^{-c\epsilon^2 N} \tag{2.61}$$

Attraverso questa proposizione è possibile quantificare la probabilità che l'interpolazione $\pi_c(t)$ di punti discreti di una successione di mutazioni sia più distante di ϵ dalla soluzione teorica $\pi(t,\pi_0)$ ottenuta mediante le equazioni di mean field, in termini della norma (2.59). Questa probabilità decresce esponenzialmente e tende a zero quando $N \to \infty$. Per valori di N molto elevati, quindi, l'andamento delle successioni $\hat{\pi}(\tau)$ che si osservano nel sistema stocastico restano vicine alla soluzione deterministica delle equazioni di mean field, in quanto la probabilità di osservarne qualcuna che si discosta più di un arbitrario valore ϵ da tali soluzioni tende a zero quando $N \to \infty$. Per questo motivo è possibile assumere che le frequenze associate alle mutazioni in [1] seguono con alta probabilità e a meno di piccoli intervalli l'andamento descritto dalla soluzione di (2.57).

Al fine di trovare la forma esplicita della soluzione deterministica (2.57) si calcoli quindi il campo vettoriale $F(\pi)$ a partire dalle probabilità condizionate per 4 urne e N elevato:

$$F(\pi) = \sum_{j=1}^{4} \mathcal{P}_{ji}(\pi) - \mathcal{P}_{ij}(\pi) = \sum_{j=1}^{4} \pi_j P_{ji} - \pi_i P_{ij} \tag{2.62}$$

dove vale l'ipotesi che la matrice delle probabilità di transizione $P_{ij} = Prob(i \to j|i)$ sia costante. A questo punto, in accordo con i passaggi visti in (2.41) si ottiene:

$$F(\pi) = (P^T - \mathbb{I})\pi. \tag{2.63}$$

Tale campo conduce allo stesso sistema differenziale ottenuto in (2.48), il quale descrive quindi in modo ottimale la dinamica delle mutazioni contenute nel database. Infatti, per N elevato queste si discosteranno di un valore maggiore di un ϵ piccolo con poca probabilità dalla soluzione deterministica dell'ODE.

2.4 Stima della matrice di transizione relativa al database

Si consideri il caso con $k = 4$ urne ed N=29903 palline, analogo al problema delle mutazioni nelle sequenze di RNA virale ove N è la lunghezza delle sequenze e k è la cardinalità di $\{A, C, G, T\}$. Riepilogando i risultati della sezione (2.3) si può concludere che la dinamica delle mutazioni è governata dalla seguente equazione autonoma:

$$\dot{\pi} = A\pi, \qquad \pi_0 = \pi(0) \tag{2.64}$$

dove $A = (P^T - \mathbb{I})$, che ha soluzione:

$$\pi(t,\pi_0) = \pi_0 e^{At}. \tag{2.65}$$

La dinamica delle mutazioni osservate nel database [1] è quindi ben descritta dal flusso dell'equazione differenziale (2.64), che a sua volta dipende dalle probabilità di transizione tra le basi e dalla frequenza iniziale, individuata da $\pi_0 = q$, ovvero la frequenza associata alla sequenza iniziale x. E' possibile

ricavare queste probabilità di transizione dal database [1] attraverso il procedimento che segue. Siano $x = (x_1, ..x_N)$ la sequenza iniziale e $y = (y_1, ..y_N)$ una mutazione osservata, si definisca n_{ij} il numero di transizioni tali che $x_\alpha = i \to y_\alpha = j$ nella sequenza mutata.

$$n_{ij}(x, y) = n(x_\alpha = i, y_\alpha = j) \tag{2.66}$$

Dividendo (2.66) per N si ottiene la stima delle probabilità congiunte che legano la transizione $x \to y$ per una generica sequenza mutata:

$$\frac{n_{ij}(x, y)}{N} = \mathcal{P}_{ij_{db}}(x, y) = P_{ij_{db}}(x, y)q_i(x) = P_{ij_{db}}(x, y)\frac{n_i(x)}{N} \tag{2.67}$$

ove il pedice *db* nelle matrici fa riferimento al fatto che sono ottenute a partire dal database. Quindi, le probabilità di transizione nell'evoluzione $x \to y$ sono date da:

$$P_{ij_{db}}(x, y) = \frac{n_{ij}(x, y)}{n_i(x)} \tag{2.68}$$

Effettuando questo conto per un grande numero k di sequenze y nel database è possibile ottenere una stima del valore medio

$$\langle n_{ij}(x, y_1, ..y_k)\rangle = \frac{n_{ij}(x, y_1) + ... + n_{ij}(x, y_k)}{k} \tag{2.69}$$

per ogni coppia di i, j, e di conseguenza la miglior stima della matrice di transizione sarà data da:

$$\langle P_{ij}\rangle_{db} = \frac{\langle n_{ij}\rangle}{n_i(x)} \tag{2.70}$$

Da questa è possibile ricavare la matrice di transizione del sistema di equazioni differenziali relativa al database e di conseguenza ottenere un unico sistema differenziale associato al problema:

$$\dot{\pi} = \hat{A}\pi, \qquad \pi_0 = q, \qquad \hat{A} := (\langle P^T\rangle_{db} - \mathbb{I}) \tag{2.71}$$

la cui relativa soluzione è data da

$$\tilde{\pi}(t) = qe^{\hat{A}t} \tag{2.72}$$

Tramite (2.72) è possibile studiare la dinamica del sistema e l'andamento di $\tilde{\pi}(t)$ in funzione del tempo. Si è in particolar modo interessati a studiare il limite, se esiste, di $\tilde{\pi}(t)$ quando $t \to \infty$, e attraverso questo osservare l'andamento dell'entropia in funzione del tempo:

$$h(\tilde{\pi}(t)) = -\sum_{i=1}^{4} \tilde{\pi}_i(t) \ln \tilde{\pi}_i(t) \tag{2.73}$$

In accordo con le previsioni fatte nel primo capitolo si dovrebbe riscontrare una diminuzione dell'entropia associata al sistema al crescere del parametro t.

Nel prossimo capitolo si utilizzerà il sistema di equazioni differenziali ottenuto in (2.71) con il fine di descrivere le mutazioni, ricavando opportunamente dai dati la matrice $\hat{A}$ attraverso il procedimento descritto e cercando le soluzioni mediante calcolo numerico. Attraverso queste soluzioni si osserverà se l'andamento dell'entropia calcolato mediante (2.73) sarà coerente con ciò che era stato previsto dall'analisi dei dati al capitolo 1, ovvero se questa diminuisce con l'evolversi delle sequenze.

Dinamica di mean field

Quest'ultimo capitolo di tesi mira ad osservare l'applicazione del modello teorico ricavato nel capitolo precedente in particolare in tre diversi casi: modello a k urne e matrice di transizione simmetrica, modello a 2 urne e matrice di transizione non simmetrica, ed, infine, modello a 4 urne e matrice di transizione non simmetrica. L'ultimo caso, in particolare, coincide con il problema delle mutazioni per le sequenze di RNA virale una volta ottenuti i coefficienti della matrice di transizione associata al problema.

3.1 k urne e matrice delle transizioni simmetrica

Data la matrice di transizione simmetrica associata al problema:

$$P = \begin{pmatrix} s & p & \cdots & p \\ p & s & \ddots & \vdots \\ \vdots & \ddots & s & p \\ p & \cdots & p & s \end{pmatrix}, \qquad P_{ii} = s \ \ \forall i = 1, ..k, \qquad P_{ij} = p \ \ i \neq j \tag{3.74}$$

sapendo che la somma di ogni riga della matrice di transizione deve restituire 1 si ha che:

$$\sum_j P_{ij} = 1 = s + (k-1)p \quad \forall i \qquad \rightarrow \qquad p = \frac{1-s}{k-1} > 0 \tag{3.75}$$

sempre positivo in quanto $s \in [0, 1]$, $k > 1$.

Volendo ora calcolare

$$\langle \Delta n_i \rangle = \sum_j P_{ij}\pi_j - \pi_i = p\sum_{j \neq i} \pi_j + s\pi_i - \pi_i = p(1-\pi_i) + s\pi_i - \pi_i = p - \pi_i(1-s+p) \tag{3.76}$$

ed utilizzando i risultati (3.75):

$$1 - s + p = kp, \qquad p = \frac{1-s}{k-1} \tag{3.77}$$

si ottiene a sua volta sostituendo p:

$$\langle \Delta n_i \rangle = \frac{1-s}{k-1} - \pi_i k \frac{1-s}{k-1} = \frac{(1-s)(1-k\pi_i)}{(k-1)} \tag{3.78}$$

Dove, sapendo che $1 - s > 0$ e $k - 1 > 0$, si ottiene che la positività di (3.78) dipende solamente dalla positività di $1 - k\pi_i$. In particolare se $\pi_i = \frac{n_i}{N}$ si ha che

$$\langle \Delta n_i \rangle > (<) 0 \qquad se \qquad n_i < (>) \frac{N}{k}. \tag{3.79}$$

Quindi le urne con $n_i < \frac{N}{k}$ tendono ad acquistare palline, al contrario, quelle con $n_i > \frac{N}{k}$ tendono a perderne. L'equilibrio del sistema sarà pertanto $n_i = \frac{N}{k}$ $\forall i$.

Da questo calcolo a salti discreti si può ottenere l'equazione differenziale mediante i passaggi riportati in (2.47):

$$\frac{d\pi_i}{dt} = \frac{1-s}{k-1} - k\frac{1-s}{k-1}\pi_i = a - ka\pi_i \qquad \forall i \tag{3.80}$$

dove è stata effettuata la sostituzione $a = \frac{1-s}{k-1}$. Questa è un'ODE a coefficienti costanti non omogenea, della quale è possibile calcolare analiticamente la soluzione:

$$\pi_i(t) = e^{-kat}(\pi_{0_i} - \frac{1}{k} + \frac{e^{kat}}{k}) \tag{3.81}$$

Effettuando il limite per $t \to \infty$ si può osservare che il sistema tende asintoticamente alla distribuzione uniforme, esattamente come ci si aspettava dal risultato (3.79):

$$\lim_{t\to\infty} \pi_i(t) = \lim_{t\to\infty} e^{-kat}(\pi_{0_i} - \frac{1}{k}) + \frac{1}{k} = \frac{1}{k} \qquad \forall i \tag{3.82}$$

Avendo la soluzione è possibile, inoltre, calcolare $\forall t$ l'entropia associata al sistema attraverso la formula (1.2) ed il suo andamento per $t \to \infty$:

$$\lim_{t\to\infty} h(\pi(t)) = \lim_{t\to\infty} -\sum_{i=1}^{k} \pi_i(t) \ln \pi_i(t) = -\sum_{i=1}^{k} \frac{1}{k} \ln \frac{1}{k} = \ln k. \tag{3.83}$$

Il sistema tende quindi ad un valore dell'entropia associata ad esso che è il massimo possibile nel limite di k urne, per la proprietà (iii) di (1.2). Dalla teoria sappiamo inoltre che per una matrice di Markov simmetrica ed ergodica l'unica distribuzione di equilibrio è quella uniforme. Nell'esempio di questo paragrafo avevamo posto $P_{ij} > 0 \ \forall ij$, quindi la matrice studiata è appunto ergodica e simmetrica. In accordo coi risultati del primo capitolo, dove si era evinto che la distribuzione di equilibrio delle sequenze si allontana dalla uniforme, risulta utile lo studio del modello applicato a casi in cui P_{ij} è matrice non simmetrica.

3.2 2 urne e matrice delle transizioni non simmetrica

Data la matrice di transizione non simmetrica associata al problema:

$$P = \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix} = \begin{pmatrix} 1-\mu & \mu \\ \rho & 1-\rho \end{pmatrix} \tag{3.84}$$

dove $\mu, \rho \in [0,1]$ e $\mu \neq \rho$ per ipotesi di non simmetria. Dalla definizione $A := (P^T - \mathbb{I})$ si ricava

$$A = \begin{pmatrix} -\mu & \rho \\ \mu & -\rho \end{pmatrix} \tag{3.85}$$

attraverso cui è possibile risolvere il sistema (2.64) per ogni frequenza iniziale π_0 data. Le soluzioni generiche sono ottenibili facilmente per via analitica. Sfruttando $\pi_2(t) = 1 - \pi_1(t) \ \forall t$, le equazioni differenziali del sistema sono linearmente dipendenti, ed è possibile ricondursi, quindi, all'ODE scalare non omogenea:

$$\frac{d\pi_1(t)}{dt} + \pi_1(t)(\mu + \rho) = \rho \tag{3.86}$$

Risolvendo tale equazione si ottengono le due soluzioni:

$$\pi_1(t) = \frac{\rho}{\mu+\rho} + (\pi_{0_1} - \frac{\rho}{\mu+\rho})e^{-\frac{\mu+\rho}{N}t}, \qquad \pi_2(t) = 1 - \pi_1(t) \tag{3.87}$$

Attraverso il limite per $t \to \infty$ delle soluzioni (3.87) si ottengono i valori di equilibrio asintotici delle frequenze, che dipendono solamente dalla scelta di μ e ρ, non dal vettore π_0 di partenza:

$$\tilde{\pi}_1 = \lim_{t\to\infty} \pi_1(t) = \frac{\rho}{\rho+\mu} \qquad \tilde{\pi}_2 = \lim_{t\to\infty} \pi_2(t) = \frac{\mu}{\rho+\mu} \tag{3.88}$$

Attraverso le soluzioni (3.88) è possibile studiare l'entropia $h(\pi(t))$ al variare di t, ed osservarne il comportamento nel limite $t \to \infty$:

$$\lim_{t\to\infty} h(\pi(t)) = h(\tilde{\pi}) = -\frac{\rho}{\mu+\rho}\ln(\frac{\rho}{\mu+\rho}) - \frac{\mu}{\mu+\rho}\ln(\frac{\mu}{\mu+\rho}) \tag{3.89}$$

che è ancora una volta indipendente dal vettore di partenza π_0. Si osservi che se valesse $\mu = \rho$, e fossimo quindi nel caso di matrice simmetrica, il sistema sarebbe evoluto verso la distribuzione uniforme $\tilde{\pi} = (\frac{1}{2}, \frac{1}{2})$, in accordo con quanto detto nel paragrafo precedente. Questo non avviene effettuando scelte che rispettano $\rho \neq \mu$.

Vogliamo ora studiare cosa accade al sistema assumendo i valori indicativi di $\mu = 0.75$ e $\rho = 0.45$ a partire da un vettore di frequenze iniziale $\pi_0 = (0.8, 0.2)$. Le soluzioni del sistema (2.64), in accordo con i parametri scelti, sono:

$$\pi_1(t) = 0.375 + 0.425e^{-1.2t} \qquad \pi_2(t) = 0.625 - 0.425e^{-1.2t} \tag{3.90}$$

come si può vedere dal grafico sottostante, il limite, dopo un certo intervallo di tempo, si stabilizza su dei valori che dipendono solamente dai parametri μ e ρ, in accordo con (3.88):

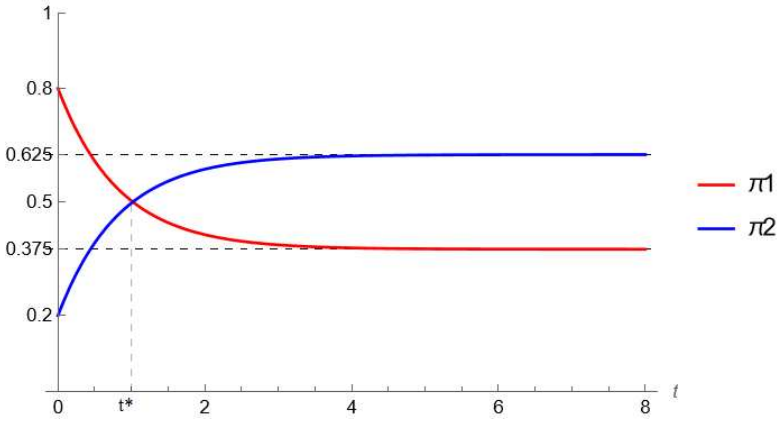

Figura 3.8: Grafico di $\pi(t)$ per la dinamica di mean field

Attraverso le soluzioni (3.90) è stato realizzato il grafico dell'entropia associata al sistema al variare di t. All'istante iniziale si ha $h(0) = 0.500$. Nella dinamica del sistema il vettore iniziale scelto $\pi_0 = (0.8, 0.2)$ evolve in modo continuo alla distribuzione d'equilibrio $\tilde{\pi} = (0.375, 0.625)$ passando per la distribuzione uniforme. Per questo motivo l'entropia associata al sistema è dapprima crescente fino al raggiungimento di un massimo in corrispondenza dell'istante t^* associato a $\pi(t^*) = (\frac{1}{2}, \frac{1}{2})$, ed in seguito decrescente nell'intervallo di tempo che segue t^*. L'entropia del sistema tende asintoticamente al valore $h(\tilde{\pi}) = 0.662$

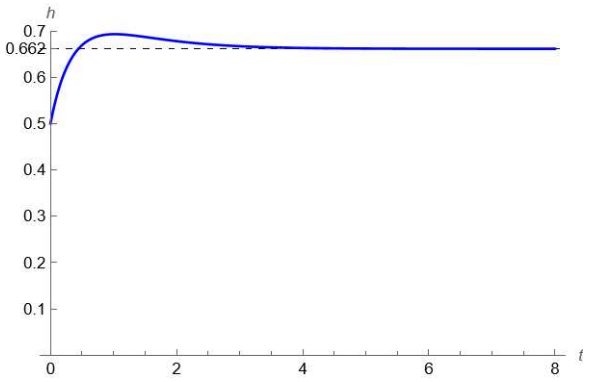

Figura 3.9: Entropia

Attraverso un grafico qualitativo che descrive l'entropia in funzione delle sole frequenze $h(\pi_1, \pi_2)$ è possibile visualizzare meglio questo tipo di andamento. Il seguente grafico rappresenta $h(\pi)$: in particolare si possono osservare $h(\pi_0)$ ed $h(\tilde{\pi})$ e più in generale l'andamento dell'entropia via via che la soluzione evolve da π_0 a $\tilde{\pi}$. Analogamente al grafico precedente, l'entropia prima sale fino a raggiungere un massimo in corrispondenza di $\pi = (\frac{1}{2}, \frac{1}{2})$, dopo il quale inizia a calare fino al valore d'equilibrio $h(\tilde{\pi})$:

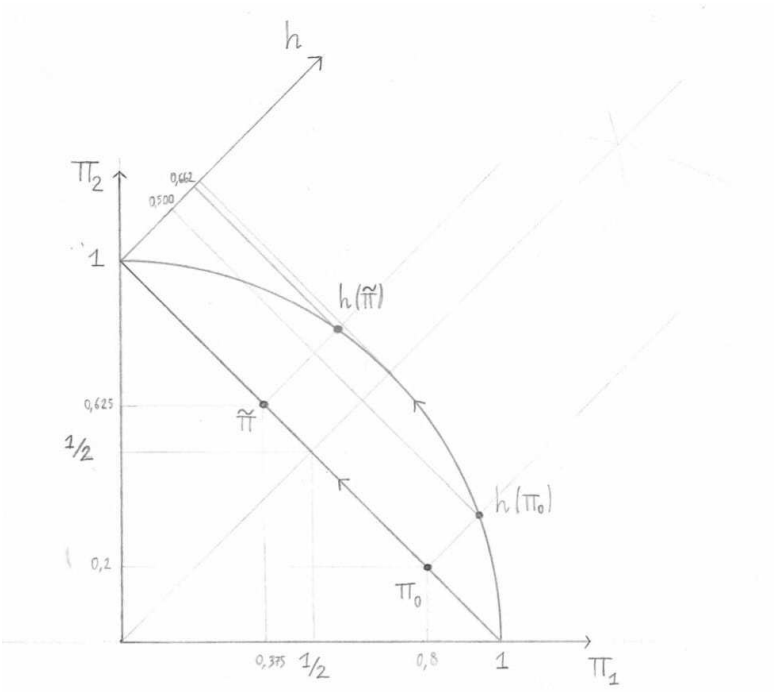

Figura 3.10: andamento di $h(\pi)$ per i valori scelti di π_0, ρ, μ

Si noti che se fosse stata scelta una diversa frequenza iniziale π_0, la quale nell'evolvere verso $\tilde{\pi}$ non fosse passata per la distribuzione uniforme (quindi, considerando la corda relativa alla curva dell'entropia in (3.10), se si fosse scelta una frequenza π_0 appartenente alla stessa metà di corda a cui appartiene la frequenza $\tilde{\pi}$), l'andamento dell'entropia sarebbe stato monotono.

3.3 Modello a 4 urne con matrice delle transizioni dal database

Affrontiamo ora il caso delle mutazioni di RNA virale. Questo problema è analogo al precedente, con l'unica differenza che il sistema ha 4 gradi di libertà. Attraverso la soluzione del sistema differenziale (2.71), è possibile capire l'andamento delle mutazioni per qualsiasi valore del parametro t. Effettuando

dal database il calcolo riportato al paragrafo 2.4, è stata ottenuta la matrice delle transizioni (2.70) associata al database, dove le colonne sono rispettivamente nell'ordine A,C,G,T:

$$\langle P\rangle_{db} = \begin{pmatrix} 0.99999 & 0 & 10^{-5} & 0 \\ 10^{-5} & 0.99993 & 0 & 6\cdot 10^{-5} \\ 0 & 0 & 0.99995 & 5\cdot 10^{-5} \\ 2\cdot 10^{-5} & 0 & 0 & 0.99998 \end{pmatrix} \tag{3.91}$$

la cui relativa matrice $A_{db} = \hat{A} := (\langle P\rangle_{db}^{T} - \mathbb{I})$ è:

$$A_{db} = 10^{-5}\begin{pmatrix} -1 & 1 & 0 & 2 \\ 0 & -7 & 0 & 0 \\ 1 & 0 & -5 & 0 \\ 0 & 6 & 5 & -2 \end{pmatrix} \tag{3.92}$$

Il vettore delle frequenze relativo alla sequenza iniziale x nell'ordine A,C,G,T ha il valore:

$$q = \pi_0 = (0.299, 0.184, 0.196, 0.321) \tag{3.93}$$

Avendo il vettore di frequenze relativo alla sequenza di riferimento x e la matrice (3.92) è possibile risolvere il sistema differenziale (2.71), ricavandone le soluzioni al variare del parametro t. E' stato realizzato un programma utilizzando il software Wolfram Mathematica per risolvere tale sistema differenziale. Le soluzioni trovate dal programma sono le seguenti:

$$\begin{aligned} A(t) &= 0.5882 + 0.0368e^{-7t\cdot 10^{-5}} + e^{-4t\cdot 10^{-5}}(-0.3260\cos(t\cdot 10^{-5}) - 0.5195\sin(t\cdot 10^{-5})) \\ C(t) &= 0.184e^{-7t\cdot 10^{-5}} \\ G(t) &= 0.1176 - 0.0184e^{-7t\cdot 10^{-5}} + e^{-4t\cdot 10^{-5}}(0.0968\cos(t\cdot 10^{-5}) - 0.4228\sin(t\cdot 10^{-5})) \\ T(t) &= 0.2941 - 0.2024e^{-7t\cdot 10^{-5}} + e^{-4t\cdot 10^{-5}}(0.2293\cos(t\cdot 10^{-5}) + 0.9423\sin(t\cdot 10^{-5})) \end{aligned} \tag{3.94}$$

Osservando le soluzioni è evidente che dopo un certo intervallo di tempo dell'ordine di grandezza di $t = 10^5$ le soluzioni, quindi le frequenze di ogni base, si assestano asintoticamente sui valori:

$$\begin{aligned} \tilde{A} &= \lim_{t\to\infty} A(t) = 0.588 \\ \tilde{C} &= \lim_{t\to\infty} C(t) = 0 \\ \tilde{G} &= \lim_{t\to\infty} G(t) = 0.118 \\ \tilde{T} &= \lim_{t\to\infty} T(t) = 0.294 \end{aligned} \tag{3.95}$$

Confrontando le frequenze del vettore q con le soluzioni asintotiche in (3.95) si nota che la distribuzione delle frequenze globalmente si allontana dalla distribuzione uniforme, ovvero dal valore $\pi_i = 0.25$, $i \in \{A, C, G, T\}$, esattamente come ci aspettavamo dall'analisi svolta in 1.3. Infatti, le frequenze relative alle basi A, C, G si allontanano da tale valore e solamente la base T vi si avvicina leggermente prima di raggiungere il suo asintoto. L'andamento delle soluzioni (3.94) è mostrato nel seguente grafico, dove si è scelto un intervallo pari a $t = 2.5\cdot 10^5$ nell'asse delle ascisse, così da poterne apprezzarne l'andamento asintotico.

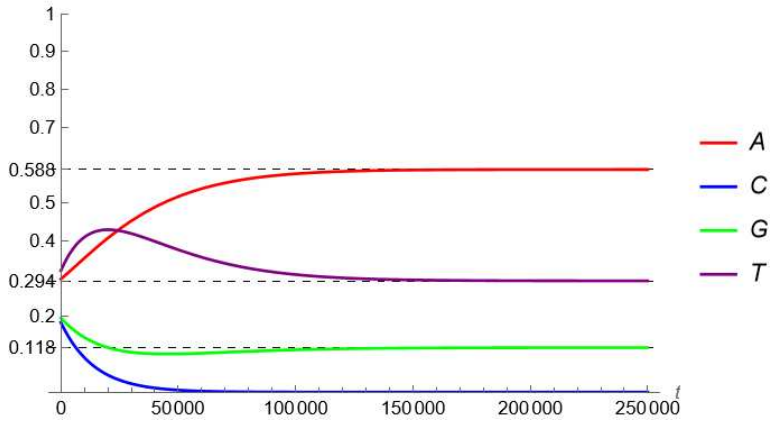

Figura 3.11: Grafico di $\pi(t)$ per la dinamica di mean field

Per capire il significato dell'unità di misura t nell'asse delle ascisse si ricorda che il valore dell'intervallo tra due eventi elementari, o mutazioni, era stato reso infinitesimale nella (2.47) ponendolo uguale a $\delta = \frac{1}{29903}$. Pertanto, l'unità di misura dell'asse delle ascisse nel grafico (3.11), del valore di $\Delta t = 1$, corrisponde a $29903 * \delta$, quindi a 29903 eventi nella sequenza. Con lo stesso ragionamento si può dire che all'istante t della scala adottata corrispondono un numero di $29903t$ mutazioni avvenute a partire dalla sequenza di rifermento. Le soluzioni si assestano quindi ai rispettivi valori di equilibrio $\tilde{\pi} = (\tilde{A}, \tilde{C}, \tilde{G}, \tilde{T})$ dopo un numero pari a circa $29903 \cdot 10^5 \sim 10^9$ mutazioni.

Come nel caso con due urne, dalle soluzioni è possibile ricavare l'entropia delle mutazioni in funzione del parametro t. Essa ha un valore di partenza di $h(0) = 1.3565$ in accordo con il grafico (1.4), e per ordini di grandezza di 10^9 mutazioni e superiori si assesta al valore asintotico di $h(\tilde{\pi}) = 0.924$, con un andamento monotono decrescente, come ci si aspettava dall'analisi svolta nel primo capitolo.

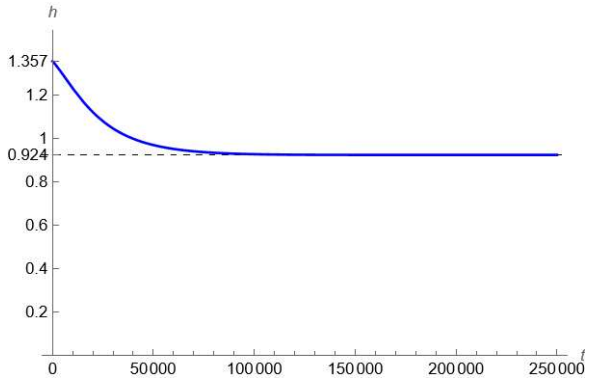

Figura 3.12: Andamento dell'entropia per la dinamica di mean field

Siamo ora interessati alla descrizione dell'entropia in funzione dell'entropia relativa associate alle soluzioni (3.94), così da poter confrontare le previsioni del modello con la corrispondente figura (1.1) ricavata dalle sequenze del database. Ci interessa dunque arrivare ad un'espressione dell'entropia del tipo $h(d)$. A tal fine, ricaviamo l'entropia relativa associata alle soluzioni in funzione del parametro t, mediante la formula (1.3), e, successivamente, ne ricaviamo l'inversa della forma $t(d)$. L'operazione è possibile in quanto $D(p(t)|q)$ è iniettiva quando q è fissato. Infine, tramite composizione di funzioni, otteniamo $h(t(d))$. Riportiamo in seguito i grafici delle funzioni appena citate. Nel caso dell'entropia relativa in funzione di t si nota $D(\pi(0)|q) = 0$ in quanto $\pi(0) = q$ e per un ordine di grandezza di $t = 10^5$ o superiore l'andamento si assesta al valore $D(\tilde{\pi}(t)|q) = 0,312$. Ne riportiamo il grafico per $t = 250000$ e per $t = 100$, quest'ultima è infatti la zona in cui i valori di $d(t)$ hanno il medesimo ordine di grandezza dell'entropia relativa delle mutazioni in figura (1.1), quindi è esattamente la zona di interesse.

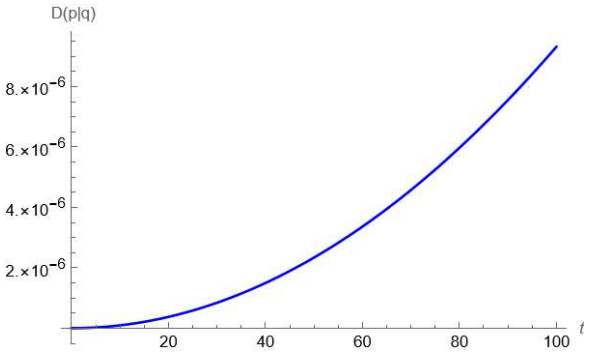

Figura 3.13: valori di $D(p|q)$ confrontabili con il database

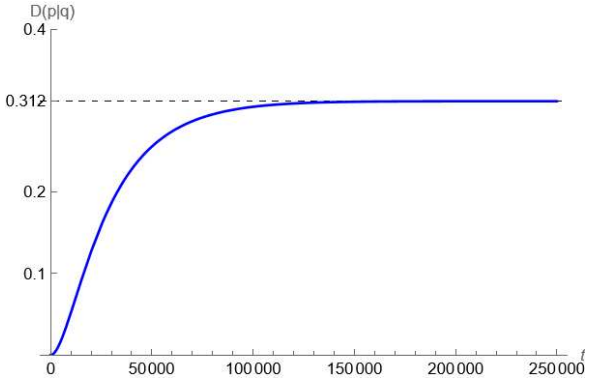

Figura 3.14: entropia relativa per $t = 250000$ associata alla dinamica di mean field

Dal grafico (3.14) emerge che esiste un limite superiore al valore dell'entropia relativa associata alle soluzioni (3.94), questa, infatti, possiede un asintoto orizzontale in $d = 0.312$. Andando quindi a rappresentare $h(t(d))$ è importante notare che questa avrà un dominio di $[0, 0.312)$. Attraverso l'inversa $t(d)$ della funzione rappresentata in (3.14) è stato ottenuto il grafico dell'entropia in funzione dell'entropia relativa $h(t(d))$ associate alle soluzioni. Ancora una volta ne riportiamo l'andamento globale e la parte relativa ai valori di d confrontabili con quelli del database.

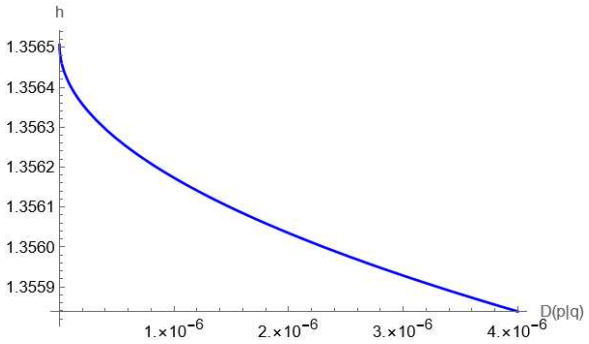

Figura 3.15: grafico di $h(d)$ per valori di d confrontabili con il database

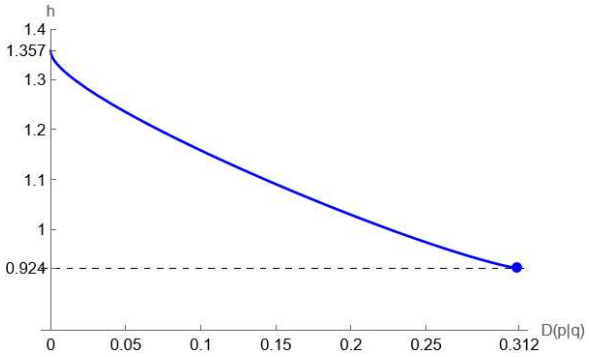

Figura 3.16: grafico di $h(d)$ associata alla dinamica di mean field

L'andamento del grafico (3.16) è monotono decrescente e presenta quindi un limite inferiore per $h(0.321) = 0.924$, ovvero in corrispondenza del limite superiore dell'entorpia relativa associata alle soluzioni (3.94). Nella zona rappresentata dal grafico (3.15) notiamo che l'andamento della funzione $h(d)$ è confrontabile con l'andamento di h_{min} del grafico (1.4), che a sua volta descriveva bene i punti della figura (1.1) vicini alla curva di minima entropia. Con lo scopo di verificare se il modo che hanno le sequenze di mutare ottenuto in (3.94) si accosti al minimo dell'entropia, e descriva quindi le sequenze del database vicine a tale minimo, riportiamo la rappresentazione della curva $h(d)$ in (3.15-3.16) sovrapposta alla curva h_{min} dei grafici (1.4-1.5).

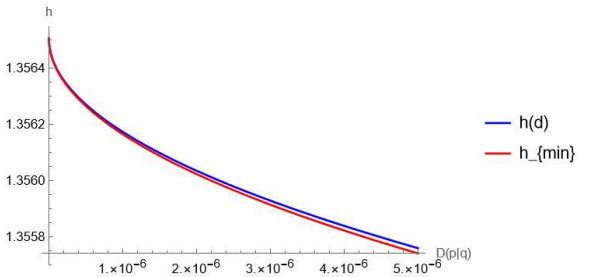

Figura 3.17: h_{min} (1.4) ed $h(d)$ (3.15)

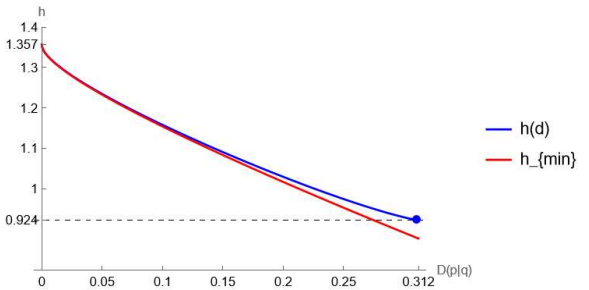

Figura 3.18: h_{min} (1.5) ed $h(d)$ (3.16)

Dai grafici (3.17-3.18) si può osservare che l'andamento dell'entropia lungo la dinamica deterministica delle soluzioni ottenute tramite l'approssimazione di mean field, rappresentato in blu, segue fedelmente la curva di minima entropia, rappresentata in rosso. Dal grafico (1.6) è possibile notare che una parte delle sequenze del database ha entropia molto vicina alla curva di minima entropia h_{min}. Attraverso la proposizione 1 sarebbe possibile stimare la frazione di sequenze che ha entropia vicina alla curva ottenuta dalla dinamica di mean field $h(d)$ a meno di un arbitrario valore ϵ, e, in seguito, stimare quanto queste siano vicine alla curva h_{min} conoscendone l'intervallo di compatibilità con $h(d)$, ma questo esula dallo scopo di questa tesi.

Bibliografia

[1] Ncbi-national center for biotechnology information. https://www.ncbi.nlm.nih.gov/search/all/?term=NC045512.2. 25/07/2023.

[2] Dimitri P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. academic Press, 2014.

[3] N. G. Van Kampen. *Stochastic processes in physiscs and chemistry*. Elsevier, 1991.

[4] Jorgen W. Weibull Michel Benaim. Deterministic approximation of stochastic evolution in games. *Econometrica*, 71(3):873–903, 2003.