Università
Cà Foscari
Venezia

Corso di Laurea magistrale
in Informatica
(ordinamento ex D.M. 270/2004)

Tesi di Laurea

—

Dorsoduro 3246
30123 Venezia

CoMeta: studio delle distanze tra vie metaboliche di organismi in KEGG.

Relatore
Chiar.ma Prof.ssa Nicoletta Cocco

Controrelatore
Chiar.ma Prof.ssa Federica Giummolè

Laureando
Anna Chiara Marabello
Matricola 824577

Anno Accademico
2011/2012

Indice

Capitolo 1

Introduzione

I processi dell'organismo che richiedono produzione, consumo o accumolo di energia e che permettono agli esseri viventi di rimanere in vita vengono indicati con il termine *metabolismo*. Il metabolismo viene in genere rappresentato da una rete di reazioni chimiche suddivisa in vie metaboliche. Le vie metaboliche sono sottoreti di reazioni enzimatiche che realizzano particolari funzioni biologiche, un esempio è la glicolisi che è la via metabolica più studiata e che consiste in una successione ordinata di reazioni che permettono la scissione del glucosio. Le vie metaboliche sono oggetto di studi da parte della comunità scientifica sia per comprenderne a fondo il funzionamento sia per lo studio delle malattie e la progettazione di nuovi farmaci. I dati delle ricerche effettuate sugli organismi sono raccolti in numerosi database biologici e messi a disposizione di tutta la comunità. Tra le tecniche di studio delle vie metaboliche proposte in letteratura, vi è la comparazione di una stessa via metabolica, o di un insieme di vie metaboliche, in organismi diversi. Ciò serve ad evidenziare similarità e differenza tra organismi rispetto ad una specifica funzione metabolica e può essere utile ad approfondirne la comprensione. In questa tesi si propone di analizzare statisticamente le distanze tra organismi rispetto a vie metaboliche utilizzate da CoMeta nei database biologici di KEGG. L'obiettivo di questo lavoro è di realizzare un tool di analisi statistica delle distanze di CoMeta tra organismi diversi rispetto a vie metaboliche. E' stato realizzato un prototipo, RCoMeta, che consente di elaborare i dati delle distanze basate su invarianti, reazioni e distanza combinata e di rappresentarle attraverso grafici di varie tipologie, istogrammi, collocazione rispetto all'indice z_{score}, analisi di un singolo organismo nei confronti di un gruppo di organismi di KEGG e di una coppia di organismi. Il lavoro è articolato nei seguenti capitoli. Il capitolo 2 introduce i concetti basilari sugli organismi viventi e sulle vie metaboliche. Il capitolo 3 descrive le reti di Petri e come le vie metaboliche possano essere da loro rappresentate. Il capitolo 4 illustra brevemente i database di KEGG. Il capitolo 5 introduce il concetto di similarità tra vie metaboliche con i principali indici di distanza tra organismi proposti in letteratura. Il capitolo 6 illustra le principali tecniche utilizzate in bioinformatica per allineamenti di sequenze e come ne venga valutata la significatività statistica. Il

capitolo 7 presenta in breve il tool CoMeta che calcola le distanze tra organismi in KEGG sulla base delle vie metaboliche. Il capitolo 8 descrive le funzionalità di RCoMeta il tool di analisi sviluppato nella tesi che guida l'utente all'uso della procedura di RCoMeta. Infine in appendice vengono fornite le analisi di alcuni gruppi di organismi secondo la classificazione NCBI adottata da KEGG. Nel capitolo 9 vengono riportate le osservazioni ricavate dalle analisi effettuate e una nota conclusiva.

Capitolo 2

Cellule e vie metaboliche

In questo capitolo si presenteranno i concetti di base sugli organismi e sulle vie metaboliche.

2.1 Cellule

Gran parte del materiale presentato nel paragrafo seguente è tratto da [3].

La cellula è considerata la struttura più piccola, sia per forma sia per funzione, in grado di svolgere tutte le funzioni vitali degli esseri viventi. Si suddivide in organismi *unicellulari* e *pluricellulari*, gli organismi *unicellulari* sono formati da una sola cellula, mentre quelli *pluricellulari*, sono composti da numerose cellule specializzate in diverse funzioni. L'insieme delle cellule costituisce un tessuto biologico le cui proprietà sono condivise da tutti gli organismi viventi:

- le cellule sono condiderate le unità strutturali, funzionali e fondamentali di tutti gli esseri viventi;
- le cellule sono portatrici di informazioni genetiche che trasmettono alla discendenza attraverso un linguaggio biologico universale;
- le cellule sono simili per composizione chimica;
- tutte le cellule sono circondate da una membrana plasmatica che le separa dall'ambiente esterno e che può essere attraversata per osmosi;
- l'ambiente interno alla membrana è costituito dal *citoplasma* dove avvengono molte reazioni chimiche.

Gli elementi chimici che caratterizzano la composizione di una cellula sono circa trenta, sette elementi: ossigeno (O), carbonio (C), idrogeno (H), azoto (N), calcio (Ca), fosforo (P) e zolfo (S), costituiscono più del 99% della massa cellulare, mentre gli altri presenti in quantità minori, sono comunque fondamentali allo svolgimento delle funzioni vitali della cellula.

Le altre sostanze che compongono la cellula si dividono in due categorie: *micromolecole* e *macromolecole* organiche, la distinzione fra le due categorie è basata esclusivamente sulla dimensione e sull'organizzazione strutturale. Le micromolecole organiche sono i carboidrati (zuccheri) i lipidi (grassi), gli aminoacidi (componenti delle proteine), i nucleotidi (acidi nucleici depositari di informazione genetica) e le vitamine; le macromolecole (polimeri) composte da molecole organiche di dimensioni uguali o più piccole sono unite da legami covalenti e formano catene di elementi di struttura più complessa. Esempi di macromolecole sono i carboidrati più complessi (glicogeno), proteine e acidi nucleici (RNA, DNA).

2.2 Classificazione delle cellule

Il materiale relativo a questo paragrafo è tratto da [3].

Le cellule si suddividono in: cellule *procariote*, formate da una struttura semplice senza nucleo e cellule *eucariote* nelle quali, il materiale genetico all'interno della cellula è raccolto in un involucro detto nucleo racchiuso da una doppia membrana esterna ed interna e circondato da una particolare sostanza chiamata citoplasma.

I biologi classificano tutti gli esseri viventi in tre domini: *archei, batteri ed eucarioti*. Gli archei e i batteri sono costituiti da cellule *procariote*. Ogni cellula è circondata da una membrana che la separa dall'ambiente esterno detta *membrana plasmatica*. Si tratta di una membrana a doppio strato formata in prevalenza da fosfolipidi il cui ruolo è:

- garantire alla cellula di conservare un ambiente interno costante;
- agire da barriera semipermeabile impedendo ad alcune sostenze di attraversarla e permettendo ad altre di entrare o uscire dalla cellula;
- separare l'ambiente cellulare da quello extracellulare;
- contenere proteine che sono responsabili dell'adesione fra cellule vicine.

Tutte le cellule *procariote* presentano la seguente struttura di base:

- la membrana plasmatica racchiude la cellula;
- all'interno della membrana plasmatica si trova il *citoplasma*, un materiale semifluido entro cui avvengono tutte le reazioni cellulari;
- una particolare componente del citoplasma, chiamata *nucleotide*, contiene il materiale ereditato (DNA) della cellula, il DNA è un polimero costituito da unità ripetute di nucleotidi.

Le cellule *eucariote* sono invece caratterizzate dalla presenza di compartimenti delimitati da particolari membrane chiamate *organuli* nei quali avvengono, specifiche reazioni chimiche in base agli enzimi presenti.

Ciascun organulo svolge, nel proprio tipo particolare di cellula, un ruolo specifico definito dalle reazioni chimiche che è in grado di attivare. Gli organuli cellulari presentano:

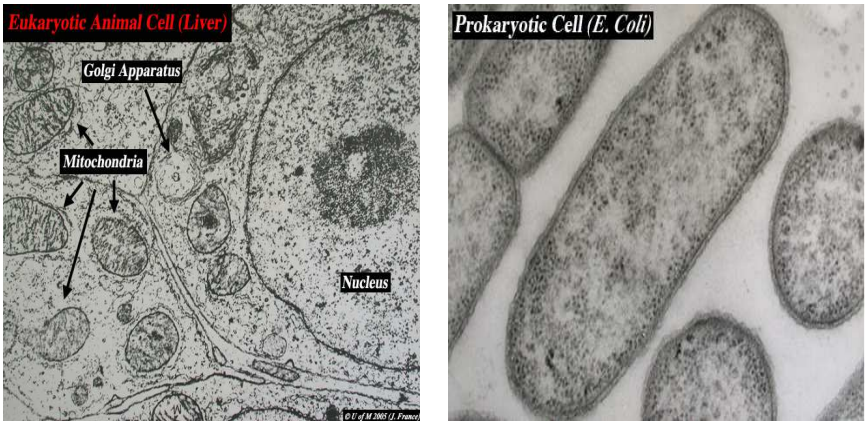
- un *nucleo* che contiene gran parte del materiale genetico della cellula. Nel nucleo hanno luogo la duplicazione del DNA e le fasi iniziali della decodificazione dell'informazione genetica;

- un *reticolo endoplasmatico e l'apparato di Golgi*, distretti nei quali avviene un'intensa attività di reazioni chimiche che elaborano continuamente materiali, alcuni dei quali vengono qui depositati, altri trasportati in altre parti della cellula, altri ancora avviati verso l'esterno;

- i *lisosomi e i vacuoli*, sistemi cellulari digerenti dove le molecole di grosse dimensioni vengono idrolizzate in monomeri facilmente utilizzabili;

- i *mitocondri*, organuli dove si svolge la respirazione cellulare, un processo in cui i carboidrati o i lipidi vengono demoliti utilizzando ossigeno;

- i *cloroplasti* presenti solo nelle cellule vegetali con funzione di fotosintesi.



(a) Cellula Eucariota (b) Cellula Procariota

Figura 2.2.1: Cellula al microscopio: Eucariota, Procariota

2.3 Enzimi

Il materiale relativo a tale paragrafo è tratto da [3].

Perchè i processi vitali delle cellule possano avvenire è necessario che si scatenino delle reazioni chimiche che liberano energia. La temperatura corporea, induce delle reazioni biochimiche troppo lente per far parte ai processi vitali degli organismi, per cui è necessario che delle sostanze particolari fungano da catalizzatori, ovvero, sostanze che accelerano una reazione senza essere modificate in modo permanente dalla reazione stessa.

I catalizzatori biologici sono particolari proteine prodotte dal *DNA* chiamate *enzimi*. Gli enzimi sono altamente selettivi, ovvero grazie ad essi i reagenti seguono una sola reazione chimica tra le tante reazioni chimiche possibili. In una reazione catalizzata da un enzima, i reagenti prendono il nome di *substrati*. Le molecole di substrato si legano a un determinato sito dell'enzima, chiamato *sito attivo* [1], dove avviene la catalisi [2]. La maggior parte degli enzimi ha dimensioni notevoli rispetto al substrato su cui agisce, il sito attivo al contrario è molto ridotto in dimensioni. Il ruolo del sito attivo è tuttavia importante: infatti la capacità dell'enzima di selezionare il substrato corretto dipende dal preciso incastro tra sito attivo e substrato e dalle interazioni tra gruppi chimici presenti nel sito di legame. Alla fine della reazione, l'enzima che può essere andato incontro anche a delle trasformazioni riacquista la stessa forma chimica che aveva all'inizio.

2.4 Metabolismo

Il materiale relativo a tale paragrafo è tratto da [3].

Il *metabolismo* è l'insieme di tutte le reazioni chimiche che avvengono in ogni essere vivente e consente alle cellule di rifornirsi dell'energia necessaria per crescere e riprodursi. Le reazioni metaboliche che avvengono nelle cellule degli organismi sono di due tipi:

- *reazioni anaboliche:* sintetizzano molecole complesse a partire da molecole semplici;
- *reazioni cataboliche:* demoliscono le molecole complesse in molecole più semplici.

Il metabolismo richiede un continuo dispendio di energia e di conseguenza necessita di un continuo apporto di energia; le reazioni che gli esseri viventi utilizzano per ricavare

[1] Il sito attivo di un enzima è la porzione di molecola direttamente implicata nel processo di catalisi e nella formazione dei legami con i reagenti.

[2] La catalisi è un termine greco che significa: sciogliere. E' un fenomeno chimico attraverso il quale la velocità di una reazione chimica subisce delle variazioni per l'intervento di una sostanza detta catalizzatore, che non viene consumata dal procedere della reazione stessa.

energia prendono il nome di *metabolismo energetico*. La sostanza più comune dalla quale le cellule ricavano energia è il glucosio, l'energia può comunque essere fornita anche da grassi e proteine. I processi più importanti per lo sfruttamento dell'energia del glucosio sono la *glicolisi*, la *fermentazione*, e la *respirazione artificiale*.

2.5 Vie metaboliche

Il materiale relativo a tale paragrafo è tratto da [3].

La vita di un organismo dipende da un sistema chimico complesso che genera amino acidi, zuccheri, lipidi e acidi nucleici, questa creazione viene indotta da insiemi di reazioni chimiche che costituiscono una *rete metabolica*.

Una rete metabolica è estremamente complessa e per dominare tale complessità la si scompone in *vie metaboliche*. Una *via metabolica* chiamata anche *metabolic pathway* è l'insieme delle reazioni chimiche coinvolte in uno o più processi di anabolismo o catabolismo all'interno di una cellula.

Le reazioni chimiche del metabolismo energetico procedono attraverso diverse tappe, in ogni tappa la reazione chimica catalizzata da uno o più enzimi trasforma alcune molecole *(reagenti o substrati)* in prodotti intermedi chiamati *metaboliti* che verranno utilizzati come reagenti nella reazione successiva, fino alla formazione del prodotto finale.

Le reazioni chimiche sono descritte da *equazioni chimiche* scritte con il formalismo di un'equazione matematica che pone al primo membro a sinistra i reagenti, e al secondo a destra i prodotti. Le equazioni chimiche seguono il postulato di Lavoisier, *nulla si crea, nulla si distrugge, tutto si conserva* che impone che la somma delle masse delle sostanze reagenti deve essere uguale alla somma delle masse dei prodotti. Le equazioni vengono quindi bilanciate associando dei coefficienti stechiometrici ai composti che partecipano alla reazione.
Un'equazione chimica è definita secondo lo schema generale:

$$\boxed{xX_{(sf)} + yY_{(sf)} \rightarrow zZ_{(sf)} + kK_{(sf)}} \tag{2.5.1}$$

dove:

- *x,y,z,k:* indicano i coefficienti stechiometrici che reappresentano il numero di molecole di ciascun composto chimico che partecipa alla reazione;
- *X,Y,Z,K:* sono i composti descritti dalla loro formula molecolare;
- *sf:* indica lo stato di aggregazione in cui si trova il composto: solido (s), gassoso (g), liquido (l), disciolto in una soluzione generica (sol), in soluzione acquosa (aq) o adsorbito su una superficie solida (ads).

Le vie metaboliche seguono alcuni principi comuni:

- ogni reazione di una via metabolica è catalizzata da un enzima specifico;
- le vie metaboliche sono simili in tutti gli organismi, dai batteri agli esseri umani;
- negli eucarioti quasi tutte le vie metaboliche sono organizzate per compartimenti in quanto le singole reazioni avvengono all'interno di un particolare organulo;
- ogni via metabolica è regolata da enzimi che determinano la velocità a cui procedono le reazioni.

Lo studio di una rete metabolica richiede di raccogliere informazioni da molte fonti biochimiche e genomiche, è di grande importanza riuscire a rappresentare le reti metaboliche tramite modelli formali che permettano l'analisi e la simulazione comportamentale. Tali modelli possono fornire una maggior comprensione dei processi delle vie metaboliche ed evidenziare come le relazioni tra vie metaboliche contribuiscano alla funzionalità e al comportamento dell'intero sistema biologico.
L'utilizzo di un modello biologico richiede:

- la traduzione delle conoscenze teoriche e sperimentali in un modello;
- la convalida del modello;
- la derivazione dal modello convalidato di ipotesi da verificare sul sistema;
- l'utilizzazione delle informazioni ricavate per affinare il modello o la teoria.

Uno dei problemi principali della modellazione del metabolismo è quello di gestire la complessità della rete metabolica, questa operazione è resa complicata dal fatto che biologi e biochimici utilizzano spesso criteri intuitivi di analisi ed è per tale motivo che frequentemente una stessa via metabolica può trovarsi in database diversi con rappresentazioni che differiscono sia nel numero di parametri sia nella struttura.

Nella pagina seguente in figura 2.5.1 è rappresentata graficamente una rete metabolica con le sue interconnessioni.

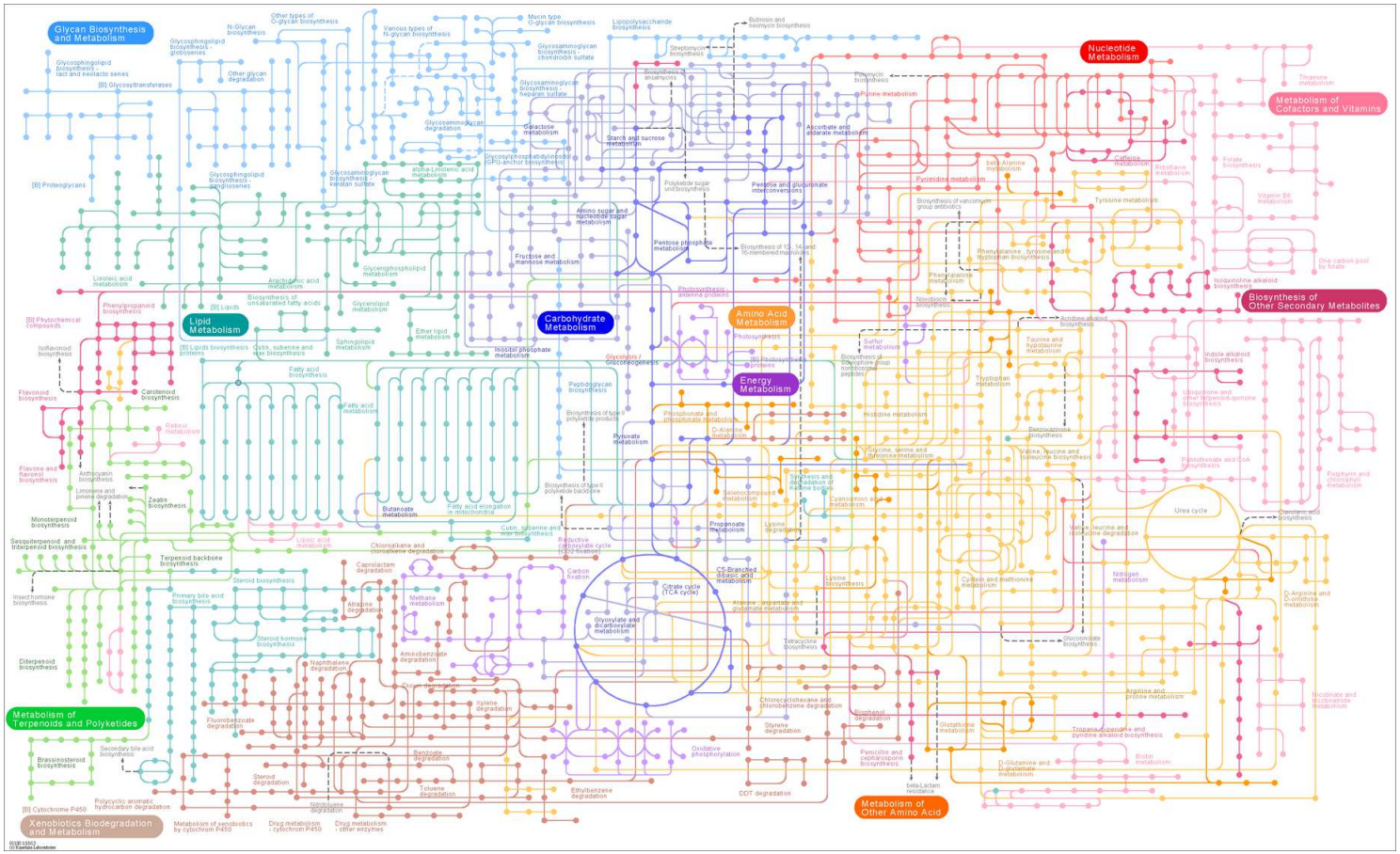


Figura 2.5.1: KEGG: Via metabolica di codice *map01100*

Capitolo 3

Reti di Petri e Sistemi Biologici

In questo capitolo vengono presentate le reti di Petri discrete e il loro utilizzo nella rappresentazione delle vie metaboliche. Le informazioni sono tratte dalle seguenti fonti: [24], [6],[26].

3.1 Reti di Petri

Il materiale relativo a questo paragrafo è tratto da [24],[6].

Le Reti di Petri sono state introdotte nel 1962 dal matematico tedesco Carl Adam Petri e rappresentano un formalismo matematico e grafico in grado di descrivere il flusso di informazioni di sistemi dinamici ad eventi discreti. Esse possono essere espresse da un formalismo grafico molto conciso e da un formalismo matematico che permette di applicare tecniche di modellazione e di analisi per lo studio di proprietà di particolare rilevanza.

Definizione 3.1.1 (Rete di Petri)
Una Rete di Petri è un grafo bipartito, orientato e pesato formato da un insieme di nodi partizionato in due sottoinsiemi denominati place e transition.

Formalmente una rete di Petri è una quadrupla:

$$\boxed{N = (P, T, Pre, Post)} \tag{3.1.1}$$

dove:

- $P = \{p_1, p_2, ..., p_m\}$ è un insieme finito di m posti, non vuoto, i cui elementi sono rappresentati graficamente con dei cerchi;
- $T = \{t_1, t_2, ..., t_n\}$ è un insieme finito di n transizioni, non vuoto, i cui elementi sono rappresentati graficamente con dei rettangoli;

- *Pre:* $P \times T \mapsto \mathbb{N}$ è la funzione di pre-incidenza;
- *Post:* $P \times T \mapsto \mathbb{N}$ è la funzione di post-incidenza;
- Inoltre indichiamo con:
  - $F \subseteq (P \times T) \times (T \times P)$ la relazione di flusso che determina l'insieme delle coppie ordinate (p,t) o (t,p), con $p \in P$ e $t \in T$, connesse da un arco con il relativo verso;
  - $W : F \mapsto \mathbb{N}$ la funzione peso che associa ad ogni arco della rete un numero intero postivo.

Si suppone che l'insieme dei posti e delle transizioni siano disgiunti e non vuoti.

$$P \cap T = \oslash$$
$$P \cup T \neq \oslash$$

Definizione 3.1.2 (Funzione di pre-incidenza)
La funzione di pre-incidenza specifica quali archi sono diretti da nodi di tipo 'place' a nodi di tipo 'transition' (chiamati archi Pre). Essa viene rappresentata da una matrice di numeri interi non negativi $Pre \in \mathbb{N}^{n \times m}$ *in cui il generico elemento* $Pre\,(p_i, t_j)$ *con* $i \in [1..m]$ *e* $j \in [1..n]$ *rappresenta il peso indicato nell'arco di collegamento tra il nodo* p_i *e la transizione* t_j*.*

Definizione 3.1.3 (Funzione di post-incidenza)
La funzione di post-incidenza specifica quali archi sono diretti da nodi di tipo 'transition' a nodi di tipo 'place' (chiamati archi Post). Essa viene rappresentata da una matrice di numeri interi non negativi $Post \in \mathbb{N}^{n \times m}$ *in cui il generico elemento* $Post\,(p_i, t_j)$ *con* $i \in [1..m]$ *e* $j \in [1..n]$ *rappresenta il peso indicato nell'arco di collegamento tra la transizione* t_j *e il nodo* p_i*.*

Definizione 3.1.4 (Matrice di incidenza)
Le informazioni contenute nelle matrici Pre Post possono essere compattate in una singola matrice chiamata 'Matrice di incidenza'.
Data una rete di Petri $N = (P, T, Pre, Post) \in \mathbb{N}^{n \times m}$ *è possibile definire la matrice di incidenza* $C : P \times T \to \mathbb{Z}$ *come:*

$$\boxed{C = Post - Pre} \tag{3.1.2}$$

Nel compattare le informazioni contenute nelle matrici Pre e Post può succedere che la matrice di incidenza perda qualche informazione sulla struttura della rete, per non aver perdita di informazioni è necessario che la rete di Petri considerara sia una rete *pura*. Una rete è pura se non esistono nodi place che siano contemporaneamente di ingresso e di uscita per uno stesso nodo transizione.

Esempio 3.1.1 (Rappresentazione grafica di una rete di Petri)
La figura 3.1.1 rappresenta la rete $N = (P, T, Pre, Post)$ *con insieme dei nodi place* $P = \{p_1, p_2, p_3, p_4, p_5\}$ *e insieme dei nodi transition* $T = \{t_1, t_2, t_3, t_4\}$ *in cui si assume che i valori dei pesi siano tutti* 1.

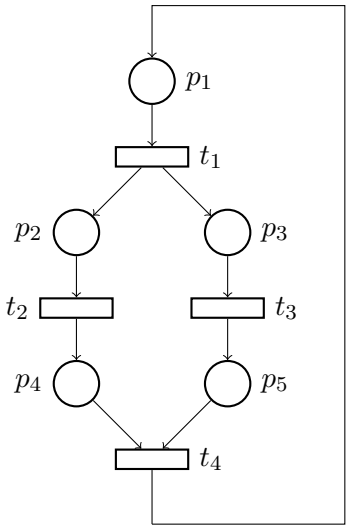


Figura 3.1.1: Rappresentazione grafica della rete di Petri N

Le matrici *Pre*, *Post* e di *Incidenza* della rete di Petri N di figura 3.1.1 sono:

$$Pre = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \quad Post = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

$$C = Post - Pre \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \end{array}$$

Definizione 3.1.5 (Marcatura)
In una rete di Petri $N = (P, T, Pre, Post)$ ad ogni nodo 'place' viene assegnato un numero intero non negativo di elementi chiamati token. La 'marcatura' di una rete di Petri permette di rappresentare l'assegnazione dei token nei nodi place e di descrivere lo stato del sistema. Essa è definita dalla seguente funzione:

$$\boxed{M = P \to \mathbb{N}} \tag{3.1.3}$$

dove: M rappresenta la rete di Petri *marcata* e P l'insieme dei nodi *place*.

La marcatura di una rete di Petri è rappresentabile tramite un vettore $M\,(m_1, m_2, ..., m_m)$ dove $m_1, m_2, ..., m_m$ rappresentano il numero di token dei nodi place di una marcatura M della rete di Petri.

Definizione 3.1.6 (Rete di Petri marcata)
Una rete di Petri $N = (P, T, Pre, Post)$ con marcatura iniziale M_0 viene chiamata rete marcata M_0. La marcatura iniziale definisce lo stato in cui la rete si trova all'inizio della sua evoluzione e viene indicata con:

$$\boxed{Rete\ di\ Petri = \{N, M_0\}} \tag{3.1.4}$$

Esempio 3.1.2 (Rete di Petri con marcatura iniziale M_0)
La figura 3.1.2 rappresenta la rete $N = (P, T, Pre, Post)$ *con insieme dei places* $P = \{p_1, p_2, p_3, p_4, p_5\}$*, insieme transition* $T = \{t_1, t_2, t_3, t_4\}$ *e marcatura* $M_0 = (3, 2, 1, 2, 0)$.

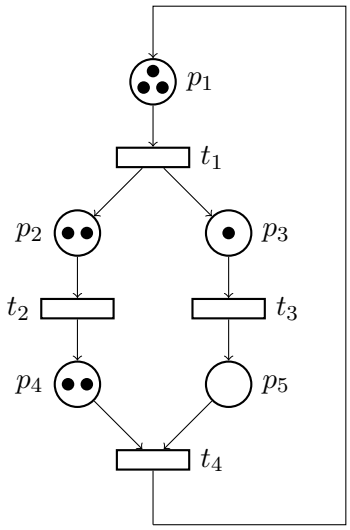


Figura 3.1.2: Rappresentazione grafica di una rete di Petri con marcatura inizale M_0.

E' possibile analizzare e descrivere il comportamento di una Rete di Petri a partire dalla sua marcatura iniziale M_0 definendo e analizzando i seguenti insiemi:

Definizione 3.1.7 (Insieme dei posti in entrata ad una transizione)

$$\boxed{\bullet t_j = \{p_i \in P | Pre\,(p_i, t_j) > 0\} \;\; con\; t_j \in T.} \tag{3.1.5}$$

Definizione 3.1.8 (Insieme dei posti in uscita da una transizione)

$$\boxed{t_j \bullet = \{p_i \in P | Post\,(p_i, t_j) > 0\} \;\; con\; t_j \in T.} \tag{3.1.6}$$

Definizione 3.1.9 (Insieme delle transizioni in entrata ad un posto)

$$\boxed{\bullet p_i = \{t_j \in T | Post\,(p_i, t_j) > 0\} \;\; con\; p_i \in P.} \tag{3.1.7}$$

Definizione 3.1.10 (Insieme delle transizioni in uscita da un posto)

$$\boxed{p_i\bullet = \{t_J \in T | Pre\,(p_i, t_j) > 0\} \;\; con\; p_i \in P.} \tag{3.1.8}$$

L'esempio della rete di Petri N di figura 3.1.2 definisce i seguenti insiemi:

Place input	Place output	Transition input	Transition output
$\bullet t_1 = \{p_1\}$	$t_1\bullet = \{p_2, p_3\}$	$\bullet p_1 = \{t_4\}$	$p_1\bullet = \{t_1\}$
$\bullet t_2 = \{p_2\}$	$t_2\bullet = \{p_4\}$	$\bullet p_2 = \{t_1\}$	$p_2\bullet = \{t_2\}$
$\bullet t_3 = \{p_3\}$	$t_3\bullet = \{p_5\}$	$\bullet p_3 = \{t_1\}$	$p_3\bullet = \{t_3\}$
$\bullet t_4 = \{p_4, p_5\}$	$t_4\bullet = \{p_1\}$	$\bullet p_4 = \{t_2\}$	$p_4\bullet = \{t_4\}$
		$\bullet p_5 = \{t_3\}$	$p_5\bullet = \{t_4\}$

Gli insiemi sono formati dai nodi di tipo (*place o transition*) a monte o a valle del nodo considerato (*place o transition*).

Definizione 3.1.11 (Abilitazione di una transizione)
Data una marcatura M_k di una rete di Petri, una transizione t_j è abilitata solo se ogni nodo di tipo place in entrata alla transizione t_j dispone di un numero di token maggiore o uguale al peso dell'arco che collega il nodo alla transizione t_j.

In una rete di Petri $N = (P, T, Pre, Post)$ con marcatura M_k una transizione $t_j \in T$ è abilitata se:

$$\boxed{M_k \geq Pre\,(\bullet, t_j) \;\; con\; t_j \in T} \tag{3.1.9}$$

dove $Pre\,(\bullet, t_j)$ rappresenta il vettore corrispondente alla colonna della transizione nella matrice *Pre*.

Lo scatto della transizione t_j genera una nuova marcatura M_{k+1} della rete di Petri che, rimuove $Pre(p_i, t_j)$ token da ogni nodo $p_i \in P$ che precede la transizione t_j e aggiunge $Post(p_i, t_j)$ token in ogni place $p_i \in P$ che segue la transizione t_j. I token rimossi e depositati sono pari al peso dell'arco che collega i nodi.

Definizione 3.1.12 (Scatto di una transizione abilitata da una marcatura M_k)
Una transizione t_j abilitata da una marcatura M_k può essere interessata da uno scatto di transizione che modifica lo stato della rete di Petri. La marcatura determinata da uno scatto di una transizione $t_j \in T$ viene espressa come:

$$\boxed{M_{k+1} = M_k - Pre\,(\bullet, t_j) + Post\,(\bullet, t_j) = M + C\,(\bullet, t_j) \;\; con\; t_j \in T} \tag{3.1.10}$$

Lo scatto di transizione t_j che porta la rete dalla marcatura M_k alla marcatura M_{k+1} viene indicato con: $M_k \xrightarrow{t_j} M_{k+1}$.

Definizione 3.1.13 (Sequenza di scatti di transizione)
Un sistema dinamico che evolve nel tempo può essere interessato da una sequenza di scatti di transizioni $\sigma = t_{x1}t_{x2}...t_{xn} \in T$ abilitata da una marcatura M_k. La sequenza a partire dalla marcatura M_k con lo scatto della transizione t_{x1} porta il sistema nella marcatura M_{k+1}, lo scatto di t_{x2} abilitato dalla marcatura M_{k+1} porta il sistema nella marcatura M_{k+2} e così via sino a raggiungere lo scatto di t_{x+n} che porta alla marcatura M_{k+n}. L'evoluzione della rete di Petri viene rappresentata dalla sequenza:

$$\boxed{M_k \xrightarrow{\sigma} M_{k+n}} \tag{3.1.11}$$

Esempio 3.1.3 (Sequenza di scatti di transizioni abilitate)
La rete di Petri rappresentata in figura 3.1.3 visualizza il risultatato della sequenza di scatti $\sigma = t_2t_3$ applicata alla rete di Petri di figura 3.1.2 in cui a partire dalla marcatura $M_0 = (3, 2, 1, 2, 0)$ si ha:

1. $M_0 \xrightarrow{t_2} M_1$ dove $M_1 = (3, 1, 1, 3, 0)$
2. $M_1 \xrightarrow{t_3} M_2$ dove $M_2 = (3, 1, 0, 3, 1)$

La sequenza di scatti $\sigma = t_2, t_3$ preleva un token dai nodi p_2 e p_3 e aggiunge un token ai nodi p_4 e p_5.

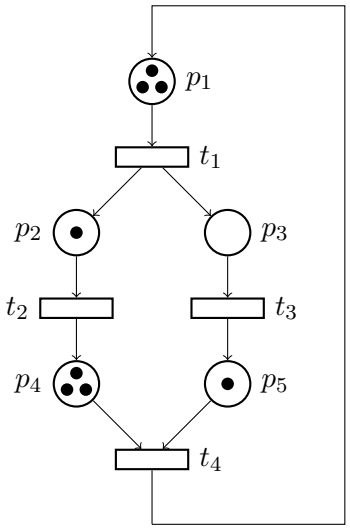


Figura 3.1.3: Rappresentazione grafica della sequenza di scatti $\sigma = t_2t_3$ applicata alla rete di Petri di figura 3.1.2.

Definizione 3.1.14 (Insieme di raggiungibilità)
L'insieme di raggiungibilità di una rete di Petri N è l'insieme di tutte le marcature raggiungibili da una sequenza di scatto di transizioni σ a partire dalla marcatura iniziale M_0. Si indica con:

$$\boxed{R\left(N, M_0\right) = \{M \in \mathbb{N}^m \mid \exists\ \sigma\ \textit{tale che}\ M_0 \xrightarrow{\sigma} M}} \tag{3.1.12}$$

3.1.1 Proprietà comportamentali delle reti di Petri

Il comportamento di una rete di Petri può essere delineato tramite proprietà che dipendono dalla marcatura iniziale M_0. Tali proprietà, indicate come *proprietà comportamentali* di una rete di Petri, sono:

- Reachability;
- Boundedness e Safeness;
- Liveness;
- Reversibility.

L'analisi viene effettuata solitamente mediante l'impiego di due strumenti: il grafo di raggiungibilità che permette di rappresentare tutti gli stati raggiungibili da una rete di Petri e le equazioni di stato.

Definizione 3.1.15 (Reachability)
Data una rete di Petri $\{N, M_0\}$, una marcatura M_k è raggiungibile dalla marcatura iniziale M_0 se esiste una sequenza di scatti σ che a partire da M_0 permette di raggiungere M_k:

$$\boxed{\exists\ \sigma\ \textit{tale che}\ M_0 \xrightarrow{\sigma} M_k} \tag{3.1.13}$$

Il problema di determinare se in una rete di Petri esiste una sequenza di scatti σ tale per cui la marcatura M_k appartenga all'insieme di raggiungibilità $R\left(N, M_0\right)$ dipende dalla cardinalità dell'insieme R. Se la cardinalità di R è:

- *Finita,* il problema è *decidibile* , ovvero esiste un algoritmo che in un numero finito di passi termina ed è in grado di determinare se la marcatura M_k è raggiungibile dalla marcatura iniziale M_0;
- *Infinita,* il problema è *semi-decidibile* ovvero esiste un algoritmo che in un numero finito di passi determina se una marcatura M_k è raggiungibile dalla marcatura

M_0, considerando la possibilità che nella rete di Petri la marcatura non sia raggiungibile ci si potrebbe trovare in presenza di un algoritmo che non termina. In questo caso la complessità computazionale dell'algoritmo e talmente elevata da renderlo non proponibile.

Definizione 3.1.16 (Boundedness e Safeness)
Una rete di Petri $\{N, M_0\}$ *è* $l-bound$ *se il numero complessivo di token di ogni nodo di tipo place non supera un valore intero positivo* l *per ogni marcatura* M_k *dell'insieme di raggiungibilità* $R\{N, M_0\}$*. La proprietà viene espressa da:*

$$\boxed{M_k(p_i) \leq l \ con \ M_k \in R \ e \ p_i \in P} \tag{3.1.14}$$

Una rete di Petri $\{N, M_0\}$ *è safe se è bound, ed il bound* l *è uguale a* 1*.*

Per poter definire la proprietà *Liveness* di una rete di Petri è necessario introdurre prima il concetto di *liveness* di una transizione.

Definizione 3.1.17 (Liveness di una transizione)
Una transizione $t_j \in T$ *in una marcatura* $M_k \in R$ *è live se esiste una sequenza* σ *in cui essa è abilitata ad effettuare uno scatto di transizione in grado di portare il sistema dalla marcatura* M_k *a una nuova marcatura* M_{k+1}*. Una transizione* t_j *può essere:*

- $L0-live$ *se non esiste nessuna marcatura* M_k *in cui* t_j *è abilitata allo scatto di transizione* σ*;*
- $L1-live$ *se esiste almeno una marcatura* M_k *in cui* t_j *è abilitata almeno una volta allo scatto di transizione* σ*;*
- $L2-live$ *se esiste almeno una marcatura* M_k *in cui* t_j *è abilitata almeno* k *volte allo scatto di transizione* σ*;*
- $L3-live$ *se esiste almeno una marcatura* M_k *in cui* t_j *è abilitata infinite volte allo scatto di transizione* σ*;*
- $L4-live$ *se per ogni una marcatura* M_k t_j *è abilitata allo scatto di transizione* σ*.*

Una rete di Petri (N, M_0) si dice:

- Morta: se tutte le transizioni t_j di una rete sono $L0-live$;
- Quasi viva: se contiene solo transizioni t_j classificate $L1$ o $L2$ o $L3-live$;
- Viva: se tutte le transizioni t_j sono $L4-live$.

Definizione 3.1.18 (Reversibility)
Una rete di Petri (N, M_0) gode della proprirtà reversibility se per ogni marcatura $M_k \in R$ raggiungibile dalla marcatura iniziale M_0 anche la marcatura iniziale M_0 è raggiungibile da M_k.

Definizione 3.1.19 (Marcatura morta e sistema bloccante)
In una rete di Petri una marcatura morta M è una marcatura che non permette lo scatto di nessuna transizione. Un sistema che raggiunge una tale marcatura si definisce un sistema bloccante.

3.1.2 Rappresentazione algebrica ed equazione di stato

L'evoluzione di una rete di Petri $\{N, M_0\}$ può essere rappresentata algebricamente dalla seguente equazione di stato:

$$\boxed{M_k = M_0 \ + C \cdot O} \tag{3.1.15}$$

dove: M_k appartiene all'insieme di raggiungibilità R, M_0 è la marcatura iniziale della rete di Petri, C è la matrice di incidenza e O è un vettore di dimensione $|T|$ di elementi che indicano il numero di volte che una transizione $t_j \in T$ è abilitata ad uno scatto di transizione.

Il significato dell'equazione di stato è che, a partire dalla marcatura iniziale M_0 e la sequenza σ di transizioni $t_j \in T$ che rappresentanta l'evoluzione della rete, è possibile determinare la nuova configurazione della rete M_k.

Esempio 3.1.4 (Equazione di stato)
Come esempio di equazione di stato viene riproposta la sequenza di scatti $\sigma = t_2 t_3$ applicata alla rete di figura 3.1.2 con vettore delle occorrenze $O[0, 1, 1, 0]^T$ e marcatura iniziale della rete di Petri $M_0 = (3, 2, 1, 2, 0)$.

$$\boxed{M_2 = M_0 \ + C \cdot O} \tag{3.1.16}$$

$$M_2 \begin{pmatrix} 3 \\ 1 \\ 0 \\ 3 \\ 1 \end{pmatrix} = M_0 \begin{pmatrix} 3 \\ 2 \\ 1 \\ 2 \\ 0 \end{pmatrix} + C \begin{pmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \cdot O \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

La sequenza di scatti σ applicata alla marcatura iniziale M_0 della rete di Petri rappresenta l'evoluzione della rete attraverso le seguenti marcature:

- $M_0 \xrightarrow{t_2} M_1$ dove $M_1 = (3, 1, 1, 3, 0)$
- $M_1 \xrightarrow{t_3} M_2$ dove $M_2 = (3, 1, 0, 3, 1)$

3.1.3 Invarianti

Definizione 3.1.20 (Invarianti)
Gli invarianti di una rete di Petri sono proprietà strutturali che rimangono conservate in ogni marcatura raggiungibile della rete. Si distinguono in:

- *P-invarianti:* insiemi di nodi place tali per cui la somma pesata dei token che contengono rimane invariata per tutte le marcature della rete appartenenti all'insieme di raggiungibilità;
- *T-invarianti:* vettori di dimensione $|T|$ i cui elementi rappresentano il numero di volte che una transizione $t_j \in T$ deve scattare per rappresentare una marcatura $M_k \in R(N, M_0)$. Essi denotano sequenze di scatti che riportano la rete di Petri nella marcatura iniziale M_0.

Definizione 3.1.21 (P-Invarianti)
Dato l'insieme dei place P si definisce P-invariante di una rete di Petri (N, M_0) un vettore γ di dimensione $|P|$ tale che il suo prodotto per una generica marcatura $M_k \in R(N, M_0)$ rimane invariato:

$$\boxed{\gamma^T \cdot M_k = \gamma^T \cdot M_0 \ \ \forall\ M_k \in R(N, M_0)} \tag{3.1.17}$$

Definizione 3.1.22 (Ricerca di P-Invarianti)
Si possono determinare i P-invarianti di una rete di Petri (N, M_0) a partire dall'equazione di stato 3.1.15:

$$M_k = M_0 \ + C \cdot O$$

moltiplicando entrambi i membri dell'equazione per γ^T

$$\gamma^T \cdot M_k = \gamma^T \cdot M_0 + \gamma^T \cdot C \cdot O$$

per definizione di P-invariante dell'equazione 3.1.17 si deduce che il vettore γ soddisfa l'equazione:

$$\gamma^T \cdot C \cdot O = 0 \ \forall O \neq 0$$

i P-invarianti rappresentano quindi le soluzioni intere di una delle due equazioni:

$$\gamma^T \cdot C = 0 \tag{3.1.18}$$

$$\gamma \cdot C^T = 0 \tag{3.1.19}$$

Le equazioni 3.1.18 e 3.1.19 hanno infinite soluzioni, per cui calcolato un P-invariante, per combinazione lineare si possono ottenere infiniti P-invarianti. Dato che è irragionevole calcolare infiniti P-invarianti, si restringe il calcolo alla ricerca di un gruppo di P-invarianti, denominato *P-invarianti minimi*, in grado di generare tutte le soluzioni delle equazioni. Perchè un P-invariante sia *minimo* deve essere *canonico* e *a supporto minimo.*

Definizione 3.1.23 (P-invariante canonico)
Un P-invariante è canonico se il M.C.D dei suoi elementi non nulli è pari a 1.

Definizione 3.1.24 (Supporto minimo di un P-invariante)
Il supporto di un P-invariante è l'insieme degli elementi non nulli del vettore γ *ed è minimo se non contiene il supporto di nessun altro P-invariante della rete.*

Un P-invariante, canonico e a supporto minimo, non è combinazione lineare di nessun altro P-invariante.

Definizione 3.1.25 (T-invarianti)
Dato l'insieme delle transizioni T *si definisce T-invariante di una rete di Petri* (N, M_0) *un vettore colonna* η *di dimensione* $|T|$ *i cui elementi individuano sequenze di scatto cicliche che riportano la rete nella marcatura iniziale* m_0*:*

$$\boxed{M_k = M_0 + C \cdot \eta \text{ e } M_k = M_0, \text{ con } M_k \in R(N, M_0)} \tag{3.1.20}$$

I T-invarianti calcolati permettono di ritornare alla sequenza iniziale solo se il vettore η rappresenta una sequenza di scatti applicabile.

Definizione 3.1.26 (Ricerca di T-Invarianti)
Data una sequenza di scatti σ *si possono determinare i T-invarianti di una rete di Petri* (N, M_0) *a partire dall'equazione di stato 3.1.15:*

$$M_k = M_0 + C \cdot O$$

per definizione di T-invariante otteniamo

$$M_0 + C \cdot \eta = M_0$$

i T-invarianti rappresentano quindi le soluzioni intere dell'equazione

$$C \cdot \eta = 0 \tag{3.1.21}$$

L'equazione 3.1.21 ha infinite soluzioni, per cui da un T-invariante, per combinazione lineare si possono ottenere infiniti T-invarianti. Dato che è irragionevole calcolare infiniti T-invarianti, si restringe il calcolo alla ricerca di un gruppo di T-invarianti denominato *T-invarianti minimi* in grado di generare tutte le soluzioni dell'equazione. Perchè un T-invariante sia *minimo* deve essere *a supporto minimo.*

Definizione 3.1.27 (Supporto minimo di un T-invariante)
Il supporto di un T-invariante γ è l'insieme delle transizioni con cardinalità maggiore di zero nell'invariante stesso, è minimo se non esistono altri T-invarianti γ' il cui supporto è strettamente incluso in γ.

Esempio 3.1.5 (P-invarianti e T-invarianti)
Determinazione dei P-invarianti e T-invarianti della rete di Petri di figura 3.1.4:

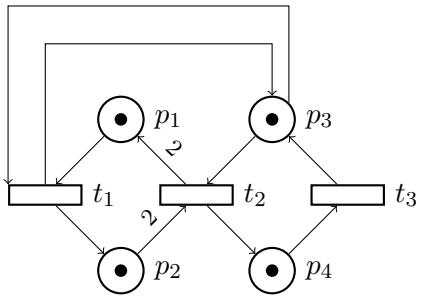


Figura 3.1.4: Esempio di P-invarianti e T-invarianti in una rete di Petri.

$$Pre = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \quad Post = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{pmatrix} 0 & 2 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \end{array} \quad C = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{pmatrix} -1 & 2 & 0 \\ 1 & -2 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{pmatrix} \end{array}$$

$$\text{P-invarianti: } \gamma \cdot C^T \Rightarrow \begin{cases} -\gamma_1 + \gamma_2 = 0 \\ 2\gamma_1 - 2\gamma_2 - \gamma_3 + \gamma_4 = 0 \\ \gamma_3 - \gamma_4 = 0 \end{cases} \Rightarrow \begin{cases} \gamma_1 = \gamma_2 \\ \gamma_3 = \gamma_4 \end{cases}$$

Ponendo $\gamma_1 = 1, \gamma_3 = 0$ e $\gamma_1 = 0, \gamma_3 = 1$ si ottengono i P-invarianti minimi: $P'_{min} = (1, 1, 0, 0)^T$ e $P''_{min} = (0, 0, 1, 1)^T$.

$$\text{T-invarianti: } C \cdot \eta = 0 \Rightarrow \begin{cases} -\eta_1 + 2\eta_2 = 0 \\ \eta_1 - 2\eta_2 = 0 \\ -\eta_2 + \eta_3 = 0 \\ \eta_2 - \eta_3 = 0 \end{cases} \Rightarrow \begin{cases} \eta_1 = 2\eta_2 \\ \eta_2 = \eta_3 \end{cases}$$

Ponendo $\eta_2 = 1$ si ottiene l'unico T-invariante minimo: $T'_{min} = (2, 1, 1)^T$.

3.2 Reti di Petri e vie metaboliche

Il materiale relativo a questo paragrafo è tratto da [24] , [15].

3.2.1 Rappresentazione delle vie metaboliche

Reti di Petri e vie metaboliche presentano delle affinità per esempio entrambe sono costituite da una serie di reazioni che consumano e producono risorse. Tale somiglianza permette quindi di utilizzare le reti di Petri nella modellazione e nell'analisi delle vie metaboliche.

Per derivare una rappresentazione di una reazione chimica tramite una rete di Petri si assegna a ciascun composto chimico un nodo place, le relazioni tra composti biochimici sono stabilite da reazioni chimiche rappresentate da nodi transition e i token definiscono la quantità di sostanza associata ai composti. Gli archi sono disposti in modo tale che i place in entrata ad un arco corrispondano ai reagenti della reazione interessata mentre quelli in uscita ai prodotti. Le molteplicità degli archi rappresentano i valori stechiometrici delle equazioni stechiometriche.

Nelle vie metaboliche le reazioni chimiche sono rappresentate da *equazioni stechiometriche* descritte in 2.5.1 che definiscono una matrice, detta *matrice stechiometrica*, in cui le righe rappresentano i composti e le colonne le reazioni.

Gli elementi di questa matrice sono chiamati *coefficienti stechiometrici* e vengono utilizzati per rappresentare sinteticamente l'insieme delle reazioni di una via metabolica.

I coefficienti stechiometrici della matrice possono assumere un valore:

- *positivo:* quando il relativo composto viene prodotto dalla reazione;
- *negativo:* quando il relativo composto viene consumato dalla reazione;
- *nullo:* quando il relativo composto non partecipa alla reazione.

La matrice stechiometrica corrisponde alla matrice di incidenza di una via metabolica analizzata nella rete di Petri e non fornisce alcuna informazione né sulle leggi cinetiche che governano le reazioni né sulle concentrazioni dei composti.

La figura 3.2.1 propone la rete di Petri che rappresentata l'equazione chimica bilanciata della creazione di una molecola d'acqua:

$$2H_2 + O_2 \xrightarrow{\sigma(t_1)} 2H_2O \tag{3.2.1}$$

La reazione chimica rappresenta per la rete di Petri una sequenza di scatto $\sigma\{t_1\}$ che estrae due token dal nodo place H_2 uno dal nodo place O_2 e ne deposita due nel nodo place H_20. Vengono poi calcolate le matrici $Pre/Post/C$ della rete di Petri e la matrice *Stechiometrica* che coincide con la matrice di incidenza C della rete di Petri.

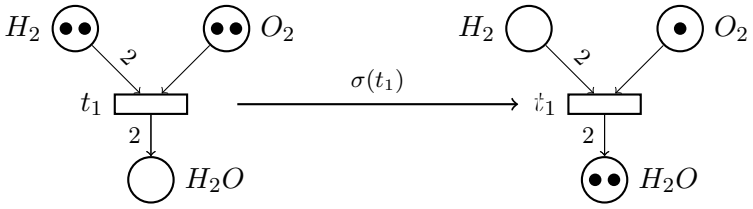


Figura 3.2.1: Rete di Petri della reazione chimica: $2H_2 + O_2 \xrightarrow{\sigma(t_1)} 2H_2O$

$$Pre = \begin{matrix} \\ H_2 \\ O_2 \\ H_2O \end{matrix} \begin{matrix} t_1 \\ \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \end{matrix} \quad Post = \begin{matrix} \\ H_2 \\ O_2 \\ H_2O \end{matrix} \begin{matrix} t_1 \\ \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \end{matrix}$$

$$C = \begin{matrix} \\ H_2 \\ O_2 \\ H_2O \end{matrix} \begin{matrix} t_1 \\ \begin{pmatrix} -2 \\ -1 \\ 2 \end{pmatrix} \end{matrix} \quad Stechiometrica = \begin{matrix} \\ H_2 \\ O_2 \\ H_2O \end{matrix} \begin{matrix} t_1 \\ \begin{pmatrix} -2 \\ -1 \\ 2 \end{pmatrix} \end{matrix}$$

3.2.2 Analisi delle vie metaboliche

Una via metabolica modellata da una rete di Petri può essere analizzata indivuandone i suoi invarianti. I P-invarianti costituiscono una rappresentazione della legge di conservazione delle masse delle sostanze chimiche. I T-invarianti assumono un ruolo decisamente importante in quanto, possono indicare la presenza di stati stazionari in cui

la concentrazione delle sostanze dei composti ha raggiunto uno stato di equilibrio. Una rete di Petri per la modellazione di una via metabolica deve essere quindi caratterizzata dalle seguenti proprietà:

- la rete deve essere coperta da T-invarianti minimi. Non deve essere possibile la presenza di transizioni che non appartengono a nessun T-invariante minimo in quanto in una via metabolica ogni reazione deve far parte di almeno un processo chimico e tali processi sono ciclici;
- la rete deve essere coperta da P-invarianti minimi, non devono essere presenti place che non appartengono ad alcun P-invariante.

Capitolo 4

La Banca dati KEGG

In questo capitolo sarà introdotto KEGG che è uno dei maggiori database per vie metaboliche, le informazioni sono tratte dal sito web di KEGG [7].

4.1 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) [7] è stato sviluppato dalla Kyoto University ed è una delle più importanti banche dati bioinformatiche che raccoglie informazioni chimiche e genetiche di molti organismi.

KEGG attualmente integra i seguenti database:

- *KEGG PATHWAY*
- *KEGG BRITE*
- *KEGG MODULE*
- *KEGG DISEASE*
- *KEGG DRUG*
- *KEGG ORTHOLOGY*
- *KEGG GENOME*
- *KEGG LIGAND*

Il database oggetto di analisi è *KEGG PATHWAY* che contiene informazioni strutturate, sotto forma di mappe interattive navigabili con browser e opportunamente catalogate, sulle vie metaboliche di tutti gli organismi finora sequenziati.

4.1.1 Accedere ai dati di KEGG

KEGG è accessibile da interfaccia web all'indirizzo *http://www.genome.jp/kegg/*. Tramite un'architettura basata su Web Service le applicazioni di KEGG sono in grado di mettere a disposizione degli utenti i dati dei suoi database nelle due modalità: consultazione e download. Per operazioni di download è possibile utilizzare o la connessione con protocollo *ftp:* che richiede una sottoscrizione ad un abbonamento a pagamento, o la connessione con protocollo *http:* per un uso esclusivamente accademico. Per gli sviluppatori di applicazioni sono inoltre disponibili delle applicazioni chiamate KEGG API [1] basate su REST-based API. [2]

La figura 4.1.1 mostra la home page di KEGG.

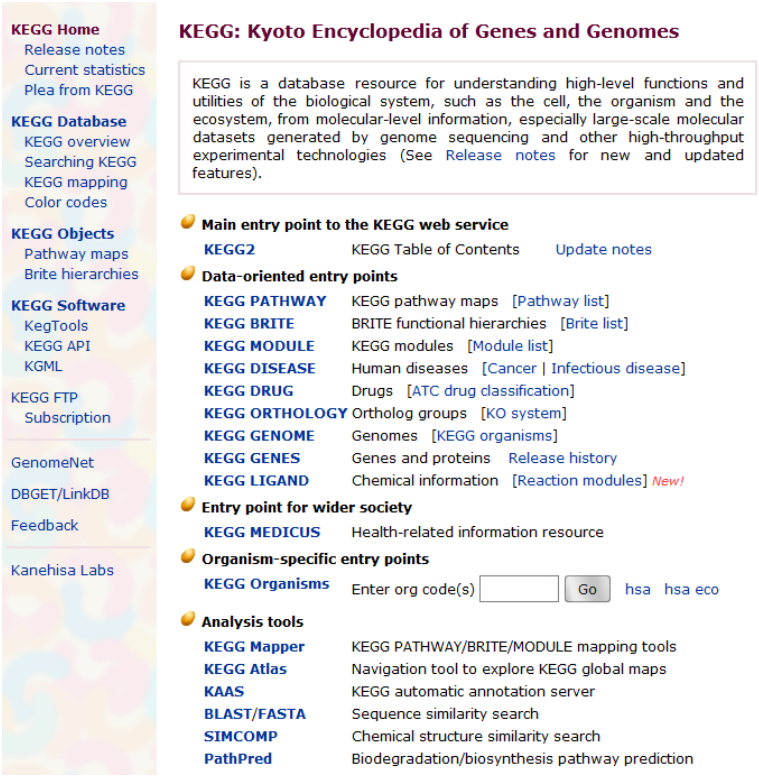


Figura 4.1.1: KEGG: Kyoto Encyclopedia of Genes and Genomes

[1] Una API (Application Programming Interface) è una particolare interfaccia che librerie, software o piattaforme possono usare per interagire con un programma.

[2] REST (REpresentation State Transfer) è un tipo di architettura software per i sistemi di ipertesto distribuiti come il World Wide Web.

4.1.2 Formato KGML di KEGG

Le vie metaboliche degli organismi presenti in KEGG PATHWAY sono generalmente rappresentati tramite mappe corredate da informazioni su proteine, geni, reazioni, enzimi reperibili anche in database diversi. KEGG mette a disposizione degli utenti, per elaborazioni personali, un ulteriore valida rappresentazione dei dati in forma testuale che utilizza il linguaggio proprietario KGML (KEGG Markup Language) basato su XML e il cui contenuto informativo è accessibile tramite query.

La figura 4.1.2 visualizza in formato KGML gli attributi della struttura delle informazioni memorizzate in una mappa grafica di una via metabolica.

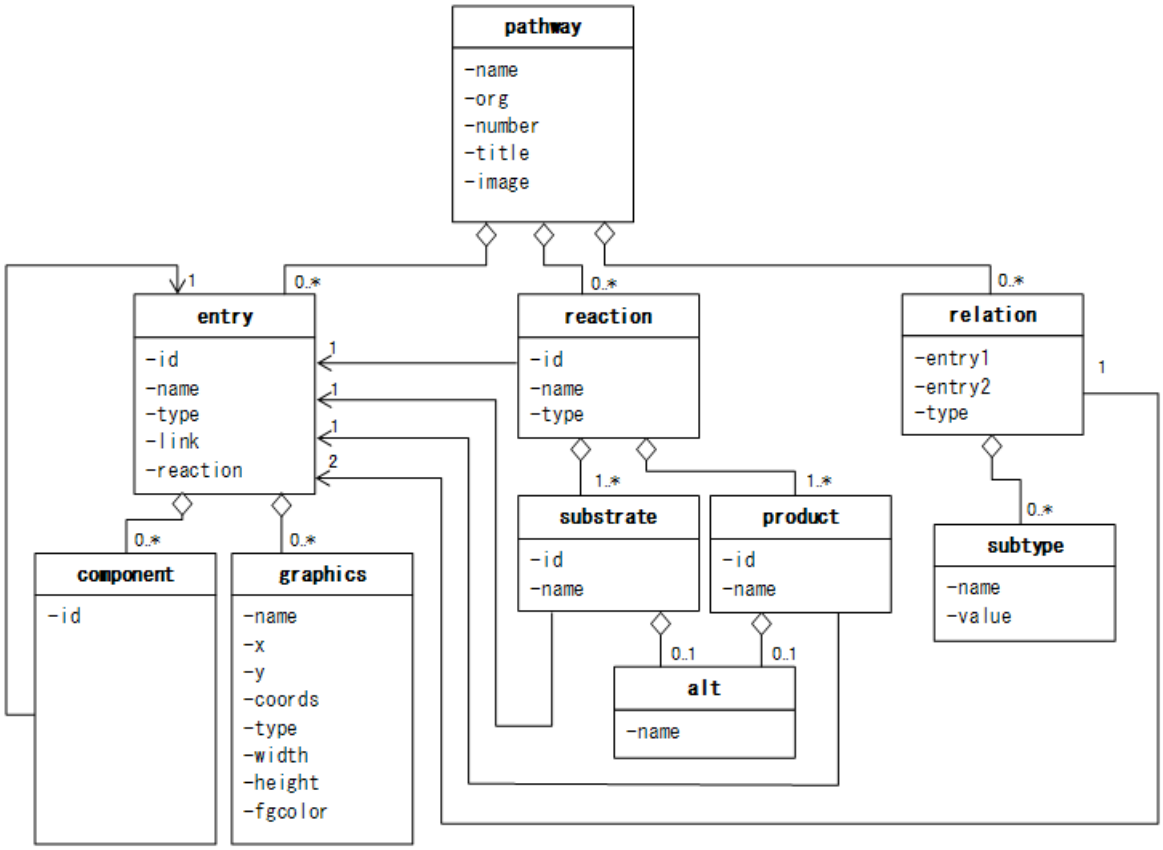


Figura 4.1.2: Rappresentazione in KGML di una via metabolica

La figura 4.1.2 visualizza come esempio di via metabolica la *Glycolysis*.

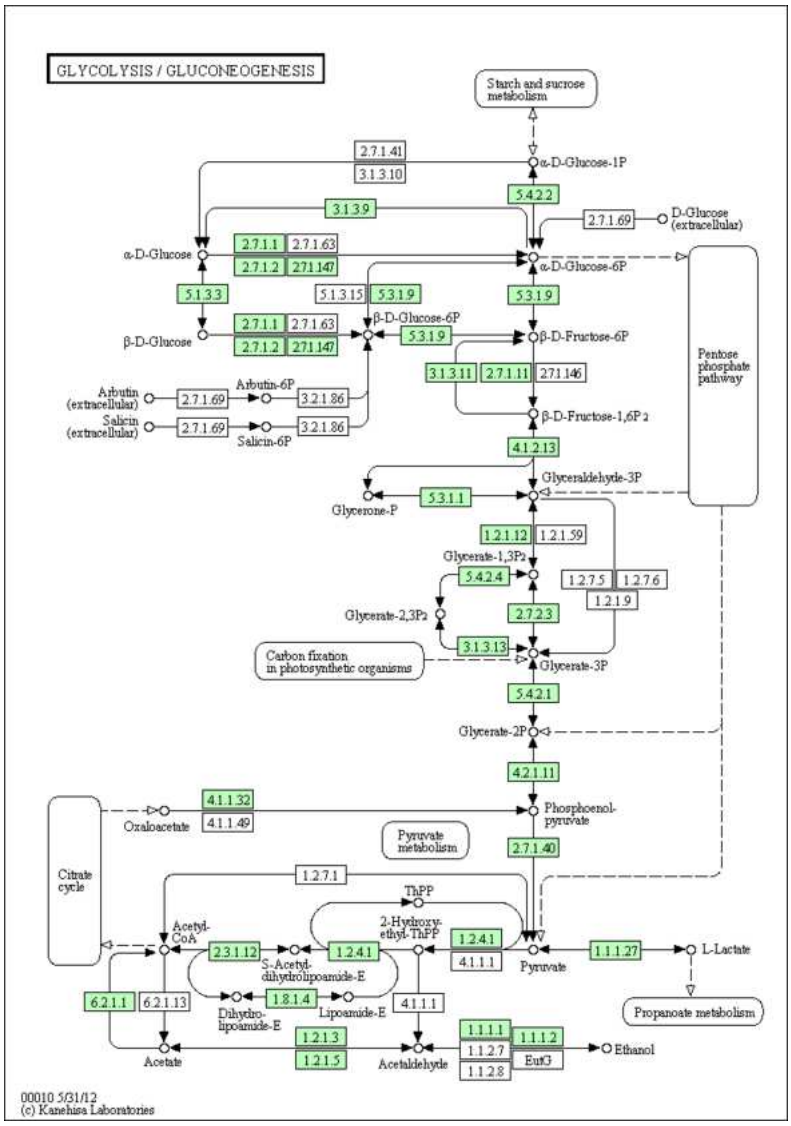


Figura 4.1.3: Via metabolica: Glycolysis Homo Sapiens

Capitolo 5

Vie metaboliche e distanze tra organismi

In questo capitolo verranno brevemente illustrate le proposte presenti in letteratura per il confronto di vie metaboliche e per il calcolo di distanze tra organismi. Le fonti informative sono tratte dai documenti [14], [28], [25], [19], [8].

5.1 Comparazione di vie metaboliche

Il confronto di vie metaboliche di specie diverse è estremamente importante in quanto ci permette di ricavare informazioni sulla loro evoluzione. In letteratura sono state proposte tecniche che si diversificano per le informazioni utilizzate, per le rappresentazioni semplificate delle vie metaboliche, per i tool di analisi specifici e per le distanze tra due vie metaboliche. Ognuna di queste tecniche si focalizza su alcuni aspetti delle vie metaboliche ma nessuna di esse riesce a fornire una descrizione esaustiva. A seconda quindi della loro rappresentazione possono essere raggruppate nelle tre classi seguenti:

- *Insiemi:* una via metabolica è definita dall'insieme degli elementi che la compongono (enzimi, e/o composti, reazioni); i confronti tra vie metaboliche di diversi organismi, sfruttano le operazioni insiemistiche di unione e intersezione;
- *Sequenze:* una via metabolica è rappresentata come un insieme di sequenze di reazioni. Le vie metaboliche sono decomposte in un insieme di sequenze di reazioni ed analizzate a partire da un componente iniziale sino a uno finale. Il confronto di vie metaboliche consiste in allineamenti di sequenze.
- *Grafi:* una via metabolica è rappresentata tramite un grafo che permette di considerare sia le componenti chimiche sia le relazioni che intercorrono tra loro. I nodi del grafo rappresentano le reazioni chimiche dei composti. I confronti di vie metaboliche richiedono confronti tra grafi che sono computazionalmente molto onerosi.

La scelta di una tecnica di rappresentazione precedentemente elencata determina la scelta del tipo di distanza tra vie metaboliche utilizzata.

5.2 Classificazione EC number degli enzimi

Quasi tutte le proposte presenti in letteratura per analizzare una via metabolica utilizzano una misura di *somiglianza tra enzimi* poichè le reazioni possono essere identificate con gli enzimi a loro associati. La più semplice misura di somiglianza tra enzimi è l'identità che associa 1 ad enzimi identici e 0 a enzimi non identici. E' comunque possibile definire delle misure di somiglianza tra enzimi più fini. Data l'importanza che gli enzimi assumo in una reazione una commissione si è occupata di elaborare una classificazione numerica sulla base di una reazione chimica catalizzata da un determinato enzima. La classificazione degli enzimi è un sistema di categorizzazione attraverso un numero chiamato *Ec number (Enzyme Commission)*. Ogni enzima viene associato a un numero e a un nome [11], il codice attribuitogli consiste delle lettere *EC* seguite da quattro numeri separati da punti. Tali numeri, letti da sinistra verso destra, rappresentano una classificazione sempre più dettagliata dell'enzima. Un numero Ec indica una specifica reazione, a cui vengono associati tutti gli enzimi in grado di catalizzarla. Se differenti enzimi catalizzano la stessa reazione essi ricevono quindi lo stesso EC number. Gli Ec number sono così classificati a partire dal primo livello (più a sinistra):

Gruppi EC degli enzimi		
Gruppo	**Reazione catalizzata**	**Enzimi tipici**
EC 1 Ossidoreduttasi	Reazioni di ossidoriduzione	Deidrogenasi, ossidasi
EC 2 Transferasi	Trasferimento di un gruppo funzionale tra substrati	Transaminasi, chinasi
EC 3 Idrolasi	Formazione di due prodotti a partire da uno.	Lipasi, amilasi, proteasi
EC 4 Isomerasi	Addizione/rimozione di gruppi funzionali dal substrato	Decarbossilasi
EC 5 Isomerasi	Riarrangiamento intramolecolare.	Isomerasi, mutasi
EC 6 Ligasi	Unione di due molecole.	Sintetasi, polimerasi

Tabella 5.2.1: Significato del numero più a sinistra nell' EC number degli enzimi

Esempio 5.2.1 (EC number: **EC 2.7.1.1**)
Il codice EC 2.7.1.1 *corrispondente a esocinasi. La sua interpretazione è la seguente:*

- 2 ⇒ *l'enzima trasferisce gruppi*
- 7 ⇒ *l'enzima trasferisce un gruppo fosforico*

- $1 \Rightarrow$ *l'enzima trasferisce un gruppo fosforico ricevente* $-OH$
- $1 \Rightarrow$ *l'enzima trasferisce un gruppo fosforico ricevente* $-OH$ *di un esoso.*

5.2.1 Similarità gerarchica tra enzimi

La similarità gerarchica tra enzimi si basa sulla classificazione gerarchica a quattro livelli dell' EC-number degli enzimi.

Definizione 5.2.1 (Similarità gerarchica tra enzimi)
Dati due enzimi $e_1 = n_1.n_2.n_3.n_4$ *e* $e_2 = n'_1.n'_2.n'_3.n'_4$ *la similarità gerarchica* S *tra* e_1 *e* e_2 *è definita dalla seguente equazione:*

$$\boxed{S\left(e_1, e_2\right) = \frac{max[i : n_i = n'_i]}{4}} \tag{5.2.1}$$

dove: $S(e_1, e_2)$ esprime la lunghezza massima dei prefissi che sono comuni agli EC number degli enzimi.

Esempio 5.2.2 (Similarità gerarchica tra l'enzima 'asparaginase' e 'glutaminase')
Dati gli EC-number di: asparaginase ($e_1 = 3.5.1.1$) *e glutaminase* ($e_2 = 3.5.1.2$)*, si determina prima il numero di classi comuni ai due EC number e poi si calcola la misura di similarità applicando la formula espressa in 5.2.1.*

Similarità gerarchica tra enzimi

EC number asparaginase	3	.	5	.	1	.	1
EC number glutaminase	3	.	5	.	1	.	2
match delle classi	=		=		=		$\neq$
$S(asparaginase, glutaminase) = 3/4 = 0.75$							

Tabella 5.2.2: Similarità gerarchica tra enzima asparaginase e glutaminase

5.3 Misure di similarità tra vie metaboliche

Sulla base della misura di somiglianza tra enzimi è possibile definire una misura di somiglianza tra vie metaboliche rappresentate come insiemi di reazioni, o meglio degli enzimi a loro associati.

5.3.1 Indici di similarità e distanze con insiemi

Definizione 5.3.1 (Identità tra enzimi)
L'identità tra enzimi è la misura di somiglianza più semplice che si possa applicare per determinare la presenza o la non presenza di un enzima in una reazione di una via metabolica. Se l'enzima è presente viene assegnata alla misura di somiglianza il valore 1 *altrimenti il valore* 0*.*

Definizione 5.3.2 (Indice di Jaccard)
L'indice di Jaccard è noto anche come coefficiente di similarità di Jaccard, è un indice statistico che è stato utilizzato per confrontare la similarità di vie metaboliche rappresentate con insiemi. L'indice è espresso da un numero compreso tra 0 *e* 1*, più è vicino a* 1 *più il grado di correlazione è alto. E' definito come:*

$$\boxed{J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}} \tag{5.3.1}$$

dove X e Y rappresentano i due insiemi da comparare.

Definizione 5.3.3 (Distanza di Jaccard)
La distanza di Jaccard è complementare al coefficiente di Jaccard e misura la dissimilarità tra due insiemi X e Y. E' definita da:

$$\boxed{d_J(X,Y) = 1 - J(X,Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|}} \tag{5.3.2}$$

Definizione 5.3.4 (Matrice delle distanze)
Data una qualsiasi misura di distanza è possibile costruire una matrice quadrata simmetrica di cardinalità pari al numero di organismi analizzati, che rappresenta le distanze tra tutte le coppie di organismi considerati. Questa tipologia di matrice viene utilizzata per rappresentare la distanza evolutiva tra gli organismi e per la costruzione degli alberi filogenetici che rappresentano in forma grafica le relazioni fondamentali di discendenza comune di gruppi di organismi.

La rappresentazione estremamente semplice delle vie metaboliche con insiemi di reazioni presenta l'inconveniente di non considerare il fatto che uno stesso enzima può essere associato a più reazioni in una via metabolica. Per poter quindi annotare tale caratteristica degli enzimi si preferisce rappresentare le vie metabolcihe non più con insiemi ma con *multi-insiemi* che permettono di associare ad ogni elemento la sua cardinalità.

Con i multi-insiemi è quindi possibile rappresentare la molteplicità degli elementi che li compongono.
Di seguito viene data la definizione formale di *multi-insieme.*

Definizione 5.3.5 (Multi-insieme)
Un multi-insieme [13] è una coppia (X, m_x) *dove* X *è un insieme e* $m_x : X \longrightarrow \mathbb{N}^+$ *è la funzione di molteplicità che associa a ciascun elemento* $x \in X$ *un numero naturale positivo che esprime il numero delle sue occorrenze nell'insieme* X*. La cardinalità di un multi-insieme è così definita:*

$$| (X, m_x) | = \sum_{z \in X} m_x (z) \tag{5.3.3}$$

Dati quindi i due multi-insiemi X, m_x *e* Y, m_y *si ha:*

$$(X, m_x) \cap (Y, m_y) = (X \cap Y, m_{x \cap y}) \tag{5.3.4}$$

dove:

$$m_{x \cap y} (z) = min (m_x (z), m_y (z)), \ \forall \ z \in X \cap Y. \tag{5.3.5}$$

Per poter valutare la somiglianza di due multi-insiemi considerando la molteplicità degli elementi che li compongono può essere utilizzato o l'indice di Sørensen o quello di Tanimoto definito su multi-insiemi. Sørensen e Tanimoto sono indici statistici in alternativa a quello di Jaccard.

Definizione 5.3.6 (Indice di Sørensen)
Dati due insiemi X, Y *l'indice di Sørensen si definisce come:*

$$\boxed{S(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}} \tag{5.3.6}$$

dove: $|X|$ *e* $|Y|$ *indicano le cardinalità dei due insiemi.*

Definizione 5.3.7 (Distanza di Sørensen)
La distanza di Sørensen è complementare al coefficiente di Sørensen e misura la dissimilarità tra due insiemi X, Y*. E' definita da:*

$$\boxed{d_S (X, Y) = 1 - S (X, Y)} \tag{5.3.7}$$

dove S *rappresenta l'indice di Sørensen.*

Le definizioni si possono estendere al caso di multi-insiemi.

Definizione 5.3.8 (Indice di Tanimoto)
L'indice di Tanimoto è un estensione dell'indice statistico di Jaccard. Dati due insiemi X, Y, l'indice di Tanimoto si definisce come:

$$\boxed{T(X,Y) = \frac{|X \cap Y|}{|X| + |Y| - |X| \cap |Y|}} \tag{5.3.8}$$

dove X, Y rappresentano i due insiemi, $|X|$ e $|Y|$ le cardinalità degli insiemi X, Y ed $|X| + |Y| - |X| \cap |Y|$ è la cardinalità degli elementi comuni agli insiemi X, Y.

Anche l'indice di Tanimoto può essere esteso al caso di multi-inisemi.

In ambito biologico se si sceglie di rappresentare due vie metaboloiche come multi-insiemi di reazioni, l'indice di Tanimoto esprime il rapporto tra il numero di reazioni comuni delle due vie metaboliche e il numero di reazioni non comuni ad esse.

Capitolo 6

Tecniche di allineamento di sequenze

Poichè il confronto tra vie metaboliche può essere inteso come un allineamento, in questo capitolo richiamiamo, i concetti fondamentali alla base dell'allineamento di sequenze, operazione fondamentale e molto studiata in bioinformatica. Il materiale di questo capitolo è principalmente tratto da [2, 10, 22].

6.1 Rappresentazione di sequenze biologiche

In biologia è possibile rappresentare una sequenza tramite delle stringhe di caratteri utilizzando la convenzione IUPAC (International Union of Pure Applied Chemistry) [5]. Secondo tale codice sia gli aminoacidi sia i nucleotidi possono essere rappresentati utilizzando delle lettere dell'alfabeto. Una generica sequenza biologica s di lunghezza n viene quindi rappresentata come una stringa di n caratteri appartenenti ad un dato alfabeto.

In biologia molecolare un'operazione fondamentale, utilizzata per affrontare la gran parte dei problemi è il confronto tra sequenze, detto *allineamento*.

6.2 Allineamento globale e locale di sequenze

Definizione 6.2.1 (Allineamento di sequenze biologiche)
L'allineamento di sequenze è una procedura bioinformatica nella quale vengono messe a confronto e allineate due o più sequenze di nucleotidi o aminoacidi.

Gli allineamenti tra sequenze permettono di risolvere problemi bioinformatici molto importanti quali: determinare la funzione di una sequenza genetica appena scoperta, comprendere la relazione evolutiva tra geni, proteine e specie, predirre la struttura. Con l'allineamento di sequenze, e opportune funzioni di distanza che possono fornire una misura di *quanto sono distanti due sequenze*, è possibile determinare il grado di

similarità tra sequenze. In questo contesto, vengono presentati i seguenti due tipi di allineamento:

- *allineamento globale:* cerca la corrispondenza ottimale delle due sequenze secondo la loro intera lunghezza, si ottine inserendo degli spazi, all'interno o agli estremi delle sequenze in modo tale da sovrapporre perfettamente le due sequenze ;

- *allineamento locale:* non richiede che il confronto delle due sequenze sia svolto per l'intera lunghezza ma solo per le zone con più alta densità di somiglianza. Lo scopo di un allineamento locale è quello di trovare nelle sequenze biologiche regioni più conservate di altre che indicano parti della sequenza in cui la struttura o la funzione biologica si è mantenuta nel tempo.

Gli esempi di tabella 6.2.1 e tabella 6.2.2 mostrano come due sequenze S, T possano essere allineate in modo globale e locale.

Allineamento globale																			
S:	T	A	G	A	R	D	S	E	D	K	P	B	W	T	M	W	D	P	E	K
	|				|		|	|	|	|			|	|							|
T:	T	P	–	I	R	L	S	E	D	K	D	C	W	T	N	D	Z	R	H	K

Tabella 6.2.1: Allineamento globale di due sequenze.

L'allineamento locale è stato fatto per una zona fortemente conservativa.

Allineamento locale						
S:		T	T	A	D	G
			|	|	|	|
T:		G	T	A	D	G

Tabella 6.2.2: Allineamento locale di due sequenze.

6.3 Allineamento di sequenze a coppie

In base al numero di sequenze coinvolte nell'allineamento possiamo avere: un allineamento di sequenze multiple o un allineamento di sequenze a coppie. Si è scelto di soffermarsi esclusivamente sull'allineamento a coppie di cui viene data una definizione formale.

Definizione 6.3.1 (Allineamento a coppie)
Date due sequenze $S = s_1 s_2 ... s_n$, $T = t_1 t_2 ... t_m$ *definite su un alfabeto di simboli* $\sum$, *un allineamento di* S *e* T *consiste in una coppia di sequenze* S', T' *sull'alfabeto* $\sum \cup \{-\}$ *che godono delle seguenti proprietà:*

- $|S'| = |T'| = l \ \ (max\,(n, m) \leq l \leq (n + m))$;
- l'eliminazione dei gap dalla sequenza S' permette di ottenere la sequenza S;
- l'eliminazione dei gap dalla sequenza T' permette di ottenere la sequenza T.

dove il simbolo $-$ è chiamato gap.
L'esempio di tabella 6.3.1 mostra l'allineamento a coppie di due sequenze S, T.

Esempio 6.3.1 (Allineamento a coppie di due sequenze S, T)
Date le sequenze $S = (ATGTATAG)$ *e* $T = (GATCTAAG)$ *un possibile allineamento è:*

	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
S	–	A	T	G	T	A	T	A	G
		|	|		|	|		|	|
T	G	A	T	C	T	A	–	A	G

Tabella 6.3.1: Allineamento a coppie di sequenze

L'allineamento delle due sequenze S, T può essere interpretato o come una trasformazione di S in T o come una trasformazione di T in S. Nell'esempio proposto la trasformazione è di S in T ed interessa simboli di S delle seguenti posizioni:

- posizione n_1 : inserimento di un simbolo G;
- posizione n_4 : sostituzione di un simbolo G con un C;
- posizione n_7 : cancellazione del simbolo T;

Una tale trasformazione in campo biologico esprime la possibilità che a seguito di mutazioni genetiche alcuni caratteri che compongono le sequenze da allineare possano essere stati interessati dai seguenti cambiamenti:

- una *mutazione:* quando dei caratteri vengono sostituiti con altri caratteri;
- un'*inserzione:* quando in una sequenza si aggiungono dei caratteri;
- una *delezione:* quando in una sequenza dei caratteri vengono eliminati.

Se tali cambiamenti sono in numero limitato allora le sequenze sono vicine da un punto di vista evolutivo e possono rappresentare sequenze di proteine con struttura e funzioni simili.

6.4 Similarità tra due sequenze biologiche

Per misurare la similarità tra due sequenze biologiche è necessario tenere conto che le sequenze genetiche subiscono nella loro evoluzione delle mutazioni dovute principalmente a eventi del tipo: sostituzioni, cancellazioni, inserzioni, inversioni e ricombinazione dei caratteri che le compongono. In questo contesto la similarità tra due sequenze può discendere da relazioni di omologia strutturale o funzionale che ha avuto origine da un antenato comune. Per esprime la relazione strutturale o funzionale tra sequenze è fondamentale quindi misurare quantitativamente la bontà di un allineamento tra due sequenze. Un metodo per la valutazione della similarità, che tiene conto di quanto detto in precedenza, è il sistema a punteggio che assegna ad ogni coppia di simboli presenti nelle due sequenze da allineare un valore numerico. Se i simboli delle due sequenze presi a coppie sono simili si assegnerà un punteggio positivo mentre nel caso di simboli non identici il punteggio costituirà una penalità. A tale punteggio si può dare un'interpretazione probabilistica, esso esprime la probabilità che le due sequenze comparate siano o meno correlate evolutivamente o funzionalmente.
Illustriamo il metodo a punteggio dandone una interpretazione probabilistica e suddividendolo nei due casi:

- modello a punteggio senza gap;
- modello a punteggio con gap.

6.4.1 Interpretazione del modello a punteggio senza gap

Consideriamo due sequenze $x = x_1...x_n$ ed $y = y_1...y_n$ definite su un alfabeto $\sum$ ed effettuiamo un allineamento globale senza gap.
Il punteggio dell'allineamento è dato dal rapporto tra i valori generati da due modelli probabilistici, il modello *random* e il modello *match*.

Definizione 6.4.1 (Modello random)
Il modello random è definito dalla seguente formula:

$$\boxed{P\left(x, y|Random\right) = \prod_{\forall_{i\in n}} f_{x_i} \prod_{\forall_{j\in n}} f_{y_j}} \tag{6.4.1}$$

dove f_{x_i} e f_{y_j} indicano la frequenza con cui l'i-esimo simbolo della sequenza x e il j-esimo della sequenza y avviene indipendentemente.

L'equazione (6.4.1) esprime la probabilità che due sequenze x e y non siano tra loro correlate. Se le due sequenze sono sequenze biologiche appartenenti a due organismi il risultato può essere interpretato come la probabilità che i due organismi non derivino da un antenato comune.

Definizione 6.4.2 (Modello match)
Il modello match è definito dalla seguente formula:

$$\boxed{P\left(x,y|Corr\right)=\prod_{\forall_{i,j\in n}} p_{x_iy_j}} \tag{6.4.2}$$

dove $p_{x_iy_j}$ indica la probabilità congiunta che i simboli x_i e y_i siano allineati.

L'equazione 6.4.2 esprime la probabilità che due sequenze x e y siano tra loro correlate, in biologia il risultato può essere interpretato come la probabilità che due sequenze appartenenti a due organismi derivino da un antenato comune non noto.

Definizione 6.4.3 (Punteggio di allineamento)
Il punteggio di allineamento di due sequenze che utilizzano i modelli random e match è ottenuto dal rapporto dei valori ottenuti dalle equazioni (6.4.1) e (6.4.2):

$$P\left(x,y\right)=\frac{P\left(x,y|Match\right)}{P\left(x,y|Random\right)}=\prod_{\forall i,j\in n}\frac{p_{x_iy_j}}{f_{x_i}f_{y_j}} \tag{6.4.3}$$

E' possibile modificare lo schema dell'equazione (6.4.3) ottenendo un punteggio additivo, passando per i logaritmi,

$$\begin{aligned}P_{Additivo}\left(x,y\right)&=log\frac{P\left(x,y|\text{Match}\right)}{P\left(x,y|\text{Random}\right)}\\&=log\prod_{\forall_{i,j\in n}}\frac{p_{x_iy_j}}{f_{x_i}f_{y_j}}\\&=\sum_{\forall i,j\in n}log\frac{p_{x_iy_j}}{f_{x_i}f_{y_j}}\\&=\sum_{\forall i,j\in n}Punteggio\left(x_i,y_j\right)\end{aligned}$$

$$\boxed{P_{Additivo}\left(x,y\right)=\sum_{\forall i,j\in n}Punteggio\left(x_i,y_j\right)} \tag{6.4.4}$$

Tale formula risulta essere la somma di una serie di punteggi di match di coppie di simboli il cui valore può essere fornito da particolari matrici denominate MATRICI DI SOSTITUZIONE.

6.4.2 Interpretazione del modello a punteggio con simboli gap

Come detto in precedenza durante la fase evolutiva la lunghezza delle sequenze biologiche degli organismi può essere stata modificata per cui, nella procedura di allineamento di due sequenze, deve essere considerata la possibilità che debbano essere inseriti dei simboli chiamati gap. L'introduzione di gap in sequenze da allineare è necessaria al fine di poter determinare allineamenti in posizioni diverse.

I gap permettono quindi di ottenere il miglior allineamento possibile. Essi poichè corrispondono a inserzioni o delezioni di elementi nelle sequenze biologiche sono dispendiosi in termini evolutivi e pertanto sono da considerarsi dei costi penalizzanti.

La maggior parte degli algoritmi di allineamento utilizza delle penalità diverse a seconda che si tratti dell'apertura di un nuovo gap o dell'estensione di un gap esistente.

Possiamo definire tale penalità tramite la seguente funzione γ:

$$\gamma(l) = \begin{cases} -le & \text{se il gap è di estensione} \\ -o-(l-1) & \text{se il gap è di nuova apertura} \end{cases} \tag{6.4.5}$$

dove:

- l: è la lunghezza del *gap*;
- e: è la penalità per un gap di estensione;
- o: è la penalità per un gap di nuova apertura.

Introdotta la funzione di penalità gap il calcolo del punteggio di allineamento tra due sequenze è simile al calcolo per allinementi a coppie senza gap del modello a punteggio random e match.

6.4.3 Matrici di sostituzione PAM

Le matrici di sostituzione sono matrici che confrontano coppie di aminoacidi o nucleotidi. Si ottengono con metodi statistici assegnando, a ciascuna delle possibili coppie, un valore che riflette la frequenza con cui un simbolo si sostituisce all'altro.

Le matrici PAM furono proposte nel 1978 da Margaret Dayhoff [12], sulla base di uno studio di filogenesi molecolare compiuto su 71 famiglie di proteine e 1572 sequenze proteiche. L'assunzione di partenza fu quella che analizzando sequenze correlate filogeneticamente, si potesse calcolare la probabilità con cui ogni aminoacido subisce un evento di sostituzione, ovvero quello che viene chiamato una Percent Accepted Mutation, da cui l'acronimo PAM.

Esistono diversi tipi di PAM, per esempio PAM120 PAM250 dove il valore numerico indica sequenze che distano $120, 250$ unità PAM. Una unità di misura di distanza

PAM denota che avviene in media, una mutazione ogni 100 aminoacidi. In aggiunta alle matrici PAM possiamo trovare le matrici BLOSUM introdotte nel 1992 da S. Heinkoff. Differiscono dalle PAM in quanto si basano su allineamenti multipli di blocchi di sequenze strettamente correlate.

Un generico elemento $p(i, j)$ di una matrice PAM è un numero intero: positivo, negativo o nullo, che indica il punteggio da attribuire all'appaiamento di una coppia di aminoacidi.

La tabella in figura 6.4.1 mostra la matrice $PAM120$.

PAM120

	A	R	N	D	C	Q	E	G	H	I	L
A	3	-3	-1	0	-3	-1	0	1	-3	-1	-3
R	-3	6	-1	-3	-4	1	-3	-4	1	-2	-4
N	-1	-1	4	2	-5	0	1	0	2	-2	-4
D	0	-3	2	5	-7	1	3	0	0	-3	-5
C	-3	-4	-5	-7	9	-7	-7	-4	-4	-3	-7
Q	-1	1	0	1	-7	6	2	-3	3	-3	-2
E	0	-3	1	3	-7	2	5	-1	-1	-3	-4
G	1	-4	0	0	-4	-3	-1	5	-4	-4	-5
H	-3	1	2	0	-4	3	-1	-4	7	-4	-3
I	-1	-2	-2	-3	-3	-3	-3	-4	-4	6	1
L	-3	-4	-4	-5	-7	-2	-4	-5	-3	1	5

Tabella 6.4.1: Matrice di sostituzione: PAM120

6.5 Algoritmi di allineamento di sequenze

Nell'allineamento a coppie di sequenze ci si pone come obiettivo la determinazione dell'allineamento ottimale, cioè quello di punteggio massimo. Il metodo più ovvio, la ricerca esaustiva tra tutti i possibili allineamenti, è generalmente non praticabile, in quanto il numero di allineamenti in funzione della lunghezza delle sequenze è esponenziale e richiede un tempo di computazione non accettabile. Si può ovviare a tale problema facendo ricorso alla *programmazione dinamica* o a metodi euristici.

Gli algoritmi di allineamento dinamici, permettono di trovare l'allineamento o gli allineamenti con il punteggio più alto tra tutti quelli possibili, mentre quelli euristici non garantiscono che l'allineamento trovato sia il migliore possibile, ma sono molto più veloci.

Verranno presentati due algoritmi di allineamento di sequenze biologiche: l'algoritmo di *Needleman-Wunsch* [22] del 1970 per allineamenti globali, e quello di *Smith-Waterman* [27], del 1981 per allineamenti locali.

6.5.1 Algoritmo di Needleman-Wunsch

L'algoritmo di Needleman-Wunsch permette di determinare l'allineamento *globale* ottimale attraverso una interpretazione computazionale della matrice dotplot. L'idea di base è quella di calcolare in modo ricorsivo l'allineamento ottimale su sequenze di lunghezza crescente sfruttando l'indipendenza e l'addivitè dei punteggi ottenuti ricorsivamente.

L'algoritmo, a partire dai seguenti dati:

- due sequenze $S = s_1 s_2 \cdots s_n$, $T = t_1, t_2 \cdots t_m$ con $|S| = n$ e $|T| = m$;
- punteggio di match o mismatch tra simboli;
- penalità per un gap.

1. crea inizialmente una matrice P di dimensione $(n+1) \times (m+1)$ in cui la prima riga e la prima colonna sono inizializzate con multipli della penalità di gap;
2. calcola gli elementi $P\,(i, j)$ della matrice in modo ricorsivo per sottosequenze via via più lunghe che corrispondono al punteggio dell'allineamento migliore tra tutte le sottosequenze $s_{1\cdots i}$ e $t_{1\cdots j}$;
3. determina l'allineamento migliore. L'elemento di posizione $P(n+1, m+1)$ rappresenta il punteggio del miglior allineamento possibile, seguendo il/i percorsi a partire dall'elemento $P(n+1, m+1)$ all'elemento $P(1, 1)$ si ottiene l'allineamento o gli allineamenti ottimali.

Descrizione formale dell'algoritmo:

Algoritmo di Needleman-Wunsch

Siano $S = s_1 s_2 c \ldots s_n$ e $T = t_1 t_2 \cdots t_m$ due sequenze da allineare e σ la funzione che esprime il costo o il beneficio della sostituzione di un simbolo così definita:

- $\sigma(s_i, t_j)$: match della coppia s_i e t_j;
- $\sigma(s_i, -)$: allineamento di s_i con un gap;
- $\sigma(-, t_j)$: allineamento di t_j con un gap.

1. Casi base per cui il valore della matrice P è direttamente calcolabile:
   (a) $P(0,0) = 0$;
   (b) $P(i,0) = \sum_{k=0}^{i} \sigma(s_k, -)$;
   (c) $P(0,j) = \sum_{k=0}^{j} \sigma(-, t_k)$.
2. Noti i valori $P(i-1, j-1)$, $P(i-1, j)$, $P(i, j-1)$ calcolare ricorsivamente l'elemento $P(i,j)$ nel seguente modo:

$$P(i,j) = max \begin{cases} P(i-1,j-1) + \sigma(s_i, t_j]) & \text{se match tra i residui } s_i, t_j \\ P(i-1,j) + \sigma(s_i, -) & \text{se } s_i \text{ si allinea con un gap} \\ P(i,j-1) + \sigma(-, t_j) & \text{se } t_j \text{ si allinea con un gap} \end{cases}$$

3. Tramite l'operazione di *traceback*, costruire l'allineamento o gli allineamenti ottimali in modo inverso a partire dall'elemento di posizione $P(n+1, m+1)$ sino all'elemento $P(1,1)$ applicando le seguenti regole:
   (a) $(i,j) \to (i-1, j-1)$ se match tra i residui s_i, t_j;
   (b) $(i,j) \to (i-1, j)$ se $s_i \in S$ è allineato con un gap;
   (c) $(i,j) \to (i, j-1)$ se $t_j \in T$ è allineato con un gap.

Esempio 6.5.1 (Algoritmo di Needleman-Wunsh)
Date due sequenze $S = TCTCA$ *e* $T = TCAGTAA$ *determinare l'allineamento ottimale applicando un valore di penalità gap* $= -1$*, un valore di match* $= +1$ *e di mismatch* $= 0$*.*

L'algoritmo di Needleman-Wunsch produce la seguente matrice di sostituzione:

		T	C	T	C	A
	0	-1	-2	-3	-4	-5
T	-1	1	0	-1	-2	-3
C	-2	0	2	1	0	-1
A	-3	-1	1	2	1	0
G	-4	-2	0	1	2	2
T	-5	-3	-1	1	1	2
A	-6	-4	-2	0	1	1
A	-7	-5	-3	-1	0	2

Figura 6.5.1: Matrice di Needleman−Wunsch

Il percorso, dall'elemento di posizione $P(n+1, m+1)$ all'elemento $P(1,1)$ della matrice di figura 6.5.1 determina l'allineamento ottimo. Ripercorrendo tale percorso, seguendo le frecce come evidenziato in figura 6.5.1, si ottiene l'allineamento ottimo.

S :	T	C	-	-	T	C	A
T :	T	C	A	G	T	A	A

Figura 6.5.2: Needleman−Wunsch allineamento globale ottimo

6.5.2 Algoritmo di Smith-Waterman

L'algoritmo, proposto da T.F. Smith e M.S. Waterman nel 1981 come variante dell'algoritmo di Needleman-Wunsch, permette di calcolare la massima similarità locale tra due sequenze. Quando si lavora con sequenze lunghe diverse migliaia, o anche milioni di nucleotidi, i metodi di allineamento locale possono identificare sottosequenze corrispondenti che sarebbero impossibili da trovare usando allineamenti globali.

L'algoritmo inizializza la prima riga e colonna della matrice P al valore zero, successivamente in base a delle regole fissate, calcola i rimanenti elementi della matrice $P(i, j)$.

Determina l'elemento massimo in P e, a partire dalla posizione in cui esso si trova procede con un'operazione di *backtracking* finchè si raggiunge un valore di soglia assegnato (usualmente la soglia viene posta a zero), determinando così tutti gli allineamenti locali di punteggio superiore alla soglia.
La descrizione formale dell'algoritmo è:

Algoritmo di Smith-Waterman

Siano $S = s_1 s_2 c \dots s_n$ e $T = t_1 t_2 ... t_m$ due sequenze da allineare e σ la funzione che esprime il punteggio così definita:

- $\sigma(s_i, t_j)$: match della coppia s_i e t_j;
- $\sigma(s_i, -)$: allineamento di s_i con un gap;
- $\sigma(-, t_j)$: allineamento di t_j con un gap.

1. Casi base per cui il valore della matrice P è direttamente calcolabile:
   (a) $P(0, 0) = 0$;
   (b) $P(i, 0) = 0 \ \forall \ i \in [1...n]$;
   (c) $P(0, j) = 0 \ \forall \ j \in [1...m]$.
2. Calcolare ricorsivamente l'elemento $P(i, j)$ nel seguente modo:

$$P(i,j) = max \begin{cases} P(i-1, j-1) + \sigma(s_i, t_j]) \\ P(i-1, j) + \sigma(s_i, -) \\ P(i, j-1) + \sigma(-, t_j) \end{cases}$$

3. Determinare la posizione in P dell'elemento massimo: $PosMax = max(P)$;
4. Costruire l'allineamento locale ottimo in modo inverso a partire dall'elemento di posizione $PosMax$ sino all primo elemento di P uguale a zero applicando le seguenti regole:
   (a) $(i, j) \rightarrow (i-1, j-1)$ se match tra i residui s_i, t_j;
   (b) $(i, j) \rightarrow (i-1, j)$ se $s_i \in S$ è allineato con un gap;
   (c) $(i, j) \rightarrow (i, j-1)$ se $t_j \in T$ è allineato con un gap.

La matrice di figura 6.5.3 presenta un esempio di allineamento locale ottimo di due sequenze S, T.

Esempio 6.5.2 (Algoritmo di Smith-Waterman)
Date due sequenze $S = AACCTATAGCT$ *e* $T = GCGATATA$ *determinare l'allineamento ottimo applicando un valore di penalità* $gap = -1$*, un valore di match* $= +1$ *e di mismatch* $= -1$*.*

L'algoritmo di Smith-Waterman produce la seguente matrice di sostituzione:

		A	A	C	C	T	A	T	A	G	C	T
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	1	0	0
C	0	0	0	1	1	0	0	0	0	0	2	1
G	0	0	0	0	0	0	0	0	0	1	0	1
A	0	1	1	0	0	0	1	0	1	0	0	0
T	0	0	0	0	0	1	0	2	1	0	0	1
A	0	1	1	0	0	0	2	0	3	2	1	0
T	0	0	0	0	0	1	1	3	2	2	1	2
A	0	1	1	0	0	0	2	2	4	3	2	1

Figura 6.5.3: Matrice di Smith−Waterman

L'algoritmo determina l'elemento di valore massimo (quattro) di P in posizione $(9, 9)$; il percorso, dall'elemento massimo sino al primo elemento di P con valore zero della matrice di figura 6.5.3 ottenuto utilizzando le stesse regole di costruzione della matrice stessa, determina l'allineamento locale ottimo. Ripercorrendo tale percorso, seguendo le frecce come evidenziato in figura 6.5.3, si ottiene:

S:	T	A	T	A
T:	T	A	T	A

Figura 6.5.4: Smith−Waterman allineamento locale ottimo

6.5.3 Complessità algoritmi di Needleman-Wunsch e Smith-Waterman

Per valutare la complessità di entrambi gli algoritmi si devono considerare i seguenti passi:

- *inizializzazione della matrice P:* complessità $O(n + m)$;
- *calcolo degli elementi interni della matrice P:* complessità $O(n \cdot m)$;
- *ricerca del massimo in P:* $O(n \cdot m)$
- *visita traceback:* complessità $O(n + m)$.

Otteniamo che la complessità degli algoritmi è $O(n \cdot m)$. Se consideriamo che molto spesso le sequenze hanno lunghezze simili ($n \cong m$) la complessità computazione cumulativa è $O(n^2)$. Gli algoritmi descritti appartengono quindi alla classe P dei problemi polinomiali trattabili.

6.6 Significatività statistica degli allineamenti a coppie

Il materiale relativo a questo paragrafo è tratto da: [1].

Il grado di similarità tra due sequenze biologiche è generalmente rappresentato da un punteggio prodotto da un allineamento a coppie. E' necessario capire se tale valore possa essere considerato interessante dal punto di vista biologico. Questa analisi prende il nome di *analisi di significatività statistica* che può essere fatta sia per allineamenti globali che per allineamenti locali tra due sequenze.

6.6.1 Significatività di allineamenti globali

Per calcolare, la significatività statistica del punteggio nel caso di allineamenti globali, si procede attraverso delle simulazioni di allineamento di una sequenza con sequenze generate permutando l'altra sequenza o estratte casualmente da database biologici di sequenze con le stessa caratteristiche di quelle da analizzare.

La significatività statistica è usualmente associata agli indici Z_{score}, P_{value} *e* E_{value}.

Definizione 6.6.1 (Z_{score})
Lo Z_{score} rappresenta il numero di deviazioni standard che separano il punteggio di allineamento globale di due sequenze che si stanno analizzando, da quello ottenuto dalla media dei punteggi degli allineamenti casuali generati da, permutazioni di una delle due sequenze iniziali o estratte casualmente da database biologici. In questo caso si assume che la distribuzione dei punteggi sia normale. L' indice è dato da:

$$\boxed{Z_{score} = \frac{P - M}{\sigma}} \tag{6.6.1}$$

dove:

- *P è il punteggio di allineamento delle due sequenze che si stanno analizzando;*
- *M rappresenta la media dei punteggi di allineamento ottenuti allineando alle permutazioni o alle sequenze casualmemte estratte dai database;*
- *σ è la deviazione standard dei punteggi ottenuti allineandoo alle permutazioni o alle sequenze casualmente estratte dai database.*

Definizione 6.6.2 (P_{value})
Il P_{value} rappresenta la probabilità di trovare per caso un punteggio uguale o superiore al valore S ottenuto come punteggio dall'allineamento che si sta analizzando. E' definito come:

$$\boxed{P_{value} = 1 - \exp\left(-k \cdot m \cdot n \cdot \exp\left(-\lambda \cdot S\right)\right)} \tag{6.6.2}$$

dove:

- *m è la lunghezza della sequenza da confrontare;*
- *n è la lunghezza della sequenza generata dalla permutazione o estratta casualmente da database biologi;*
- *K e λ sono parametri che dipendono dalla banca dati biologica utilizzata nell'operazione di allineamento tra sequenze e dalla matrice dei punteggi utilizzata.*

Definizione 6.6.3 (E_{value})
L' E_{value} rappresenta il numero atteso di sequenze che danno per caso un punteggio maggiore o uguale a un valore S dato. E' definito da:

$$\boxed{E_{value} = k \cdot m \cdot n \cdot \exp\left(-\lambda \cdot S\right)} \tag{6.6.3}$$

dove:

- m è la lunghezza della sequenza da confrontare;
- n è la lunghezza della sequenza estratta in modo casuale da database biologici o generata da funzioni di distribuzioni statistiche;
- K e λ sono parametri che dipendono dalla banca dati biologica utilizzata nell'operazione di allineamento tra sequenze.

L'E_{value} è una misura statistica il cui calcolo dipende da parametri propri di ogni database biologico. Per database biologici che utilizzano BLAST si assume che tutte le sequenze abbiano la stessa lunghezza n per cui l' E_{value} è dato da:

$$\boxed{E_{value} = k \cdot n \cdot l \cdot \exp\left(-\lambda \cdot S\right)} \tag{6.6.4}$$

dove l è la somma delle lunghezze di tutte le sequenze nel database.

Per database biologici che utilizzano FASTA, si assume che la ricerca non dipenda dalla lunghezza delle sequenze. L'E_{value} è quindi dato da:

$$\boxed{E_{value} = P \cdot N} \tag{6.6.5}$$

dove P è il P_{value} di S e N è il numero di sequenze nel database.

6.6.2 Significatività di allineamenti locali

Assumere in un allineamento locale che la distribuzione di probabilità sia normale non è vero. In uno studio corretto di allineamenti casuali di sequenze fatto da *Karlin and Altschul del 1990* [21] si è dimostrato che la probabilità che un punteggio p di un allineamento locale di due sequenze casuali ecceda la media x della distribuzione del punteggio ottenuto con sequenze reali può essere stimata con:

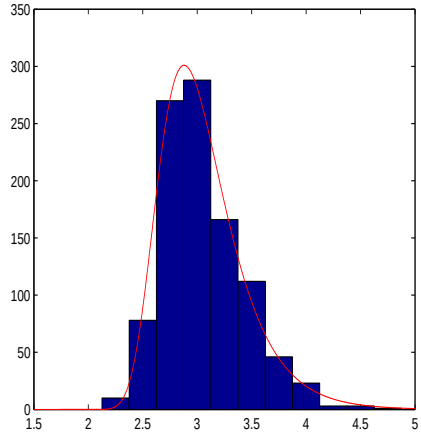
$$\boxed{P(p > x) \approx 1 - \exp(-K \cdot m \cdot n \cdot e^{-\lambda x})} \tag{6.6.6}$$

dove K è una costante positiva e m, n sono le lunghezze delle sequenze allineate.

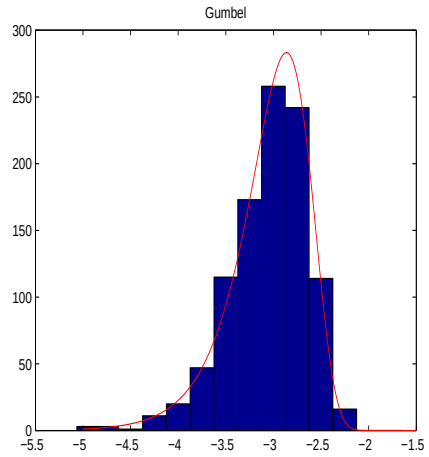
Il valore stimato dell'allineamento viene rappresentato dalla *distribuzione di Gumbel*. In teoria delle probabilità la *distribuzione di probabilità di Gumbel* viene usata per descrivere i valori estremi di una serie stocastica continua. La funzione di densità di probabilità per la rappresentazione di una distribuzione di valori estremi di parametri μ e σ è formalmente data da:

$$\boxed{y = f\left(x|\mu,\sigma\right) = \sigma^{-1} \exp\left(\frac{x-\mu}{\sigma}\right) \exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right)} \tag{6.6.7}$$

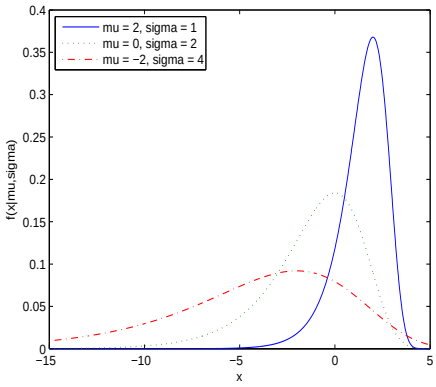
Un esempio di *distribuzione Gumbel* è rappresentato in figura 6.6.2



(a) Esempio n.1 di distribuzione Gumbel



(b) Esempio n.2 di distribuzione Gumbel



(c) Esempio n.3 di distribuzione Gumbel

Figura 6.6.1: Funzione di distribuzione Gumbel

Capitolo 7

CoMeta: un tool per il confronto di vie metaboliche

In questo capitolo illustriamo brevemente le caratteristiche del tool CoMeta che utilizza distanze per il confronto di vie metaboliche. Il materiale è tratto da [15, 13, 14, 20].

7.1 Introduzione

CoMeta (COmputing METAbolic pathways) [13] è un tool per la comparazione di vie metaboliche di diversi organismi. Esso calcola una distanza tra coppie di organismi diversi considerando una o più vie metaboliche selezionate. Consente inoltre di rappresentare il legame evolutivo tra gli organismi oggetto di comparazione con alberi filogenetici. La distanza può essere calcolata a partire da indici di similarità basati sull' omologia delle reazioni delle vie metaboliche e sull' analisi dei T-invarianti delle reti di Petri che modellano le vie metaboliche. Tutte le informazioni che il tool elabora provengono dal database PATHWAY di KEGG [7]. In CoMeta le vie metaboliche sono rappresentate come reti di Petri il cui formalismo matematico si presta efficacemente alla rappresentazione e all'analisi di vie metaboliche. Il tool è eseguibile su piattaforma Windows e Linux, ed è stato implementato in linguaagio Java.

7.2 Tool utilizzati in CoMeta

All'interno di CoMeta vengono utilizzati due tool: MPath2Pn per la traduzione delle vie metaboliche in reti di Petri ed INA per l'analisi delle reti di Petri e la individuazione dei T-invarianti minimi, vediamo ora una breve descrizione di tali tool.

7.2.1 MPath2PN

Le informazioni relative alle vie metaboliche degli organismi sono memorizzate in numerosi database biologici ciascuno dei quali utilizza un proprio formato di rappresentazione dati. MPath2Pn [15] è un tool che fornisce la traduzione automatica in rete

di Petri di una via metabolica le cui informazioni sono presenti nei database biologici. Riconosce diversi formati di rappresentazione di una via metabolica come: KGML utilizzato in KEGG per la rappresentazione delle vie metaboliche e SBML (Systems Biology Markup Language) [23]. In CoMeta i dati delle vie metaboliche provenienti da KEGG vengono tradotti da MPath2PN in rete di Petri nel formato PNML [17] che è un formato proposto come standard per reti di Petri.

7.2.2 INA

INA (Integrated Net Analyzer) [4] è un tool Open Source sviluppato dal Prof. Peter. Starke che permette, in modalità interattiva, di considerare modelli diversi di reti di Petri e varie proprietà, ad esempio consente il calcolo dei T-invarianti minimi di una rete di Petri. Il tool è eseguibile su piattaforme Unix, Linux e Windows. La modalità interattiva di utilizzo ha reso necessario in CoMeta lo sviluppo di una funzione che simula l'interazione: invoca il tool INA e recupera le proprietà dei T-invarianti minimi associati alle vie metaboliche da analizzare.

7.3 Distanze in CoMeta

In una via metabolica le reazioni possono essere identificate dagli enzimi che le catalizzano. Una misura di distanza tra enzimi (reazioni), può essere espressa come:

- *Identità*: due enzimi identici hanno punteggio di similarità uguale a 1 mentre due diversi hanno punteggio uguale a 0.
- *EC gerarchico*: la similarità viene determinata comparando l'EC number associato a ciascun enzima. 5.2.1.

CoMeta propone due distanze tra organismi [13, 14, 15] rappresentate da vie metaboliche come reti di Petri. Una distanza d_R (basata su reazioni) cattura l'omologia tra reazioni nelle vie metaboliche confrontate e l'altra distanza d_I (basata su invarianti) cattura i comportamenti delle vie metaboliche degli organismi espressi con i T-invarianti. Un ulteriore distanza d_C viene calcolata come combinazione delle distanze d_R e d_I

Per rappresentare le reazioni di una via metabolica in CoMeta sono stati usati i multi-insiemi poichè uno stesso enzima può essere usato in più reazioni diverse. Ogni elemento di un multi-insieme ha associata una cardinalità che identifica il numero di occorrenze di un dato enzima all'interno di una stessa via metabolica.

Il punteggio di similarità tra vie metaboliche in due diversi organismi rappresentate da due multi-insiemi X e Y si ottiene in CoMeta applicando a scelta o l'indice di Sørensen o l'indice di Tanimoto.

Siano P_1, P_2 due vie metaboliche rappresentate da reti di Petri, X e Y i due multi-insiemi degli $ECnumber$ associati alle vie metaboliche P_1 e P_2 le distanze d_R e d_I sono calcolate a partire dalla scelta di uno dei due seguenti indici:

- indice di Sørensen esteso ai multi-insiemi X e Y

$$S(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{7.3.1}$$

- indice di Tanimoto esteso ai multi-insiemi X e Y

$$T(X,Y) = \frac{|X \cap Y|}{|X| + |Y| - |X| \cap |Y|} \tag{7.3.2}$$

dove $\cap$, $|X|, |Y|$ indicano intersezione e cardinalità dei due multi-insiemi X, Y. L'indice R_{score} viene calcolato comparando le reazioni delle due vie metaboliche P_1 e P_2 espresso in funzione dell'indice di Sørensen o di Tanimoto:

$$R_{score}(X,Y) = S(X,Y) \text{ con Sørensen} \tag{7.3.3}$$

$$R_{score}(X,Y) = T(X,Y) \text{ con Tanimoto} \tag{7.3.4}$$

la distanza tra le due vie metaboliche P_1 e P_2 basata sulle reazioni è data da:

$$d_R(P_1,P_2) = 1 - R_{score}(X,Y) \tag{7.3.5}$$

La distanza d_I che permette di catturare i comportamenti delle due vie metaboliche P_1, P_2, è ottenuta comparando gli insiemi dei T-invarianti minimi delle reti di Petri P_1 e P_2. Ogni invariante è rappresentata da un multi-insieme di *EC number* che corrisponde alle reazioni verificatesi nell'invariante considerato. La similarità tra due invarianti è espressa come in precedenza in funzione dell'indice scelto (Sørensen o Tanimoto). Un algoritmo euristico basato sui T-invarianti minimi permette di calcolare l'indice I_{score} associato agli invarianti e la distanza descritta d_I è data da:

$$d_I(P_1,P_2) = 1 - I_{score}(X,Y) \tag{7.3.6}$$

Le distanze d_R e d_I possono essere combinate tra loro in base ad un peso α, con $\alpha \in [0,1]$, permettendo di calcolare la distanza combinata basata su reazioni e invarianti seguente:

$$d_C(P_1,P_2) = \alpha\, d_R(P_1,P_2) + (1-\alpha)\, d_I(P_1,P_2) \tag{7.3.7}$$

In CoMeta è inoltre possibile comparare due organismi utilizzando la distanza combinata d_C considerando più vie metaboliche. Dati due organismi o_1 e o_2 ed n vie metaboliche $P_1, ..., P_n$ la distanza combinata tra i due organismi rispetto ad una via metabolica P_i, $i \in [1..n]$ è calcolata nel seguente modo:

$$d_C(o_1,o_2) = \frac{\sum_{i=1}^{n} d_C(P_i^1, P_i^2)}{n} \tag{7.3.8}$$

Se una via metabolica P_i è presente in uno solo dei due organismi considerati si attribuisce alla distanza $d_C(P_i^1, P_i^2)$ il valore massimo 1 .

Nel tool CoMeta le distanze descritte in precedenza vengono memorizzate in tre matrici distinte in formato *txt*, esportabili nell'ambiente di analisi statistica *R* [18] i cui dati saranno successivamente elaborati. Le tre matrici prodotte sono:

- *matrixI.txt*: matrice di distanza basata su invarianti;
- *matrixR.txt*: matrice di distanza basata su reazioni;
- *matrixRI.txt*: matrice di distanza basata su reazioni e invarianti.

7.4 Funzionalità di CoMeta

CoMeta è fornito di un' interfaccia grafica (*GUI*) vedi figura 7.4.1 che mette a disposizione le seguenti funzionalità:

1. Recupero degli organismi e delle vie metaboliche da analizzare in due modalità:
   (a) da server KEGG tramite protocollo http;
   (b) in locale, da una lista di organismi e vie metaboliche esistenti.
2. Traduzione in reti di Petri delle vie metaboliche selezionate con il tool MPath2Pn;
3. Calcolo dei T-invarianti delle reti di Petri con il tool INA;
4. Calcolo delle distanze tra gli organismi selezionati scegliendo l'indice di Sørensen o di Tanimoto vedi figura 7.4.2;
5. Costruzione delle seguenti matrici di distanza tra gli organismi selezionati:
   (a) *matrixI.txt*: matrice di distanza tra organismi basata su invarianti;
   (b) *matrixR.txt*: matrice di distanza tra organismi basata su reazioni;
   (c) *matrixRI.txt*: matrice di distanza combinata basata su reazioni e invarianti).
6. Rappresentazione delle distanze tra gli organismi considerati mediante alberi filogenetici costruiti con i metodi UPGMA e Neigbor Joining;
7. Esportazione dei dati in formato *txt* delle matrici delle distanze calcolate per elaborazioni statistiche.

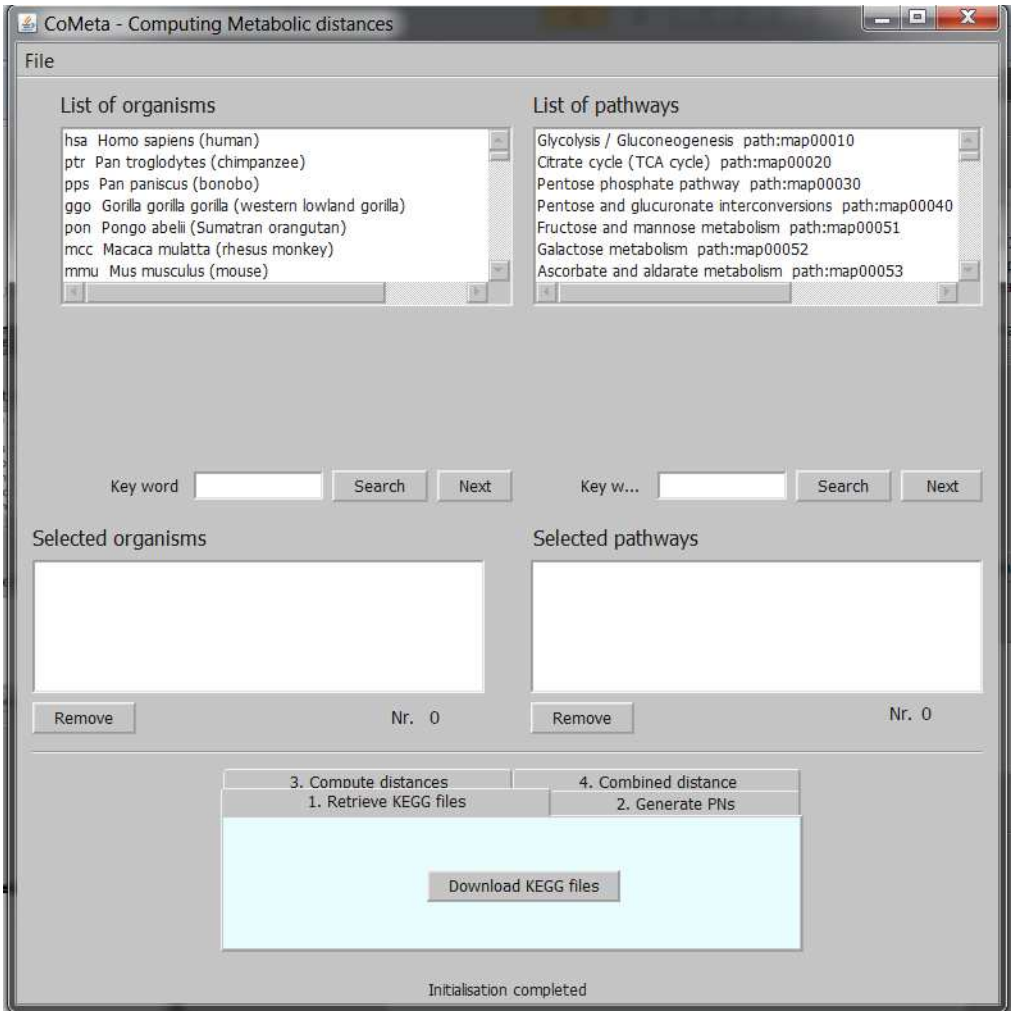


Figura 7.4.1: Schermata principale di attivazione del tool CoMeta

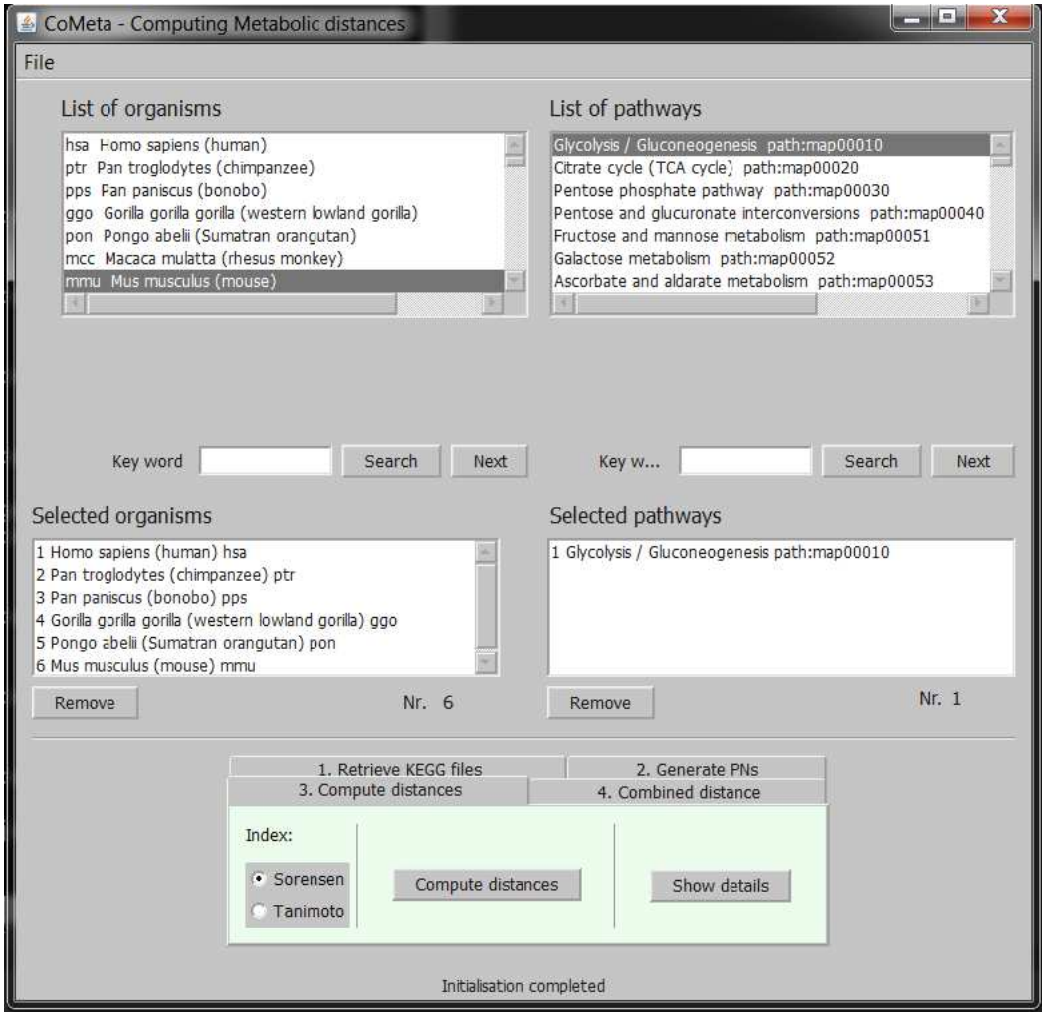


Figura 7.4.2: Schermata per la scelta dell'indice di Sørensen o di Tanimoto

Capitolo 8

RCoMeta

In questo capitolo viene illustrata la struttura e le scelte implementative del tool RCoMeta che offre la possibilità di esplorare e analizzare dal punto di vista statistico le distanze calcolate in CoMeta tra vie metaboliche di organismi memorizzati nei database di KEGG.

8.1 RCoMeta

Lo scopo del tool RCoMeta è quello di analizzare le distanze tra vie metaboliche di diversi organismi calcolate da CoMeta e di fornire per esse una valutazione di significatività statistica relativamente agli organismi memorizzati in KEGG. L'analisi si riferisce ad organismi presenti nel database PATHWAY di KEGG [7]. Le informazioni relative alle distanze tra organismi vengono recuperate dalle matrici delle distanze degli invarianti, delle reazioni e della distanza combinata reazioni e invarianti prodotte dal tool CoMeta. RCoMeta consente di elaborare i dati di tali matrici e di fornire una serie di rappresentazioni grafiche utili ad analizzare e comparare tra loro le distanze tra organismi. L'analisi statistica delle distanze viene effettuata con il software di analisi statistica *R*.

8.1.1 *R*: software per l'analisi statistica

R [18] è un ambiente statistico per la manipolazione e la rappresentazione grafica dei dati. E' distribuito in modalità Open Source sotto licenza GPL (General Public License) ed è disponibile per i sistemi operativi Linux, Windows e Macintosh. *R* utilizza un linguaggio di programmazione object-oriented derivato direttamente dal software *S* sviluppato da John Chambers presso i laboratori Bell della AT&T ed interagisce con l'utente tramite una interfaccia a linea di comando. Esso fornisce una vasta gamma di metodi statistici di base a cui vengono continuamente integrate nuove funzionalità.

8.2 Struttura di RCoMeta

RCoMeta è un tool modulare composto dalle seguenti funzioni:

- Main
- Loading data to be processed
- Analysis of distances in pairs
- Analysis of an organism respect all the other organisms
- Graphs of distance matrices
- Extraction of a sample of organisms in random mode

8.2.1 Modulo Main

Il modulo *Main* si occupa della gestione dell'applicazione e delle componenti dell'interfaccia. L'interfaccia è semplice, lineare ed è stata sviluppata interamente in ambiente di programmazione *R*.

8.2.2 Modulo: Loading data to be processed (modulo obbligatorio)

Il modulo provvede a caricare nella RAM le matrici delle distanze tra organismi prodotte da CoMeta. Le matrici da analizzare sono:

- la matrice invarianti d_I identificata con il nome *matrixI.txt*;
- la matrice reazioni d_R identificata con il nome *matrixR.txt*;
- la matrice combinata reazioni invarianti d_C identificata con il nome *matrixRI.txt*.

Completata la fase di caricamento delle matrici, i dati contenuti in esse vengono reindirizzati verso specifici sottomoduli indipendenti, (sviluppati utilizzando l'ambiente di programmazione di *R*) che elaborano le distanze e forniscono una serie di dati intermedi che in seguito verranno utilizzati dagli altri moduli di RCoMeta. Il sistema richiede che le matrici dati da elaborare, vengano preventivamente salvate in una directory indicata dall'utente che va posizionata all'interno della directory principale del tool *RCoMeta*. Tutti i dati di output prodotti verranno salvati nella stessa directory di input delle matrici indicata dall'utente.

8.2.3 Modulo: Analysis of distances in pairs (modulo facoltativo)

Il modulo permette di analizzare la distanza di una singola coppia di organismi presente nelle matrici delle distanze rispetto alle distanze delle altre coppie. Per la coppia di organismi indicata dall'utente viene estratta la corrispondente distanza e vengono identificate quante coppie di organismi nelle matrici prese in considerazione hanno distanza uguale, minore o maggiore. La lista dei codici degli organismi che l'utente può inserire per selezionare la coppia viene visualizzata a video.

8.2.4 Modulo: Analysis of an organism respect all the other organisms (modulo facoltativo)

Il modulo permette di analizzare le distanze di un singolo organismo rispetto a tutti gli altri organismi presenti nelle matrici considerate. All'utente si chiede di introdurre:

- Codice dell'organismo da analizzare. La lista dei codici degli organismi che l'utente può inserire viene visualizzata a video secondo la codifica gerarchica KEGG;
- Il livello di raggruppamento degli organismi ($Reign = 0, Category = 1, 2, 3$) definito in base alla codifica gerarchica KEGG che prevede la classificazione degli organismi nei quattro livelli interni:

  Livello 0: $Reign$;
  Livello 1: $Category_1$;
  Livello 2 : $Category_2$;
  Livello 3: $Category_3$.

Il livello di raggruppamento selezionato permette di visualizzare i codici degli organismi assieme alla descrizione del gruppo di appartenenza.

La tabella 8.2.1 mostra come vengono classificati gerarchicamente in KEGG gli organismi appartenenti al regno degli *Eucarioti*.

Esempio 8.2.1 (Esempio di raggruppamento in base ai livelli di appartenenza degli organismi in KEGG.)
Supponiamo che le matrici delle distanze considerate siano relative agli organismi con codice: oaa, hsa, mmu, pps, ptr, tru e dre. La scelta del raggruppamento $Category_3$ permetterà nei grafici di visualizzare i dati degli organismi ordinati in base a tale livello assieme alla descrizione del loro gruppo di appartenenza:

Gruppo 1 Mammals
Organismi: hsa, mm, oaa, pps.
Gruppo 2 Fishes
Organismi: dre, tru.

8.2.5 Modulo: Graphs of distance matrices (modulo facoltativo)

Il modulo permette di generare gli istogrammi delle matrici di distanza d_I d_R e d_C degli organismi e i grafici dell'indice z_{score}.

8.2.6 Modulo: Extraction of a sample of organisms in random mode (modulo facoltativo)

Il modulo permette di generare una lista di n organismi estratti in modo casuale in base ad un valore di probabilità denominato *peso* ed i cui elementi appartengono agli organismi presenti in KEGG. Nel database KEGG alcuni gruppi di organismi, ad esempio i *Firmicutes* o i *Streptococcus* appartenenti al gruppo dei *Bacteria* sono composti da centinaia di organismi. Invece di selezionare gli oltre 2000 organismi del gruppo dei Bacteria per calcolare le distanze con il tool CoMeta e successivamente analizzare tali distanze in RCoMeta, è possibile effettuare analisi statistica campionando un numero di Bacteria molto minore di 2000 opportunamente.

Il modulo in esame provvede ad espletare questa funzionalità richiedendo all'utente le seguenti informazioni:

- Scelta tra download della lista completa di organismi, effettuando un collegamento via http direttamente al sito web di KEGG o, se presente in locale, utilizzo dalla lista degli organismi esistente;

- Livello del gruppo di appartenenza degli organismi da estrarre che coincide con i livelli di classificazione degli organismi di KEGG descritti in precedenza ($Reign = 0, Category = 1, 2, 3$);

- Valore n indicante il numero di organismi da estrarre. Il modulo calcola la cardinalità del gruppo di organismi scelto controllando che il valore n sia compreso nel range dei valori ammissibili.

Negli esempi d'uso di RCometa, illustrati nella sezione 8.3 sarà u lteriormente approfondita la descrizione dei moduli precedenti.

KEGG: organismi appartenenti al regno degli Eucarioti

Reign	**Category$_1$**	**Category$_2$**	**Category$_3$**
Eukaryotes	Animals	Vertebrates	Mammals
			Birds
			Reptiles
			Amphibians
			Fishes
		Lanceletes	
		Ascidians	
		Echinoderms	
		Arthtopods	Insects
			Mites and ticks
		Nematodes	
		Flatworms	
		Cnidarians	
		Placozoans	
		Poriferans	
	Plants	Eudicotes	Mustard family
			Pea family
			Willow family
			Spurge family
			Grape family
		Monocotos	Grass family
		Ferns	
		Mosses	
		Green algae	
		Red algae	
	Fungi	Ascomycetes	Saccharomycetes
			Sordariomycetes
			Leotiomycetes
			Eurotiomycetes
			Dothideomycetes
			Pezizomycetes
			Schizosaccharomycetes
		Basidiomycetes	
		Microsporidians	
	Protists	Choanoflagellates	
		Amoeboflagellate	
		Amoebozoa	Dictyostelium
			Entamoeba
		Alveolates	Apicomplexans
			Ciliates
		Euglenozoa	kinetoplasts
		Diplomanads	
		Parabasalids	
		Diatoms	
		Oomycetes	

Tabella 8.2.1: Classificazione KEGG del regno di organismi degli *Eucarioti*

8.3 Manuale d'uso del software RCoMeta

In questa sezione verrà fornito un semplice manuale d'uso in grado di guidare l'utente all'utilizzo del software RCoMeta. Le spiegazioni riguardano:

- Installazione del software *R*;
- Visione della struttura ad albero della directory del tool RCoMeta;
- Esempio di analisi: il gruppo di organismi *Vertebrates* di KEGG;
- Esempo dii estrazione di 20 organismi nel gruppo dei *Bacteria* di KEGG.

8.3.1 Installazione del software *R*

Il software *R* è scaricabile dal sito ufficiale di *R*, http://www.r-project.org, nella sezione *Download, Packages* disponibile per multipiattaforme Linux, MacOS X e Windows. Al termine dell'installazione di *R* è necessario caricare il package *stringr* non integrato nelle librerie di default di *R*. L'installazione di un package si effettua o attivando la funzione *Installa pacchetti* presente nel menù funzioni di *R* indicando:

- *CRAN Mirror*: CRAN (Comprehensive R Archive Network) è una rete di siti web che mette a disposizione numerosi package aggiuntivi per *R*;
- *Nome del package da installare*: per RCoMeta si dovrà installare il package *string*.

oppure da linea di comando con il metodo: *install.packages()* se la libreria *string* è presente localmente.

8.3.2 Visione della struttura ad albero della directory del tool RCoMeta

Il software RCoMeta è memorizzato nella directory di nome *RCoMeta* la cui struttura ad albero è rappresentata graficamente in figura 8.3.1. La directory *RCoMeta* è così strutturata:

- *RCoMeta*: rappresenta la root del tool *RCoMeta*;

- *file*: è la directory utilizzata dal modulo *Extraction of a sample of organisms in random mode* in cui sono memorizzati i seguenti file:

    - *organism.txt*: lista di tutti gli organismi presenti nei database di KEGG generata tramite collegamento http al sito web di KEGG;
    - *organismCount.txt*: lista di tutti gli organismi del punto precedente a ciascuno dei quali è stato aggiunto un valore numerico chiamato *peso*. Il *peso* costituisce il valore di *probabilità* di estrazione associato all'organismo. I pesi permettono di estrarre in modo casuale una sottolista di n organismi

da sottoppre al tool CoMeta per la generazione delle matrici delle distanze da analizzare.

source : è la directory in cui sono presenti i codici sorgente in R del tool RCoMeta;

experiment$_1$: rappresenta la directory in cui l'utente memorizzerà i dati delle matrici *matrixI.txt, matrixR.txt e matrixRI.txt* prodotte dal tool CoMeta. Ogni analisi su matrici diverse richiede all'utente di memorizzare i dati in distinte directory: $experiment_1, experiment_2, ..., experiment_n$. I grafici, di ogni singola elaborazione di RCoMeta vengono memorizzati all'interno della directory in cui sono contenute le matrici delle distanze d_I, d_R e d_C, che in questo caso si chiama $experiment_i$. A seconda della tipologia i grafici vengono salvati nelle sotto directory seguenti:

- *bidimensional*: contiene i grafici dei punti che rappresentano le matrici delle distanze d_I e d_R in uno spazio a p dimensioni. Alla fine di questa sezione verrà illustrato dettagliatamente il metodo cha va sotto il nome di *Scale dimension*. I grafici di tale sezione vengono salvati in formato *pdf*;
- *histograms*: contiene gli istogrammi e i poligoni di frequenza delle matrici delle distanze d_I, d_R e d_C. I grafici di tale sezione vengono salvati in formato *pdf*;
- *pairs*: contiene gli istogrammi delle matrici delle distanze d_I, d_R e d_C. I codici degli organismi della coppia selezionata nel modulo *Analysis of distances in pairs* vengono posizionati all'interno dell'istogramma nella posizione corrispondente alla loro distanza. I grafici di tale sezione vengono salvati in formato *pdf*;
- *zscore*: contiene i grafici dell'indice statistico *z-score* di ogni coppia di organismi presente nelle matrici considerate. La rappresentazione grafica viene proposta in due modalità vedi figura 8.3.8 o 8.3.9 e 8.3.5, in una si evidenziano solo i codici degli organismi mentre nell'altra si evidenziano anche le classi di appartenenza degli organismi. I grafici di tale sezione vengono salvati in formato *pdf*;
- *rows*: contiene i grafici delle distanze di un singolo organismo rispetto a tutti gli altri organismi presenti nelle matrici considerate. Anche in questo caso abbiamo una duplice rappresentazione grafica, una propone le coppie di organismi e la classe di appartenenza mentre l'altra solo i codici degli organismi. I grafici di tale sezione vengono salvati in formato *pdf*;
- *eps*: contiene tutti i grafici prodotti nelle sezioni precedenti in formato *eps*.

La figura (8.3.1) visualizza la struttura ad albero della directory RCoMeta.

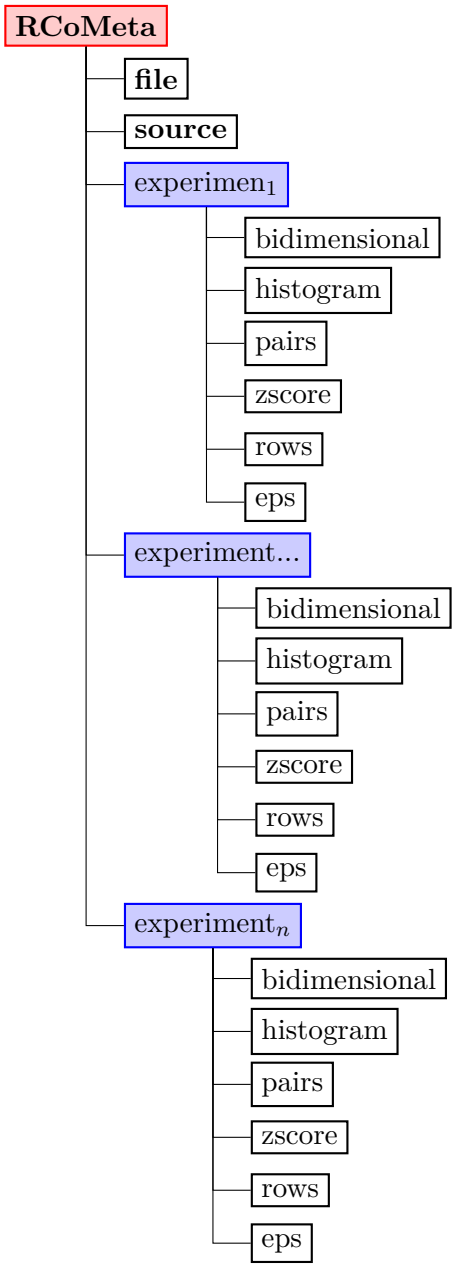


Figura 8.3.1: Struttura ad albero della directory RCoMeta

Scale dimension

Il materiale di questa sezione è tratto dalla fonte [9]. Si supponga di avere a disposizioni dei valori che rappresentano le distanze di un insieme di n punti di cui non si conoscono le coordinate. Si desidera ricostruire le distanze dei punti in uno spazio specifico di p dimensioni in cui mapparli. In R esiste un metodo tradizionale denominato *Scale multidimension* che tenta di assegnare ai punti una configurazione $p-dimensionale$ in modo tale che le distanze tra i punti rispecchino nel modo migliore quelle definite dalla matrice di distanza. In R la funzione preposta a tale compito si chiama *cmdscale* che accetta un numero minimo di due parametri: la matrice di distanze su cui operare e il numero minimo di dimensioni in cui mappare i punti. In output si ottengono le coordinate dei punti generati. Sono disponibili inoltre altre funzioni che rappresentano varianti alla funzione *cmdscale* per esempio la funzione *isoMDS* che permette di ricondurre l'analisi ad uno scaling non metrico.

8.3.3 Esempio di analisi del gruppo di organismi *Vertebrates* di KEGG

Questa sezione mostra l'utilizzo di RCometa di un gruppo di organismi in KEGG. Si sono scelti i *Vertebrates*. L'utilizzo di RCoMeta si articola in una serie di fasi alcune obbligatorie e altre a discrezione dell'utente, per ognuna viene indicato se *obbligatoria o se facoltativa*.

Attivazione tool RCoMeta (fase obbligatoria)

Per utilizzare RCoMeta l'utente dovrà:

- Attivare l'ambiente di lavoro R;
- Posizionarsi nella directory *source* della directory *RCoMeta* in uno dei due seguenti modi:
  - con la funzione *Cambia directory* del menù File di R posizionarsi nella directory *source* di *RCoMeta*;
  - da linea di comando digitando il comando: *setwd(path)* dove path indica il percorso in cui si trova la directory *source di RCoMeta*.
- Digitare da linea di comando: **source('main.r')**.

Verrà visualizzato a video il menù funzioni di tabella 8.3.1 e l'utente potrà iniziare l'analisi dell'esempio proposto.

MAIN	
1)	Extraction of a sample of organisms in random mode
2)	Loading data to be processed
3)	Analysis of distances in pairs
4)	Analysis of an organism respect all the other organisms
5)	Graphs of distance matrices
6)	END
Enter the number of the function in [1..6]	

Tabella 8.3.1: Menù funzioni del tool RCoMeta

8.3.4 Funzione 1: Extraction of a sample of organisms in random mode (fase facoltativa)

Il modulo è preposto alla generazione di una lista di n organismi selezionati dalla lista completa di organismi di KEGG. Gli organismi vengono selezionati in modo casuale in base ad un valore di probabilità calcolato dal modulo stesso.

La procedura verifica se nella directory *file* è presente la lista di organismi di KEGG, in caso affermativo propone all'utente la possibilità di scegliere se estrarre gli organismi da tale lista o se effettuare un collegamento via http con KEGG.

Si consiglia di attivare la procedura di download a KEGG in quanto i database KEGG vengono periodicamente aggiornati con nuovi organismi sequenziati.

L'utente viene quindi invitato ad introdurre i dati indicati in tabella 8.3.2.

Extraction of a sample of organisms in random mode
The list of Organism is present in locally,use these? Press (y/n) n provo con l'URL 'http://rest.kegg.jp/list/organism' Content type 'text/plain' length unknown URL aperto downloaded 212 Kb Enter a class to be sampled: Vertebrates Enter the number of organisms to be selected: [1..26] 20 Enter the name of the folder where to save the output *organism.txt*: Vertebrates Operation has been completed successfully

Tabella 8.3.2: Funzione 1 del MAIN: Extraction of a sample of organisms in random mode

Il file *organism.txt*, contenente la lista degli organismi selezionati da sottoporre al tool CoMeta viene automaticamente salvato nella directory il cui nome è stato indicato dall'utente.

Se durante l'operazione di download la sessione *http con KEGG* si interrompe è necessario eseguire la sequenza di operazioni:

- chiudere tutte le connessioni aperte digitando *closeAllConnections()*;
- rieseguire la funzione indicata alla sezione *Attivazione tool RCoMeta*.

Esempio 8.3.1 (Campionamento casuale di 10 organismi della classe Bacteria in KEGG)
Di seguito vediamo un esempio di campionamento che prevede l'estrazione casuale di 10 *organismi della classe dei Bacteria. All'utente vengono richieste le seguenti informazioni:*

- *classe di appartenenza degli organismi: Bacteria;*
- *numero di organismi da inserire nella lista:* 10.

Dalla lista prodotta e memorizzata nel file *organism.txts* estraimo il primo organismo e vediamo il peso calcolato.

- *Organismo*: T00831 eum;
- *Regno*: Prokaryotes, cardinalità della classe 2260;
- *Livello 1*: Bacteria, cardinalità della classe 2115;
- *Livello 2*: Gammaproteobacteria, cardinalità della classe 467;
- *Livello 3*: Escherichia, cardinalità della classe 60;

Se la classe di analisi richiesta dell'organismo *eum* è quella dei Gammaproteobacteria associata al livello 2 il peso è dato dal raporto: $\frac{467}{2115} = 0.22080378250591$. Ovvero il rapporto tra il numero dei Gammaproteobacteria sul numero totale dei Bacteria.

Funzione 2 del MAIN: Loading data to be processed (fase obbligatoria)

Per procedere con l'analisi statistica è obbligatorio caricare le matrici delle distanze prodotte dal tool CoMeta. In questa fase l'utente viene invitato ad indicare:

- La *directory* in cui sono state salvate le matrici delle distanze tra organismi che deve essere contenuta obbligatoriamente all'interno della directory *RCoMeta*. Il sistema controlla l'esistenza di tale directory, nel caso in cui essa non sia presente visualizzerà a video un messaggio di errore e verrà richiesto un nuovo input come illustrato in tabella 8.3.3. Nell'esempio proposto la directory contenente i dati delle matrici si chiama *vertebrates*.

```
Enter the number of the function in [1..6] 2
Enter the name of the data directory: vertttebrates
Error! vertttebrates not exists: try again
Enter the name of the data directory: vertebrates
```

Tabella 8.3.3: Funzione 2 del MAIN: richiesta directory di lavoro

- L' *indice* utilizzato in CoMeta per costruire le matrici delle distanze d_I e d_R, che può essere o quello di Sørensen o quello di Tanimoto, ed il valore del parametro α utilizzato per produrre la matrice delle distanze combinata d_C. I dati qui introdotti sono puramente indicativi e servono a RCoMeta per visualizzare nei grafici tali informazioni. Nell'esempio proposto l'indice è quello di Sørensen con $\alpha = 0.5$. La tabella (8.3.4) mostra la sequenza delle richieste.

```
Index
1) Sørensen
2) Tanimoto

Enter the code of the index used for the analysis [1..2] 1
Enter the value of alpha used for the analysis: 0.5
```

Tabella 8.3.4: Funzione 2 del MAIN: richiesta parametri di analisi

Funzione 3 del MAIN: Analysis of distances in pairs (fase opzionale)

Questa funzione, data una coppia di organismi, determina quante coppie di organismi nelle matrici delle distanze considerate, hanno distanza uguale, minore o maggiore della coppia considerata. Viene visualizzata a video la lista degli organismi presenti nelle matrici delle distanze in input e richiesto all'utente di indicare la coppia di organismi tramite il codice KEGG. Il sistema controlla la validità dell'input informando l'utente di eventuali errori tramite opportuna messaggistica di errore. Nell'esempio proposto la coppia di organismi è *(hsa,pon)*. La tabella 8.3.5 visualizza l'interazione con la funzione e gli output prodotti.

List of organisms
hsa ptr pps pon mcc mmu rno cfa aml bta ssc ecb mdo shr oaa gga mgp tgu acs xla xtr dre tru
Enter the code of first organism without quotes: hsa Enter the code of second organism without quotes: pon
Invariant: Distance between hsa and pon : 0 Number of pairs of organisms with equal distance: 153 Number of pairs of organisms with a smaller distance: 0 Number of pairs of organisms with greater distance: 100 **Reaction:** Distance between hsa and pon : 0 Number of pairs of organisms with equal distance: 153 Number of pairs of organisms with a smaller distance: 0 Number of pairs of organisms with greater distance: 100 Operation has been completed successfully **Combined distance: Reaction + Invarianti:** Distance between hsa and pon : 0 Number of pairs of organisms with equal distance: 153 Number of pairs of organisms with a smaller distance: 0 Number of pairs of organisms with greater distance: 100 Operation has been completed successfully

Tabella 8.3.5: Funzione 3 del MAIN: Analysis of distances in pairs

Inoltre viene generato il grafico di figura 8.3.2 e memorizzato nella directory *pairs*. Il grafico di figura 8.3.2 visualizza gli istogrammi delle matrici della distanza basata su invarianti, reazioni e combinata evidenziando con il colore la collocazione della coppia di organismi (hsa, pon) analizzata. In didascalia è riportata anche la numerosità delle coppie. Sull'asse x è rappresentato il valore della distanza, sull'asse y la *densità*.

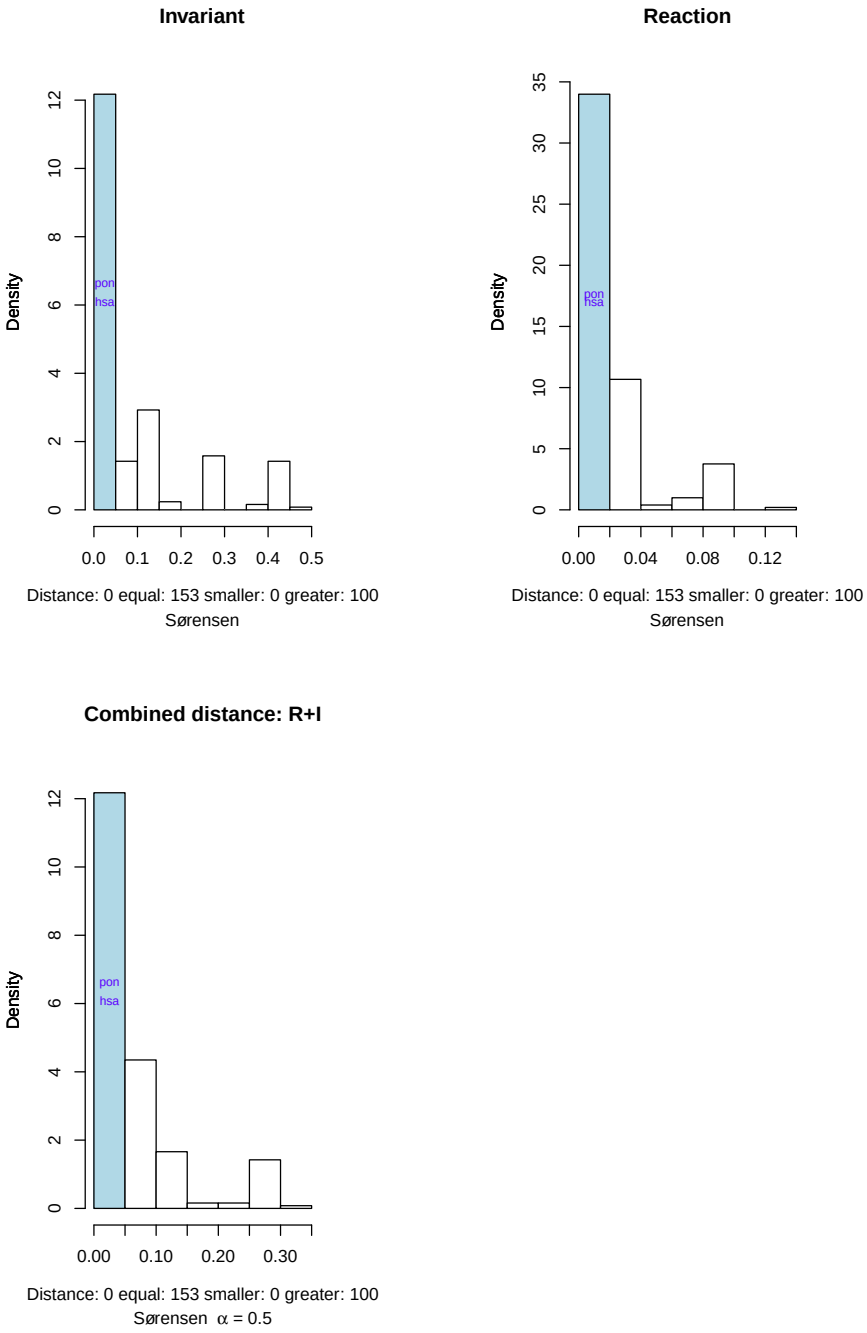


Figura 8.3.2: Istogrammi e coppia di organismi (hsa, pon)

Funzione 4 del MAIN: Analysis of an organism respect all the other organisms (fase opzionale)

Questa funzione, dato un singolo organismo, permette di compararlo con tutti gli altri organismi presenti nelle matrici considerate. Viene visualizzata a video la lista degli organismi presenti nelle matrici date e richiesto all'utente di indicare il codice dell'organismo da analizzare ed il codice di raggruppamento. Il sistema controlla la validità dell'input informando l'utente di eventuali errori tramite opportuna messaggistica. Nell'esempio proposto l'organismo da analizzare ha codice *hsa* e codice di raggruppamento 3 che corrisponde alla $Category_3$ della classificazione degli organismi di KEGG. La tabella 8.3.6 visualizza l'operatività della funzione e gli output prodotti.

List of organisms
hsa ptr pps pon mcc mmu rno cfa aml bta ssc ecb mdo shr oaa gga mgp tgu acs xla xtr dre tru
Enter code of organism without quotes: hsa Enter the level of grouping organisms [0..3] (Reign = 0, Category = 1,2,3) 3 The graphs of the analysis of an organism respect to all the other organisms have been saved in the folder -> rows Operation has been completed successfully

Tabella 8.3.6: Funzione 4 del MAIN: Analysis of an organism with all other organisms

Vengono prodotti i seguenti grafici e memorizzati nelle directory sotto indicate:

- Grafici delle distanze tra l'organismo *hsa* e tutti gli altri organismi presenti nelle tre matrici d_I, d_R e d_C suddivisi per classi di appartenenza. I grafici vengono memorizzati nella directory *rows*.
- Grafici dell'indice z_score della distanza tra l'organismo *hsa* e tutti gli altri organismi presenti nelle matrici delle distanze considerate. I grafici vengono memorizzati nella directory *zscore*.

- Il grafico di figura 8.3.3 mostra la distanza basata su invarianti e su reazioni dell'organismo *hsa* da tutti gli altri organismi vertebrati in KEGG. Gli organismi sono rappresentai sull'asse y rispetto alla classe;
- Il grafico di figura 8.3.4 mostra la distanza basata su invarianti e su reazioni dell'organismo *hsa* da tutti gli altri vertebrati in KEGG. Ogni organismo viene visualizzato con il relativo codice di classificazione KEGG;
- Il grafico di figura 8.3.5 mostra i valori calcolati per l' l'indice z_{score}. Sull'asse x sono riportate le distanze mentre la linea colorata in rosso rappresenta la media dei dati rappresentati.

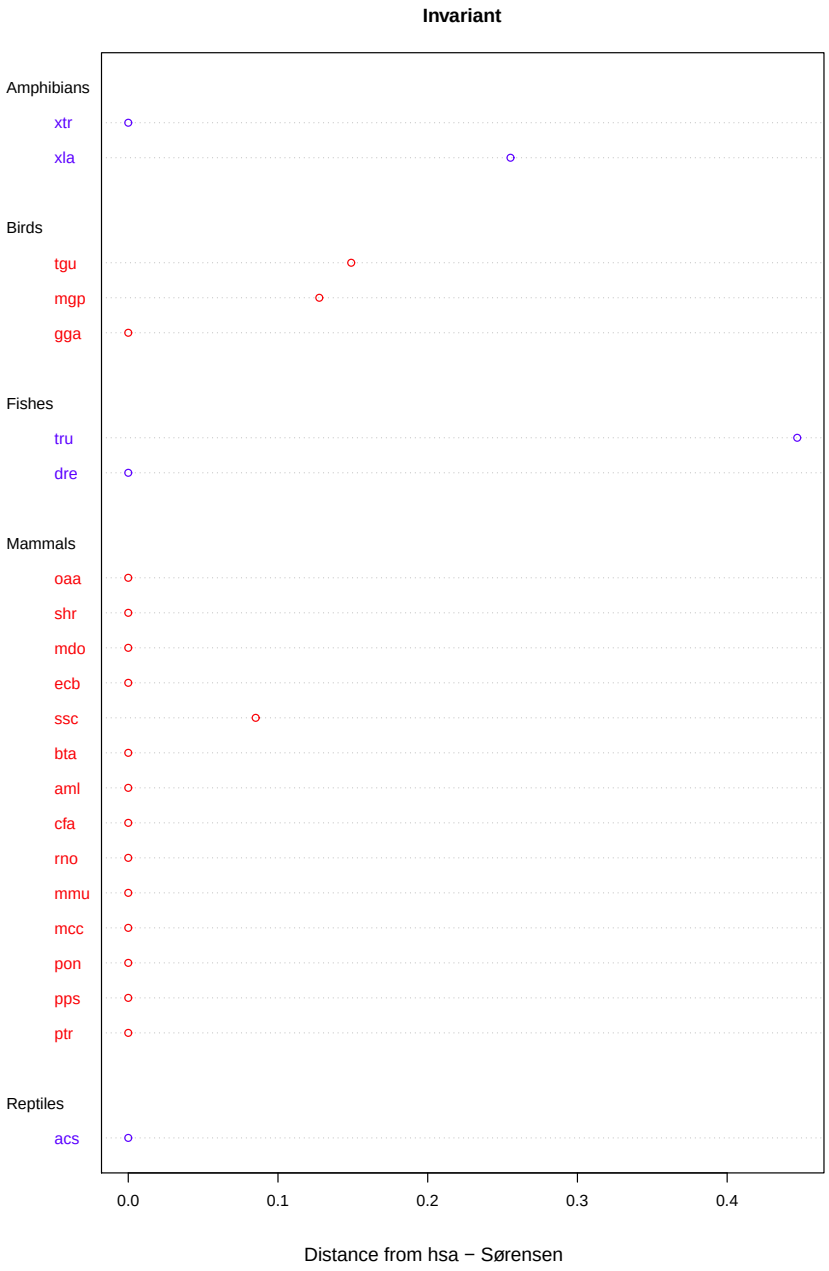


Figura 8.3.3: Organismo: hsa con classificazione KEGG

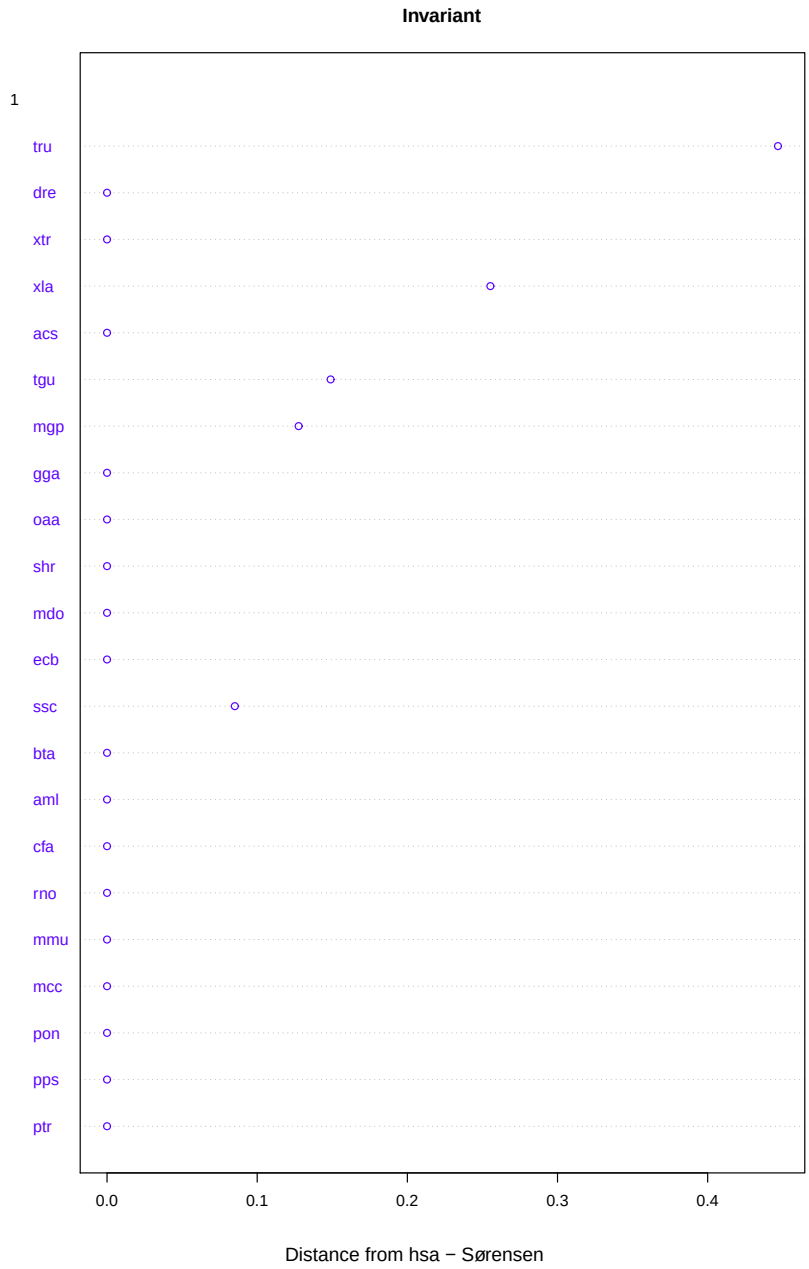


Figura 8.3.4: Organismo: hsa

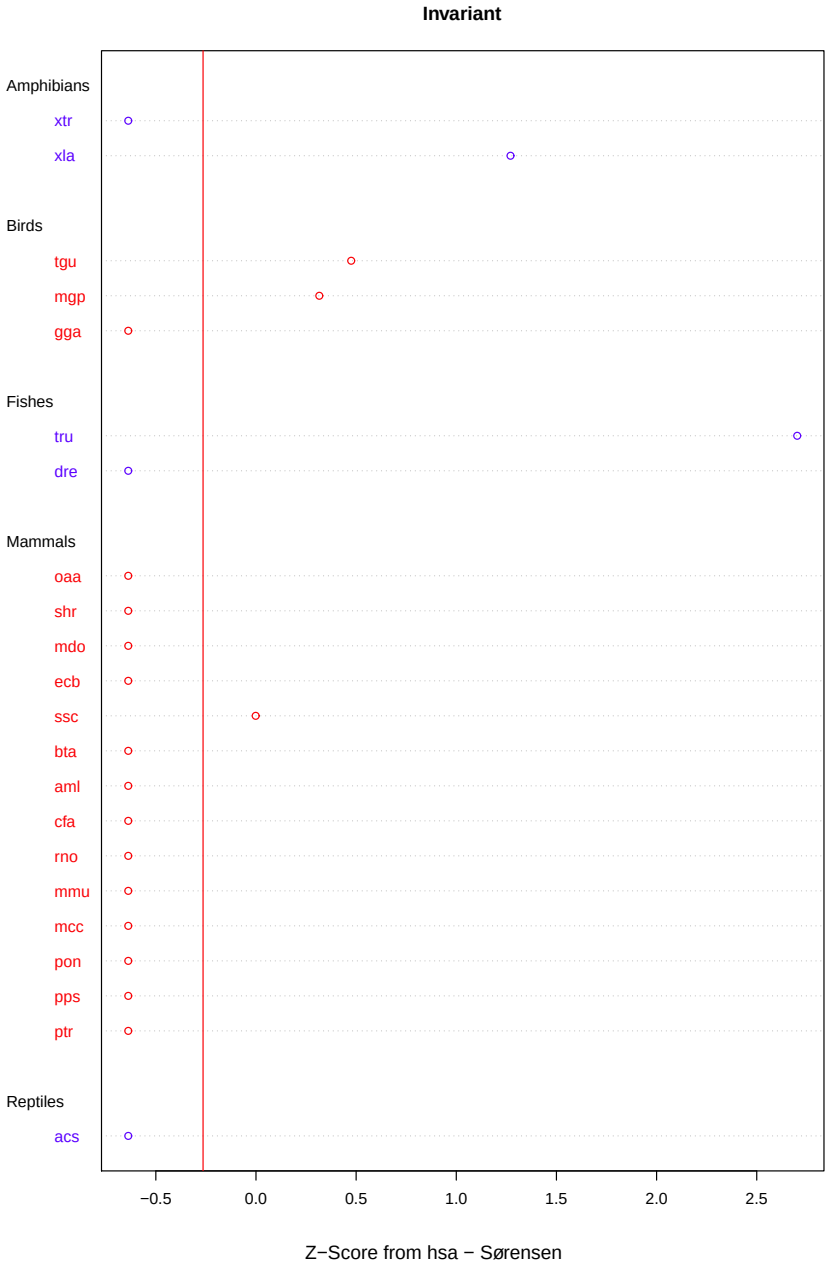


Figura 8.3.5: Indice z_{score}

Funzione 5 del MAIN: Graphs of distance matrices (fase opzionale)

Questa funzione specificata nel seguito permette di generare e di memorizzare nella directory i seguenti grafici:

- Istogrammi delle matrici d_I, d_R e d_C memorizzati nella directory *histograms* vedi figura 8.3.6.
- Grafico in bidimensione degli organismi della matrice d_I memorizzato in *bidimensional* vedi figura 8.3.7.
- Indice z_{score} delle distanze delle matrici d_I e d_R memorizzati in *zscore*.

Data la numerosità dei dati da rappresentare graficamente vengono mostrati solo alcunu grafici.

- Per i grafici *bidimensional* verrà visualizzato solo il grafico della matrice invarianti vedi figura 8.3.7.
- Per lo z_{score} data la numerosità delle coppie di organismi i dati plottati sono suddivisi a blocchi di 35 coppie per grafico. Verranno visualizzati solo i grafici numero 1 e numero 5 della matrice d_I vedi figura 8.3.8 e figura 8.3.9.

```
Enter the number of the function in [1..6] 5
1) The graphs histogram and frequency polygon of: matrix.I, matrix.R, matrixRI
have been saved in the folder -> histogram
2) The graphs of the Z-Score
have been saved in the folder –> zscore
3) The bidimensional graphs of: matrix.I, matrix.R, matrixRI
have been saved in the folder -> bidimensional

Operation has been completed successfully
```

Tabella 8.3.7: Funzione 5 del MAIN: Graphs of distance matrices

La figura 8.3.6 visualizza gli istogrammi delle matrici della distanza basata su invarianti, reazioni, combinata con $\alpha = 0.5$ e dello scarto tra le due distanze basate su invarianti e reazioni. Gli elementi che compongono gli istogrammi sono:

- Asse x: rappresenta le distanze delle matrici suddivise in basi la cui ampezza viene determinata da una scala automatica direttamente dal software R;
- Asse y rappresenta la densità con frequenze relative;
- Curva di colore rosso: sovrappone gli istogrammi e rappresenta la curva del poligono delle frequenze relative. Il poligono delle frequenze si ottiene unendo i segmenti rettilinei dei punti medi dei lati superiori dei rettangoli che formano l'istogramma.

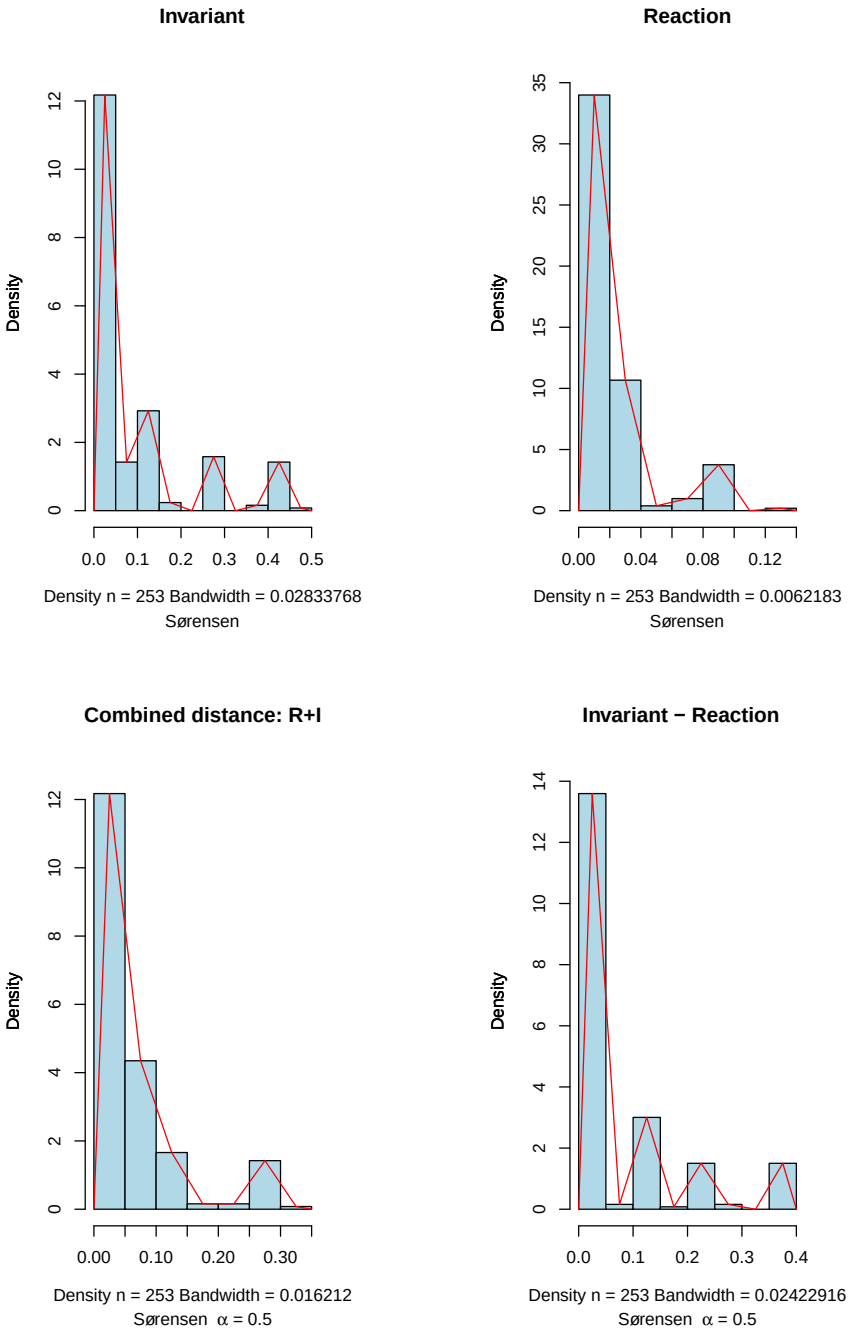


Figura 8.3.6: Istogrammi delle matrici: d_I, d_R, d_C e differenza $d_I - d_R$

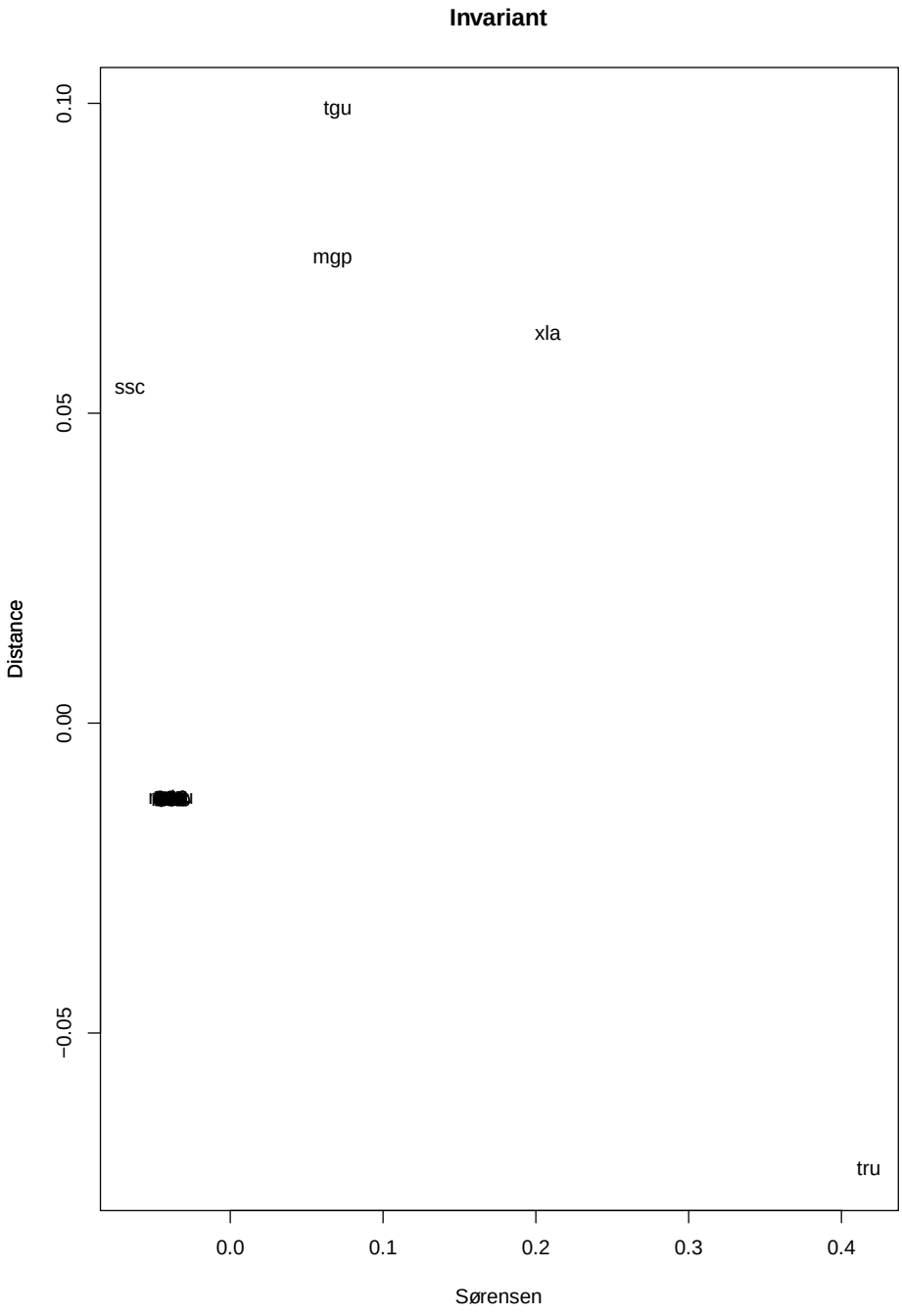


Figura 8.3.7: Grafico dei punti le cui distanze sono rappresentate dalla distanza d_I

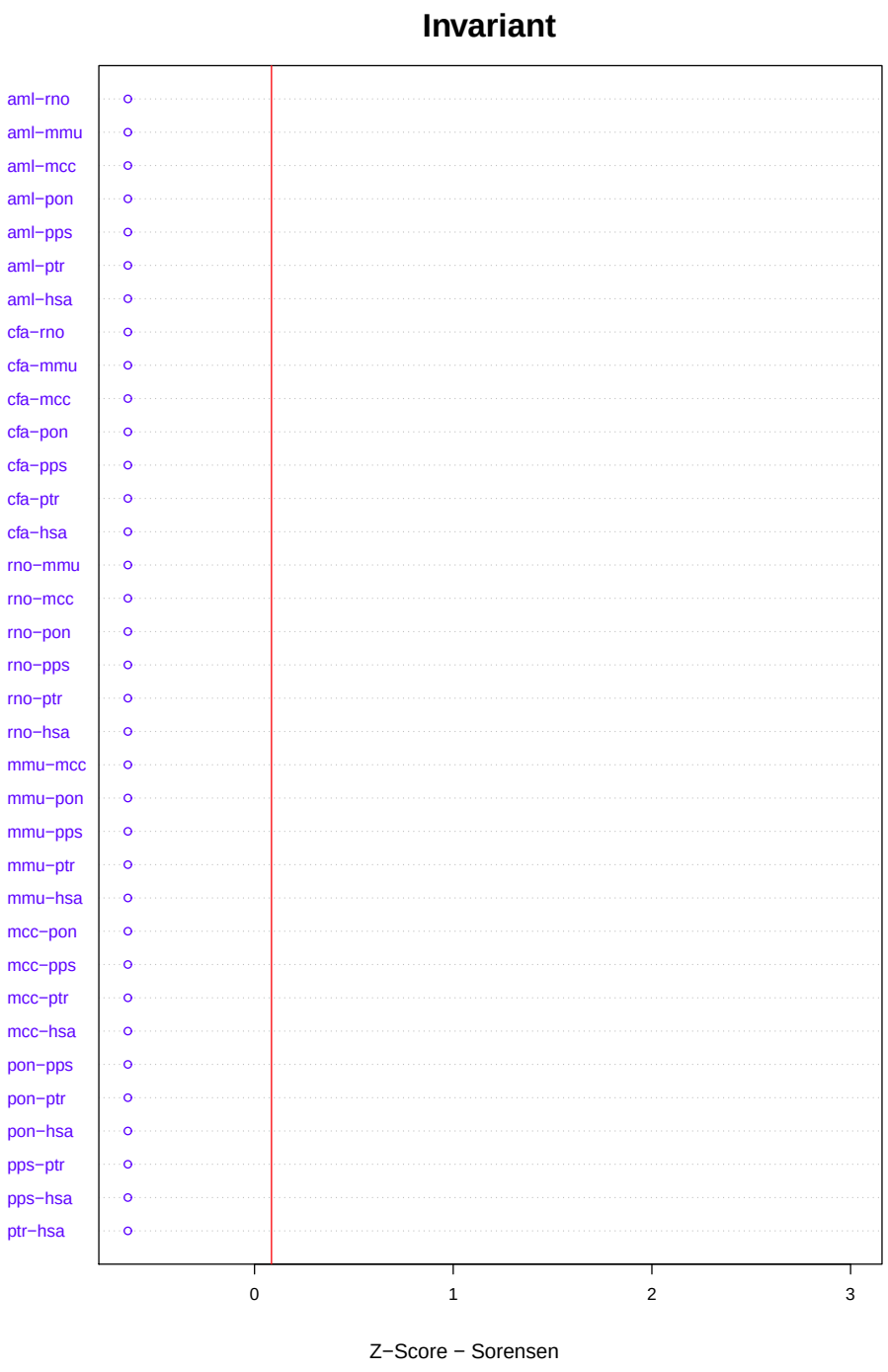


Figura 8.3.8: Indice z_{score} della matrice basata su invarianti d_I di pagina 1

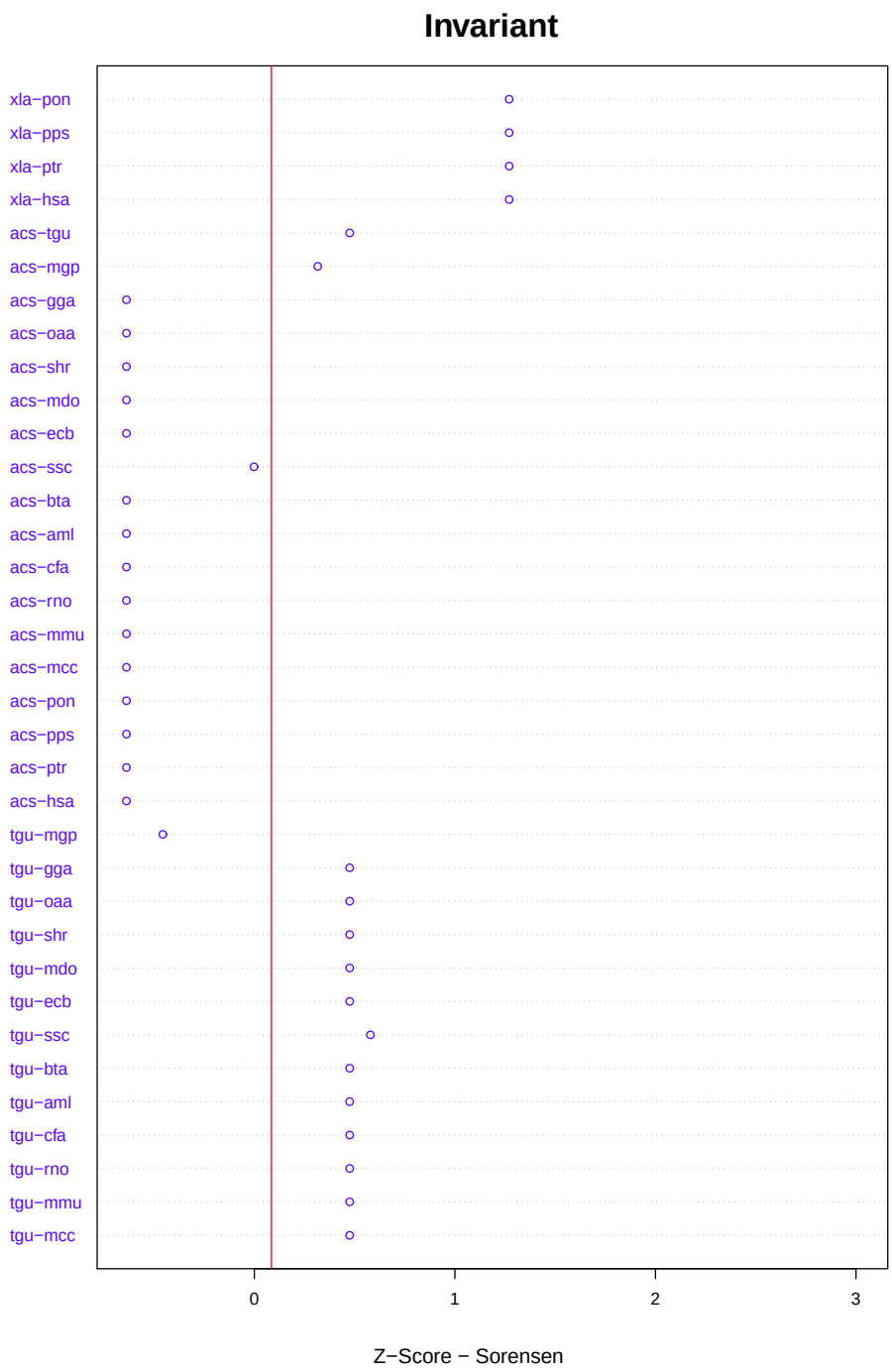


Figura 8.3.9: Indice z_{score} della matrice basata su invarianti d_I di pagina 5

Capitolo 9

Conclusioni

In questa tesi si sono analizzate le distanze utilizzate dal tool CoMeta per confrontare organismi diversi rispetto ad una o più vie metaboliche. Per l'analisi si sono considerati vari gruppi di organismi del database KEGG. A tale scopo si è sviluppato il tool RCoMeta, sviluppato in R, che consente di effettuare diverse elaborazioni statistiche e visualizzazioni dei dati. Nel capitolo 6 si è brevemente trattato dell'analisi di significatività per gli allineamenti di sequenze. Una tecnica prevede l'esplorazione ed il campionamento di sequenze *simili* a quelle allineate in un database di riferimento. In modo simile, RCoMeta può essere usato per effettuare campionamenti di organismi in KEGG e per analizzare le distanze di CoMeta su tali campioni rispetto ad una o più vie metaboliche. Con il tool RCoMeta sono state svolte le sperimentazioni di seguito indicate.

9.1 Analisi con Sørensen e via metabolica glicolisi

L'analisi è stata fatta rispetto alla via metabolica della *glicolisi* e all'indice di *Sørensen* ed è stata suddivisa in:

- analisi delle distanze d_I, d_R e d_C;
- analisi dell'organismo *hsa* nei confronti di più classi di organismi;
- analisi di coppie di organismi diversi;
- calcolo indice z_{score}.

Le sperimentazioni hanni interessato i seguenti gruppi di organismi:

- Eukaryotes;
- Animals;
- Vertebrates;
- Mammals;
- Insects;

- Plants;
- Fungi;
- Protists;
- Archaea;
- Bacteria.

Tali analisi sono presentate nell'Appendice A.

9.2 Analisi con Tanimoto e via metabolica glicolisi

L'analisi è stata fatta rispetto alla via metabolica della *glicolisi* e all'indice di *Tanimoto* ed è stata suddivisa in:

- analisi delle distanze d_I, d_R e d_C;
- analisi dell'organismo *hsa* nei confronti di più classi di organismi;
- analisi di coppie di organismi diversi;
- calcolo indice z_{score}.

Le sperimentazioni hanni interessato i seguenti gruppi di organismi:

- Eukaryotes;
- Animals;
- Vertebrates;
- Mammals;
- Insects;
- Plants;
- Fungi;
- Protists;
- Archaea;
- Bacteria.

Tali analisi sono presentate nell'Appendice B.

9.3 Risultati delle analisi

I risultati delle analisi effettuate vengono brevemente riportati di seguito.

9.3.1 Analisi in appendice A.1

1. In tutti gli esperimenti la distanza basata su invarianti varia in un intervallo più ampio rispetto a quella basata su reazioni. Talvolta si tratta di un intervallo di

ampiezza almeno doppia (Animals, Mammals, Plants), in alcuni casi (Vertebrates, Insects) l'intervallo è quasi un ordine di grandezza superiore. Solo per Fungi, Protists e Archaea le ampiezze dell'intervallo delle due distanze è simile. Osservando le distanze degli organismi del regno degli Eukaryotes la distanza reazioni sembra tendere alla nota *distribuzione di Gumbel*. Le ampiezze degli intervalli delle due distanze per gli Eukaryotes sono quasi simile. Per la classe dei Bacteria possiamo notare che tutte e due le distanze sembrano tendere alla *distribuzione di Gumbel*, le ampiezze dei loro intervalli sono pressochè identiche.

2. Nella maggior parte dei casi la distanza basata su invarianti discrimina maggiormente rispetto a quella basata su reazioni. Ciò puo' essere determinato esaminando l'istogrammma che rappresenta la differenza tra le due distanze. In particolare ciò si evidenzia negli Animals e Insects. Invece lo scarto è minore nel caso dei Fungi e Protists. Per Plants e Archea non vi è una chiara prevalenza tra le due distanze e nel caso dei Mammals le due distanze sono coincidenti per la maggior parte delle coppie di organismi.

3. Se consideriamo tutti gli organismi appartenenti al regno degli Eukaryotes si può dire che la distanza basata su invarianti non differisce molto da quella basata su reazioni, entrambe hanno una densità quasi normale. La distanza basata su invarianti è in genere superiore ma non sempre, vedi le distanze tra alghe verdi e uomo, dove la distanza reazioni distingue di più, lo stesso avviene anche per i microsporidians-uom. Questo pare indicare che ha senso usare la distanza combinata.

4. Il gruppo dei mammals in KEGG consiste in 15 organismi tutti molto simili rispetto alla glicolisi. Ciò si puo' vedere dal grafico in figura B.2.15 che riporta la distanza di hsa (homo sapiens) dagli altri organismi del gruppo: solo ssc (sus scrofa) ha distanza diversa da zero dagli altri organismi. Entrambe le distanze rappresentano fedelmente questa situazione come si può notare dalla figura A.1.4.

5. Il gruppo degli Archea in Kegg consiste in 178 organismi distribuiti in varie sottoclassi. L'istogramma della distanza basata su reazioni rappresenta tale distribuzione con un andamento che approssima una normale con distanza massima corrispondente al valore 0.2, quindi ad un valore di distanza abbastanza basso. Ciò sembra indicare che gli Archea non si differenziano molto rispetto alla glicolisi per quel che riguarda le reazioni. Una situazione simile si evidenzia nel caso dei Protists. La distanza basata su invarianti, anche in questi due casi, distingue maggiormente tra gli organismi. La distanza combinata, con $\alpha = 0.5$, ha un andamento simile a quella basata sulle sole reazioni.

9.3.2 Analisi in appendice A.13

1. L'analisi dello z_{score} dell'organismo (hsa) per le classi (Vertebrates, Mammals) si colloca rispetto alla media dei valori attorno al valore -0.7 valore che evidenzia una forte concentrazione degli organismi. Se estendiamo la classe di appartenenza di (hsa) a quella di livello superiore di classificazione KEGG (Animals) possiamo notare che i valori si distribuiscono in numerosità quasi uniformente attorno al valore della media evidenziando la vicinanza degli organismi appartenenti al gruppo dei Vertebrati. Le considerazioni fatte sono valide sia per la distanza basata su invarianti sia per quella su reazioni.

9.3.3 Analisi in appendice B.1

1. Il passaggio all'indice di Tanimoto per il calcolo delle distanze non sembra evidenziare grossi cambiamenti rispetto a quanto detto per l'indice di Sørensen.

2. Il gruppo di organismi Archaea evidenzia per tutte le distanze considerate una distribuzione che si approssima alla normale pur mantenendo una certa differenza nell'ampiezza dei valori assunti dalla densità.

3. Per il gruppo dei Bacteria le due distanze invarianti e reazioni si identificano quasi perfettamente e lo scarto tra le due è decisamente identificabile da una curva di distribuzione normale.

Terminata l'analisi ci si può soffermare brevemente sugli eventuali sviluppi futuri e sulle problematiche che non si sono potute affrontare data la ristrettezza dei tempi a disposizione. In futuro il tool può essere ampliato nelle sue funzionalità integrandolo con:

- Costruire un campione casuale su cui valutare la significatività statistica, per esempio generando le vie metaboliche utilizzando la tecnica delle permutazioni degli Ec-Number o enzimi delle vie metaboliche considerate per la distanza basata su reazioni, e delle colonne delle matrici stechiometriche se trattasi di distanza basata su invarianti.

- Valutare la significativtà statistica su numerosi campionamenti utilizzando gli intervalli di confidenza.

Ringraziamenti

Finalmente posso iniziare a scrivere questo capitolo e dire *Ci sono!*.

Questa avventura è iniziata un giorno caldo di un'estate che non ricordo quale sia, quando la mia più cara amica Donatella mi chiese di prendere il treno con lei e di fare un giro a Mestre. Aveva preso un appuntamento con una Docente della Facoltà di Informatica di CA' Foscari per riprendere gli studi universitari abbandonati qualche anno addietro. Destino o casualità l'ufficio era proprio quello della Professoressa Nicoletta Cocco, quel giorno si parlò di molte cose con il risultato che decisi di seguire la mia amica in questa grande scelta: ritornare all'università. Le nostre strade per esigenze personali si separarono ma in tutti questi anni abbiamo sempre condiviso i nostri pensieri.

In questa pagina desidero ringraziare la Professoressa Nicoletta Cocco che mi ha seguito in tutto il percorso di realizzazione di questa tesi in momenti felici e in momenti tristi, in cui oltre ad essere stata una Docente di grande valenza professionale è stata anche una persona di grande umanità. Sono felice di avere iniziato questo mia percorso con la Professoressa Cocco e di averlo terminato con Lei. Un ringraziamento anche alla Dott.ssa Marta Simenoni che ha collaborato al progetto di tesi.

Voglio ringraziare Donatella che non mi ha mai fatto perdere la fiducia nel portare avanti questa nostra decisione.

Ringrazio la mia famiglia, mia madre che con i suoi 90 anni per motivi di salute non potrà venire alla Laurea, mia figlia Alessia che è tutta la mia vita e mio marito Walter per avermi sopportato anche nei momenti di sconforto.

Un soddisfatto ringraziamento lo ripongo anche nei confronti di tutti quei colleghi di lavoro che mi hanno tacitamente augurato di non arrivare al mio obiettivo in quanto sono stati lo stimolo più grande per arrivare alla fine di tale percorso.

Non posso dimenticare di ringraziare di cuore Rosaria che con la sua gentilezza e disponibiltà quotidianamente gestisce la biblioteca di dipartimento di Informatica.

Infine ringrazio me stessa per essere riuscita ad ottenere questo nuovo traguardo affrontando le difficoltà incontrate senza mai perdere di vista la vita e l'obiettivo finale.

Grazie a tutti per essermi stati vicini.
Annachiara

Appendice A

Analisi della glicolisi in KEGG

In questa sperimentazione sono riportati i risultati ottenuti da alcune prove eseguite con il tool RCoMeta. Sono state analizzate 10 classi di organismi del database KEGG. Gli organismi sono classificati secondo la tassonomia di riferimento di NCBI [16]. Le distanze provengono da elaborazioni con il tool CoMeta. Lo scopo è quello di esplorare una specifica via metabolica in classi di organismi diversi. Per ciascuna classe viene analizzata: la distribuzione dei valori delle distanze di tutte le coppie di organismi appartenenti alla classe la distribuzione dei valori delle singole coppie di organismi campionando un organismo rispetto a tutti gli altri organismi presenti nella classe.
I parametri della sperimentazione sono i seguenti:

- Classi di organismi:
  - Eucaryotes;
  - Animals;
  - Vertebrates;
  - Mammals;
  - Insects;
  - Plants;
  - Fungi;
  - Protists;
  - Archaea;
  - Bacteria (campione di 150 batteri).
- Via metabolica analizzata: glicolisi;
- Indice di similarità: Sørensen;
- Distanze
  - Distanza invarianti d_I;
  - Distanza Reazioni d_R:
  - Distanza combinata d_C con coefficiente α: 0.5.

Il risultato dell'esperimento è riportato nei seguenti grafici suddivisi per categoria.

A.1 Istogrammi delle distanze: d_I, d_R e d_C

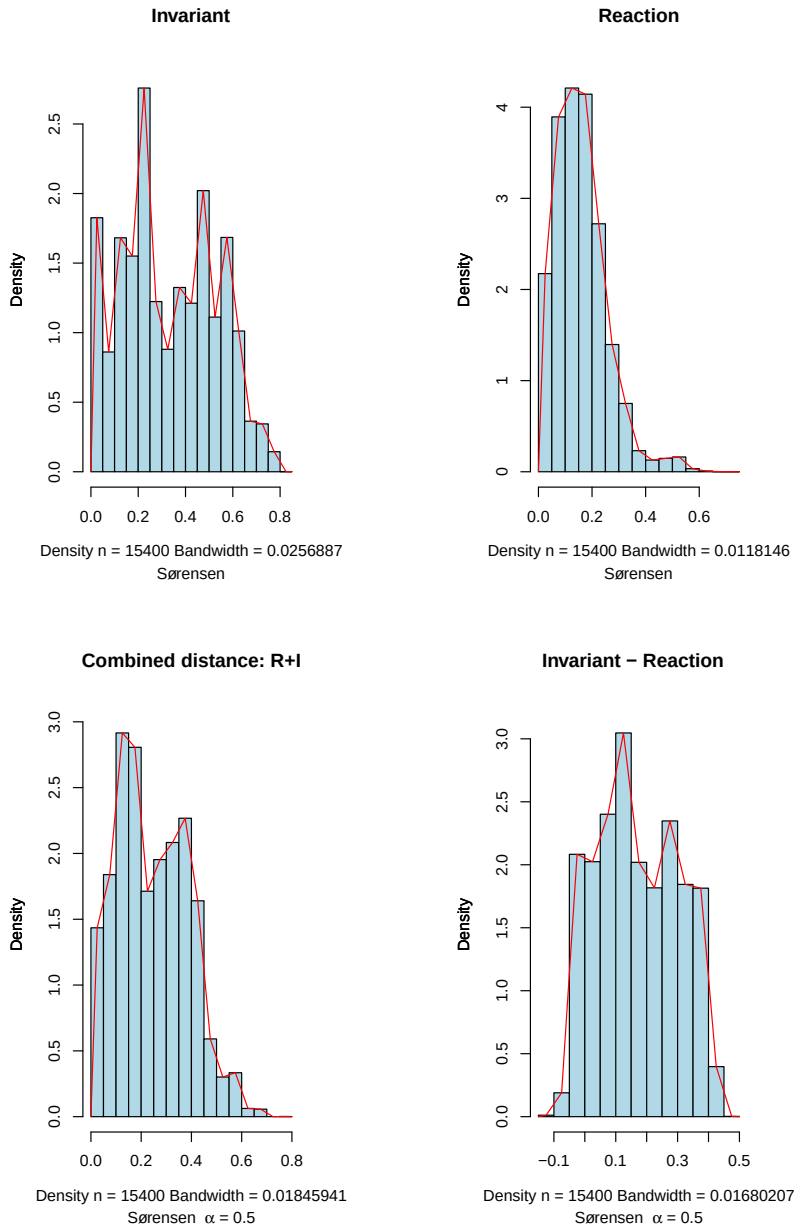


Figura A.1.1: Classe di organismi: *Eukaryotes*

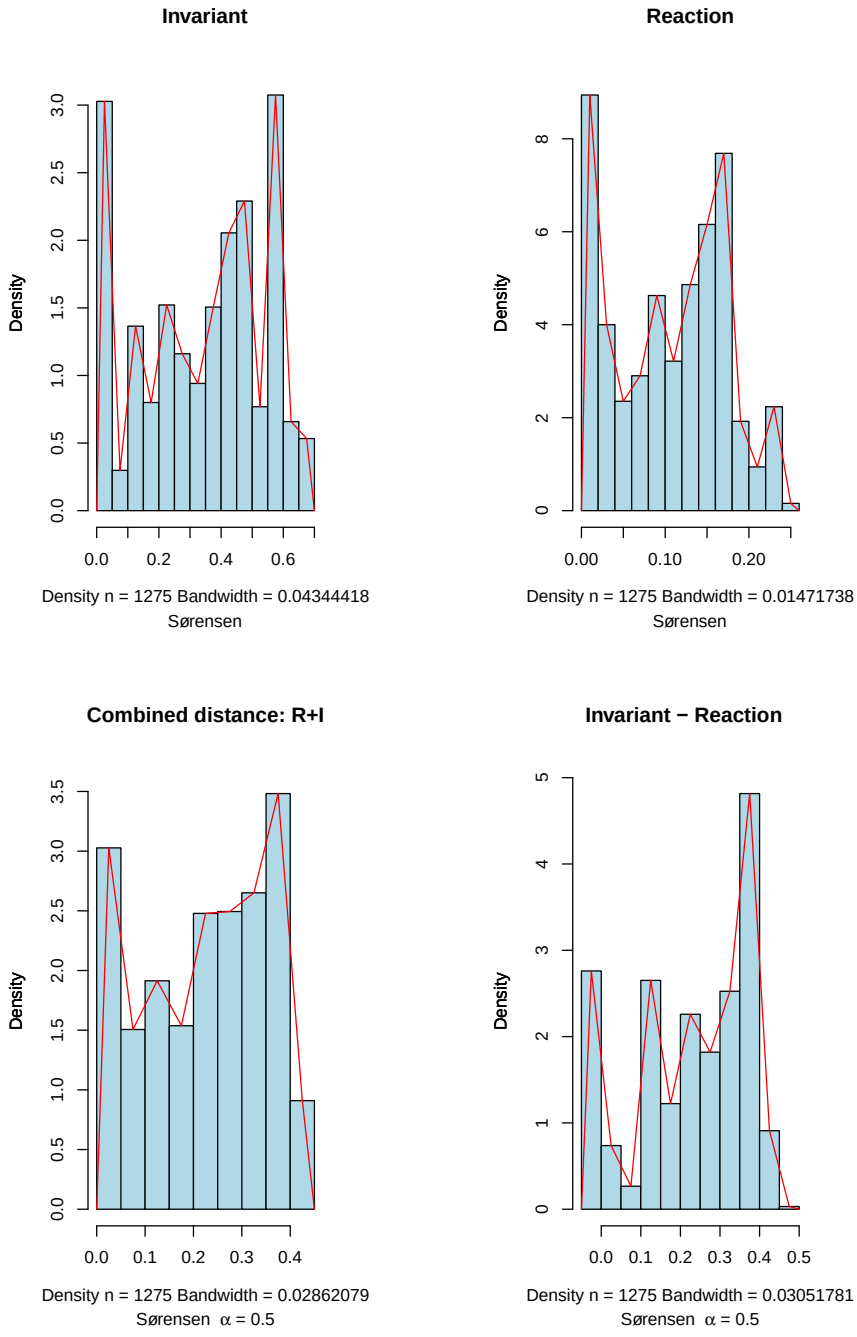


Figura A.1.2: Classe di organismi: *Animals*

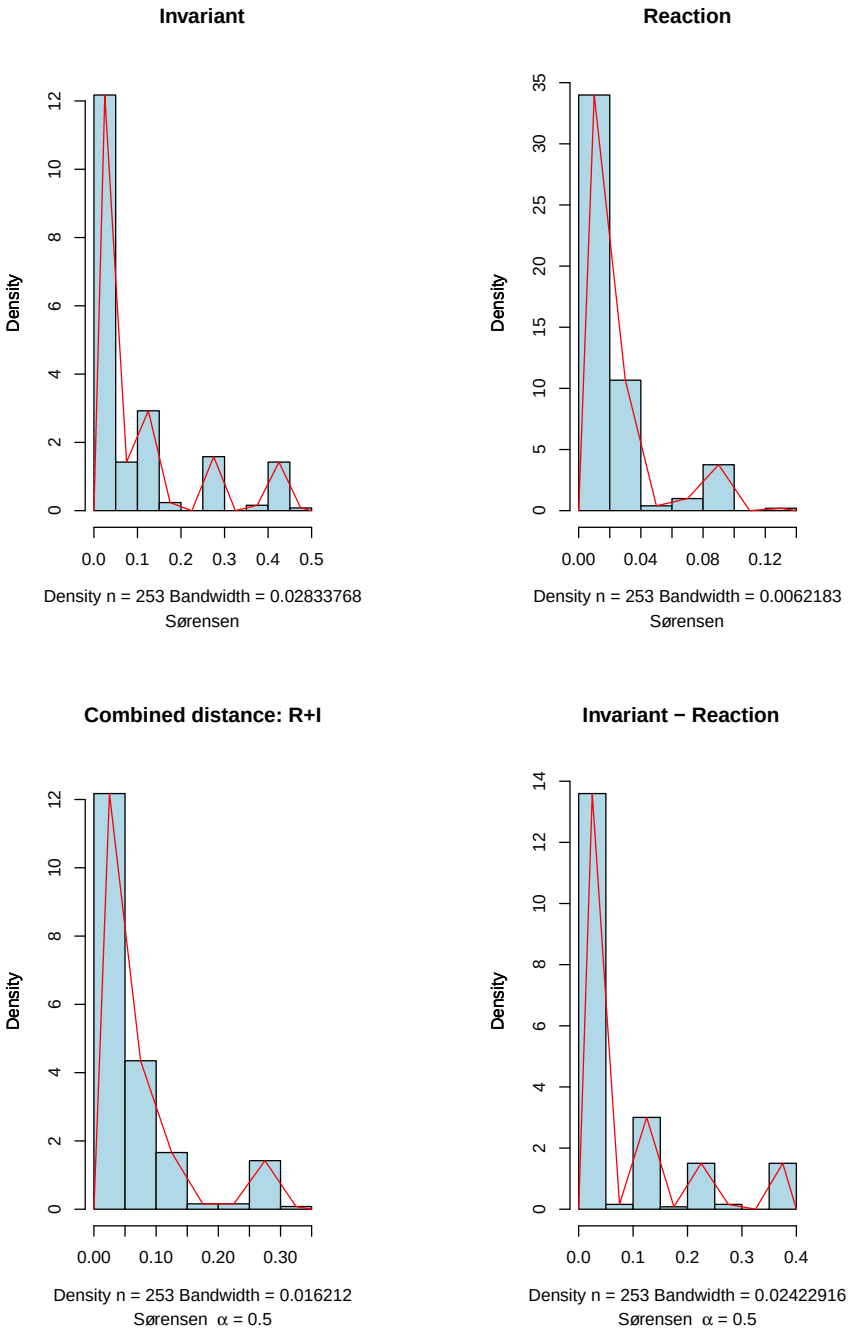


Figura A.1.3: Classe di organismi: *Vertebrates*

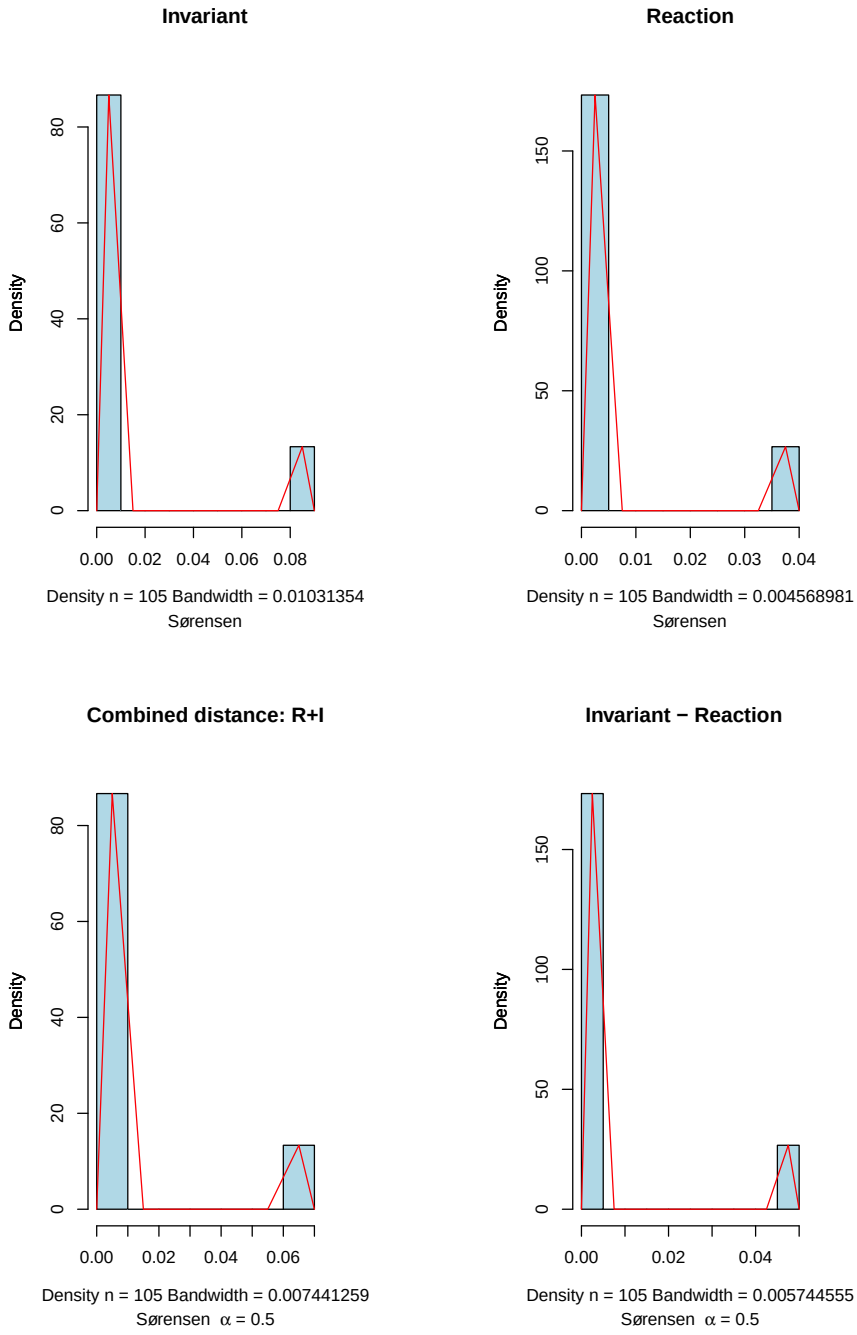


Figura A.1.4: Classe di organismi: *Mammals*

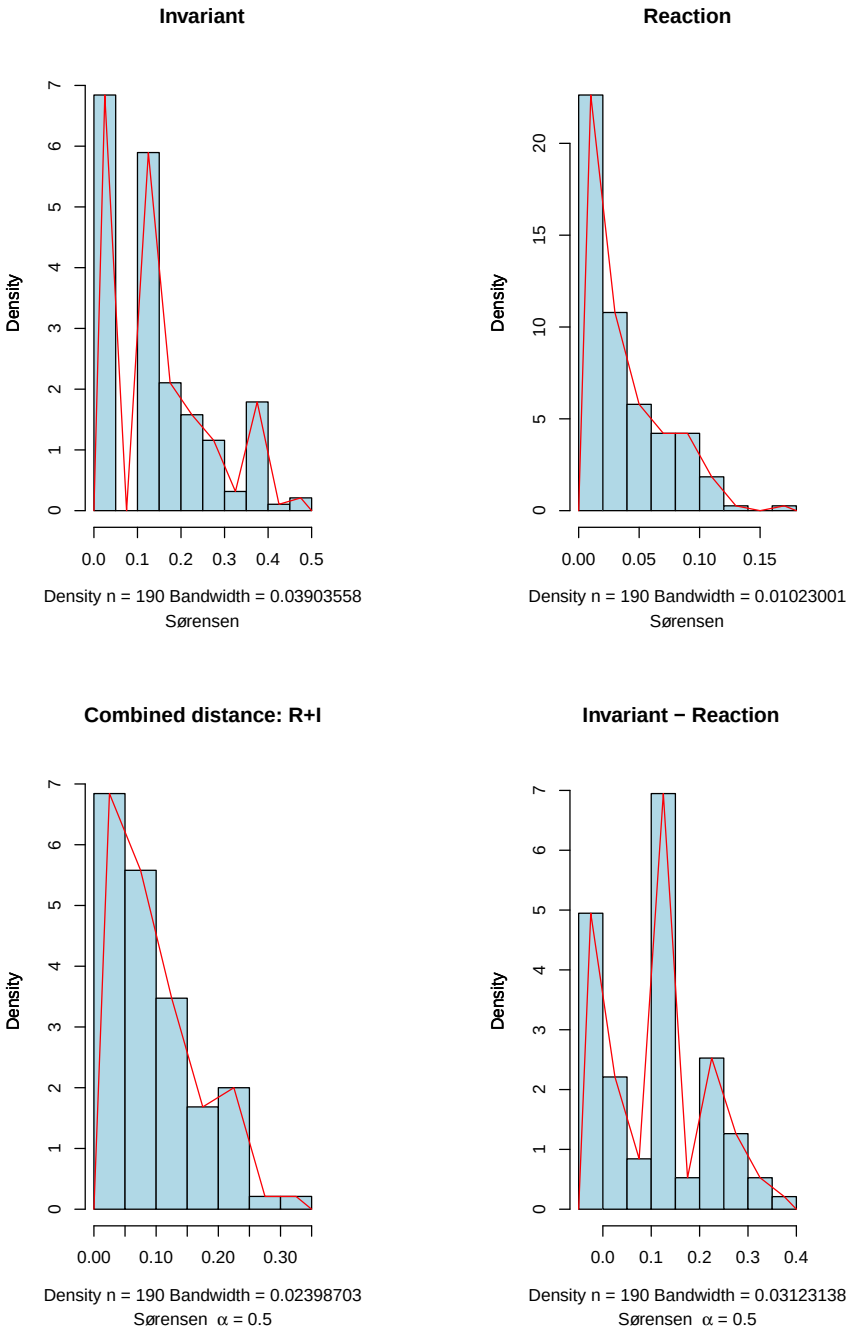


Figura A.1.5: Classe di organismi: *Insect*

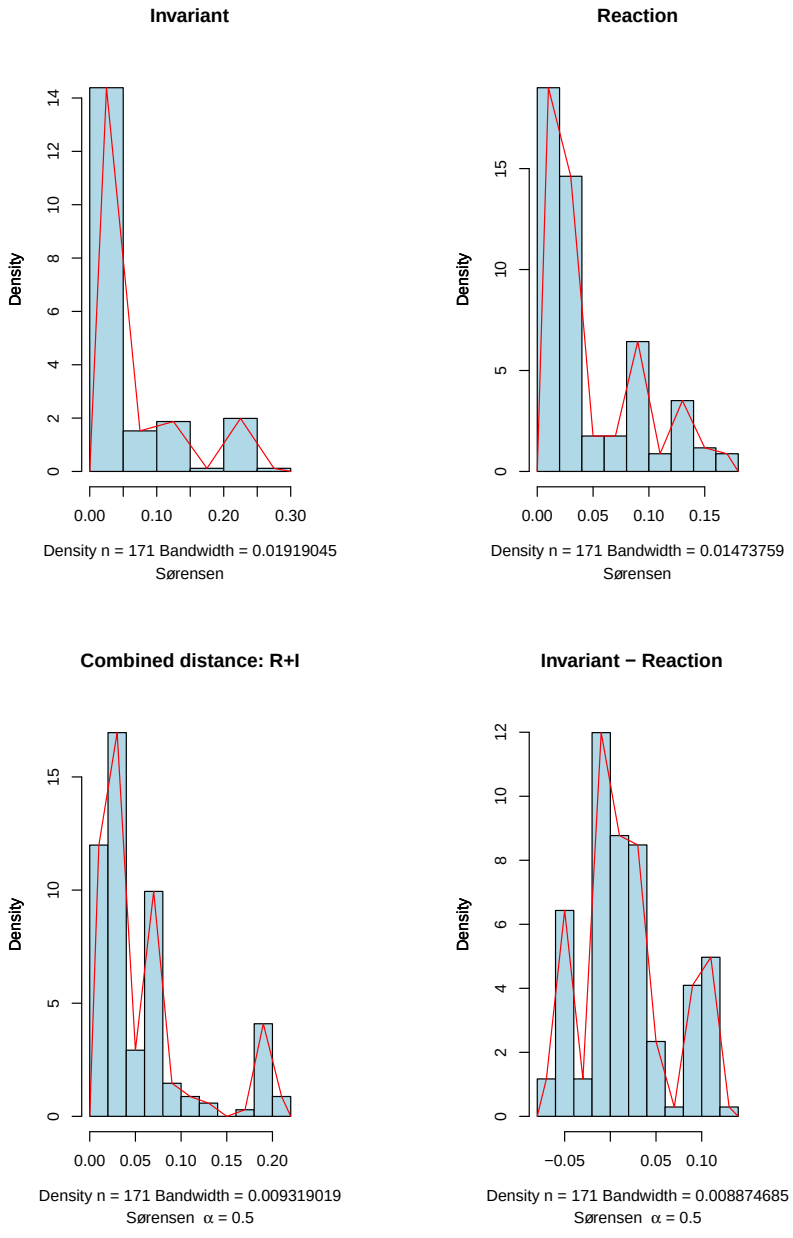


Figura A.1.6: Classe di organismi: *Plants*

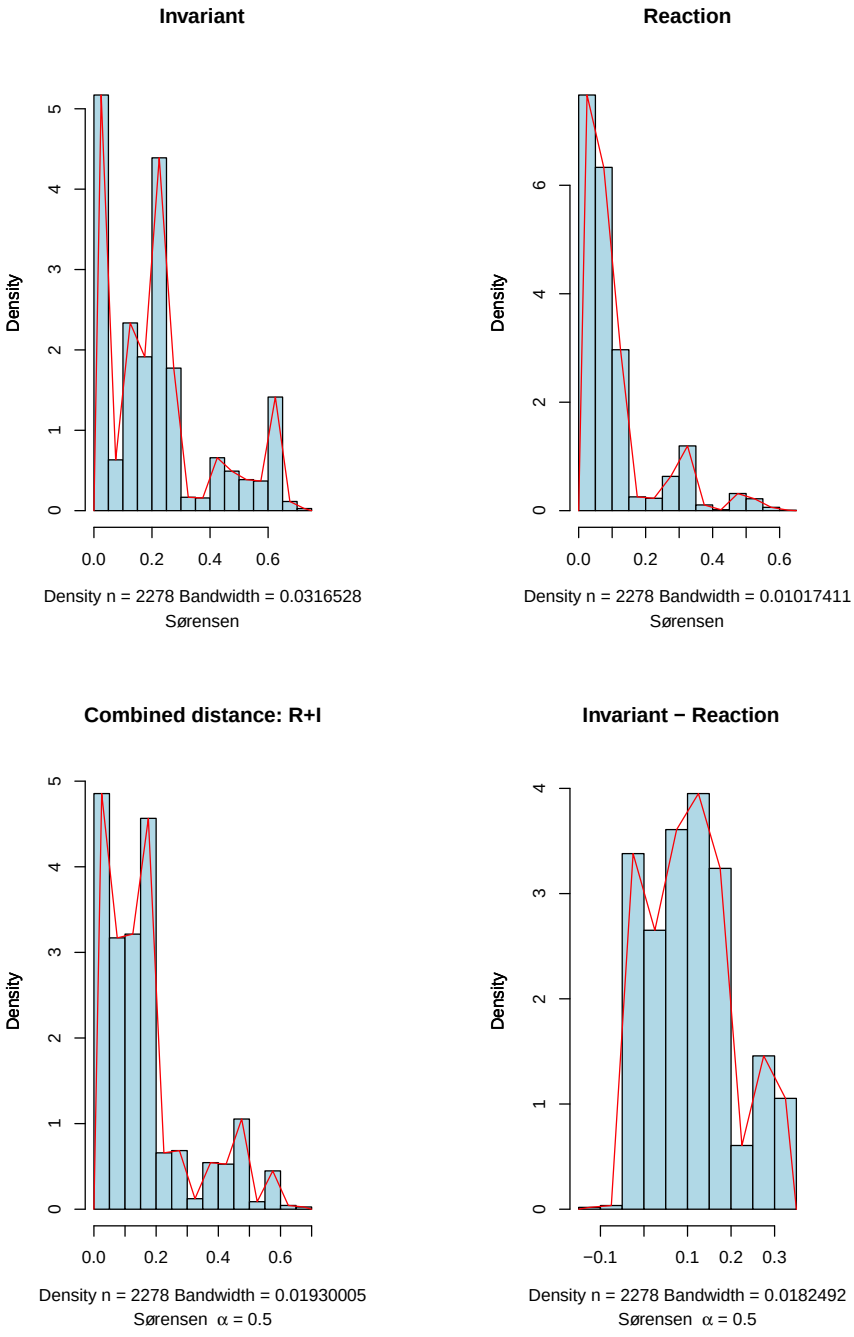


Figura A.1.7: Classe di organismi: *Fungi*

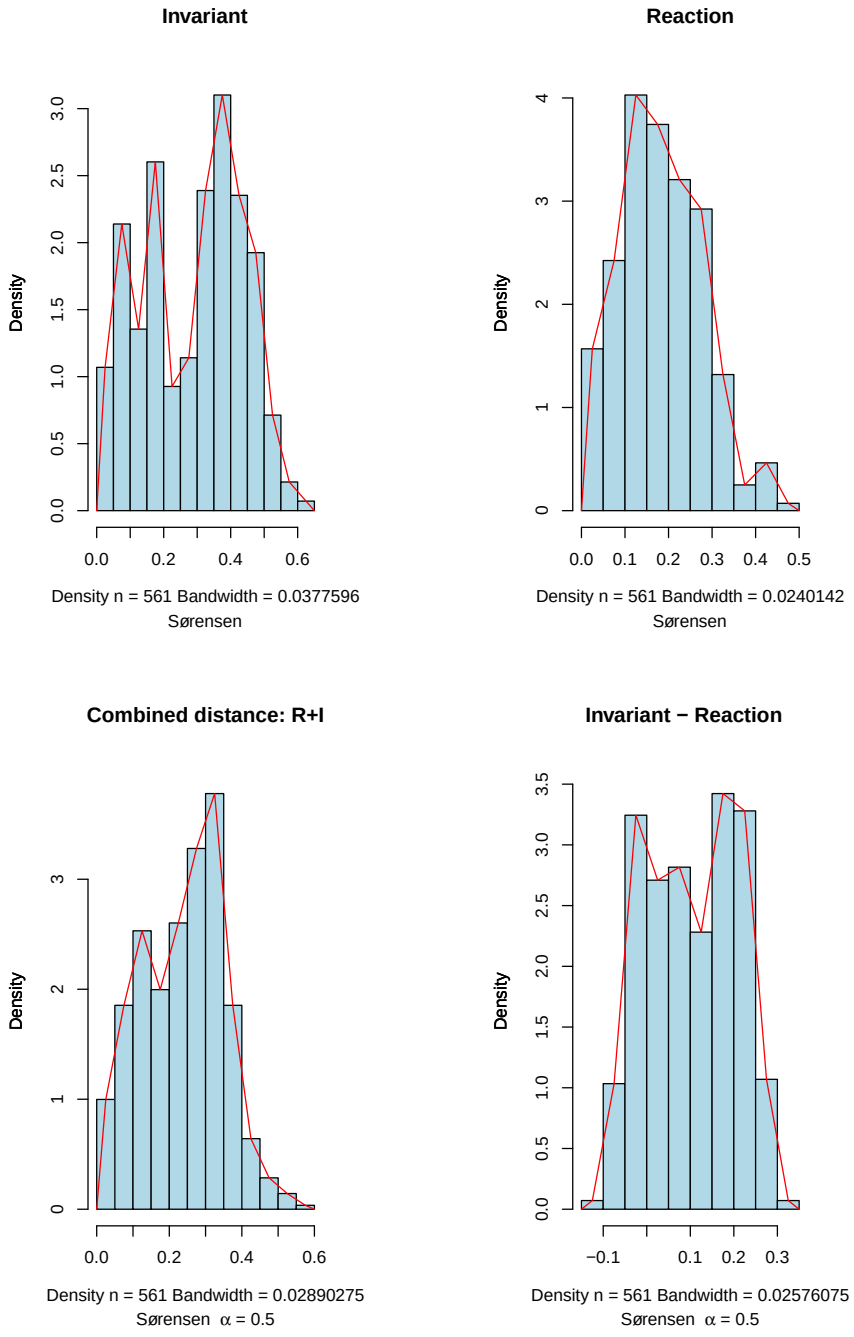


Figura A.1.8: Classe di organismi: *Protists*

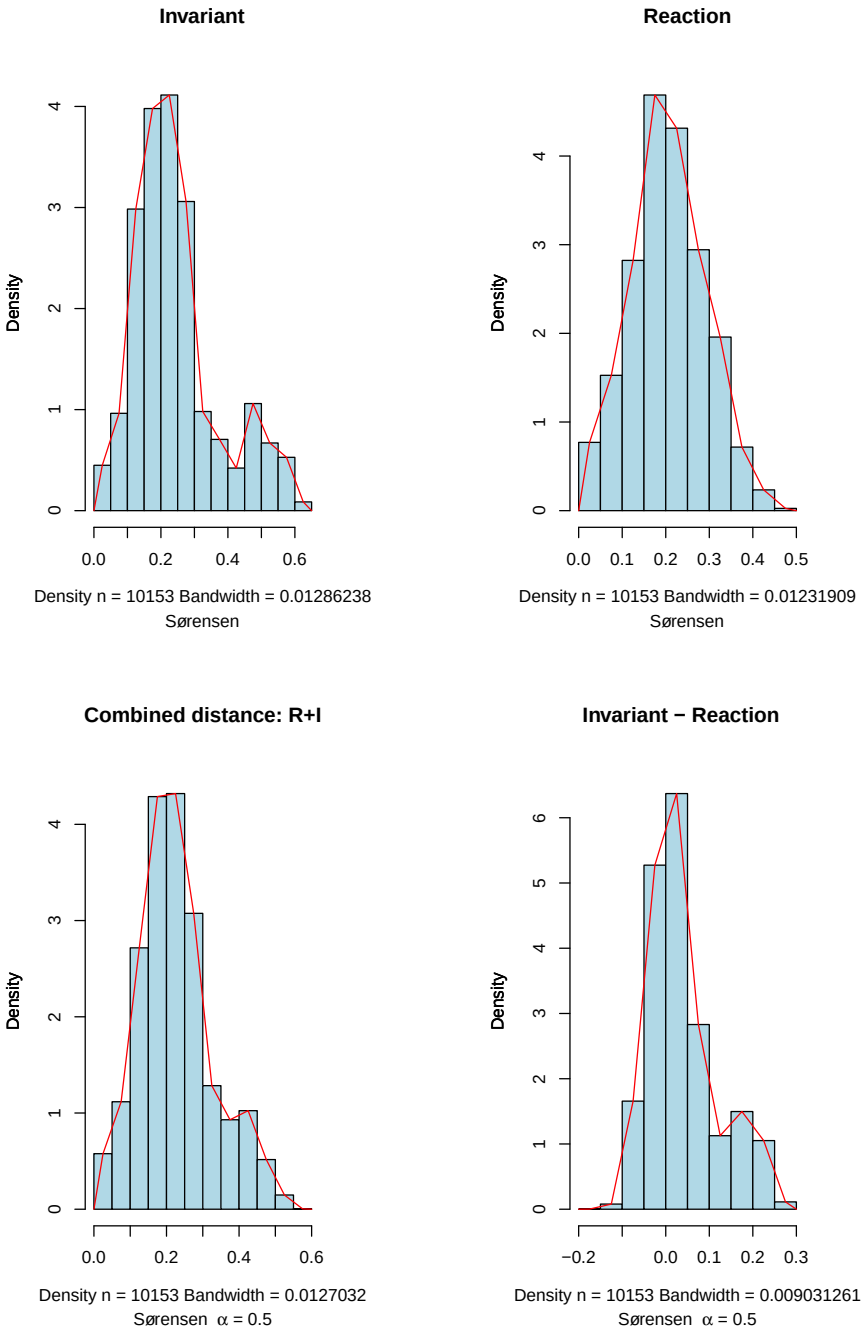


Figura A.1.9: Classe di organismi: *Archaea*

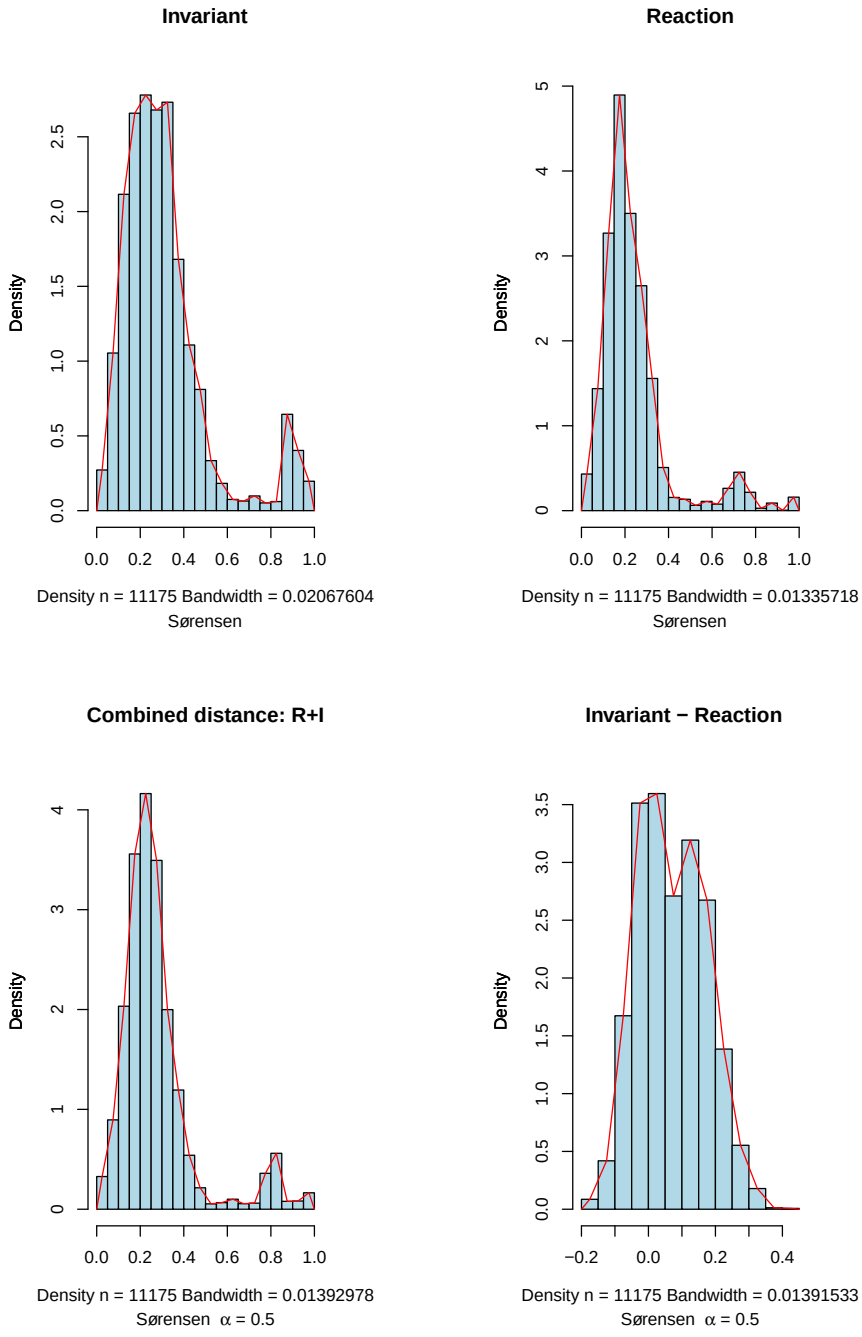


Figura A.1.10: Classe di organismi: *Bacteria*

A.2 L'organismo *hsa.*

A.2.1 L'organismo *hsa* rispetto alla classe *Eukaryotes*

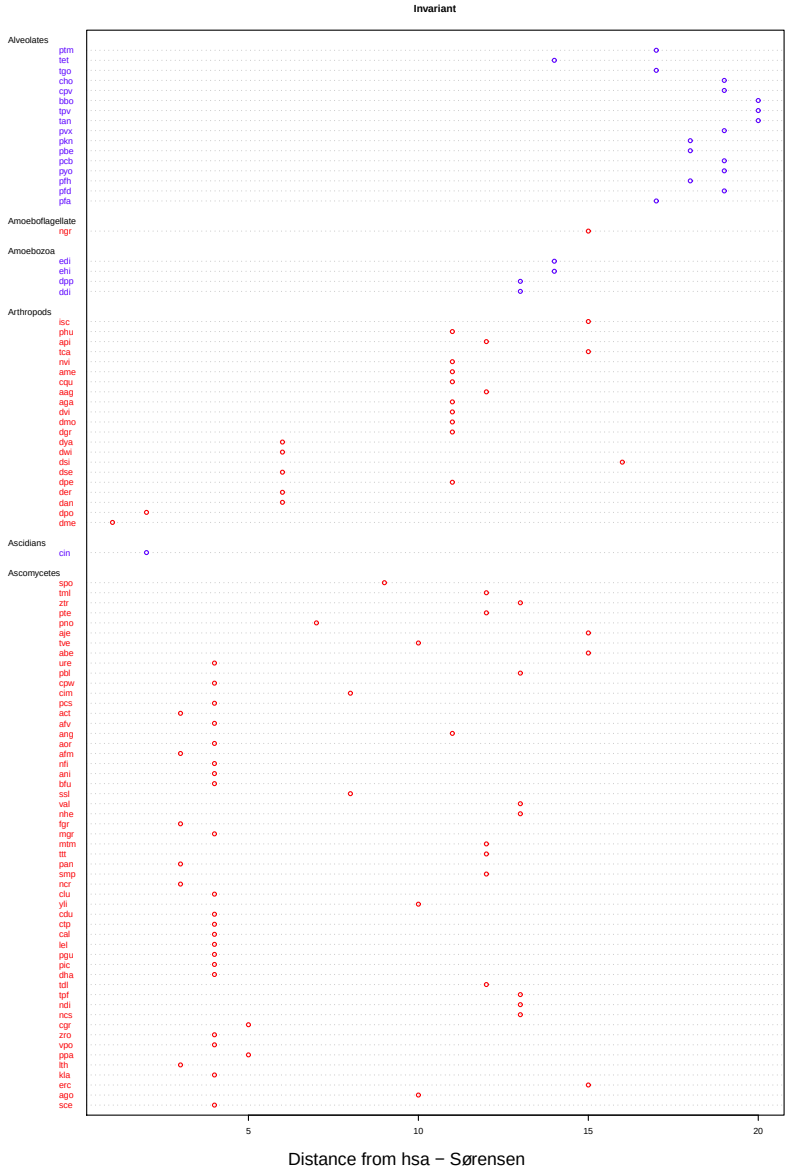


Figura A.2.1: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 1

A.2.2 L'organismo *hsa* rispetto alla classe *Eukaryotes*

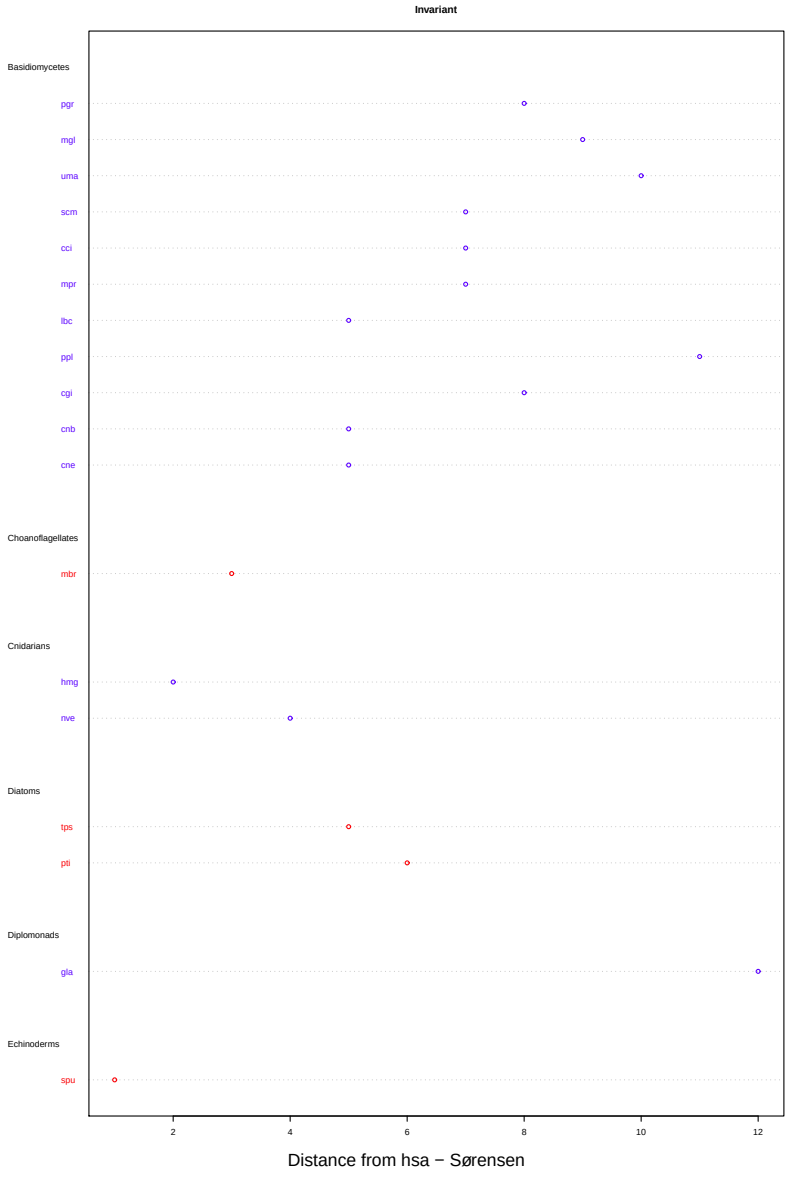


Figura A.2.2: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 2

A.2.3 L'organismo *hsa* rispetto alla classe *Eukaryotes*

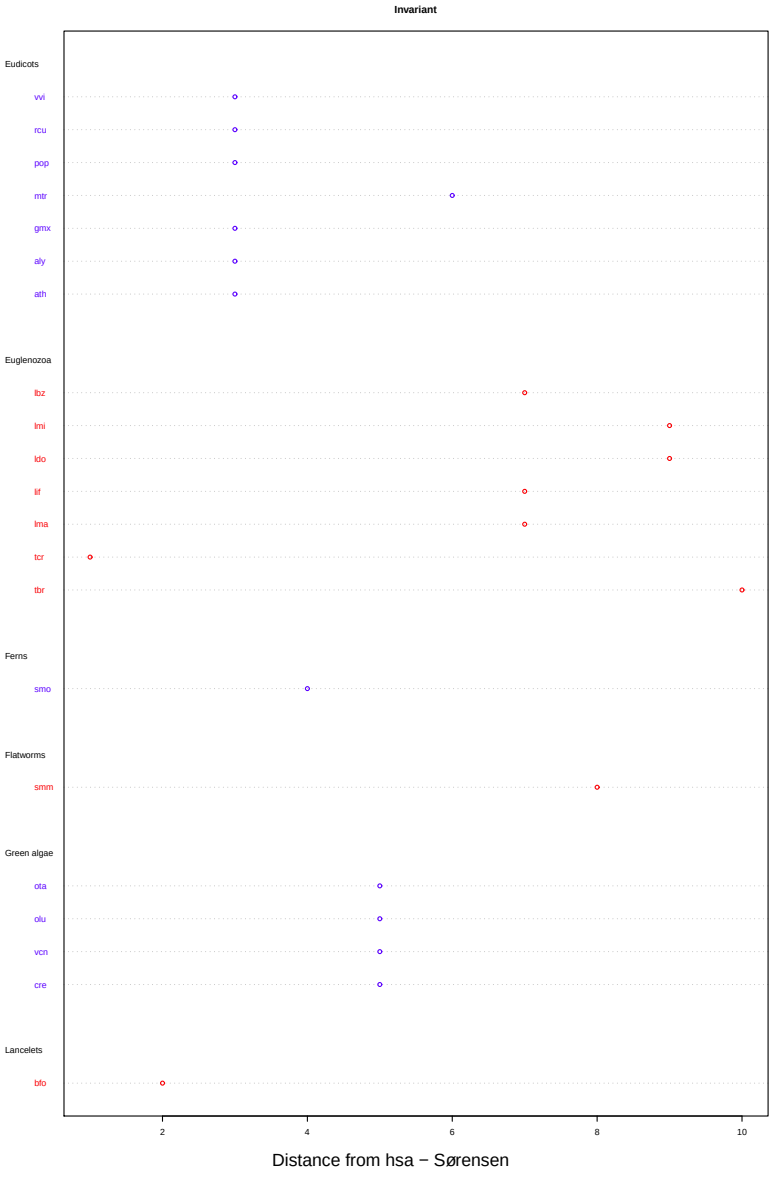


Figura A.2.3: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 3

A.2.4 L'organismo *hsa* rispetto alla classe *Eukaryotes*

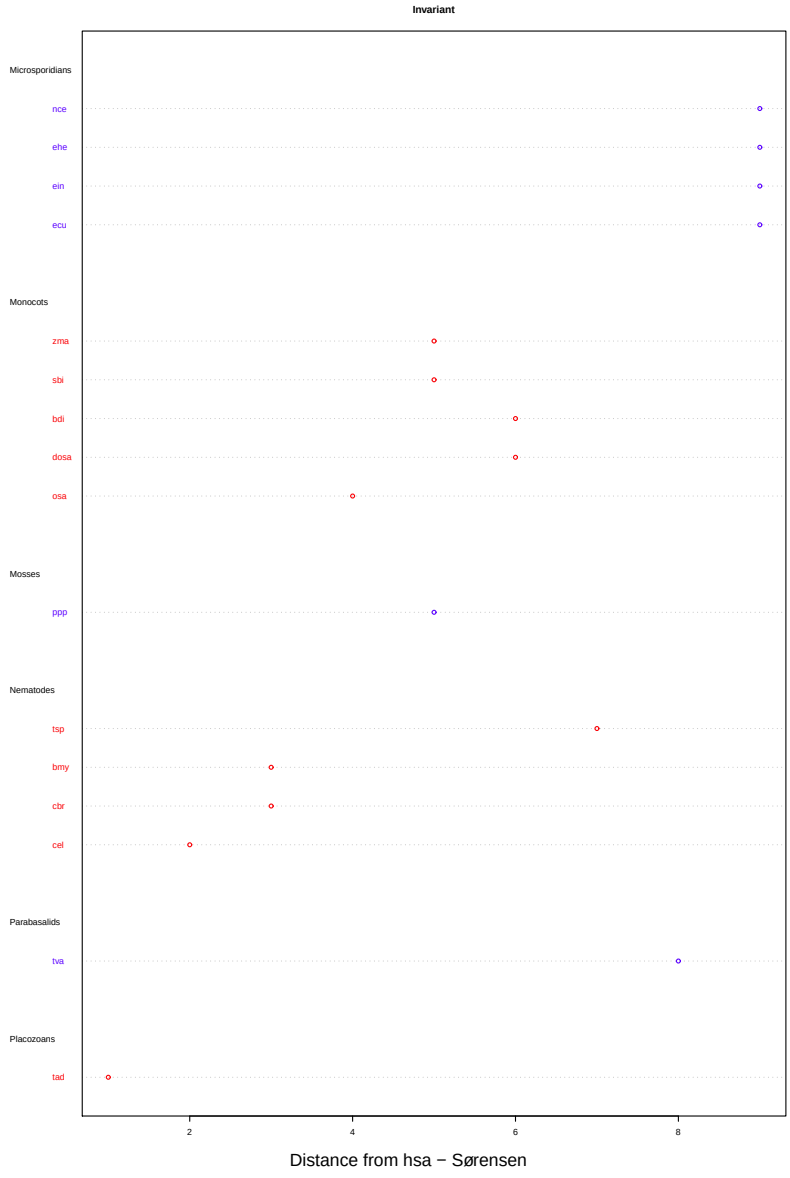


Figura A.2.4: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 4

A.2.5 L'organismo *hsa* rispetto alla classe *Eukaryotes*

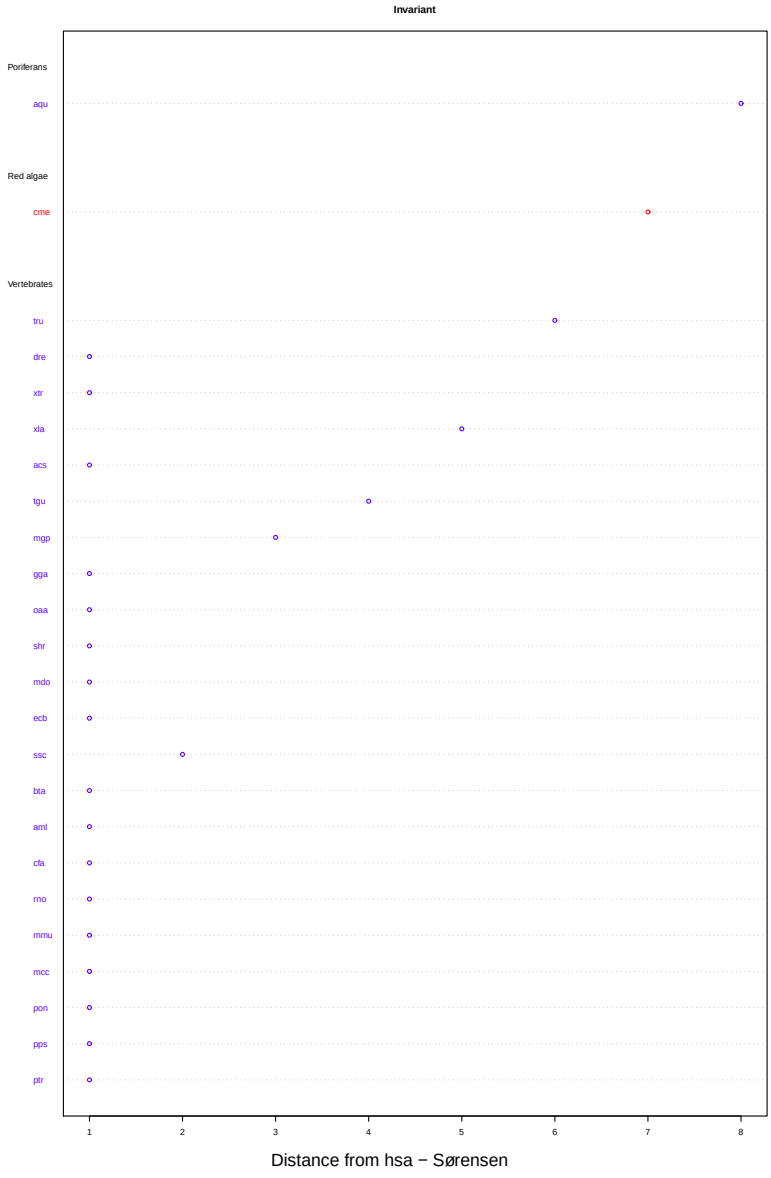


Figura A.2.5: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 5

A.2.6 L'organismo *hsa* rispetto alla classe *Eukaryotes*

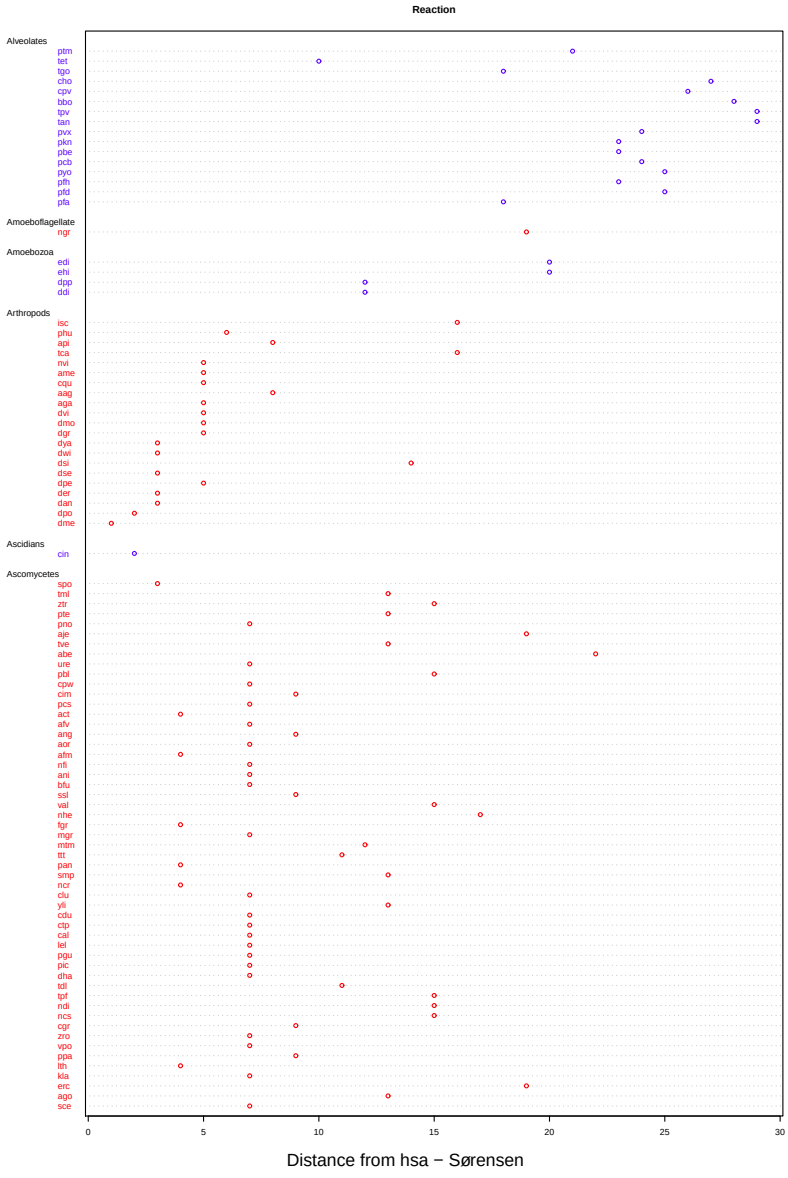


Figura A.2.6: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 1

A.2.7 L'organismo *hsa* rispetto alla classe *Eukaryotes*

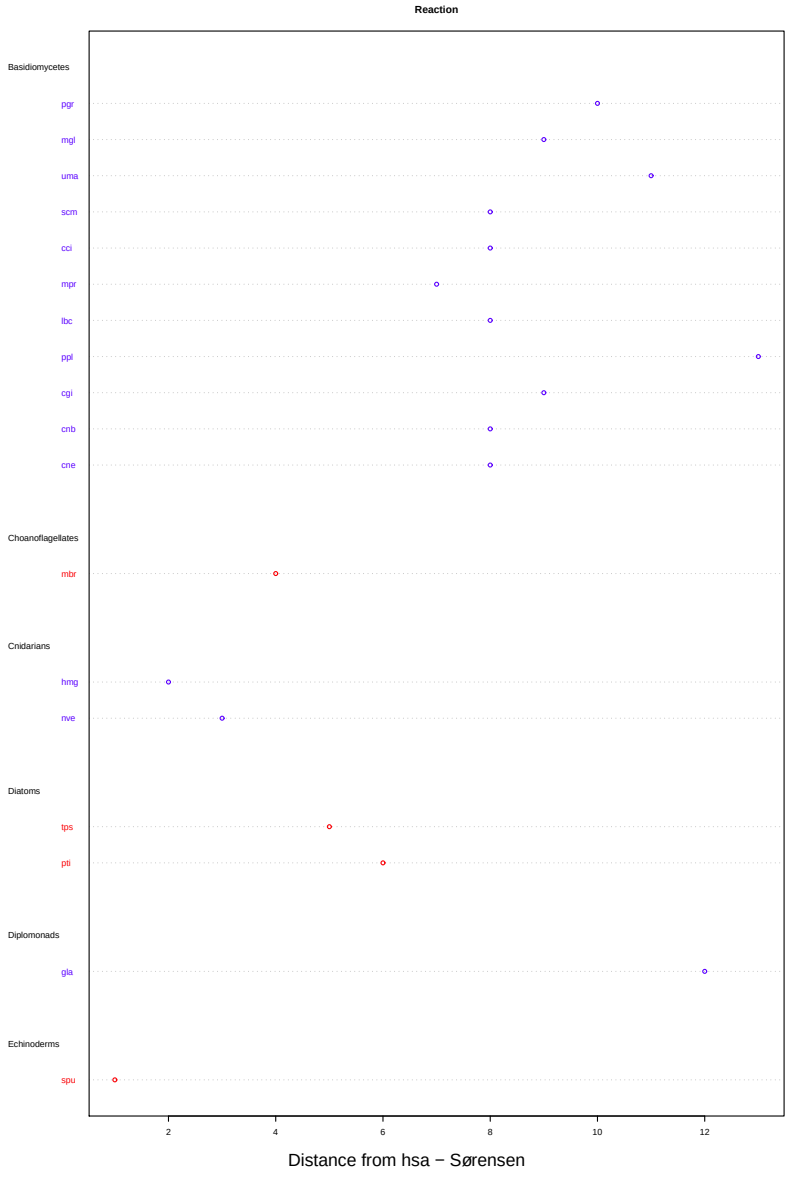


Figura A.2.7: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 2

A.2.8 L'organismo *hsa* rispetto alla classe *Eukaryotes*

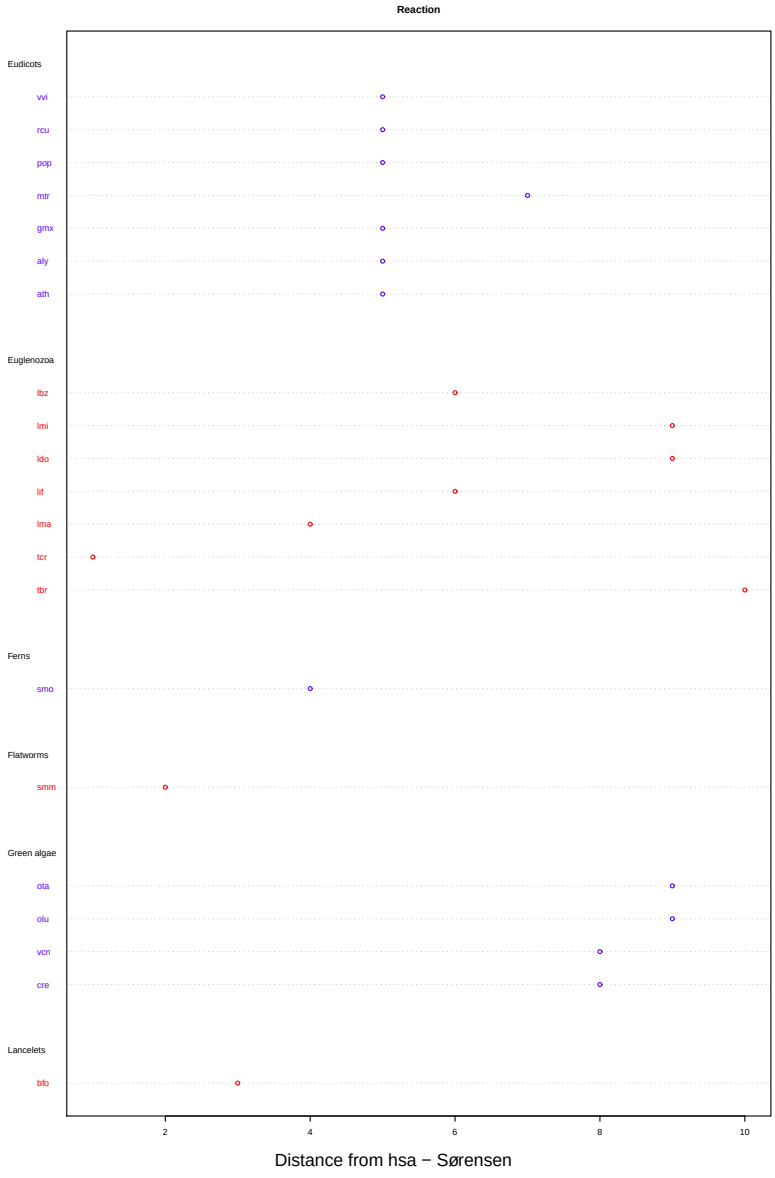


Figura A.2.8: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 3

A.2.9 L'organismo *hsa* rispetto alla classe *Eukaryotes*

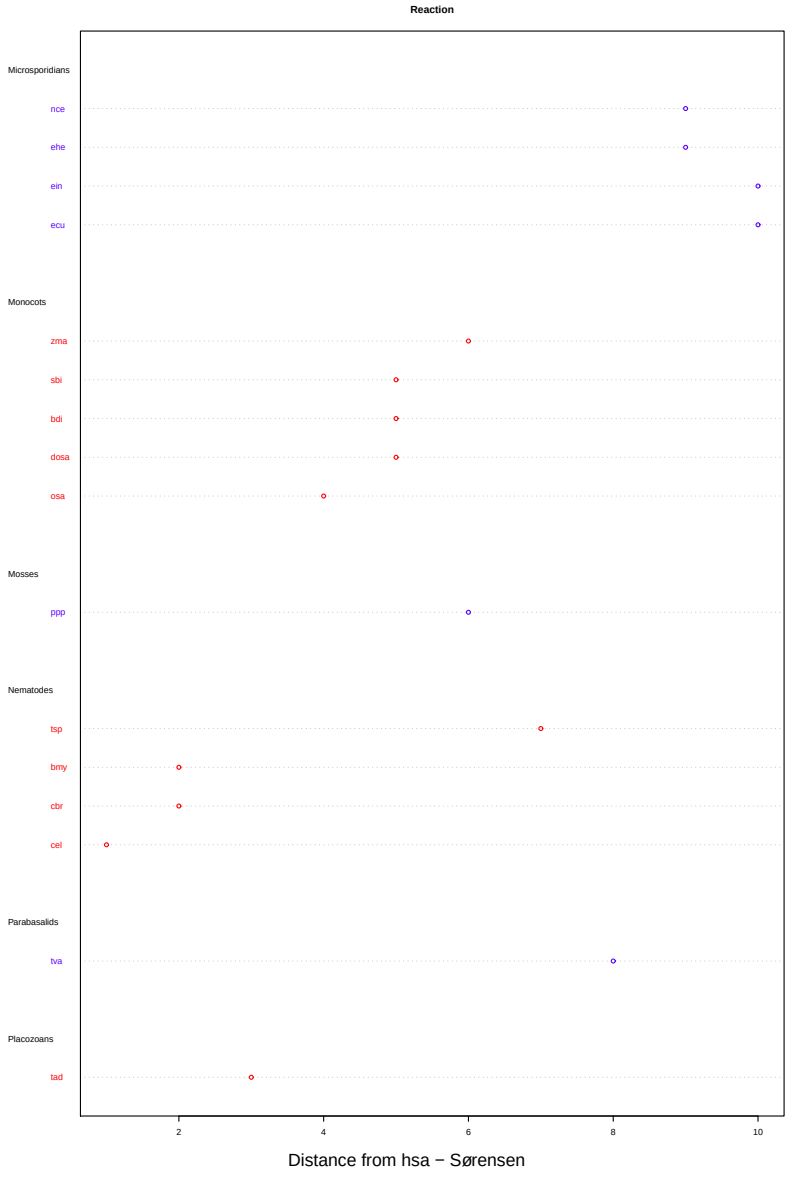


Figura A.2.9: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 4

A.2.10 L'organismo *hsa* rispetto alla classe *Eukaryotes*

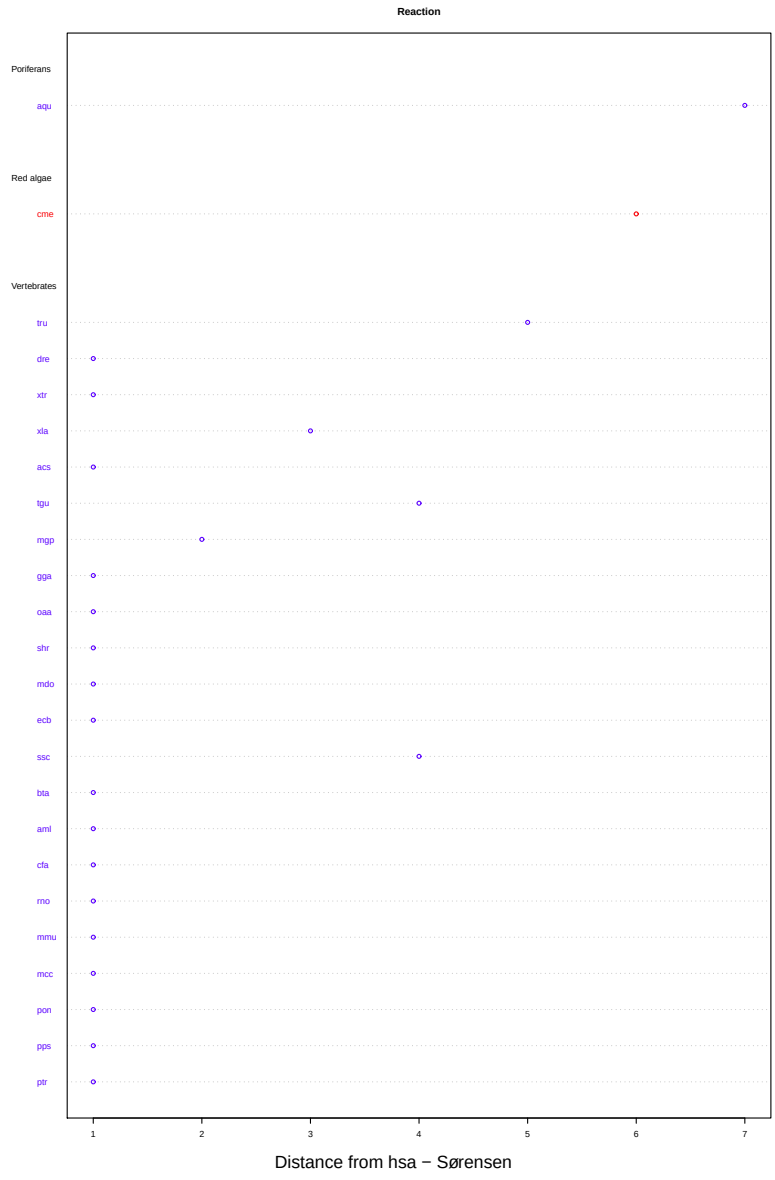


Figura A.2.10: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 5

A.2.11 L'organismo *hsa* rispetto alla classe *Animals*

Figura A.2.11: Distanza: d_I l' organismo *hsa* nella classe *Animals*

A.2.12 L'organismo *hsa* rispetto alla classe *Animals*

Figura A.2.12: Distanza: d_R l' organismo *hsa* nella classe *Animals*

A.2.13 L'organismo *hsa* rispetto alla classe *Vertebrates*

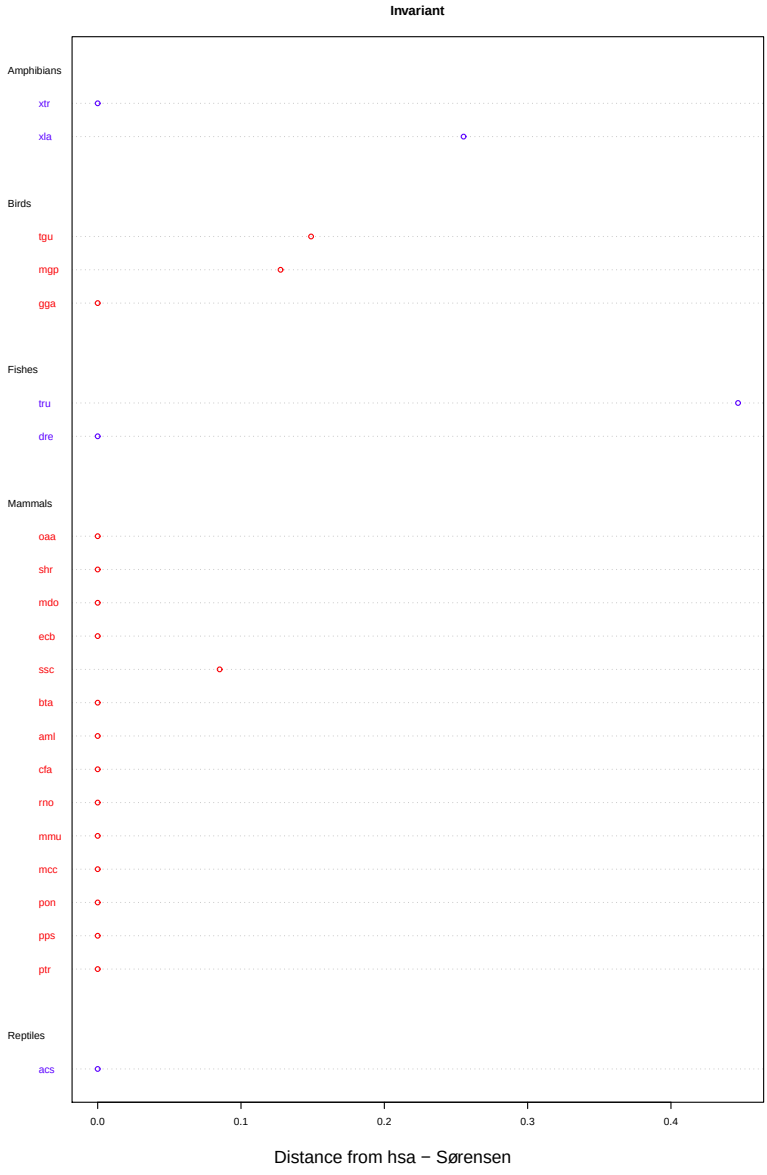


Figura A.2.13: Distanza: d_I l' organismo *hsa* nella classe *Vertebrates*

A.2.14 L'organismo *hsa* rispetto alla classe *Vertebrates*

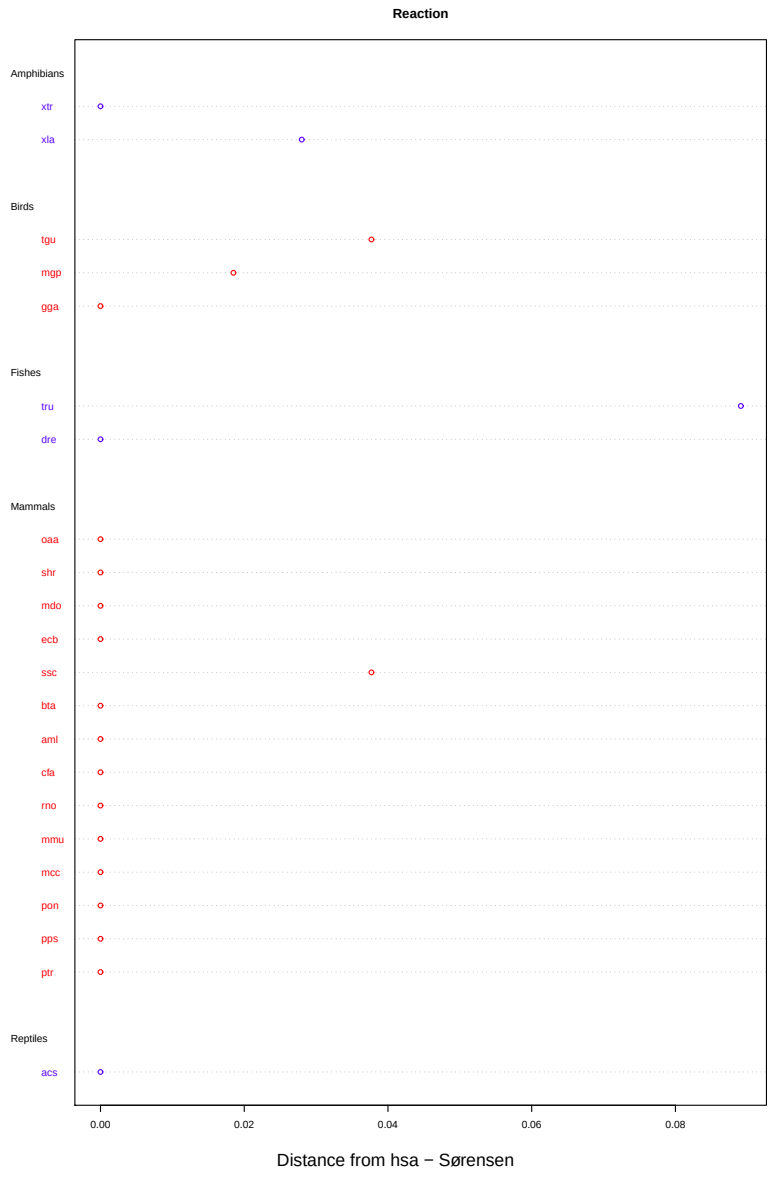


Figura A.2.14: Distanza: d_R l' organismo *hsa* nella classe *Vertebrates*

A.2.15 L'organismo *hsa* rispetto alla classe *Mammals*

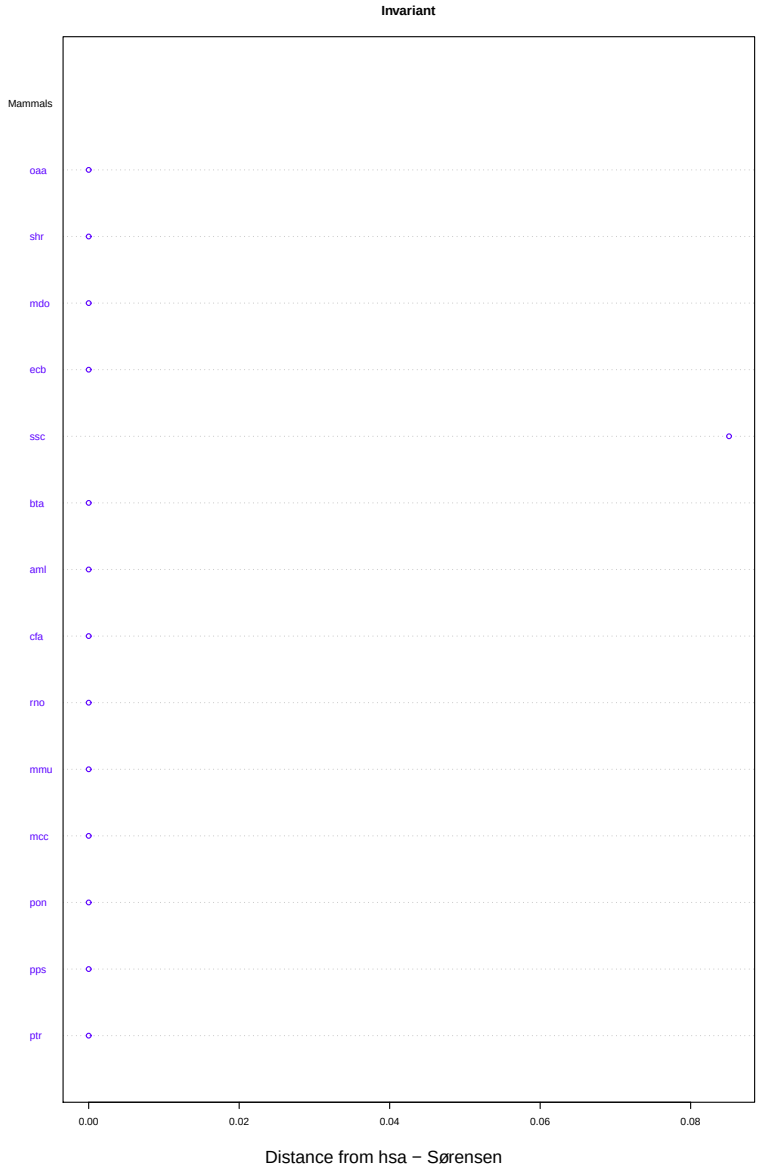


Figura A.2.15: Distanza: d_I l' organismo *hsa* nella classe *Mammals*

A.2.16 L'organismo *hsa* rispetto alla classe *Mammals*

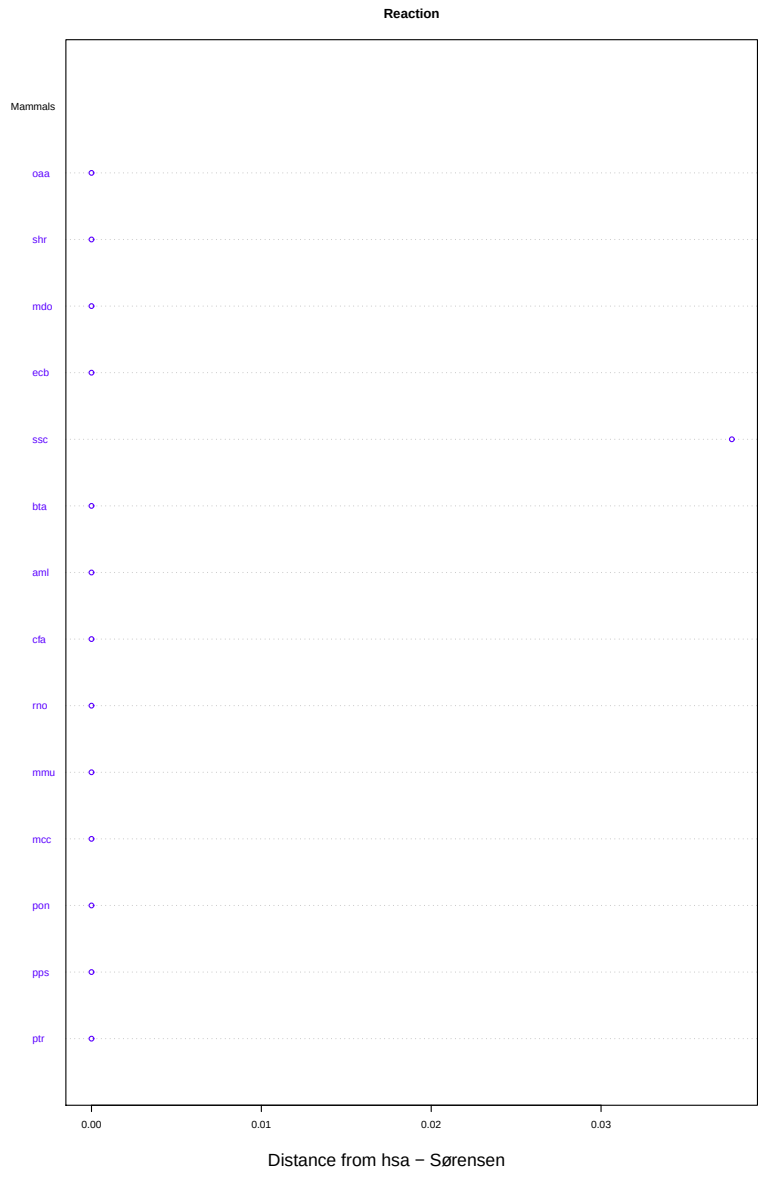


Figura A.2.16: Distanza: d_R l' organismo *hsa* nella classe *Mammals*

A.3 La coppia *(hsa,ldo)* nella classe *Eukaryotes*

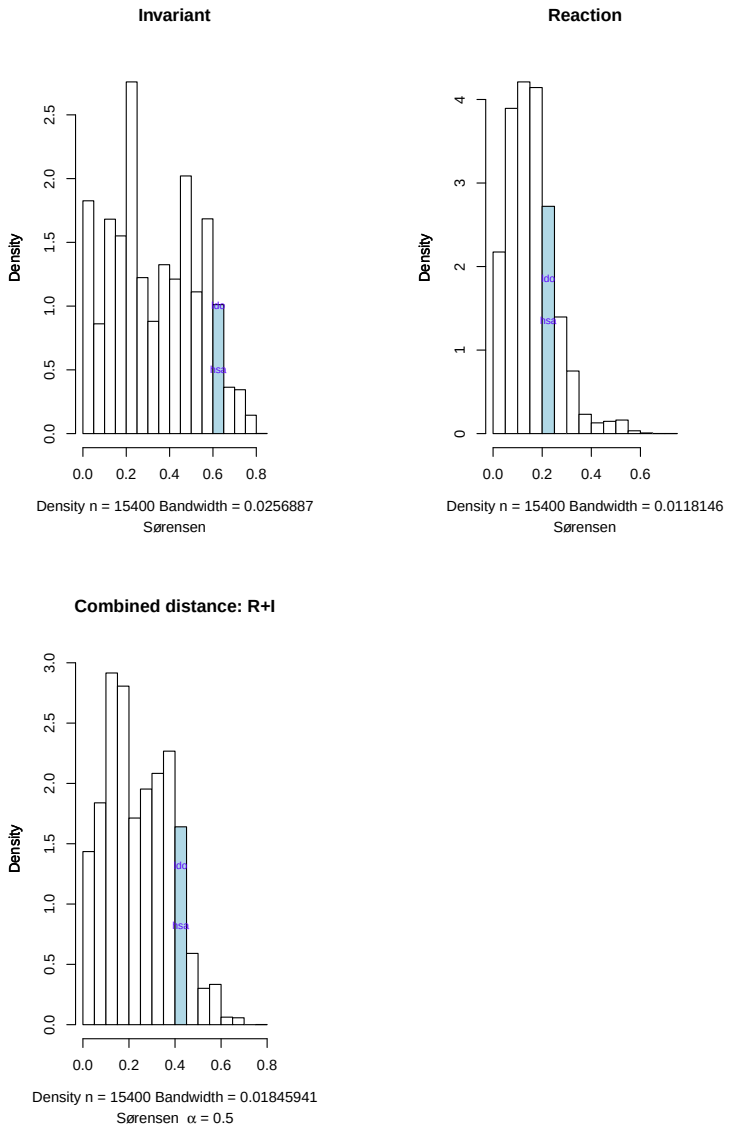


Figura A.3.1: Coppia: (hsa, ldo)

A.4 La coppia *(hsa,nve)* nella classe *Animals*

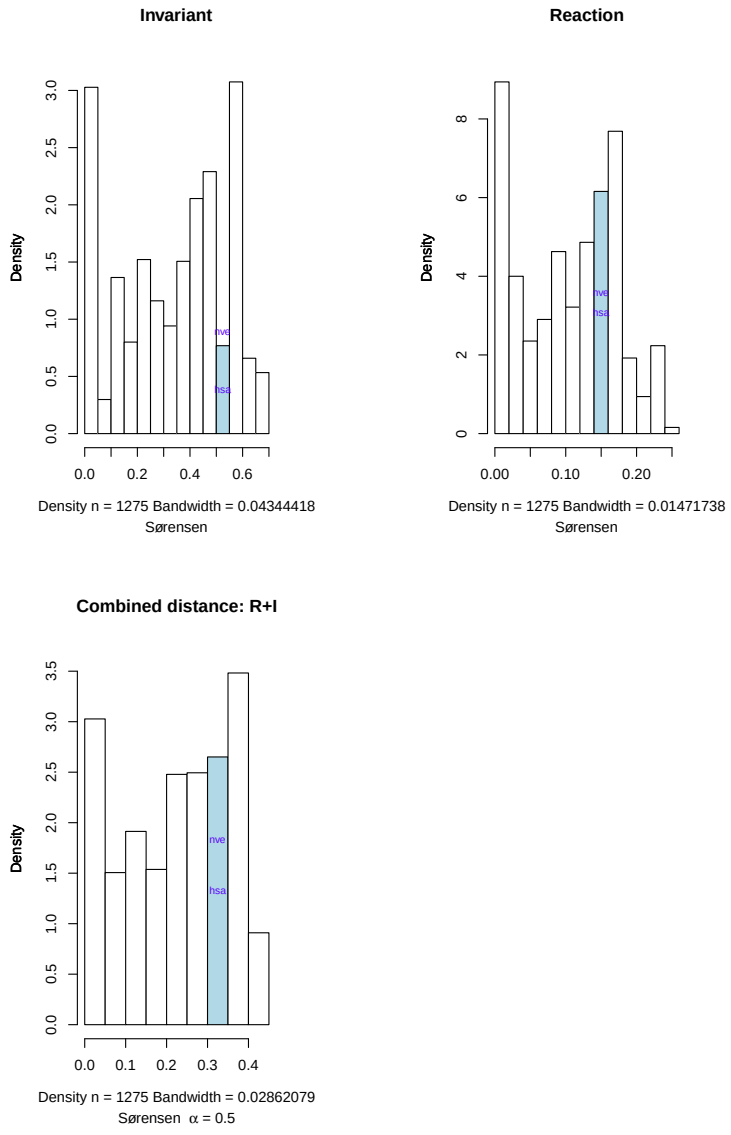


Figura A.4.1: Coppia: (hsa, nve)

A.5 La coppia *(hsa,pon)* nella classe *Vertebrates*

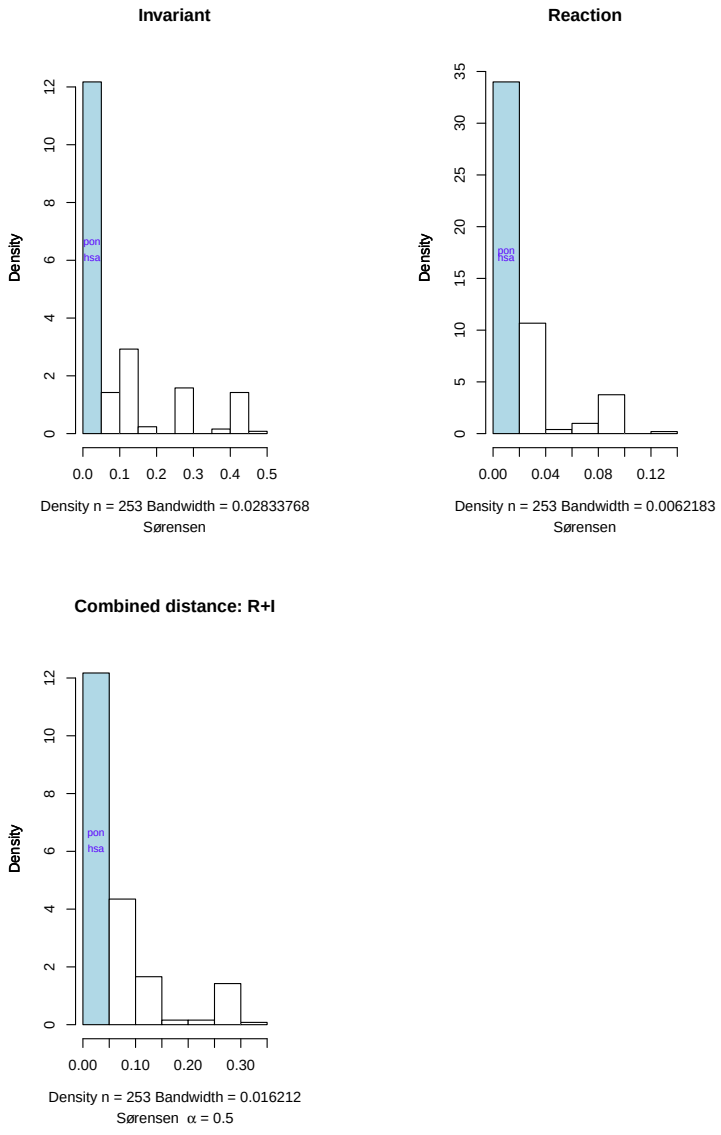


Figura A.5.1: Coppia: (hsa, pon)

A.6 La coppia *(hsa,oaa)* nella classe *Mammals*

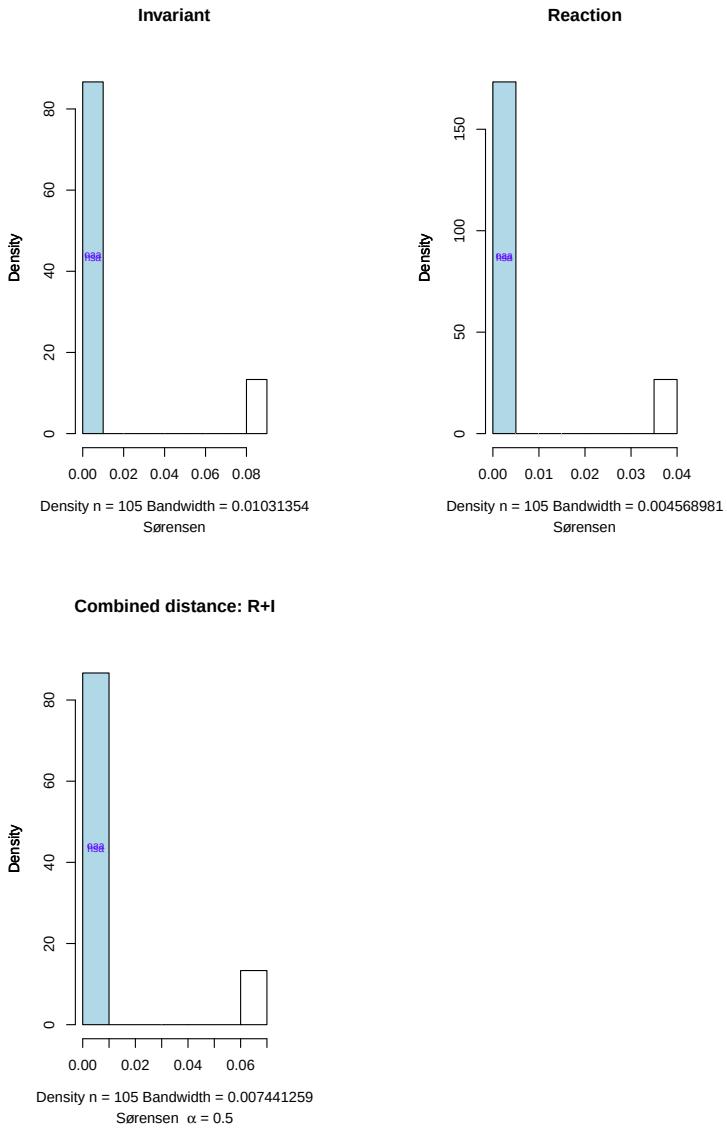


Figura A.6.1: Coppia: (hsa, oaa)

A.7 La coppia *(dme,phu)* nella classe *Insects*

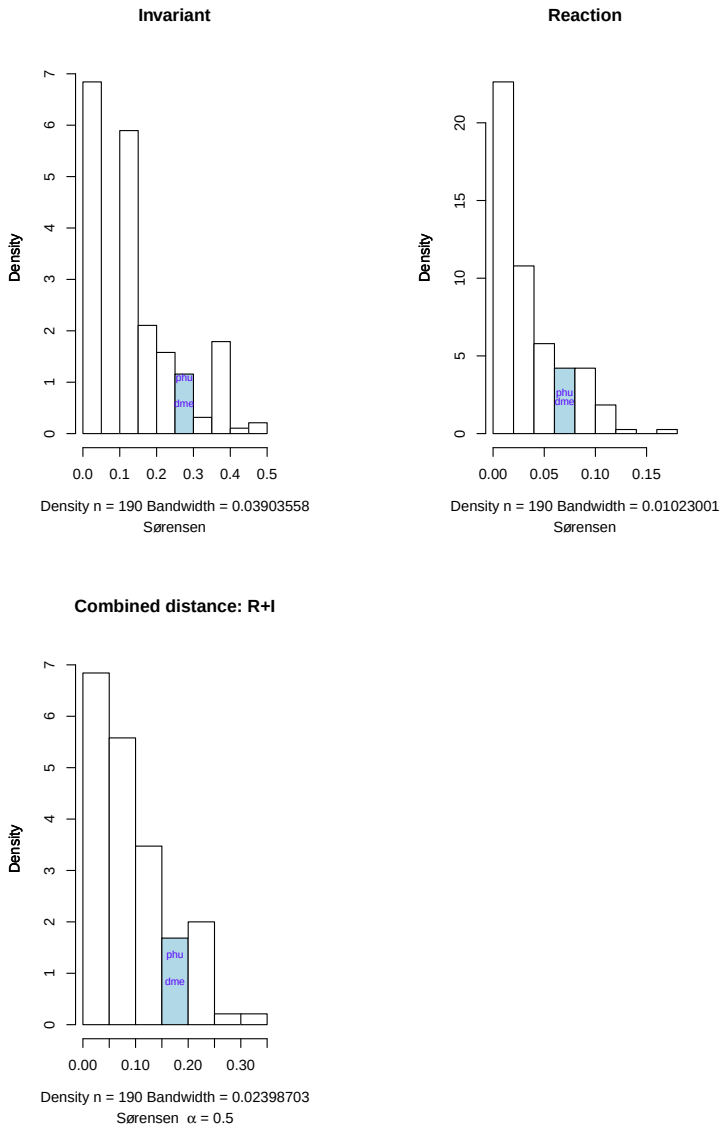


Figura A.7.1: Coppia: (dme, phu)

A.8 La coppia *(ath,cme)* nella classe *Plants*

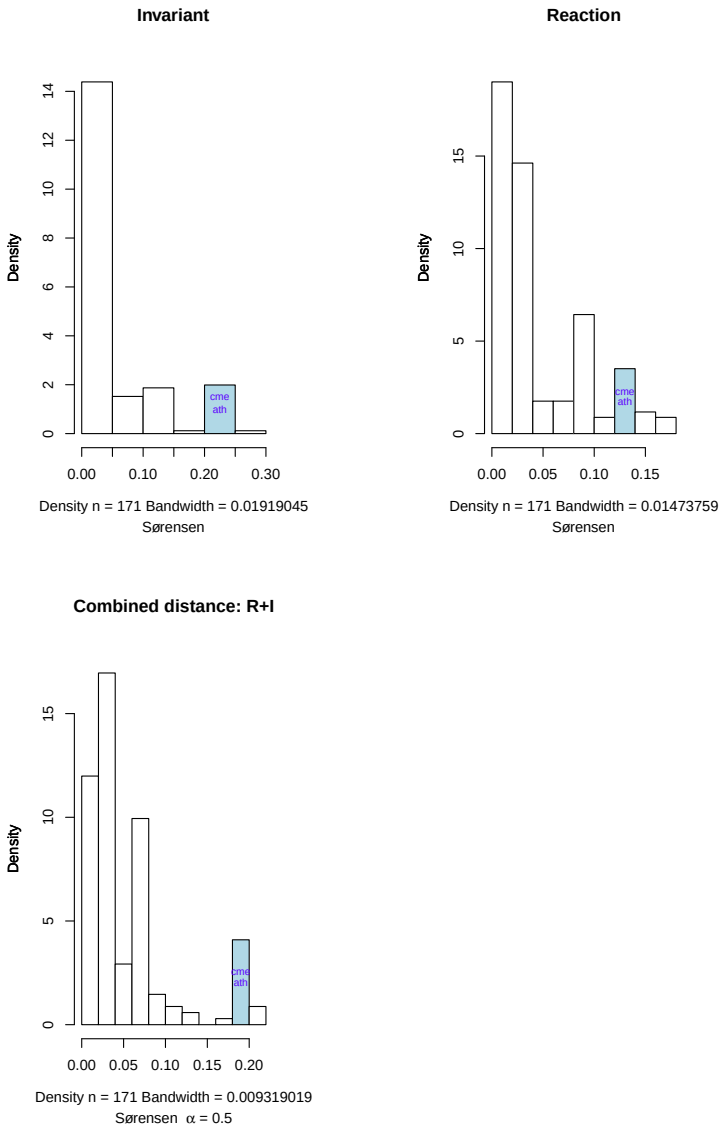


Figura A.8.1: Coppia: (ath, cme)

A.9 La coppia *(sce,nce)* nella classe *Fungi*

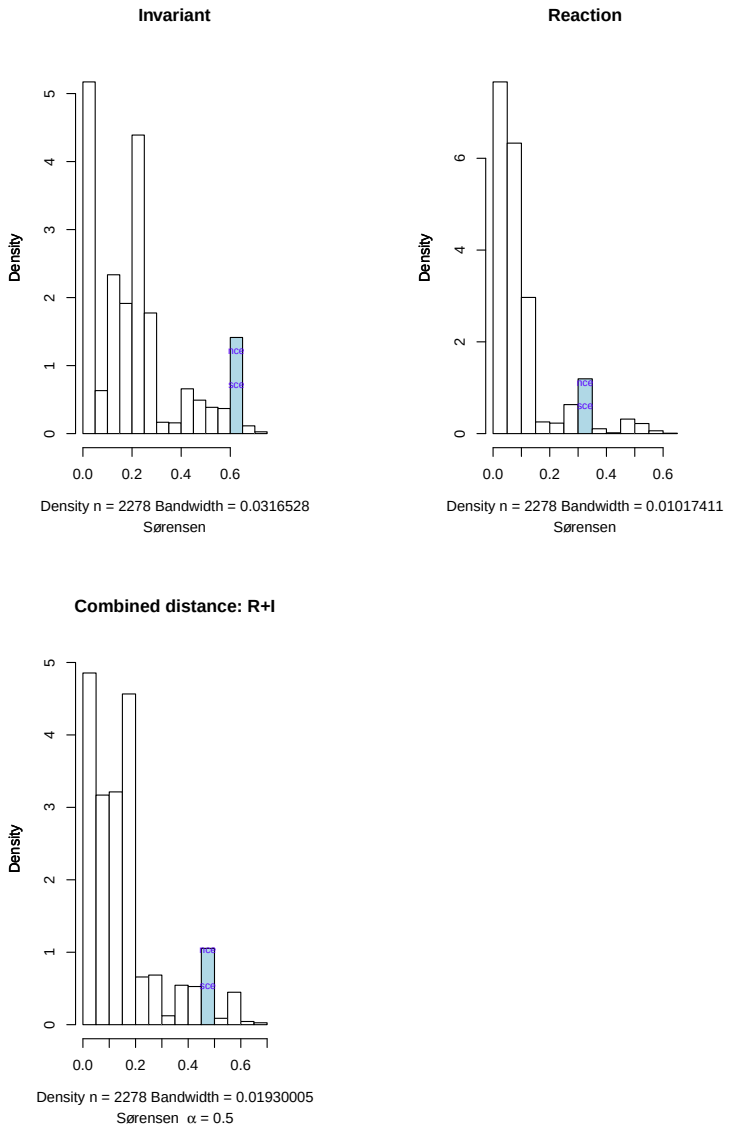


Figura A.9.1: Coppia: (sce, nce)

A.10 La coppia *(mbr,tcr)* nella classe *Protists*

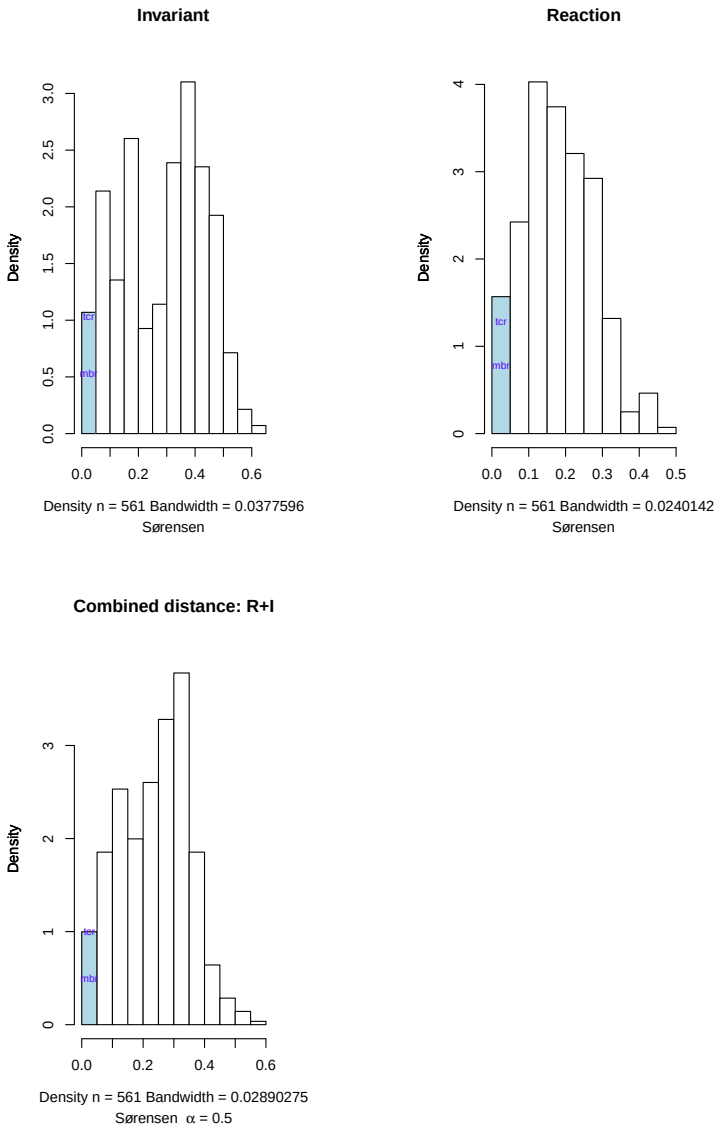


Figura A.10.1: Coppia: (mbr, tcr)

A.11 La coppia *(mja,hah)* nella classe *Archaea*

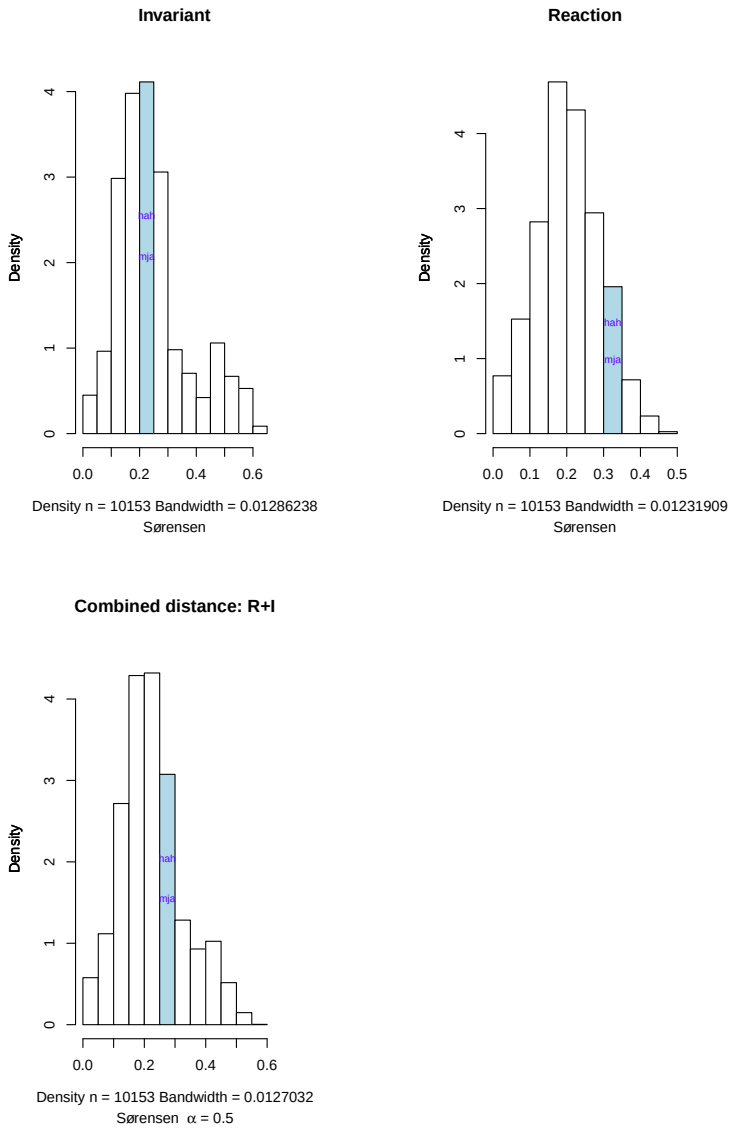


Figura A.11.1: Coppia: (mja, hah)

A.12 La coppia *(ecl,kva)* nella classe *Bacteria*

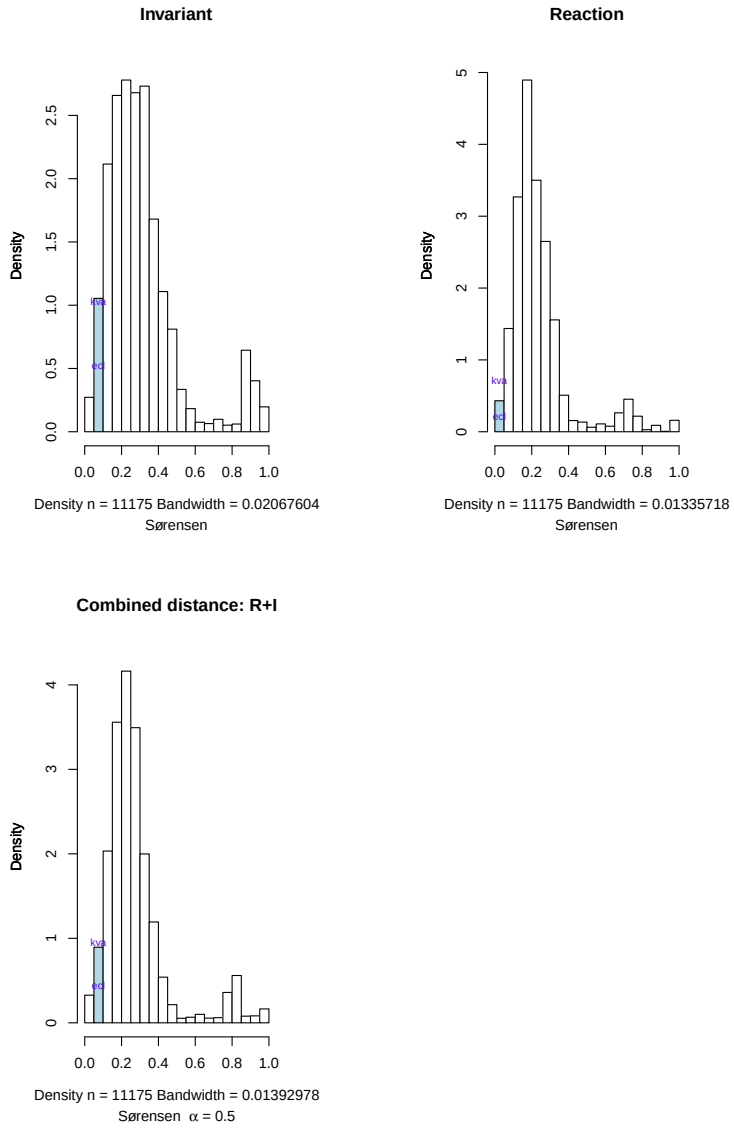


Figura A.12.1: Coppia: (ecl, kva)

A.13 Indice z_{score}

A.13.1 L'organismo *hsa* rispetto alla classe *Vertebrates*

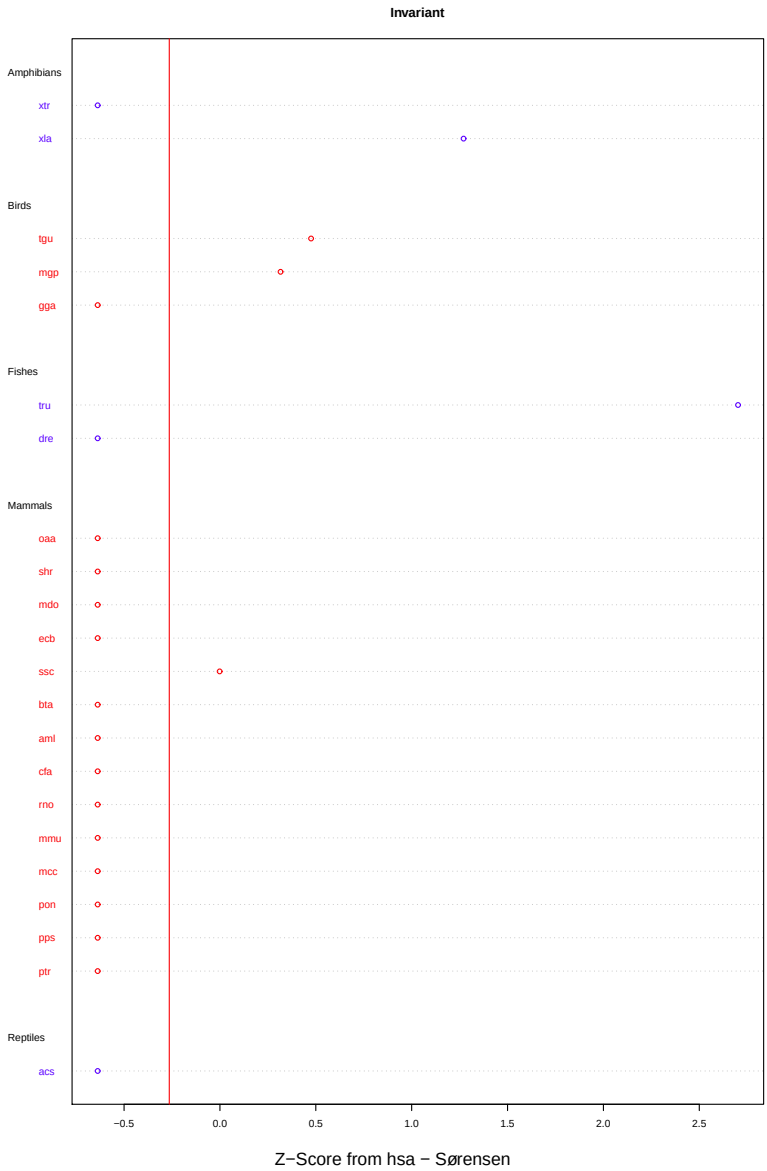


Figura A.13.1: z_{score} : I l' organismo *hsa* nella classe *Vertebrates*

A.13.2 L'organismo *hsa* rispetto alla classe *Vertebrates*

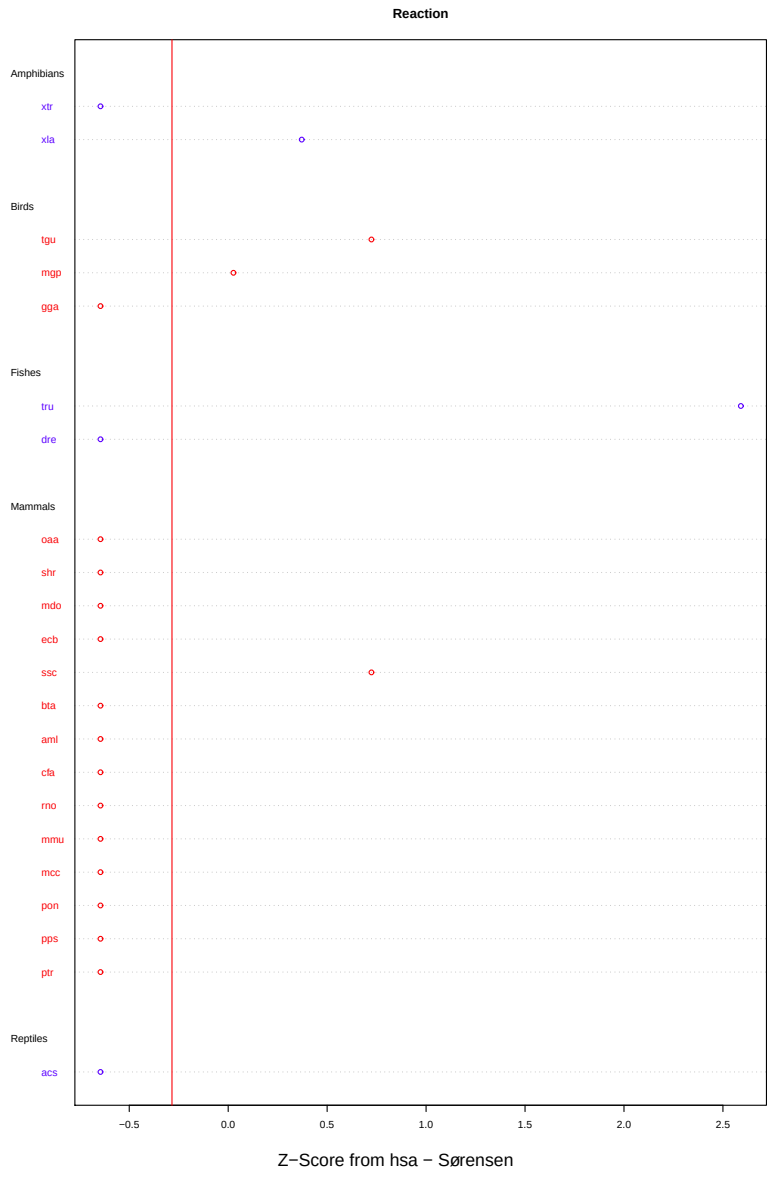


Figura A.13.2: z_{score} : R l' organismo *hsa* nella classe *Vertebrates*

A.13.3 L'organismo *hsa* rispetto alla classe *Animals*

Figura A.13.3: z_{score} : I l' organismo *hsa* nella classe *Animals*

A.13.4 L'organismo *hsa* rispetto alla classe *Animals*

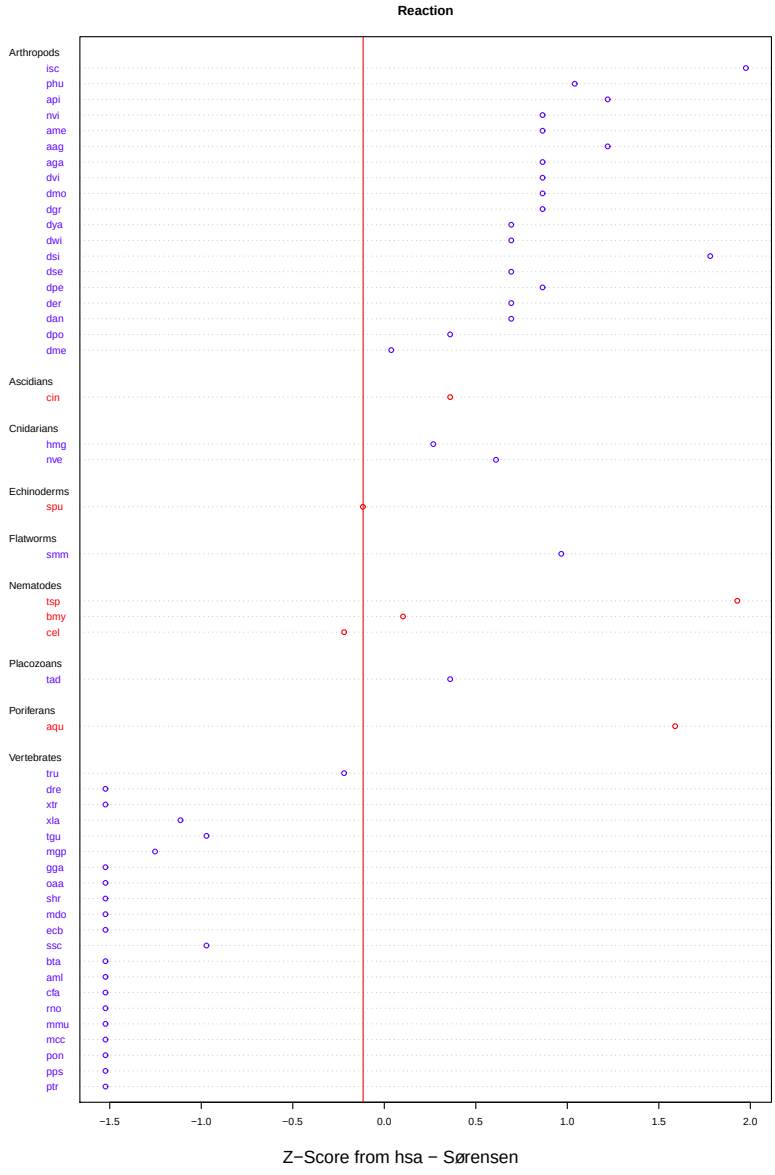


Figura A.13.4: z_{score} : R l' organismo *hsa* nella classe *Animals*

A.13.5 L'organismo *hsa* rispetto alla classe *Mammals*

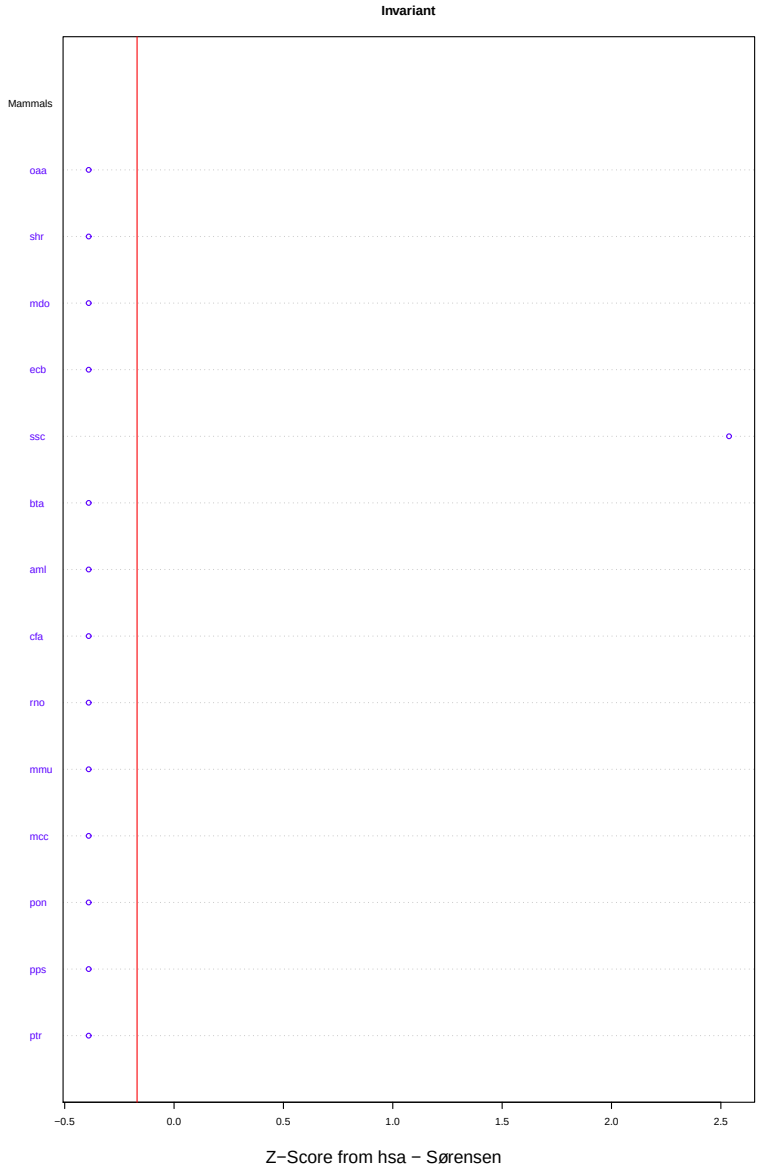


Figura A.13.5: z_{score} : I l' organismo *hsa* nella classe *Mammals*

A.13.6 L'organismo *hsa* rispetto alla classe *Mammals*

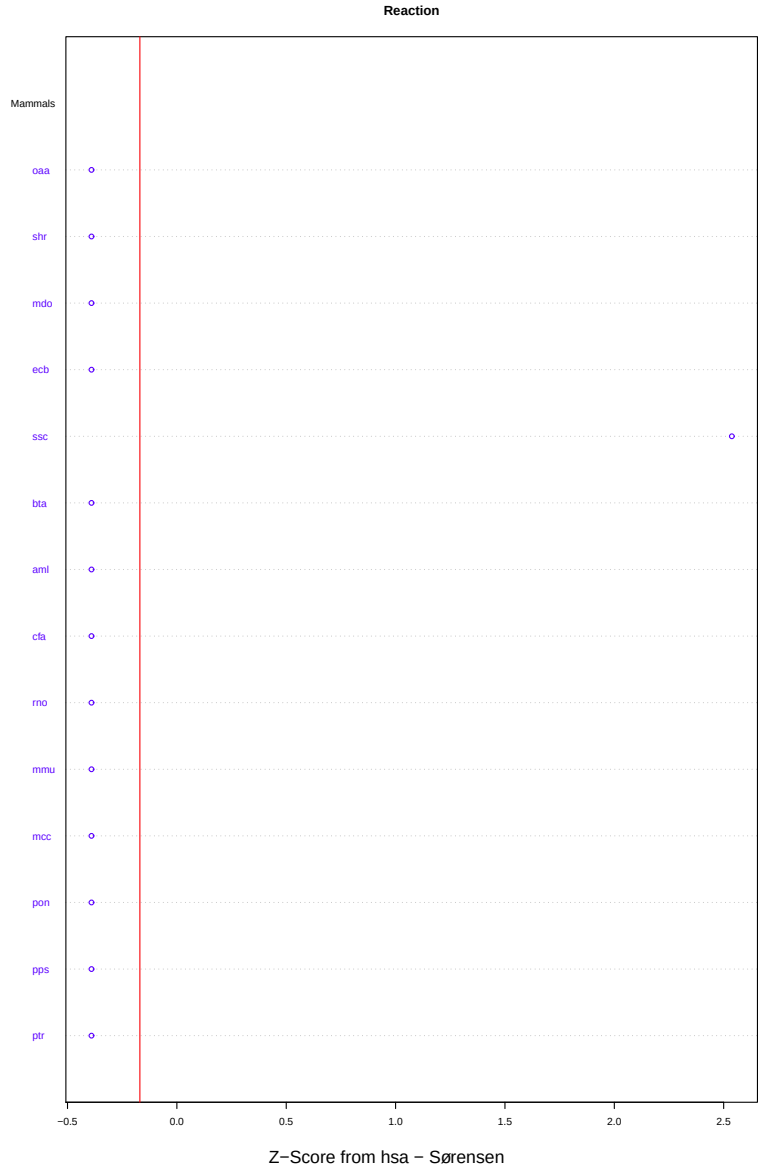


Figura A.13.6: z_{score} : R l' organismo *hsa* nella classe *Mammals*

Appendice B

Analisi della glicolisi in KEGG

In questa sperimentazione sono riportati i risultati ottenuti da alcune prove eseguite con il tool RCoMeta. Sono state analizzate 10 classi di organismi del database KEGG. Gli organismi sono classificati secondo la tassonomia di riferimento di NCBI [16]. Le distanze provengono da elaborazioni con il tool CoMeta. Lo scopo è quello di esplorare una specifica via metabolica in classi di organismi diversi. Per ciascuna classe viene analizzata: la distribuzione dei valori delle distanze di tutte le coppie di organismi appartenenti alla classe la distribuzione dei valori delle singole coppie di organismi campionando un organismo rispetto a tutti gli altri organismi presenti nella classe.
I parametri della sperimentazione sono i seguenti:

- Classi di organismi:
  - Eucaryotes;
  - Animals;
  - Vertebrates;
  - Mammals;
  - Insects;
  - Plants;
  - Fungi;
  - Protists;
  - Archaea;
  - Bacteria (campione di 150 batteri).
- Via metabolica analizzata: glicolisi;
- Indice di similarità: Tanimoto;
- Distanze
  - Distanza invarianti d_I;
  - Distanza Reazioni d_R:
  - Distanza combinata d_C con coefficiente α: 0.5.

Il risultato dell'esperimento è riportato nei seguenti grafici suddivisi per categoria.

B.1 Istogrammi delle distanze: d_I, d_R e d_C

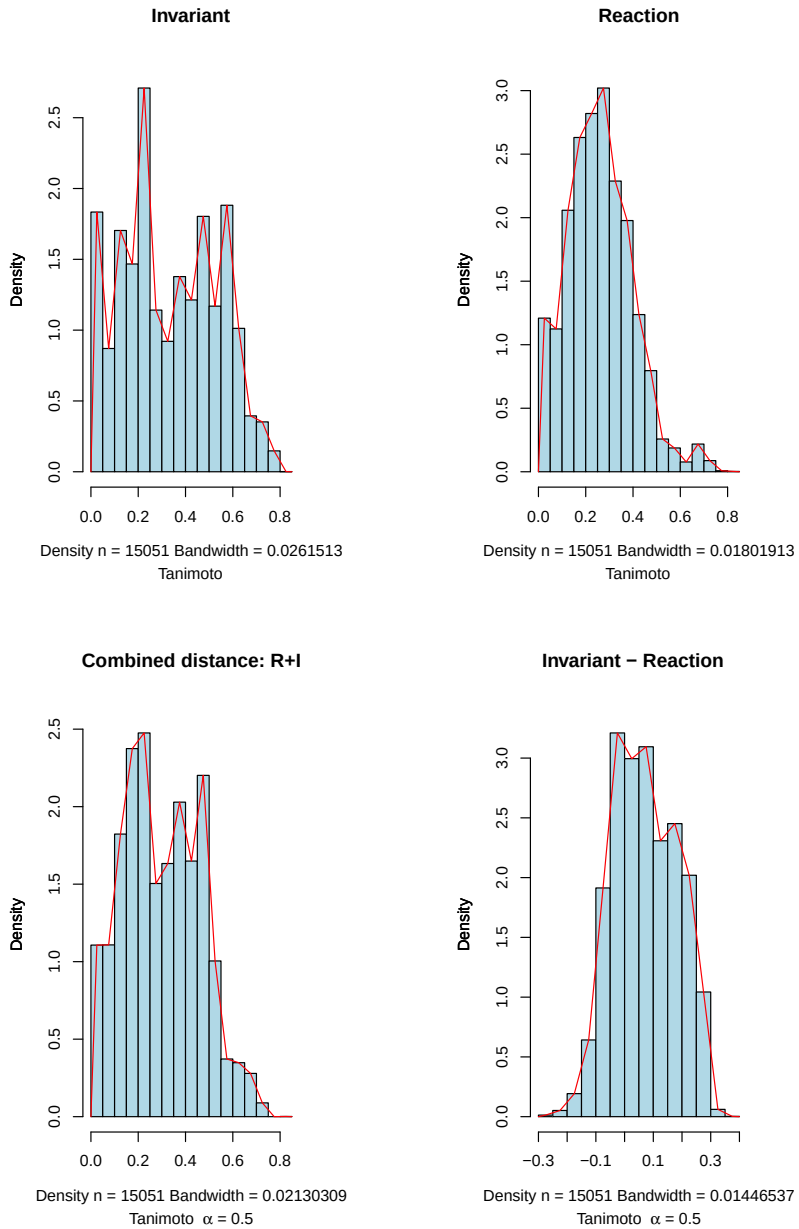


Figura B.1.1: Classe di organismi: *Eukaryotes*

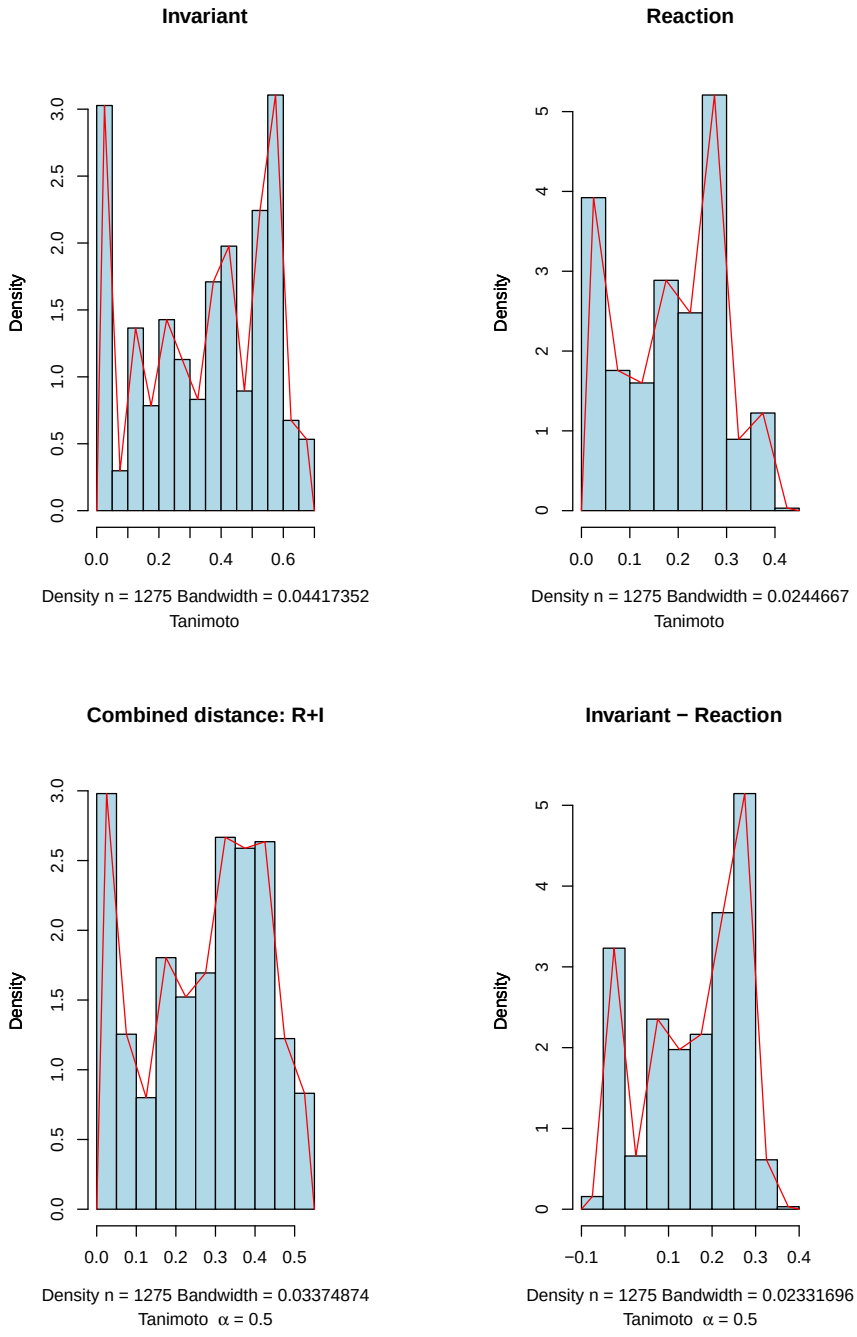


Figura B.1.2: Classe di organismi: *Animals*

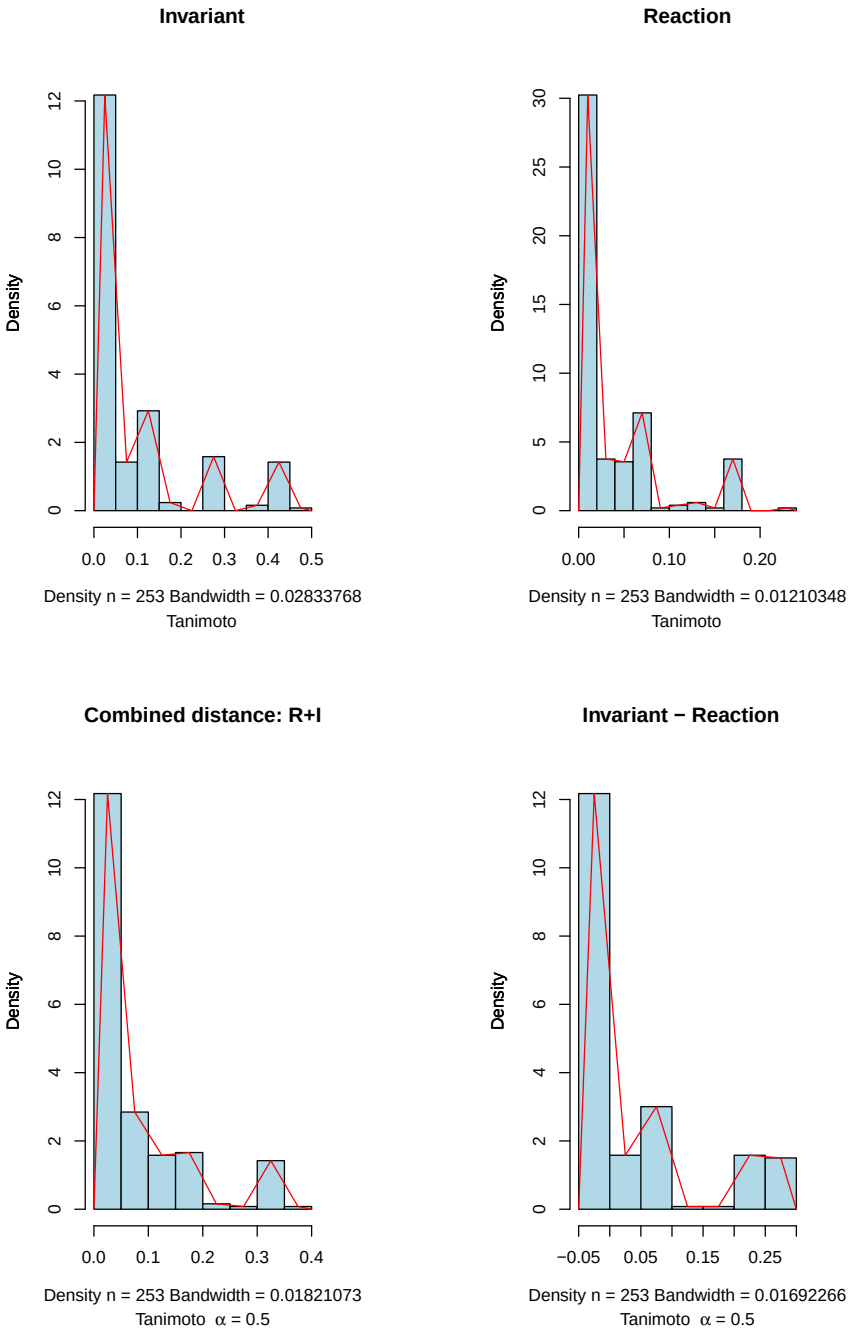


Figura B.1.3: Classe di organismi: *Vertebrates*

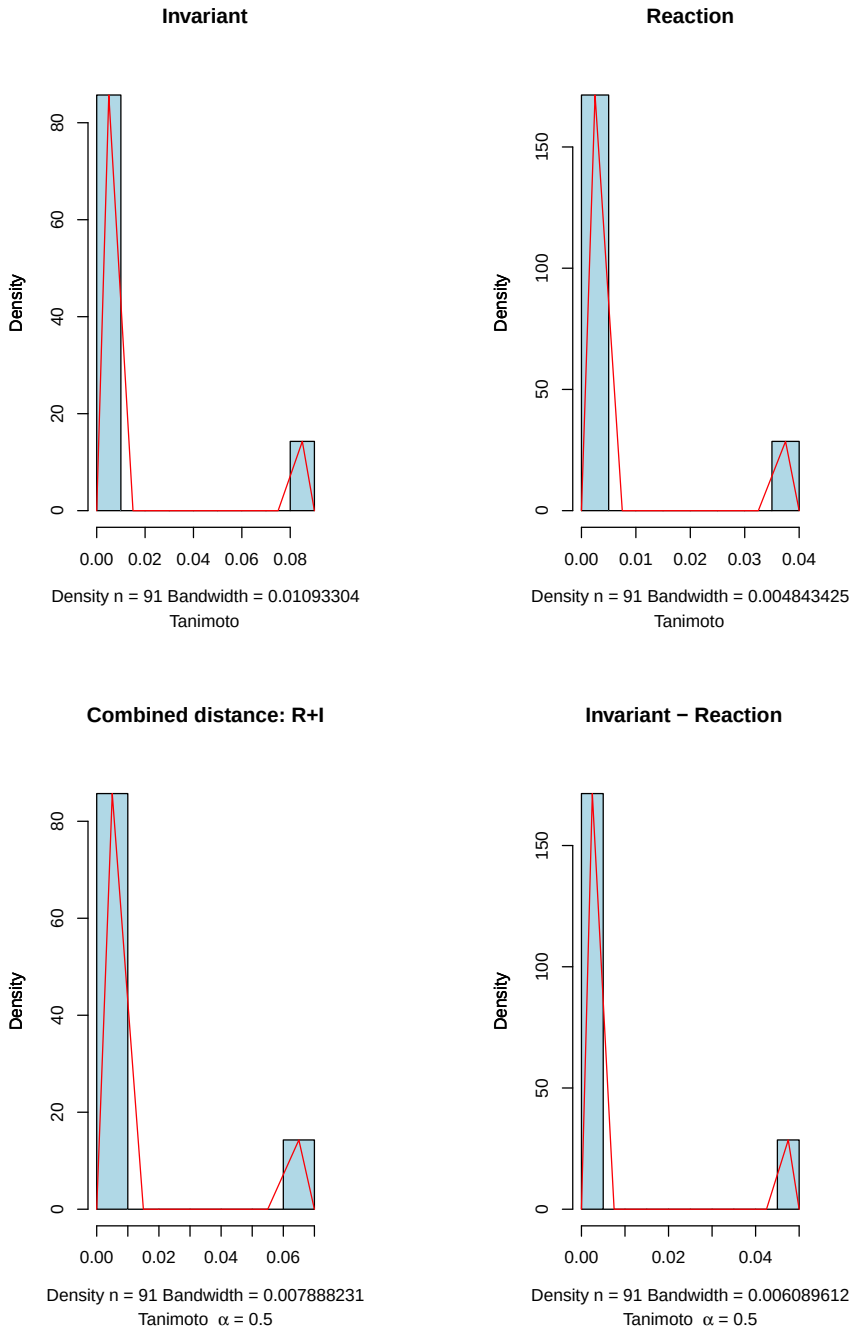


Figura B.1.4: Classe di organismi: *Mammals*

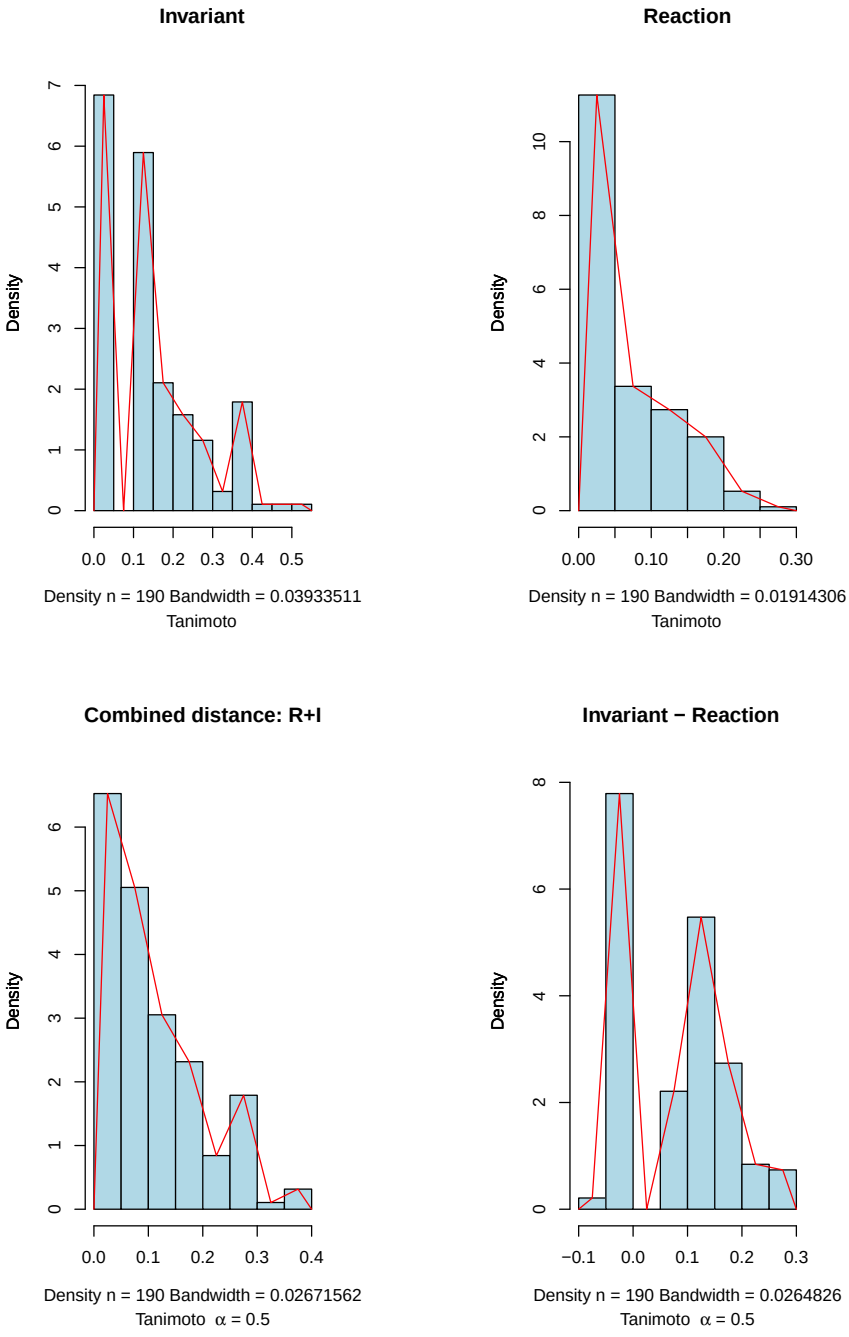


Figura B.1.5: Classe di organismi: *Insect*

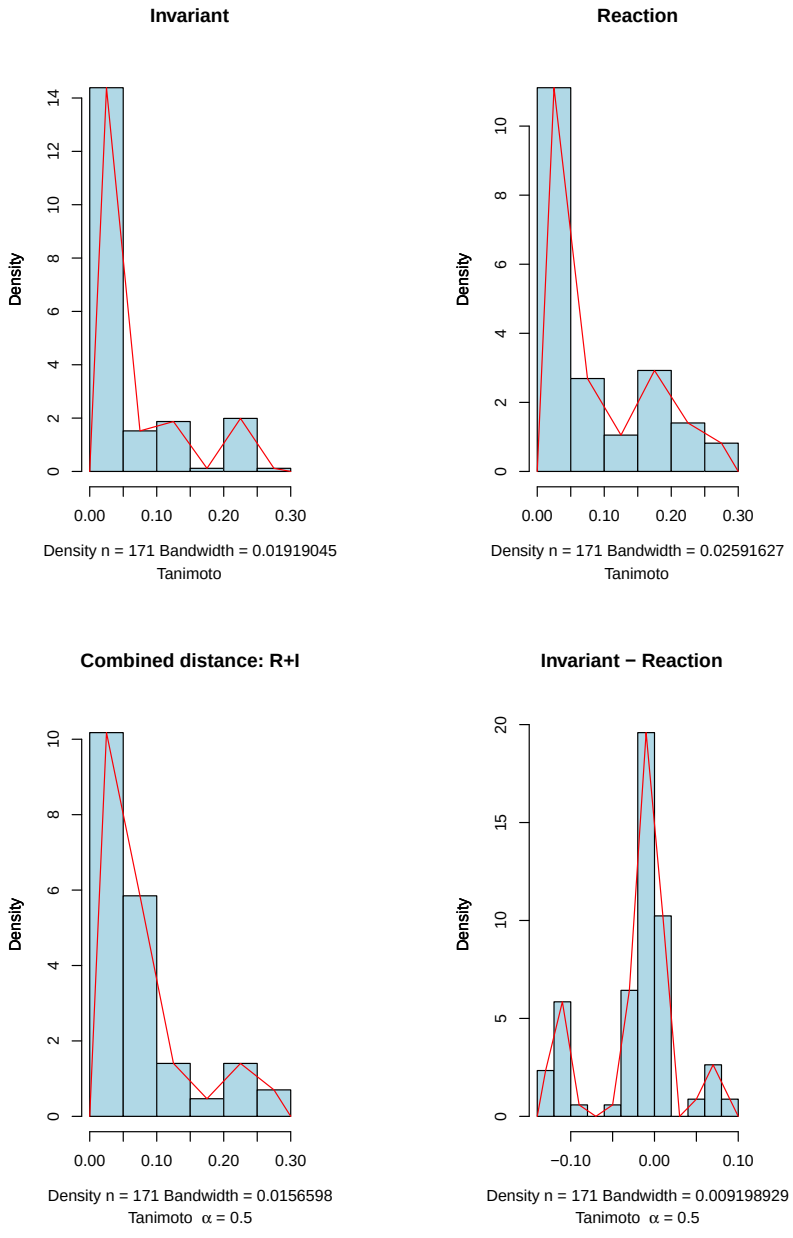


Figura B.1.6: Classe di organismi: *Plants*

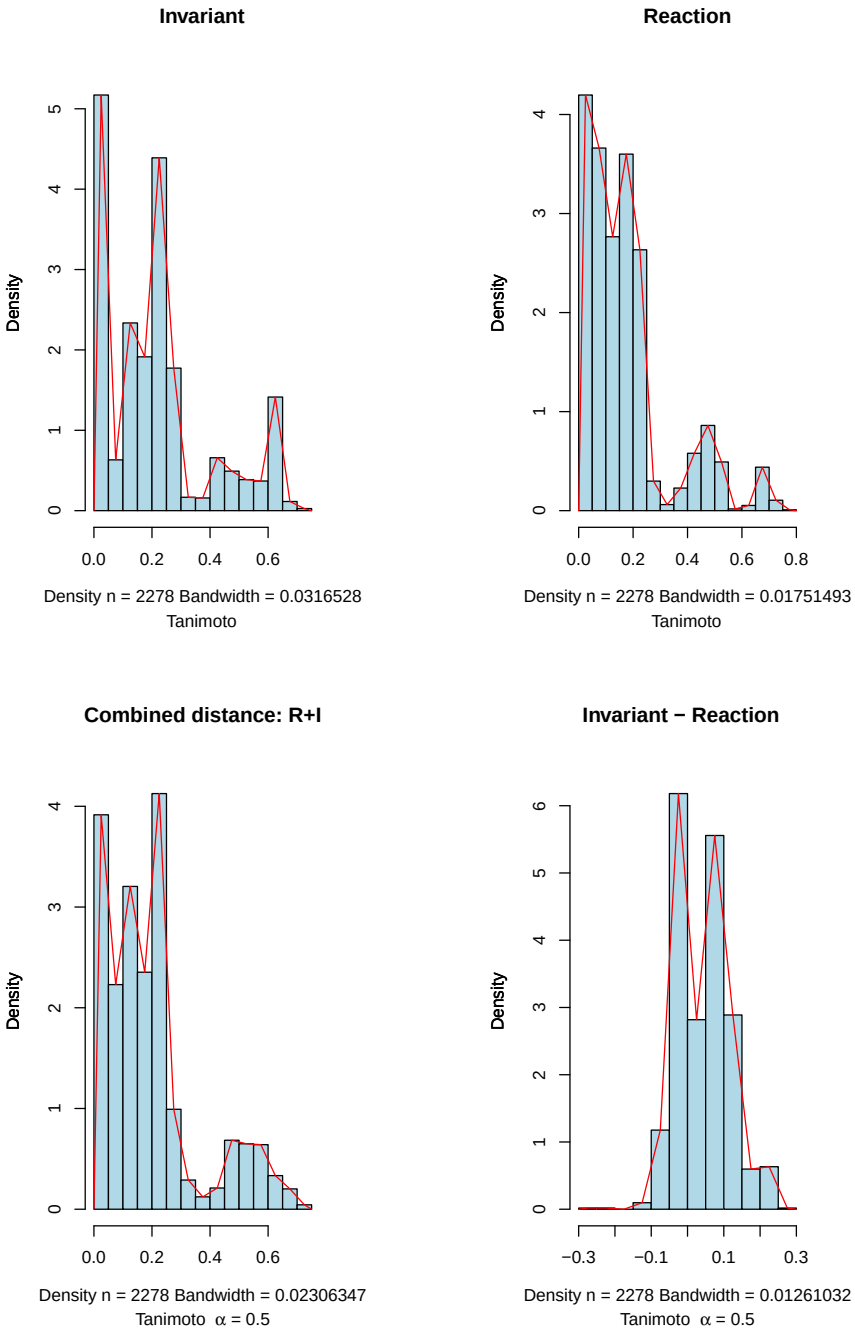


Figura B.1.7: Classe di organismi: *Fungi*

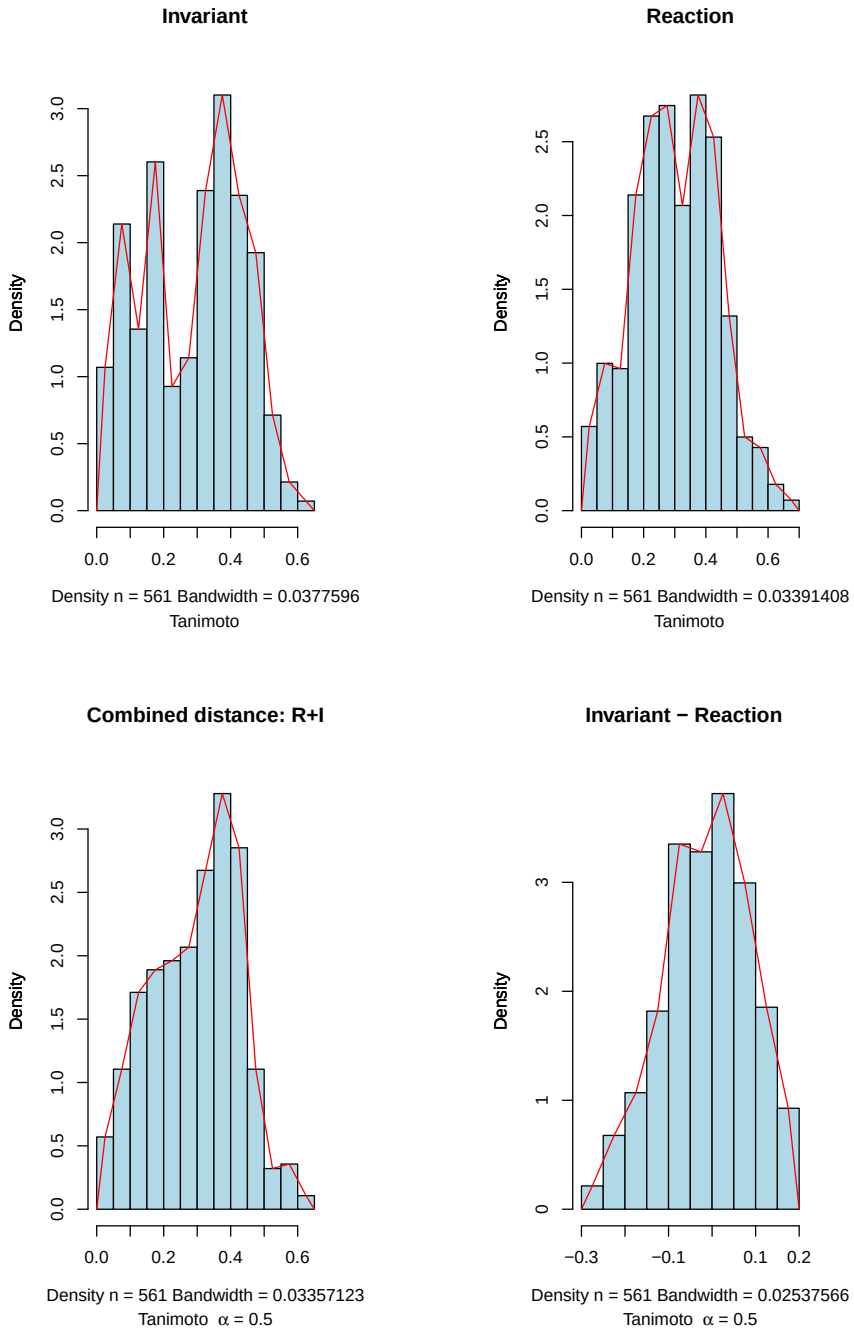


Figura B.1.8: Classe di organismi: *Protists*

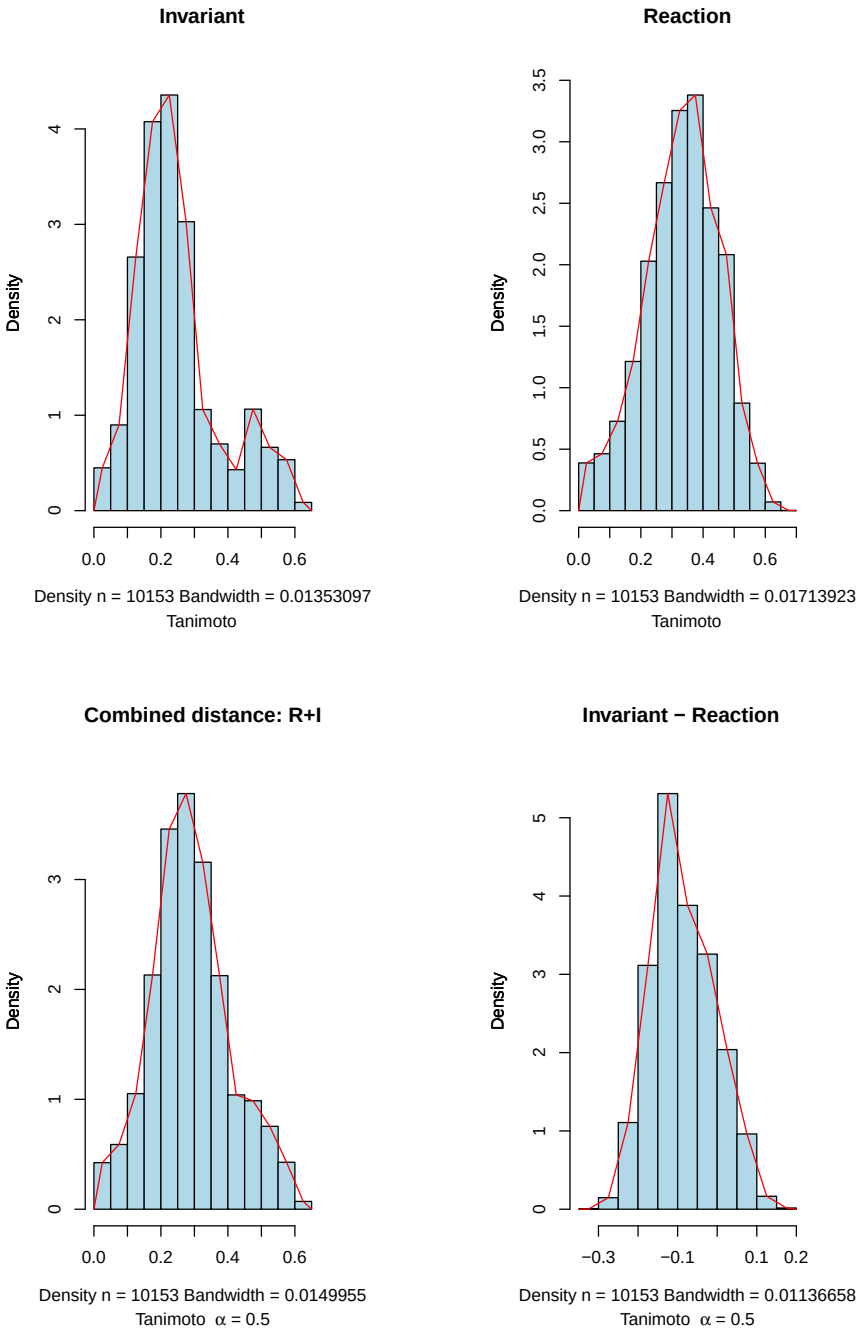


Figura B.1.9: Classe di organismi: *Archaea*

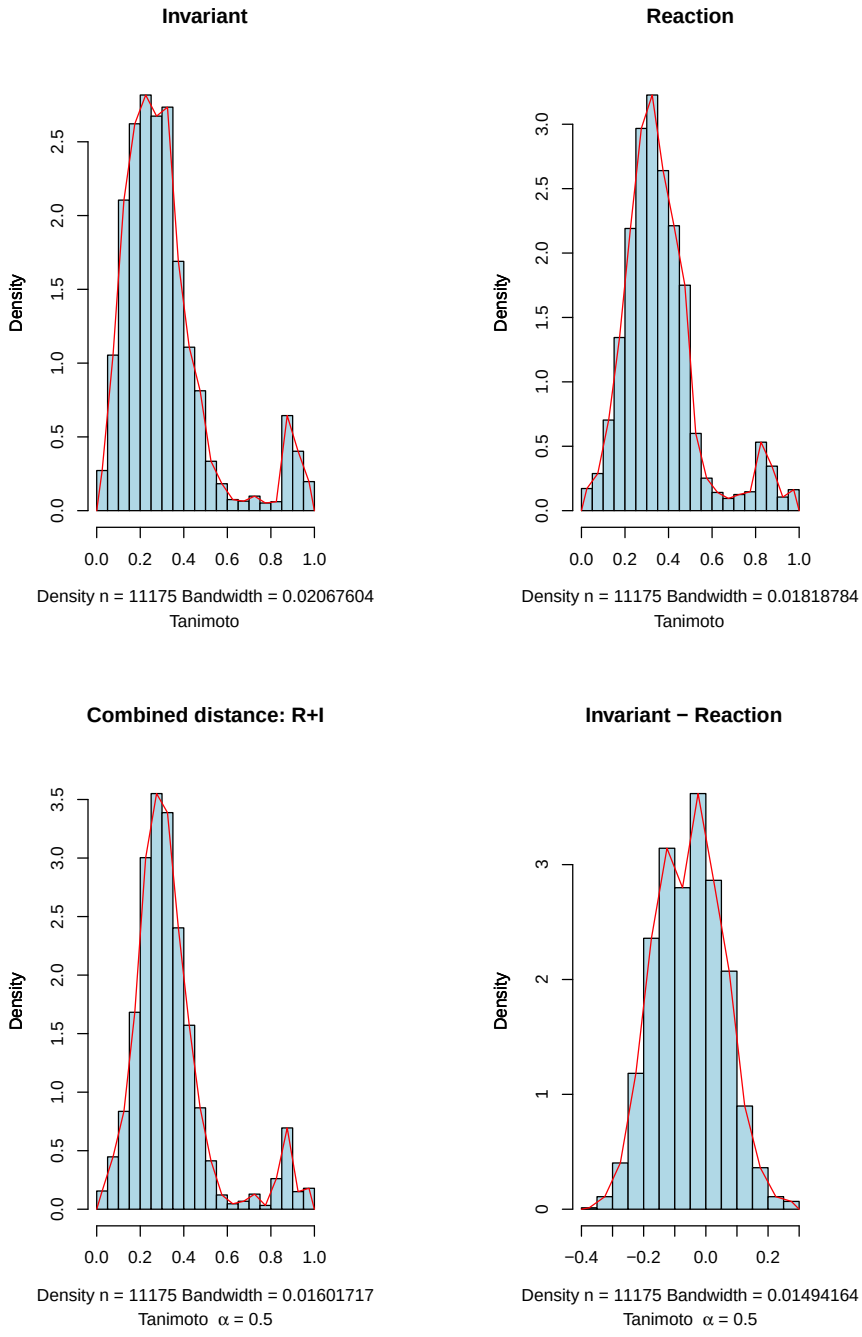


Figura B.1.10: Classe di organismi: *Bacteria*

B.2 L'organismo *hsa*.

B.2.1 L'organismo *hsa* rispetto alla classe *Eukaryotes*

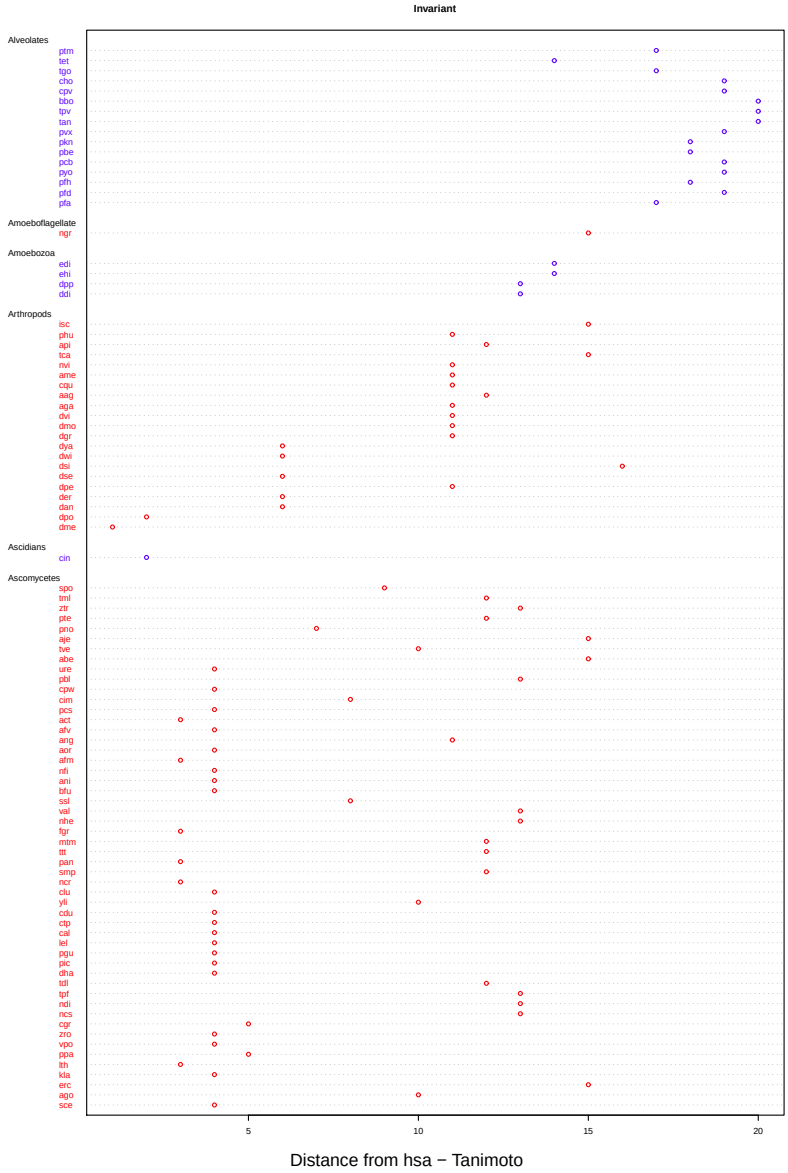


Figura B.2.1: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 1

B.2.2 L'organismo *hsa* rispetto alla classe *Eukaryotes*

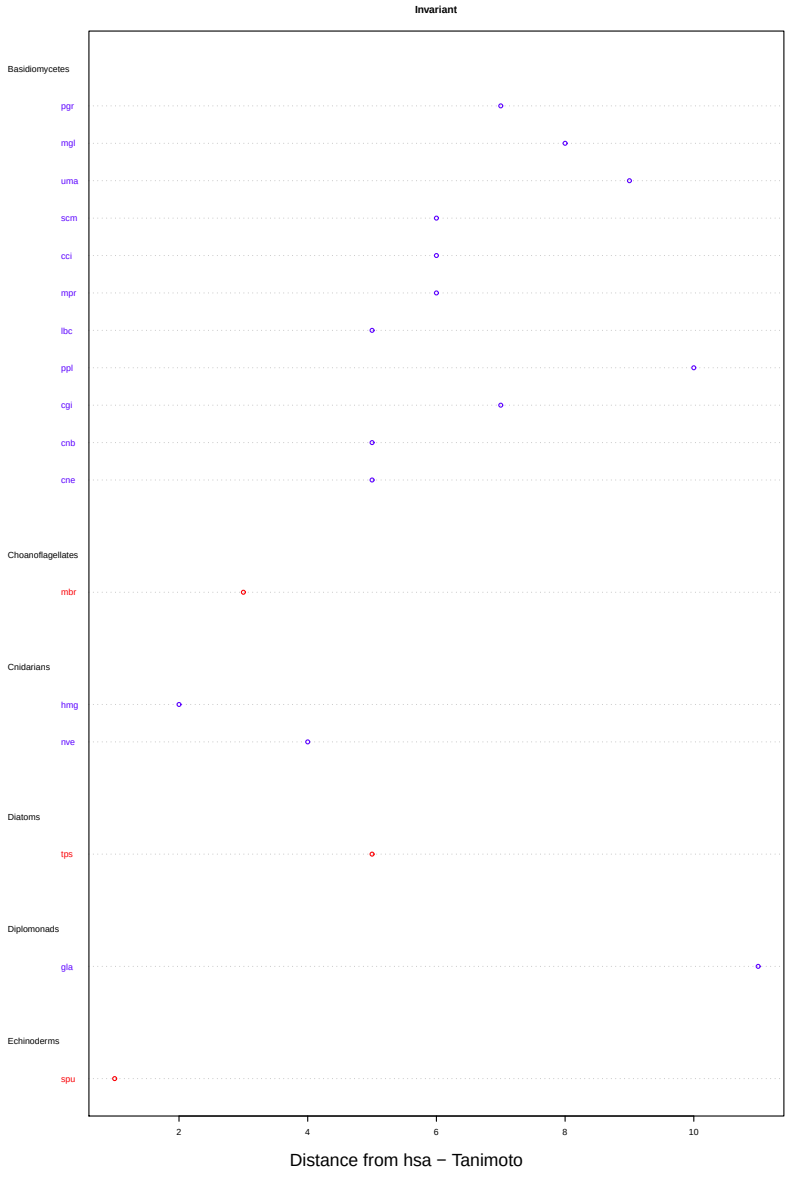


Figura B.2.2: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 2

B.2.3 L'organismo *hsa* rispetto alla classe *Eukaryotes*

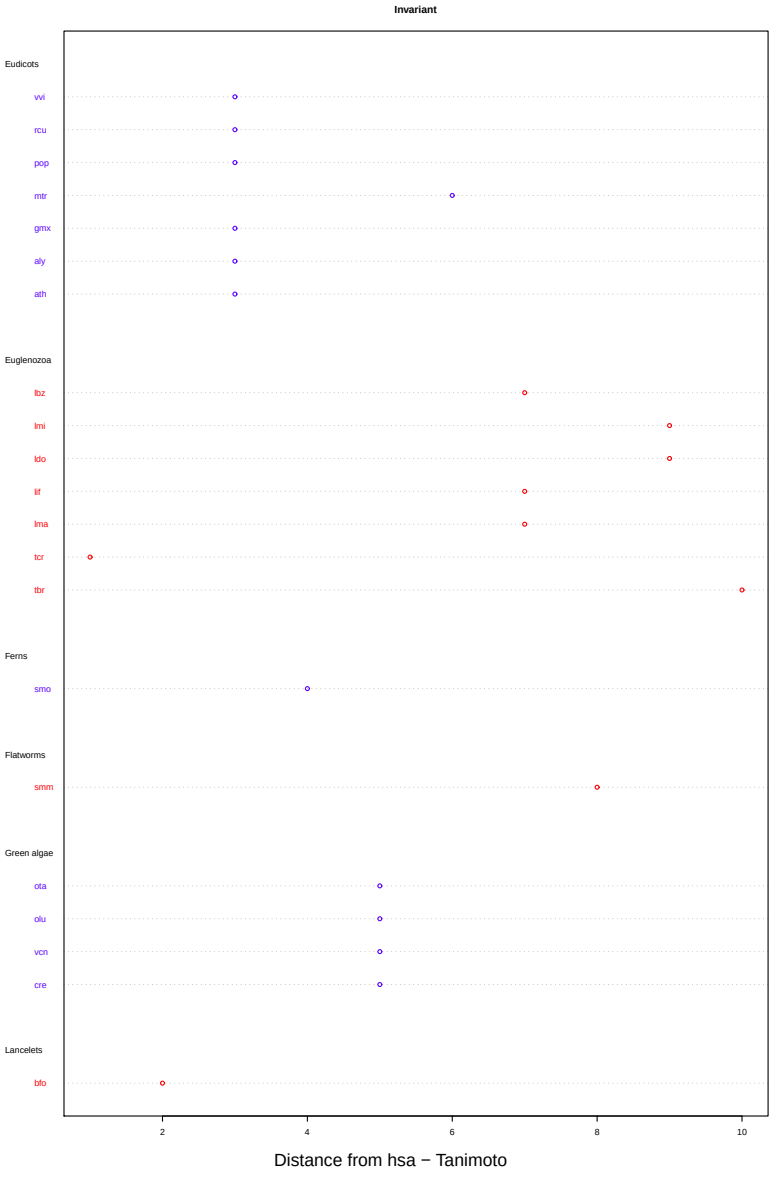


Figura B.2.3: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 3

B.2.4 L'organismo *hsa* rispetto alla classe *Eukaryotes*

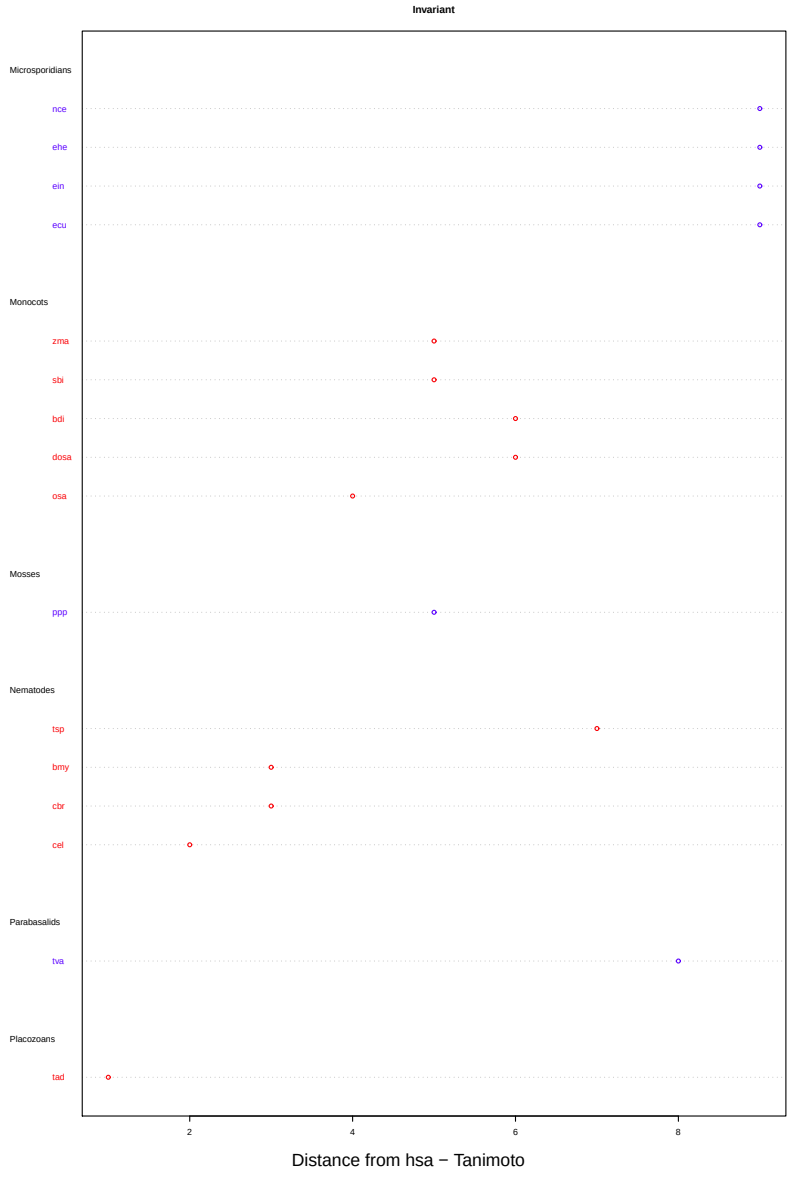


Figura B.2.4: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 4

B.2.5 L'organismo *hsa* rispetto alla classe *Eukaryotes*

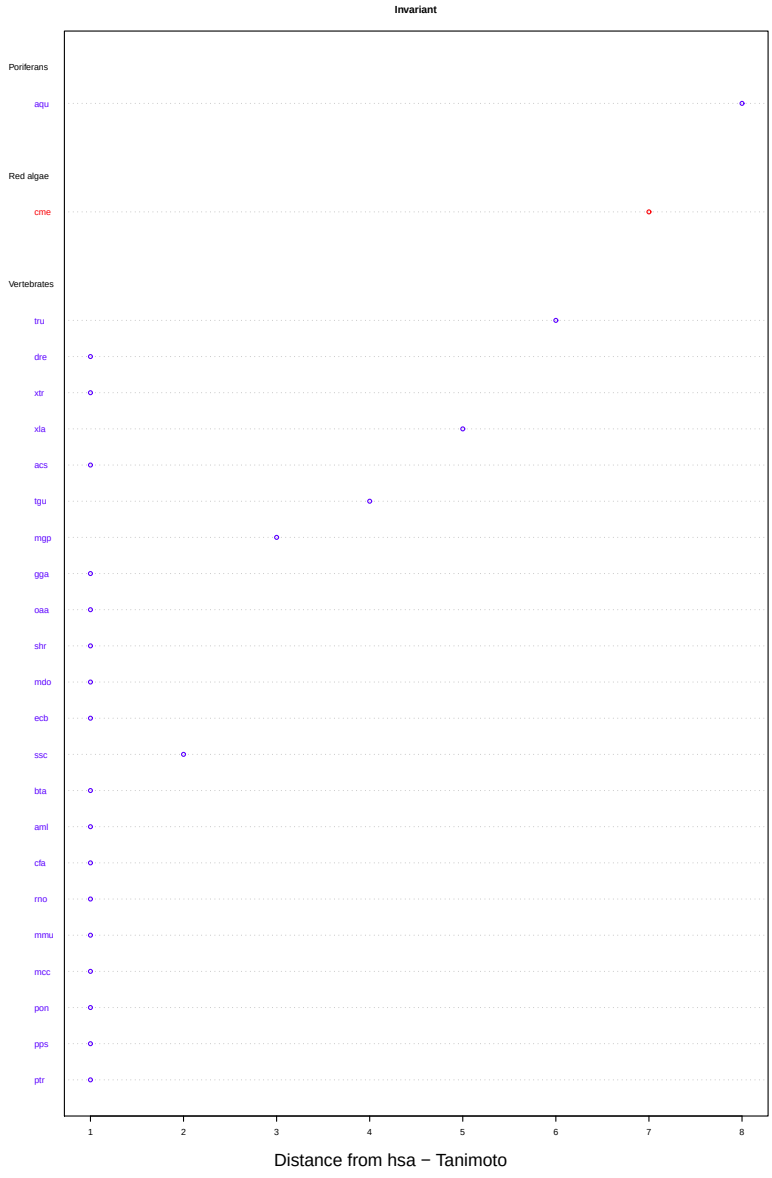


Figura B.2.5: Distanza: d_I l' organismo *hsa* nella classe *Eukaryotes* pagina 5

B.2.6 L'organismo *hsa* rispetto alla classe *Eukaryotes*

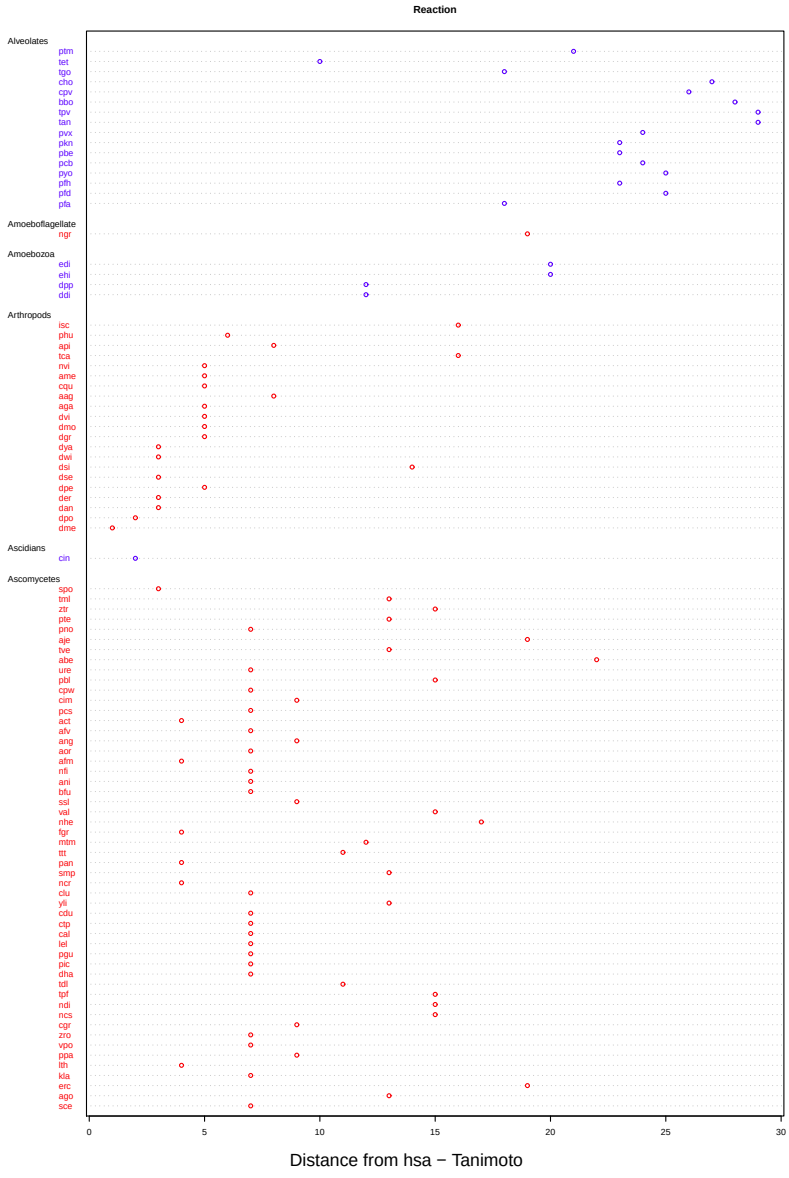


Figura B.2.6: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 1

B.2.7 L'organismo *hsa* rispetto alla classe *Eukaryotes*

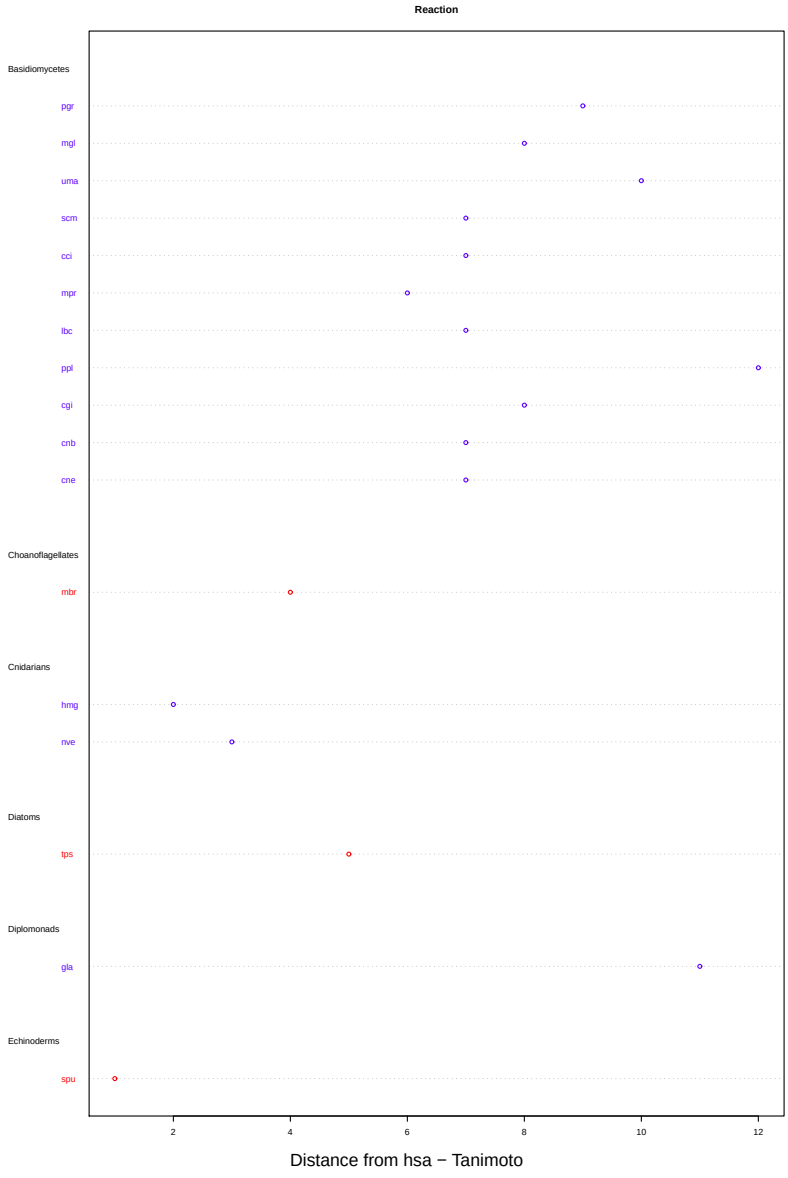


Figura B.2.7: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 2

B.2.8 L'organismo *hsa* rispetto alla classe *Eukaryotes*

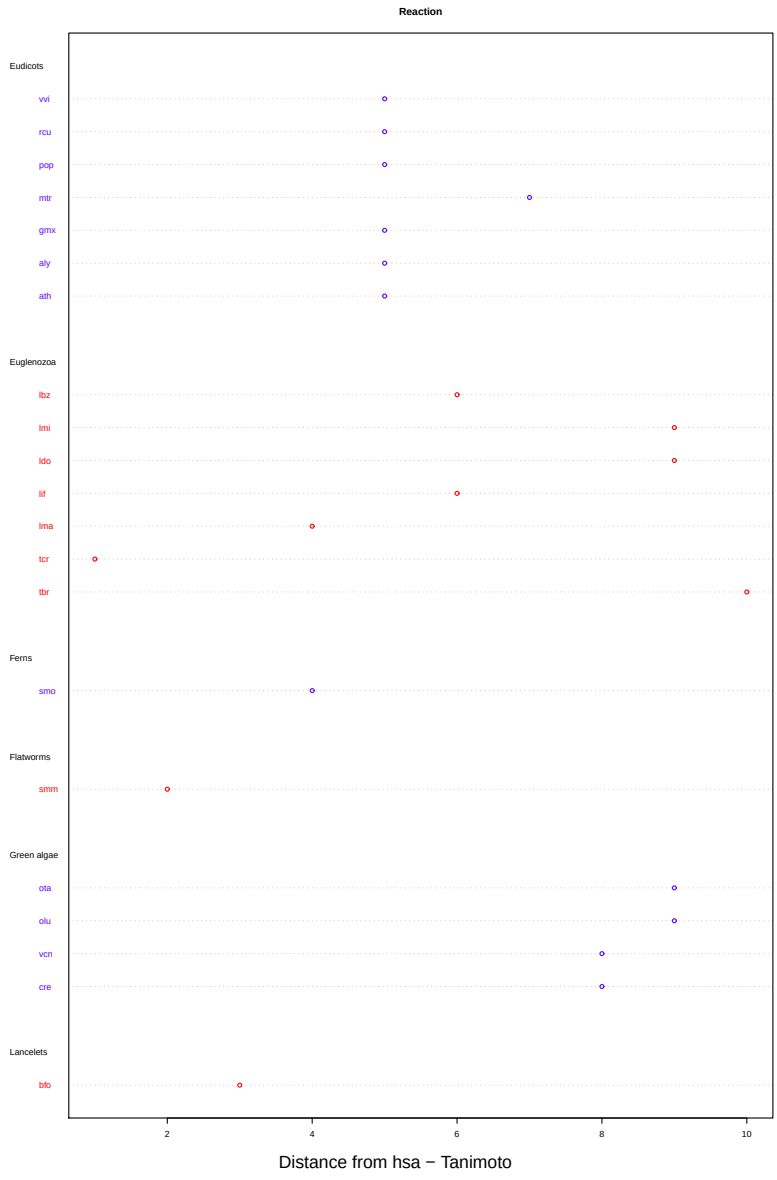


Figura B.2.8: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 3

B.2.9 L'organismo *hsa* rispetto alla classe *Eukaryotes*

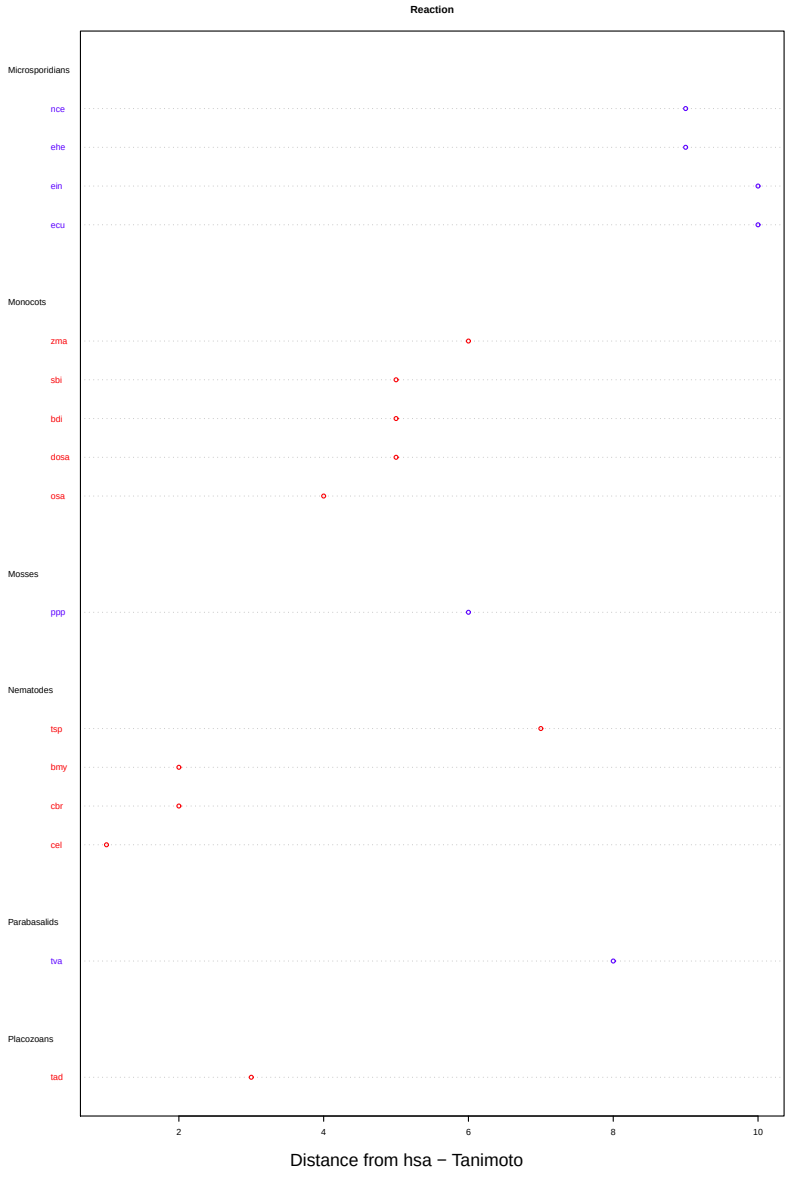


Figura B.2.9: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 4

B.2.10 L'organismo *hsa* rispetto alla classe *Eukaryotes*

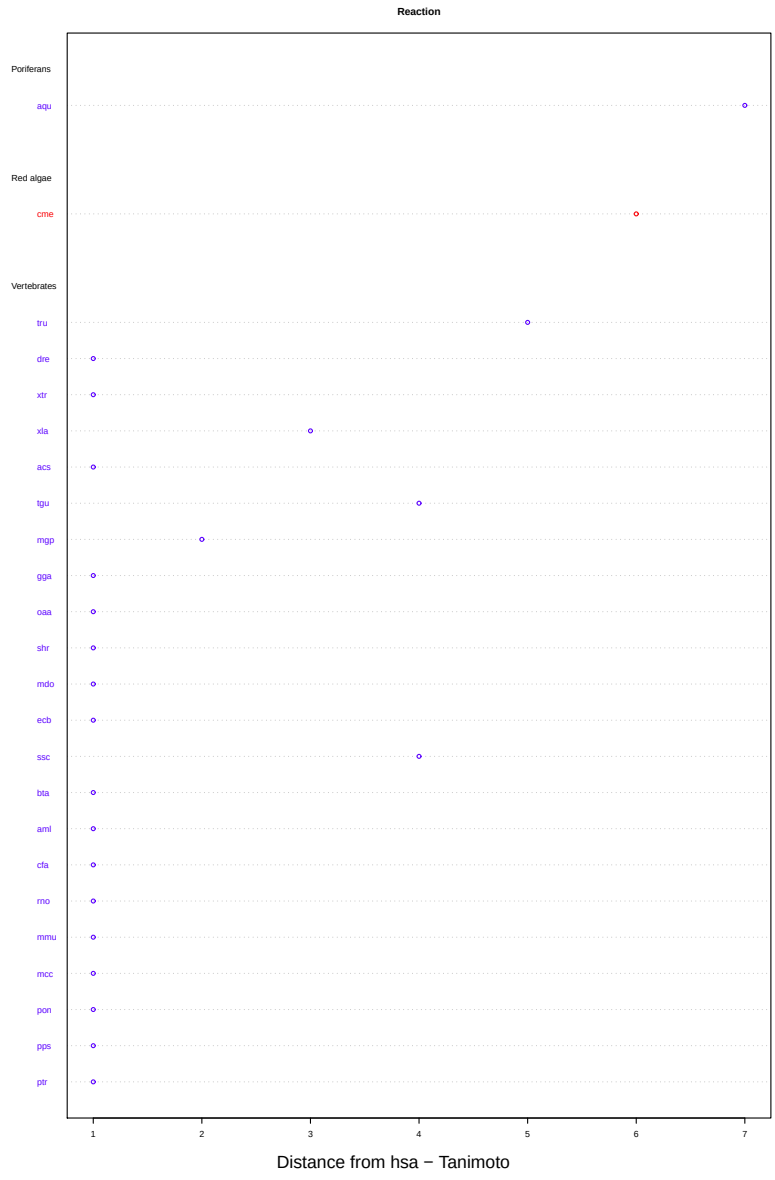


Figura B.2.10: Distanza: d_R l' organismo *hsa* nella classe *Eukaryotes* pagina 5

B.2.11 L'organismo *hsa* rispetto alla classe *Animals*

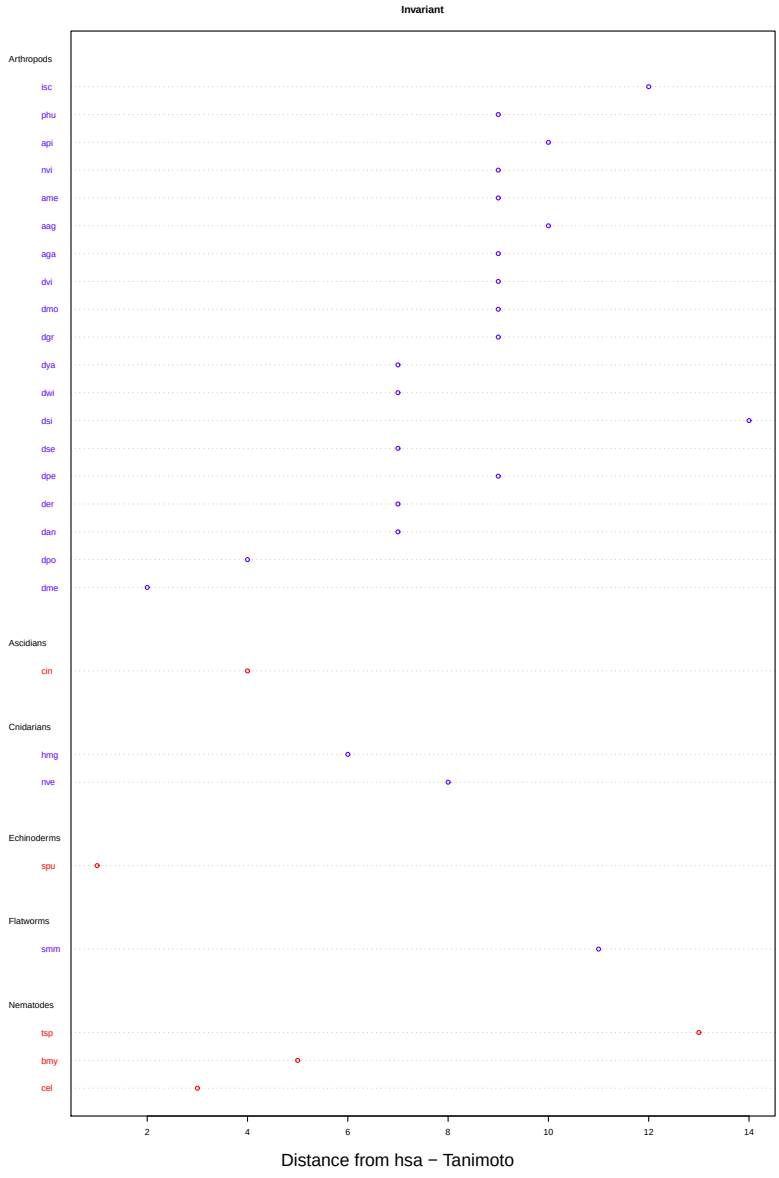


Figura B.2.11: Distanza: d_I l' organismo *hsa* nella classe *Animals*

B.2.12 L'organismo *hsa* rispetto alla classe *Animals*

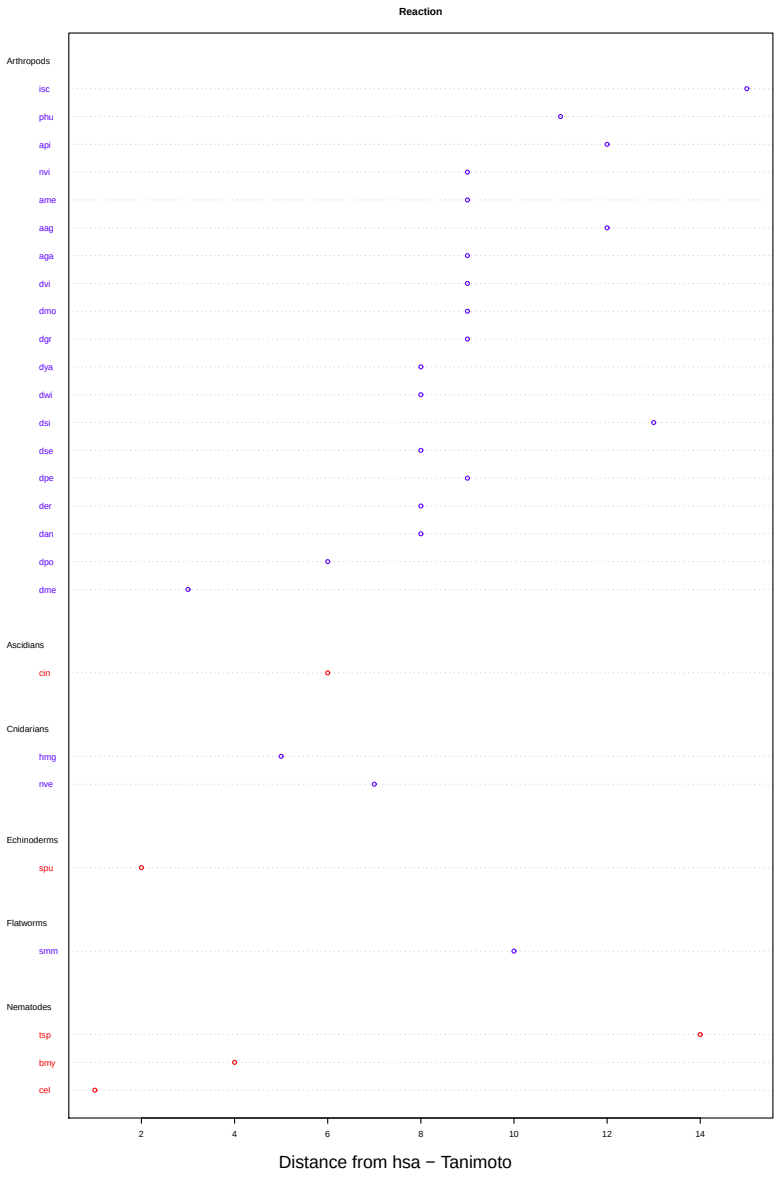


Figura B.2.12: Distanza: d_R l' organismo *hsa* nella classe *Animals*

B.2.13 L'organismo *hsa* rispetto alla classe *Vertebrates*

Figura B.2.13: Distanza: d_I l' organismo *hsa* nella classe *Vertebrates*

B.2.14 L'organismo *hsa* rispetto alla classe *Vertebrates*

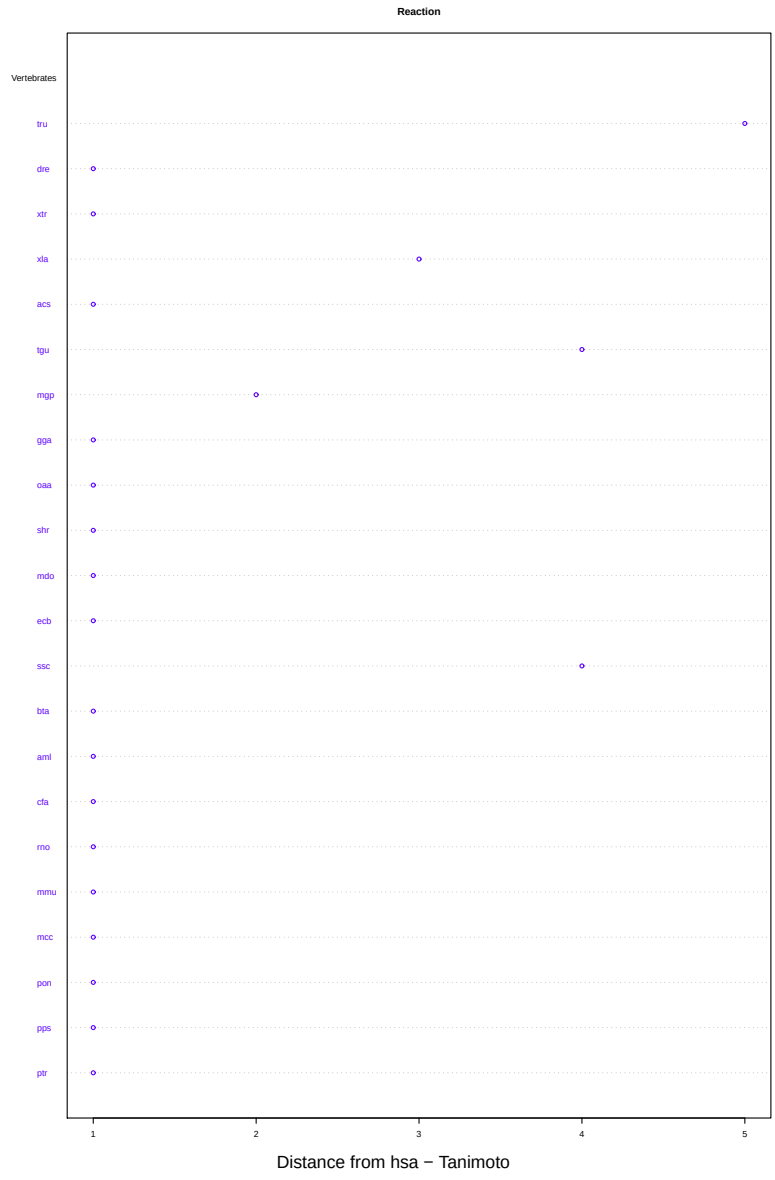


Figura B.2.14: Distanza: d_R l' organismo *hsa* nella classe *Vertebrates*

B.2.15 L'organismo *hsa* rispetto alla classe *Mammals*

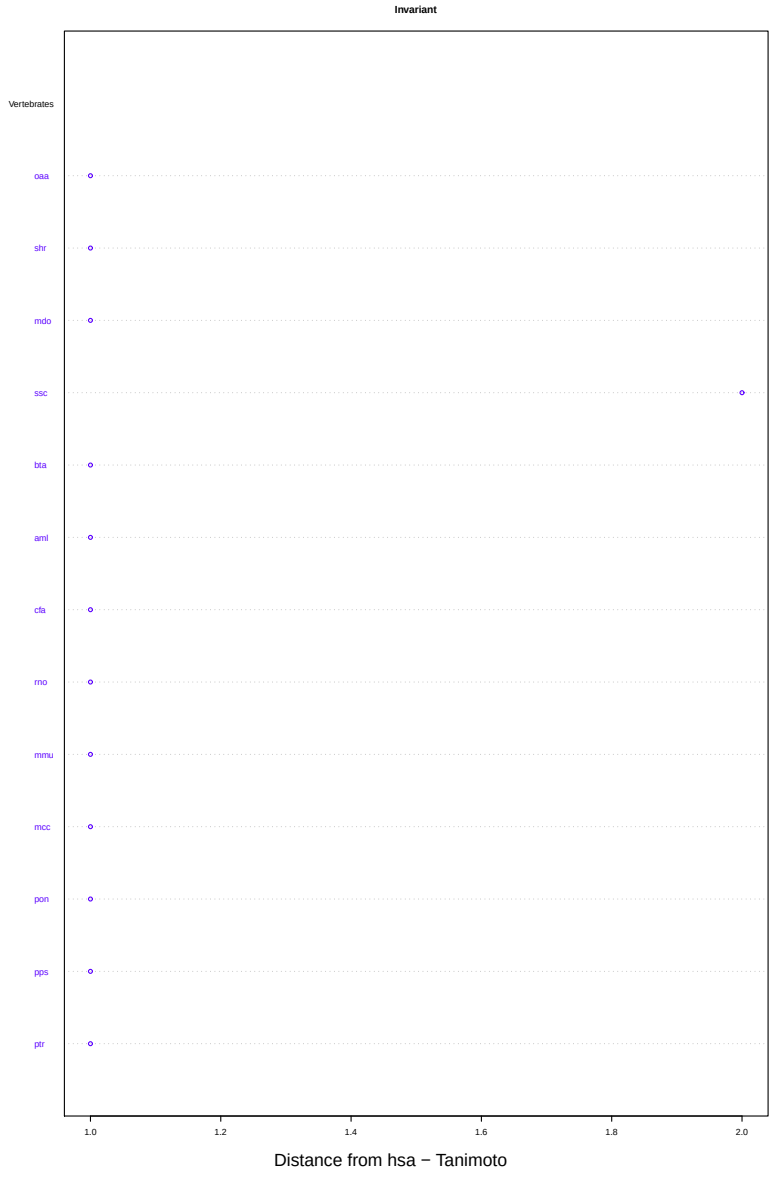


Figura B.2.15: Distanza: d_I l' organismo *hsa* nella classe *Mammals*

B.2.16 L'organismo *hsa* rispetto alla classe *Mammals*

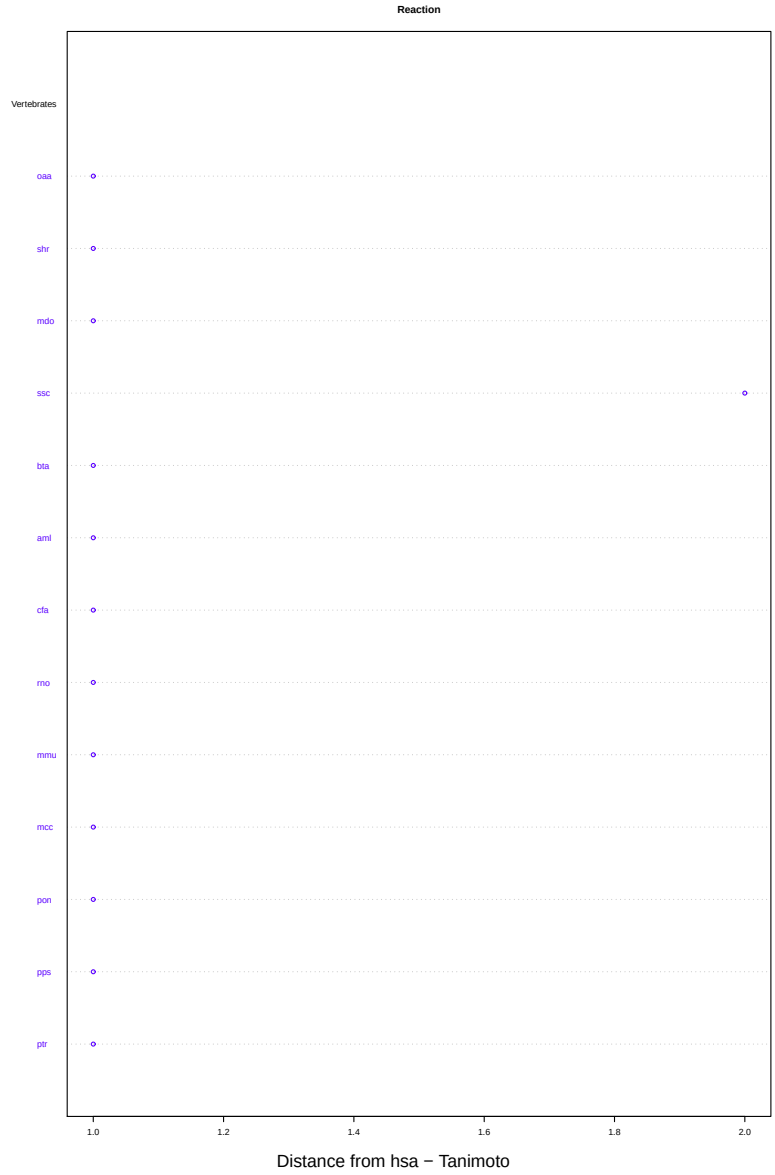


Figura B.2.16: Distanza: d_R l' organismo *hsa* nella classe *Mammals*

B.3 Indice z_{score}

B.3.1 L'organismo *hsa* rispetto alla classe *Vertebrates*

Figura B.3.1: z_{score} : I l' organismo *hsa* nella classe *Vertebrates*

B.3.2 L'organismo *hsa* rispetto alla classe *Vertebrates*

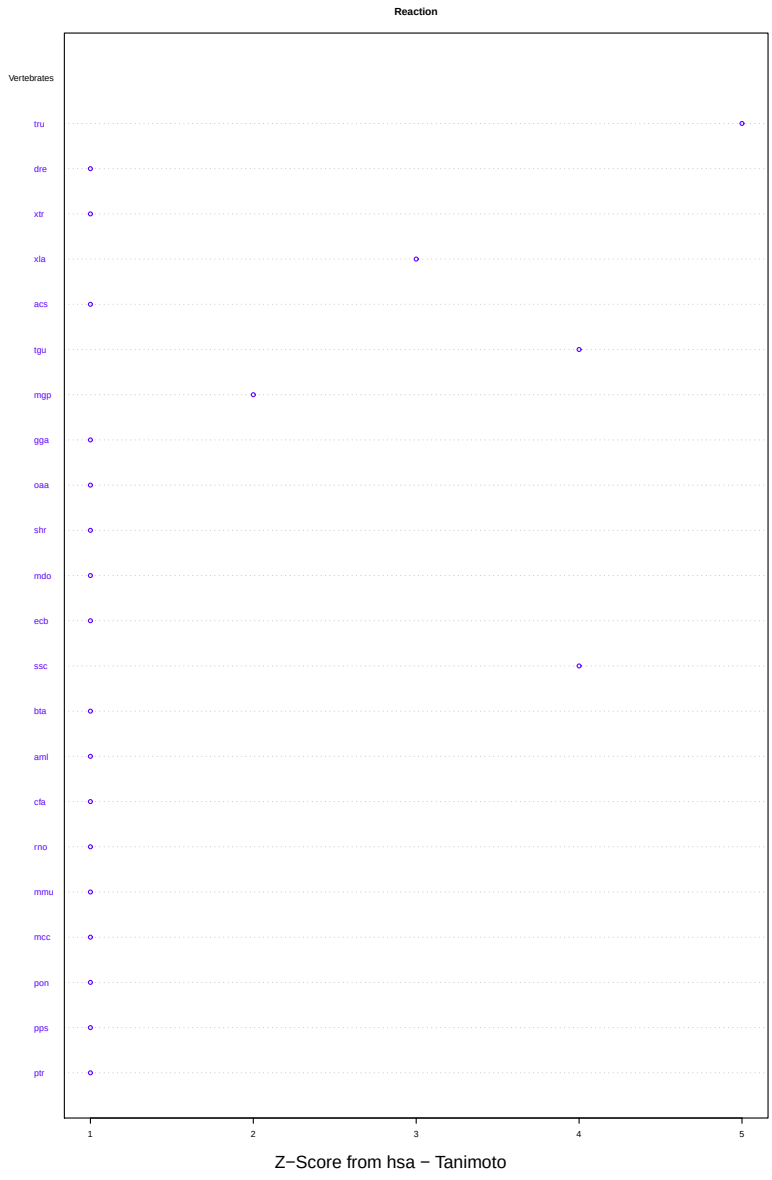


Figura B.3.2: z_{score} : *R* l' organismo *hsa* nella classe *Vertebrates*

B.3.3 L'organismo *hsa* rispetto alla classe *Animals*

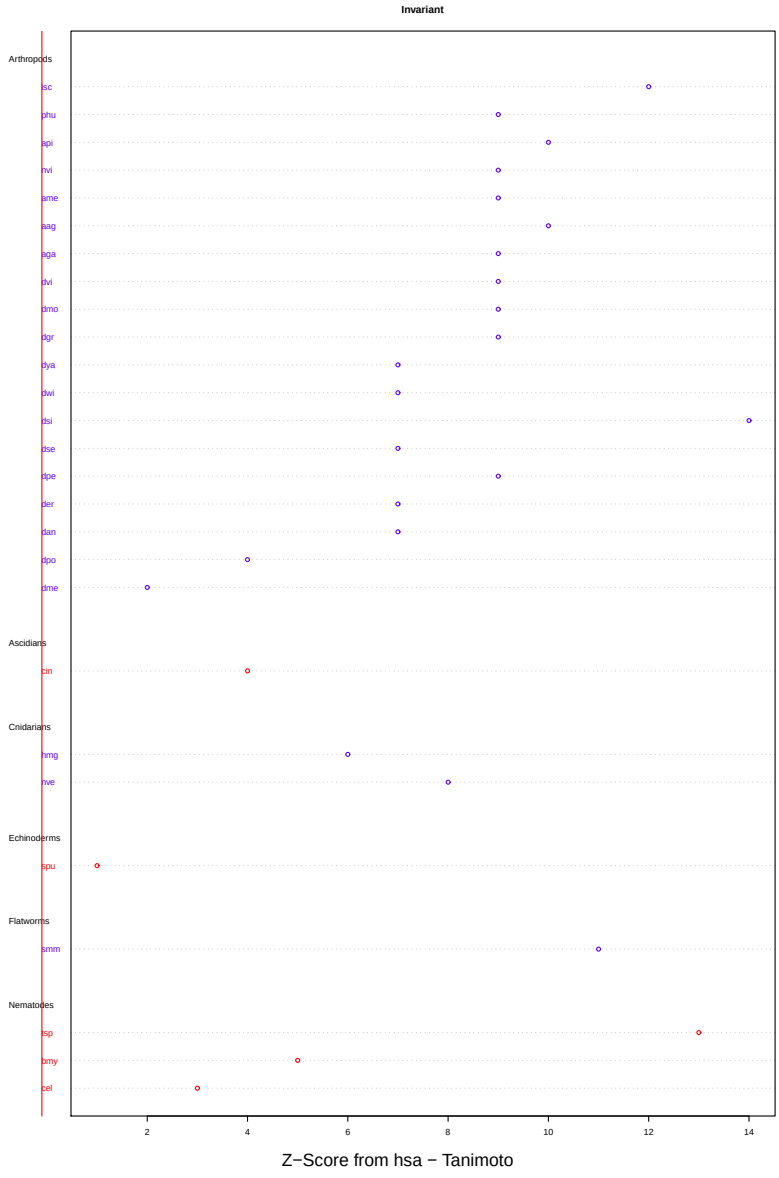


Figura B.3.3: z_{score} : *I* l' organismo *hsa* nella classe *Animals*

B.3.4 L'organismo *hsa* rispetto alla classe *Animals*

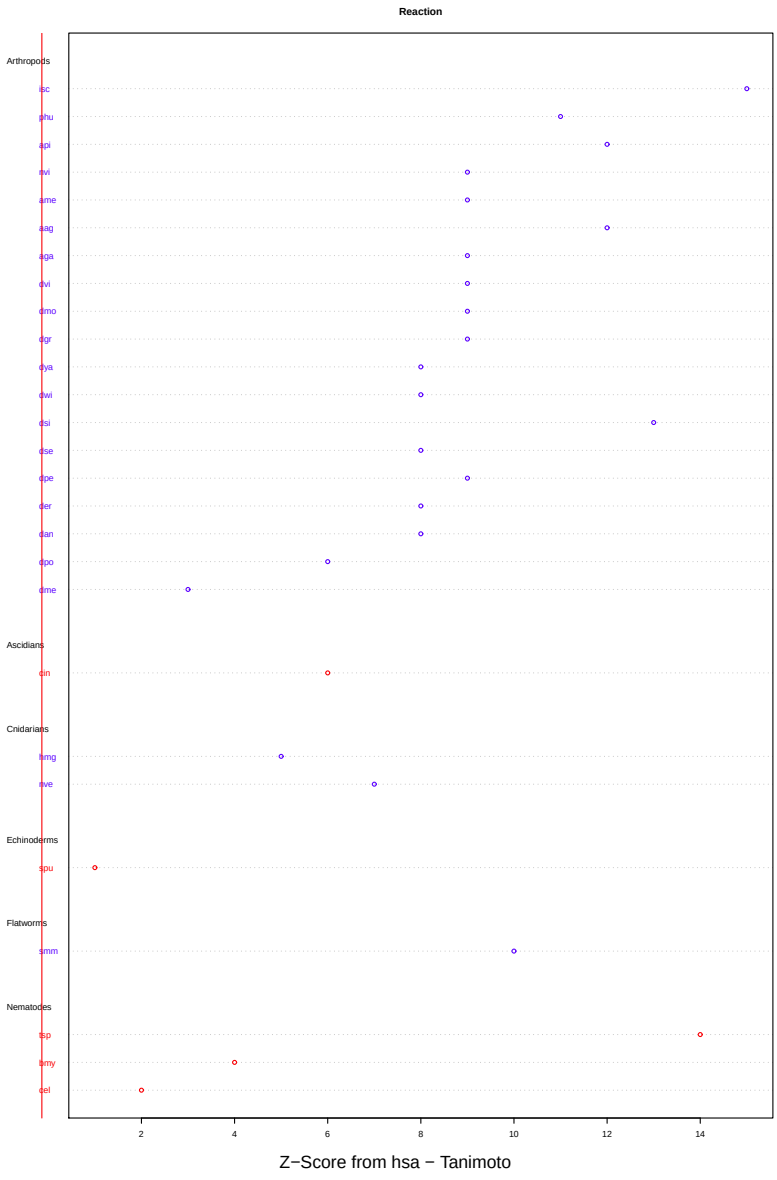


Figura B.3.4: z_{score} : *R* l' organismo *hsa* nella classe *Animals*

B.3.5 L'organismo *hsa* rispetto alla classe *Mammals*

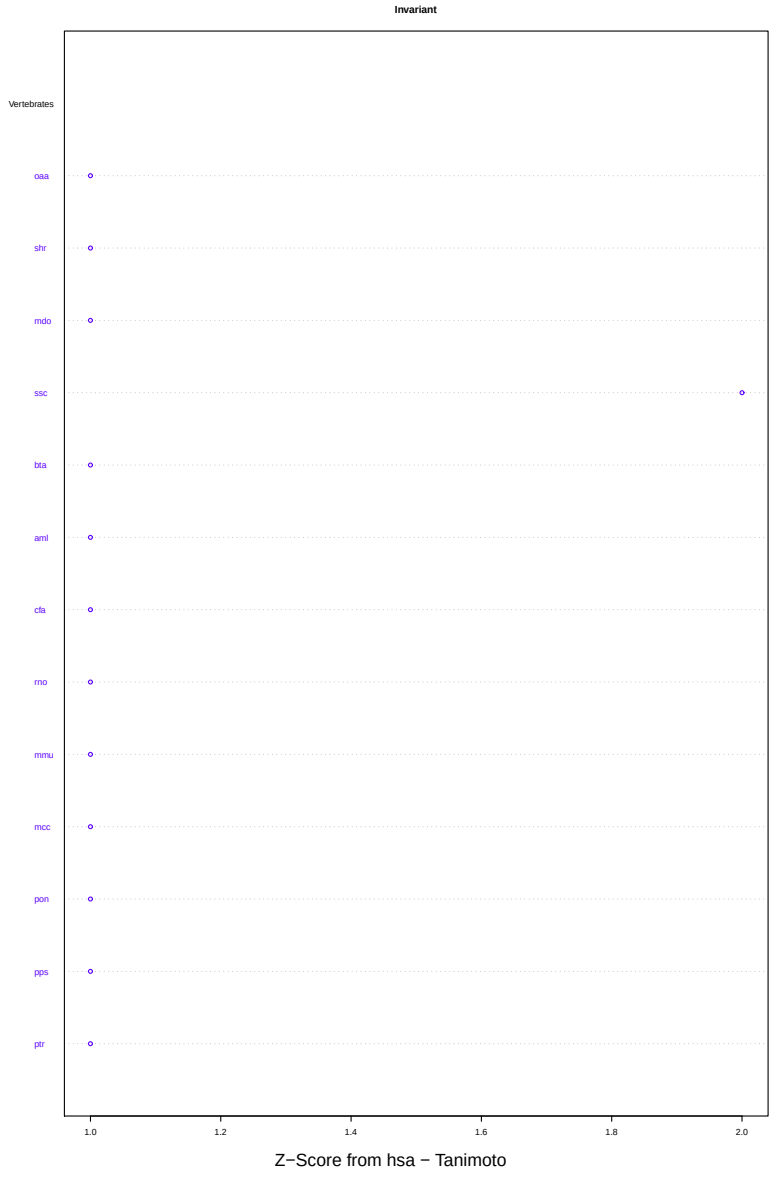
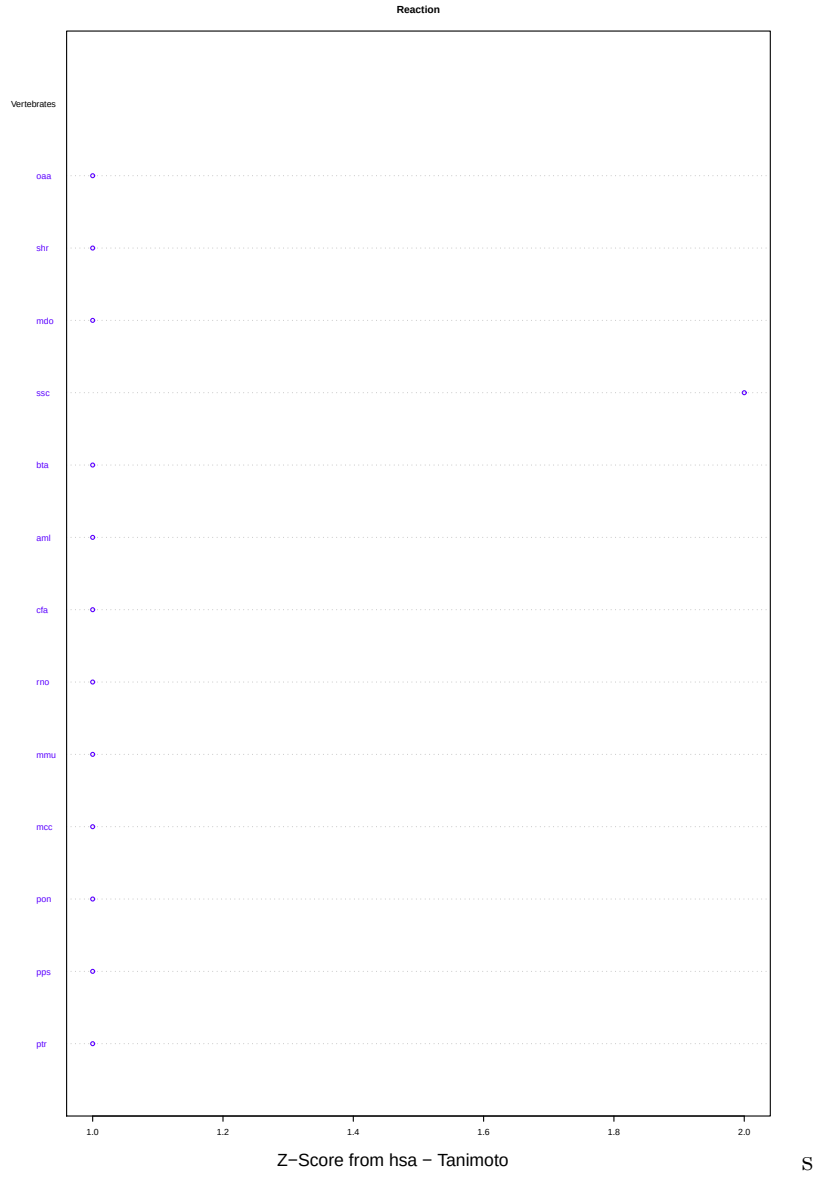


Figura B.3.5: z_{score} : I l' organismo *hsa* nella classe *Mammals*

B.3.6 L'organismo *hsa* rispetto alla classe *Mammals*



S

Figura B.3.6: z_{score} : R l' organismo *hsa* nella classe *Mammals*

Bibliografia

[1] A.Y. Mitrophanov - M. Borodovsky, M. Statistical significance in biological sequence analysis. *Briefings in Bioinformatics*, 7(1):2–24, 2006.

[2] D. E. Krane - M. L. Raymer. *Fondamenti di Bioinformatica.* Pearson, Milano, Aprile 2007.

[3] D. Sadova - H. Craig Helle - H. Gordon - W. Horions - D. Hills. *Biologia: La scienza della vita.* Zanichelli, Padova, Aprile 2011.

[4] INA. Integrated Net Analyzer (Petri Nets). http://www2.informatik.hu-berlin.de/ starke/ina.html, 2009.

[5] IUPAC. International Union of Pure and Applied Chemistry. http://www.iupag.org, 2012.

[6] J.L. Peterson. Petri Net Theory and the Modeling of Systems. *Pretince-Hall, Inc., Englewood Cliffs, NJ 07632*, Aprile 1981.

[7] Kyoto Encyclopedia of Genes and Genomics. KEGG PATHWAY Database. http://www.genome.jp/kegg/pathway.html, 2012.

[8] L. Leydesdorff. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2007.

[9] M. Dell Omodarve. Esercitazioni di statistica biomedica. , pages 157–159, 2012.

[10] M. Heiner - I. Koch - J. Will. Model validation of biological pathways using Petri nets-demonstrated for apoptosis. *Biosystems*, 75(1):15–28, 2004.

[11] M. Kotera - Y. Okuno- M. Hattori - S. Goto - M. Kanehisa. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *Journal of the American Chemical Society*, 126(50):16487–16498, 2004.

[12] M.O. Dayhoff. A model of evolutionary change in proteins. In *In Atlas of protein sequence and structure*, volume 5, pages 315–335. Citeseer, National Biomedical Research Foundation, 1978.

[13] N. Cocco - M. Simeoni - P. Baldan. Comparison of Metabolic Pathways by Considering Potential Fluxes. *Biological Processes & Petri Nets*, 2012.

[14] N. Cocco - M. Simeoni - P. Baldan - A. Marin. Petri nets for modelling metabolic pathways: a survey. *Natural Computing*, 9(4):955–989, 2010.

[15] N. Cocco - M. Simeoni - P. Baldan - F. De Nes - M.L. Segura - A. Marin. MPath2PN-Translating metabolic pathways into Petri nets. In *BioPPN2011 Int. Workshop on Biological Processes and Petri Nets, CEUR Workshop Proceedings*, pages 102–116, 2011.

[16] NCBI. The NCBI Taxonomy. http://www.ncbi.nlm.nih.gov/taxonomy, 2013.

[17] PNML. Petri Net Markup Language. http://www.pnml.org/, 2012.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[19] R. Real - J.M. Vargas. The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, 45(3):380–385, 1996.

[20] S. Alaimo. *Estensioni di Cometa per il confronto tra vie metaboliche in organismi diversi*. PhD thesis, Ca' Foscari di Venezia Dip. di Informatica, Venezia Mestre ITALY, 2011.

[21] S. Karlin - SF. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA*, 87:2264–68, 1990.

[22] S.B Needleman - C.D Wunsch, C.D. et. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[23] SBML. System Biology Markup Language. http://www.sbml.org, 2012.

[24] T. Murata. Petri nets: Properties, analysis and applications. *Proceeding of the IEEE*, 77(4):541–580, Agosto 2002.

[25] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. skr*, 5:1–34, 1948.

[26] V.N. Reddy - ML. Mavrovouniotis - MN. Liebman. Petri net Representation in Metabolic Pathways. In *Proc Int Conf Intell Syst Mol Biol*, volume 1, pages 328–336, 1993.

[27] W.R. Pearson. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11(3):635, 1991.

[28] Y. Tohsato. A method for species comparison of metabolic networks using reaction profile. *IPSJ Digital Courier*, 2(0):685–690, 2006.

Elenco delle figure

Elenco delle tabelle