UNIVERSITÀ DEGLI STUDI DI CATANIA

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

CORSO DI LAUREA SPECIALISTICA IN FISICA

Roberta Sinatra

ANALISI STATISTICA DEL PROTEOMA MEDIANTE STUDIO DI MOTIVI SIGNIFICATIVI

TESI DI LAUREA SPECIALISTICA

RELATORI:

CHIAR.MI PROF. V. LATORA

PROF. D. CONDORELLI

ANNO ACCADEMICO 2006/2007

A Mamma e Papà

Indice

Prefazione **1**

Introduzione **4**

1 Valori osservati e valori attesi: come nascono i *motivi* **9**

1.1 Dalle *parole* ai *motivi*: l'analogia linguistica 9

1.2 Proteomi e k-peptidi 11

1.3 Probabilità dei k-peptidi 12

1.3.1 Probabilità osservata 13

1.3.2 Probabilità attesa 14

Approssimazione di ordine zero (nessuna correlazione) . . . 14

Catene di Markov (correlazioni corte) 15

Oltre le catene di Markov (correlazioni a $k-1$ corpi) . . . 17

2 Proteomi reali **20**

2.1 La scelta dei dati 21

2.1.1 Proteomi di *homo sapiens* e di *mus musculus* 22

2.1.2 Proteomi batterici 23

2.2 Analisi statistica dei k-peptidi 23
2.2.1 Composizione aminoacidica 24
2.2.2 Confronto tra valori osservati e valori attesi 26
2.3 Equivalenza dei database RefSeq, IPI e CCDS 26
2.4 Definizione di motivi e nascita dei dizionari 28
2.4.1 Peptidi sovrarappresentati, sottorappresentati e nullomeri . 28
2.5 Complessità di sequenza 34
2.6 Specie a confronto: il diverso utilizzo delle parole 36
2.6.1 I k-peptidi: correlazione tra specie 36
2.6.2 I k-motivi: copertura del proteoma 40

3 Grafi di motivi 43

3.1 Cos'è un dominio? 44
3.1.1 Probabilità di cooccorrenza di motivi 45
3.2 Grafi di motivi 47
3.2.1 Cos'è un grafo? 47
3.2.2 Costruzione del grafo 48
3.3 Presenza di comunità nella struttura di un grafo 49
3.3.1 Markov Clustering Algorithm 50
L'algoritmo 50
Implementazione e test dell'algoritmo 52
3.4 Alcuni risultati 53
3.4.1 Il parametro r 54
3.4.2 Come si trovano i domini? 55

Conclusioni e sviluppi futuri **57**

A Processi stocastici e catene di Markov **59**

B Ulteriori risultati e grafici **63**

Ringraziamenti **67**

Bibliografia **70**

Prefazione

Quit now, you'll never make it.
If you disregard this advice,
you'll be halfway there.

David Zucker

Il concetto di intellettuale, e di conseguenza quello di scienziato, si è evoluto in maniera molto varia nel corso dei secoli fino ad assumere ai giorni nostri una fisionomìa spiccatamente diversa da quella che lo caratterizzava, ad esempio, durante il Medioevo. Se infatti sei o sette secoli fa era ancora possibile poter immaginare un unico individuo dal sapere enciclopedico (si pensi agli illustri casi di Dante Alighieri e Leonardo Da Vinci), oggi si assiste ad una spiccata specializzazione delle conoscenze dello scienziato. Ciò è del tutto normale e plausibile se si pensa all'incredibile ampliamento della Conoscenza sviluppatasi nel corso dei secoli. Se quindi un tempo addirittura non si distingueva tra letterato e scienziato, ma si parlava semplicemente di sapiente, il quale poteva essere contemporaneamente conoscitore di entrambe le artes, oggi non basta più nemmeno distinguere, ad esempio, tra fisici, chimici e matematici, per poter ritenere di aver una conoscenza globale e completa di un determinato ambito di studio. Si assiste sempre ad una maggiore specializzazione, le competenze diventano sempre più specifiche e tecni-

che con una conseguente maggiore difficoltà nello scambio e nella comunicazione tra settori diversi.

Tuttavia, oggigiorno si sta attuando, almeno in alcuni ambiti, un'inversione di tendenza. Trattando diverse problematiche, non è più possibile continuare a fare ricerca ricorrendo solo ad un unico settore di specializzazione. Al contrario, in molti casi è necessario che addetti ai lavori di discipline diverse interagiscano al fine di analizzare da più punti di vista un problema, dandone un quadro più variegato e completo. Si parla di *sistemi complessi*: espressione non a caso efficace nel suggerire la necessità di un'analisi a diversi livelli e con approcci di discipline differenti.

In questa categoria rientra senza dubbio lo studio del DNA e del sistema proteico da esso codificato[1]. In un primo momento dopo la sua scoperta vi era la convinzione secondo la quale si sarebbe conquistata una forte capacità predittiva nell'individuare le caratteristiche biologiche di un organismo, soprattutto dell'uomo, non appena tutti i 'messaggi' contenuti nel DNA e nelle proteine fossero stati decifrati, compito assunto dalla biologia molecolare. Tuttavia, i dati forniti dalla ricerca hanno presto smentito un'ipotesi di lavoro tanto semplicistica, mentre hanno delineato la necessità di analizzare il sistema con diversi approcci e dal punto di vista di varie discipline. In seguito sono stati coniati nomi appositi per questi studi, in un certo senso, 'ibridi': nascono per esempio la biofisica, la biologia computazionale, la bioinformatica.

In questo scenario di interdisciplinarietà gioca un ruolo importante anche la figura del fisico. Ciò che si vuole fare è analizzare sistemi biologici in maniera

[1] Una piccola percentuale del DNA, mediante i processi di trascrizione e traduzione, codifica per le proteine.

sostanzialmente diversa da quella 'standard', usando appunto metodi tipici della fisica. Il lavoro qui presentato ha proprio queste caratteristiche: viene formulato un modello che mira a studiare le proprietà statistiche del proteoma usando la teoria dei processi stocastici e l'approccio della teoria dei grafi.

La problematica affrontata ha tra l'altro richiesto lo studio di argomenti che esulano dal background culturale di un fisico, ma che sono probabilmente di base in ambito biologico. Per questo motivo il lavoro di tesi che segue introduce innanzitutto il problema biologico: la necessità di uno studio linguistico e statistico del proteoma. Solo successivamente sarà presentato il modello matematico utilizzato e i risultati ottenuti applicandolo a proteomi reali.

Introduzione

Wir müssen wissen, wir werden wissen.

David Hilbert

È stato scritto:

> *Sei scimmie che si mettessero a battere a caso su macchine da scrivere per milioni di anni finirebbero inevitabilmente per scrivere tutti i libri contenuti nel British Museum.*[2]

Ad una mente che non si sia mai cimentata nell'attribuire un "significato fisico" alla parola *mai*, l'affermazione può sembrare plausibile. Eppure credere ad una tale affermazione equivalrebbe a credere che un oggetto, una bottiglia vuota ad esempio, possa improvvisamente lievitare e fluttuare in aria.

Vediamo meglio il perché l'affermazione sia falsa e ingannevole[3]. Riduciamo le pretese del problema e vediamo quanto tempo sia necessario perché sia non nulla la probabilità che tutte le scimmie del pianeta Terra possano produrre battendo a caso su una macchina da scrivere un dato libro, per esempio l'*Amleto* di

[2] Jeans, J., *Mysterious Universe*, Cambridge University Press (1930). L'affermazione è attribuita a Huxley.

[3] Le osservazioni che presentiamo traggono spunto da uno degli esercizi proposti nel primo capitolo del testo: Kittel, C. e Kroemer, H., *Termodinamica statistica*, Boringhieri (1985).

Shakespeare. Ammettiamo che esistano 3 scimmie per ogni uomo sulla Terra. In tal caso il numero di individui del popolo di primati sarebbe dell'ordine di 10^{10}. Supponiamo che ogni scimmia rimanga seduta a scrivere per 10^{18} secondi e riesca a premere 10 caratteri al secondo su una tastiera che ne contiene 44. Poiché l'Amleto è un testo costituito da 10^5 caratteri, senza fare alcuna differenza tra caratteri minuscoli e maiuscoli, la probabilità che le lettere battute a caso risultino nella sequenza corretta è dell'ordine di:

$$\left(\frac{1}{44}\right)^{100000} \simeq 10^{-164\,345}$$

Il risultato si ottiene notando che il numero di tutte le possibili stringhe lunghe 10^5 caratteri che si possono formare con un alfabeto di 44 simboli è dato proprio da $44^{(10^5)}$. Ora, considerato il numero di scimmie, la loro velocità di scrittura e la lunghezza dell'*Amleto*, sarà necessario che trascorrano $\frac{10^{164\,345}}{10^{10}*10/10^5} = 10^{164\,339}$ secondi affinché ogni possibile sequenza di caratteri sia rappresentata almeno una volta. Considerando che l'età dell'universo è stimata essere 10^{17} secondi circa, si intuisce come la probabilità sia praticamente zero a tutti gli effetti. L'affermazione riportata all'inizio è dunque falsa: nella produzione letteraria delle scimmie non comparirà mai nemmeno un libro, né tantomeno una biblioteca.

Un calcolo simile si può effettuare per stimare, ad esempio, quanto sia probabile che tutte le molecole dell'aria all'interno di una bottiglia si trovino nella sua metà superiore (eventualità che se occorresse per un tempo abbastanza lungo, permetterebbe alla bottiglia di sollevarsi in aria da sola). Il calcolo porterebbe ad un risultato simile a quello mostrato precedententemente per le scimmie-scrittrici.

Questi due esempi possono facilmente tracciare un'idea del significato fisico di *mai*.

Abbiamo voluto iniziare con queste considerazioni per capire quanto si possa essere lontani dalla casualità quando si studia e si analizza quello che sarà il protagonista assoluto del nostro lavoro: il *proteoma*.

Normalmente con questo termine si indica l'insieme di tutte le proteine di un organismo; lo si può figurare, con un'immagine in un certo senso un po' naïf, come un enorme libro di 10^7 caratteri scelti da un alfabeto di 20 simboli, l'"alfabeto aminoacidico"[4]. L'immagine può essere considerata riduttiva perché ci si riferisce soltanto alla sequenza aminoacidica delle proteine, definita struttura primaria, e non si menziona la conformazione spaziale (struttura secondaria e terziaria) o l'associazione con altre proteine e catene peptidche (struttura quaternaria). Tuttavia, già nella sequenza aminoacidica è contenuta tutta l'informazione necessaria alla proteina per svolgere la sua funzione biologica. Ed è proprio questa l'informazione cui siamo interessati e che vogliamo carpire.

La metafora di libro per il proteoma è particolarmente stringente: così come un gruppo di scimmie-scrittrici non comporrà mai un libro battendo a caso su una tastiera, è impossibile pretendere di ricomporre anche solo una piccola frazione di proteoma assemblando aminoacidi in maniera random. Infatti, le sequenze

[4] A volte sono utilizzati alfabeti aminoacidici che contengono meno di 20 simboli, per via di una degenerazione del codice che assegna lo stesso simbolo ad aminoacidi che hanno caratteristiche chimico-fisiche simili.

peptidiche che osserviamo non sono nate per caso, bensì sono state "scritte", plasmate e modificate dall'evoluzione fino ad assumere la forma e l'ordine manifestato, forma e ordine che contengono tutte le informazioni necessarie perché le proteine svolgano il ruolo cui sono state destinate [3, 2].

Tuttavia, molte di queste informazioni rimangono ancora sconosciute, o meglio, non decifrate. Non sempre per esempio si conosce la funzione biologica di una proteina, nonostante se ne possa determinare la sequenza [1].
Il codice genetico può essere quindi considerato a tutti gli effetti un linguaggio. Non ha le stesse caratteristiche (almeno fino ad oggi!) di un linguaggio "formale", nel senso che non è definita una "grammatica", né è definita la "parola" come intesa usualmente, cioé come breve sequenza di caratteri delimitata da due spaziature. Tuttavia è possibile individuare delle unità semantiche da cui estrarre una parte del contenuto informativo ricorrendo a metodi che si basano sull'analisi di brevi tratti di sequenza. Tali metodi prendono il nome di *metodi linguistici* [4, 5, 6].

Pietrokovski [6] scrive:

> *Linguistic analysis does not examine a sequence as a whole but by its oligomer makeup. The attribute of each oligomer is the difference between its observed occurrence in the sequence and its expected occurrence. This difference is termed the oligomer's contrast value, and the contrast values of all the oligomers make up the contrast vocabulary of the sequence.*

Sarà proprio questo il leit-motiv nonché ipotesi di partenza del lavoro: cercare di

individuare quei peptidi[5] che, in quanto over-espressi, svolgono la stessa funzione che la "parola" svolge nel linguaggio umano e che divengono il punto da cui partire per definire veri e propri "vocabolari" proteici.

Per determinare questo genere di peptidi, denominati *motivi*, è necessario confrontare un numero osservato, ossia l'evidenza sperimentale di occorrenza, con un numero atteso, calcolato secondo un certo modello. Il primo capitolo di questo lavoro focalizzerà proprio sul problema della previsione del numero di occorrenze, che è naturalmente vincolato ad alcune ipotesi biologiche relative a come un proteoma nasce e si forma.

Nel capitolo successivo si passerà allo studio dei vocabolari di peptidi significativi in proteomi reali di specie diverse e differentemente evolute. I vocabolari di motivi saranno caratterizzati e si analizzeranno le differenze e le analogie tra i vari organismi.

Nell'ultimo capitolo si è interessati a capire come i motivi si combinino tra loro, costituendo strutture che preludono all'equivalente della frase e che in ambito biologico vengono denominati *domini funzionali*. Per poter studiare questo tipo di strutture si è ricorso allo strumento dei network e alla teoria dei grafi [7].

[5] I termini peptidi e oligomeri vengono utilizzati come sinonimi.

Capitolo 1

Valori osservati e valori attesi: come nascono i *motivi*

Prediction is very difficult,
especially of the future.

Niels Bohr

Come già accennato nell'introduzione, siamo interessati ad individuare quei brevi tratti aminoacidici che risultano sovraespressi nel proteoma rispetto ad una certa previsione. In questo capitolo tratteremo proprio il problema della probabilità di occorrenza dei k-peptidi, partendo da alcune ipotesi biologiche e dall'analisi di dati sperimentali.

1.1 Dalle *parole* ai *motivi*: l'analogia linguistica

Il confronto tra valori attesi e valori osservati permette di capire quali k-peptidi svolgano un ruolo chiave nel proteoma perché particolarmente devianti dalla ca-

sualità. Questi peptidi saranno denominati *motivi* e svolgono, in un certo senso, un ruolo equivalente a quello che svolge la parola in un linguaggio.

Per esemplificare meglio questo concetto si può riflettere su come le lettere dell'alfabeto italiano, ad esempio, si combinino a formare una parola. Innanzitutto, ad una prima analisi si osserverà che non tutte le lettere nel vocabolario della lingua italiana hanno la stessa distribuzione. Intuitivamente è evidente, per esempio, come le vocali siano più frequenti delle consonanti, o che la lettera "c" compare più della lettera "z". Per di più, la "formazione" della parola va oltre la sua composizione in lettere. Prendendo un dato insieme di lettere, non tutte le permutazioni creano una sequenza con un senso. Ad esempio dall'insieme di lettere $\{A,\ M,\ O,\ R\}$ si compongono le sequenze "mora","orma","ramo","roma","amor", che nella nostra lingua hanno un significato e sono quindi *parole*; tuttavia si possono anche formare combinazioni come "mroa","amro","rmao","omra", che, a meno di trovarsi davanti un testo futurista, non saranno mai presenti in nessun libro scritto in italiano. Se quindi ci si basa sulla composizione in lettere per prevedere l'occorrenza di una parola, tutte le combinazioni precedentemente elencate avrebbero stesso "diritto di esistenza"; tuttavia solo alcune di esse, quelle che appunto costituiscono dei termini del nostro vocabolario, risulteranno sovraespresse e potranno quindi essere distinte da quelle senza senso. Si potrebbe essere interessati ad un'analisi più profonda, dove termini più consueti o particelle molto frequenti, come articoli e preposizioni, vengano distinti da quelli meno usati anche se dotati comunque di significato. In tal caso un modello di previsione basato esclusivamente sulla frequenza delle lettere sarebbe probabilmente inadeguato e ci si dovrebbe quantomeno ricondurre,

per avere un'idea, ad un'analisi basata sul concetto di sillaba.

Si è voluto introdurre con un esempio propriamente linguistico il problema della previsione dell'occorrenza di una parola per poter meglio comprendere la formulazione matematica del problema e le varie ipotesi che possono essere assunte alla base.

1.2 Proteomi e k-peptidi

Sia P un proteoma costituito da L caratteri suddivisi in N_P proteine. Indichiamo con L_i il numero di caratteri dell'i-esima proteina, ossia la sua lunghezza. Si ha ovviamente:

$$\sum_{i=1}^{N_P} L_i = L$$

Facendo scorrere su P di un carattere alla volta una finestra di lunghezza k, individuiamo dei peptidi formati da k caratteri (k-parole o k-peptidi). Si richiede che la finestra non si sovrapponga tra due proteine. Il numero totale $N^{(k)}$ di k-parole contenute in P si calcola notando che in una proteina di lunghezza L_i se ne hanno$(L_i - k + 1)$, e poi sommando su tutte le proteine:

$$\begin{aligned} N^{(k)} &= \sum_{i=1}^{N_P} (L_i - k + 1) = \\ &= L - kN_P + N_P = \\ &= L - (k-1)N_P \end{aligned} \tag{1.1}$$

Dato un alfabeto di 20 caratteri, quale l'alfabeto aminoacidico, è possibi-

le formare 20^k parole diverse[1] di lunghezza k. D'ora in poi indicheremo con $X_1, X_2, \ldots, X_k$ una generica parola di k caratteri, dove X_i rappresenta l'i-esimo aminoacido della parola. Il suo numero di occorrenze osservate all'interno di P sarà invece indicato con $N^{obs}(X_1, X_2, \ldots, X_k)$.

1.3 Probabilità dei k-peptidi

Indichiamo con $p(X_1, X_2, \ldots, X_k)$ la probabilità di trovare la sequenza $X_1, X_2, \ldots, X_k$ in P. Questa probabilità può essere espressa in termini di probabilità condizionata, ossia quanto è probabile che la variabile casuale X_k segua la sequenza di variabili $X_1, X_2, \ldots, X_{k-1}$. Formalmente:

$$\begin{aligned} p(X_1, X_2) &= p(X_1)p(X_2|X_1) \\ p(X_1, X_2, X_3) &= p(X_1, X_2)p(X_3|X_1, X_2) \\ ... &= .... \\ p(X_1, X_2, ..., X_k) &= p(X_1, X_2, ..., X_{k-1})p(X_k|X_1, X_2, ..., X_{k-1}) \end{aligned} \tag{1.2}$$

dove denotiamo con $p(X_k|X_1, X_2, ..., X_{k-1})$ la probabilità condizionata di avere X_k come k-esimo carattere dopo una data sequenza di $k-1$ caratteri $X_1X_2...X_{k-1}$.

La scrittura $p(X_k|X_1, X_2, ..., X_{k-1})$ è del tutto generale e non aggiunge nulla alla nostra conoscenza sulla occorrenza dei k-peptidi. Vi sono due possibilità per stimarne il valore: basarsi sui dati sperimentali (il proteoma P in sè) e iden-

[1] Disposizioni con ripetizione di 20 elementi presi a k a k.

tificare $p(X_1, X_2, ..., X_k)$ con la frequenza della stringa, ottenendo in tal caso il valore esatto di occorrenza, oppure fare delle approssimazioni e dare una stima, una previsione della probabilità di occorrenza basandosi su altre evidenze sperimentali, come la composizione aminoacidica del proteoma o le correlazioni di ordine più basso. Il primo tipo di stima sarà chiamato *probabilità osservata* e sarà indicato con $p^{obs}(X_1, X_2, ..., X_k)$; il secondo tipo sarà invece identificato come *probabilità attesa* e sarà indicato con $p^{exp}(X_1, X_2, ..., X_k)$. La deviazione tra questi due valori sarà debitamente analizzata al fine di capire quale sia l'ipotesi di base migliore per la previsione di occorrenza dei peptidi e per individuare quei peptidi che deviano particolarmente dalla probabilità attesa.

1.3.1 Probabilità osservata

La probabilità di occorrenza di una stringa $X_1, X_2, ..., X_k$ può essere estratta dai conteggi dei k-peptidi nel proteoma. In tal caso si tratta di un'osservazione puramente sperimentale; si parlerà di *probabilità osservata* della stringa $X_1, X_2, \ldots, X_k$ e sarà indicata con $p^{obs}(X_1, X_2, \ldots, X_k)$.

Si ha pertanto:

$$
\begin{aligned}
p^{obs}(X_1) &= \frac{N^{obs}(X_1)}{\sum_{X_1} N^{obs}(X_1)} = \\
&= \frac{N^{obs}(X_1)}{L} \\
p^{obs}(X_1, X_2) &= \frac{N^{obs}(X_1, X_2)}{\sum_{X_1,X_2} N^{obs}(X_1, X_2)} = \\
&= \frac{N^{obs}(X_1, X_2)}{L - N_P} \\
... &= ... \\
p^{obs}(X_1, X_2, ..., X_k) &= \frac{N^{obs}(X_1, X_2, ..., X_k)}{\sum_{X_1,X_2,...,X_k} N^{obs}(X_1, X_2, ..., X_k)} \\
&= \frac{N^{obs}(X_1, X_2, ..., X_k)}{L - (k-1)N_P}
\end{aligned}
\tag{1.3}
$$

Il valore di probabilità così calcolato è una vera e propria frequenza di occorrenza del k-peptide $X_1, X_2, ..., X_k$.

1.3.2 Probabilità attesa

Approssimazione di ordine zero (nessuna correlazione)

Assumiamo che non ci sia nessun tipo di correlazione tra i simboli del proteoma. Ciò equivale a dire che nella formazione del proteoma, la probabilità di aggiungere un certo aminoacido non è in alcun modo influenzata dagli aminoacidi precedenti.

In tal caso si avrà:

$$\begin{aligned} p(X_2|X_1) &= p(X_2) \\ p(X_3|X_1, X_2) &= p(X_3) \\ ... &= .... \\ p(X_k|X_1, X_2, ..., X_{k-1}) &= p(X_k) \end{aligned} \tag{1.4}$$

Per $k = 1$ si ha $p(X_i) = p^{exp}(X_i) = p^{obs}(X_i)$, che altro non è che la frequenza dei singoli aminoacidi, al variare di i da 1 a 20. Partendo da questo dato, si possono calcolare le probabilità attese $p^{exp}(X_1, X_2, ..., X_k)$ per qualunque stringa $X_1, X_2, ..., X_k$, con $k \geq 2$. Si ha:

$$p^{exp}(X_1, X_2, ..., X_k) = \prod_{i=1,...,k} p^{obs}(X_i) \tag{1.5}$$

Per ogni $k \geq 2$, si può confrontare il valore di $p^{obs}(X_1, X_2, ..., X_k)$, ricavato dai conteggi (1.3), con quello stimato secondo la (1.5).
L'approssimazione di ordine zero è una delle più diffuse in letteratura [8, 9, 10].

Catene di Markov (correlazioni corte)

L'approssimazione si basa sull'idea che nella formazione del proteoma un aminoacido sia aggiunto conservando una certa memoria della parte di sequenza precedente. Se questa quantità di memoria è minima, la sequenza costituisce una catena di Markov (vedi appendice A). Di fatto, quel che si suppone è la presenza di correlazioni tra un aminoacido e quello immediatamente precedente. L'ipotesi

dal punto vista biologico è più che plausibile: gli aminoacidi sono infatti costituenti molecolari dalle proprietà chimico-fisiche[2] molto diverse. La presenza di un aminoacido, per via di queste proprietà, non può non influenzare l'aggiunta dell'aminocido successivo. Si fa un'ipotesi di stazionarietà per le catene di Markov: l'aggiunta dell'aminoacido i dipende soltanto dal tipo di aminoacido $i-1$-esimo, e non dalla posizione i occupata all'interno del proteoma o della proteina.

Formalmente, l'approssimazione si traduce come segue:

$$\begin{aligned} p(X_2|X_1) &= p(X_1)p(X_2) \\ p(X_3|X_1,X_2) &= p(X_3|X_2) \\ ... &= .... \\ p(X_k|X_1,X_2,...,X_{k-1}) &= p(X_k|X_{k-1}) \end{aligned} \tag{1.6}$$

Da cui si pone:

$$\begin{aligned} p^{exp}(X_1) &= p^{obs}(X_1) \\ p^{exp}(X_1,X_2) &= p^{obs}(X_1,X_2) = p^{obs}(X_1)p^{obs}(X_2|X_1) \\ p^{exp}(X_1,X_2,X_3) &= p^{obs}(X_1)p^{obs}(X_2|X_1)p^{obs}(X_3|X_2) \\ ... &= ... \\ p^{exp}(X_1,X_2,...,X_k) &= p^{obs}(X_1)p^{obs}(X_2|X_1)...p^{obs}(X_k|X_{k-1}) = \\ &= p^{obs}(X_1)\prod_{i=2,...,k} p^{obs}(X_i|X_{i-1}) \end{aligned} \tag{1.7}$$

Conoscendo le probabilità dei singoli aminoacidi $p^{obs}(X_i)$ e le probabilità condi-

[2]Ne sono esempi l'idrofobicità o la carica elettrica.

zionate a due corpi $p^{obs}(X_k|X_{k-1})$, è possibile stimare le probabilità attese dei k-peptidi per $k \geq 3$. Infatti dalle (1.3) conosciamo già la forma per $p^{obs}(X_i)$, mentre $p^{obs}(X_k|X_{k-1})$ va dedotta dai conteggi osservando che si richiede la probabilità che fissato l'aminoacido X_{k-1} segua proprio X_k. Tale probabilità è data da:

$$p^{obs}(X_i|X_{i-1}) = \frac{N^{obs}(X_{i-1}, X_i)}{\sum_{X_i} N^{obs}(X_{i-1}, X_i)} \tag{1.8}$$

Oltre le catene di Markov (correlazioni a $k-1$ corpi)

Modellizzando le sequenze aminoacidiche come catene di Markov, si tengono in considerazione soltanto correlazioni a due corpi. Può però rivelarsi importante tenere in conto anche la possibilità che si conservi maggiore memoria della parte di sequenza precedente. Infatti, il ripiegamento spaziale delle proteine e le proprietà biochimiche degli aminoacidi danno evidenza dell'esistenza di correlazioni che vanno oltre il legame peptidico.

A livello biologico, si suppone che i dipeptidi preesistano ai tripeptidi, i tripeptidi ai tetrapeptidi e così via. Quando quindi si forma un tripeptide, esso può essere visto come unione di due dipeptidi che abbiano un aminoacido in comune, mentre i tetrapeptidi nascono dalla combinazione di due tripeptidi che abbiano un dipeptide in comune, e così via. Pertanto, nella formazione di una sequenza a k corpi intervengono correlazioni a $k-1$ corpi. Ciò implica che nell'aggiungere un aminoacido ad una sequenza di $k-1$ elementi, abbiano influenza solo i $k-2$ aminoacidi precedenti, ossia tutti gli aminoacidi escluso il primo. Da un punto

di vista matematico l'approssimazione si esprime nel modo seguente:

$$\begin{aligned} p(X_3|X_1,X_2) &= p(X_3|X_2) \\ p(X_4|X_1,X_2,X_3) &= p(X_4|X_2,X_3) \\ ... &= .... \\ p(X_k|X_1,X_2,...,X_{k-1}) &= p(X_k|X_2,...,X_{k-1}) \end{aligned} \tag{1.9}$$

Si noti che le approssimazioni (1.9) si applicano solo all'ultimo termine moltiplicativo di ciascuna delle (1.2). Riconducendosi quindi ai valori osservati delle sequenze di ordine inferiore, si ottiene:

$$\begin{aligned} p^{exp}(X_1) &= p^{obs}(X_1) \\ p^{exp}(X_1,X_2) &= p^{obs}(X_1)p^{obs}(X_2) \\ p^{exp}(X_1,X_2,X_3) &= p^{obs}(X_1,X_2)p^{obs}(X_3|X_2) = \\ &= p^{obs}(X_1)p^{obs}(X_2|X_1)p^{obs}(X_3|X_2) \\ ... &= ... \\ p^{exp}(X_1,X_2,...,X_k) &= p^{obs}(X_1)p^{obs}(X_2|X_1)\cdots p^{obs}(X_k|X_2,\cdots,X_{k-1}) \end{aligned} \tag{1.10}$$

Per ogni $k > 2$ i valori osservati $p^{obs}(X_1,X_2,...,X_k)$ possono essere confrontati con le (1.10). Le $p^{obs}(X_k|X_2,...,X_{k-1})$ si valutano in base ai conteggi nel modo seguente:

$$p^{obs}(X_k|X_2,...,X_{k-1}) = \frac{N^{obs}(X_2,X_3,...,X_k)}{\sum_{X_k} N^{obs}(X_2,X_3,X_k)} \tag{1.11}$$

o anche in maniera equivalente (vedi appendice A per dettagli):

$$p^{obs}(X_k|X_2, ..., X_{k-1}) = \frac{p^{obs}(X_2, ..., X_k)}{p^{obs}(X_2, ..., X_{k-1})} \tag{1.12}$$

Usando quest'ultima espressione, le 1.10 possono essere scritte in una forma più compatta [11]:

$$\begin{aligned}
p^{exp}(X_1) &= p^{obs}(X_1) \\
p^{exp}(X_1, X_2) &= p^{obs}(X_1, X_2) \\
p^{exp}(X_1, X_2, X_3) &= p^{obs}(X_1, X_2, X_3) \\
p^{exp}(X_1, X_2, X_3, X_4) &= p^{obs}(X_1, X_2, X_3)\frac{p^{obs}(X_2, X_3, X_4)}{p^{obs}(X_2, X_3)} \\
... &= .... \\
p^{exp}(X_1, X_2, ..., X_k) &= p^{obs}(X_1, X_2, ..., X_{k-1})\frac{p^{obs}(X_2, ..., X_k)}{p^{obs}(X_2, ..., X_{k-1})}
\end{aligned} \tag{1.13}$$

Nei capitoli successivi per stimare $p^{exp}(X_1, X_2, ..., X_k)$ verrà utilizzata proprio quest'ultima approssimazione. Infatti, come sarà mostrato in seguito, tenendo in conto le correlazioni a $(k-1)$ corpi si ottiene un valore previsto di occorrenze in buon accordo con il valore osservato. In particolare, faremo vedere che le fluttuazioni dei valori attesi dai conteggi sperimentali sono di tipo gaussiano (vedi paragrafo 2.2.2).

Capitolo 2

Proteomi reali

In the drama of life on a molecular scale, proteins are where the action is.

A.M. Lesk

Negli ultimi anni, le nuove tecniche della biologia molecolare e l'incremento delle capacità di storage di informazioni digitali hanno consentito la produzione e la conservazione di un enorme quantità di dati biologici. La maggior parte di questi dati risultano per certi aspetti ancora inesplorati.

La digitalizzazione di sequenze nucleotidiche e proteiche e l'accesso libero tramite internet alle banche dati hanno reso agevole la possibilità di effettuare analisi statistiche su questo genere di dati, fornendo uno strumento in più all'interpretazione dell'informazione genetica.

La possibilità di disporre di dati proteomici relativi ad organismi diversi sta proprio all'origine del lavoro che presentiamo in questo capitolo e nel seguente. Parleremo di come si derivino i vocabolari di motivi significativi dai proteomi e di come si caratterizzino, di quali siano le differenze analizzando specie diffe-

rentemente evolute e di come i motivi si organizzino a dare strutture proteiche chiamate domini.

2.1 La scelta dei dati

Per poter effettuare un'analisi statistica su sequenze biologiche è innanzituto necessario districarsi nell'enorme quantità di dati e informazioni presenti nelle banche dati. Una serie di database tra i più aggiornati e completi è organizzata e strutturata nel *National Center for Biotechnology Information* (NCBI) [13] e nello *European Bioinformatics Institute* (EBI) [14]; ci siamo riferiti ad entrambi per le nostre ricerche.

Per il nostro studio si sono resi necessari i proteomi[1] delle seguenti specie:

- *Homo Sapiens* (HS), ossia l'uomo;
- *Mus Musculus* (MM), ossia il topo;
- 15 specie di *Archeabacteria* (AB), o archibatteri;
- 12 specie di *Proteobacteria* (PB), o protobatteri.

In tabella 2.1 sono riassunte le informazioni generali relative a questi dati.

[1] Come già anticipato nell'introduzione, per proteoma di una data specie intendiamo esclusivamente l'insieme delle sequenze delle proteine dello stesso, dimenticandone la conformazione spaziale e l'eventuale combinazione con altri metaboliti. Quando si citano database proteici, si fa riferimento a dati che consistono soltanto nella struttura primaria di proteine.

2.1.1 Proteomi di *homo sapiens* e di *mus musculus*

I proteomi di topo e uomo, tra gli organismi più studiati dal punto di vista genetico e per i quali per primo si è reso disponibile l'intero genoma e proteoma, sono stati prelevati da tre diverse banche dati: il *Reference Sequence* (RefSeq) [15], l'*International Protein Index* (IPI) [16] e il *Consensus Coding Sequence* (CCDS) [17].

Il RefSeq, collezione costruita e curata dall'NCBI, consiste in una serie di dati "ripuliti" da ridondanze e da errori di sottomissione[2]. L'IPI è un database proteomico strutturato nell'EBI; mediante un sistema di riferimenti incrociati esso cerca contemporaneamente di ridurre al minimo le sequenze ridondanti e di massimizzare il numero di proteine (una sequenza per ogni trascritto). È aggiornato mensilmente. Il CCDS è un progetto nato in seno all'NCBI che mira ad inviduare un core set di sequenze esclusivamente di uomo e topo che siano di alta qualità, escludendo tutti quei dati che contengono caratteri ambigui[3], presenti invece in altri database. I tre database sono stati scelti tra i numerosi disponibili in rete per via delle caratteristiche di affidabilità e frequente aggiornamento.

[2]È particolarmente frequente trovare nei database biologici più versioni dello stesso gene o proteina, dati affetti da diversi tipi di errore o falsi positivi.

[3]Talvolta si può riscontrare in una sequenza un carattere che non appartiene all'alfabeto dei 20 aminoacidi. Ciò è spesso dovuto ad un'incertezza sperimentale su quale sia il residuo nella relativa posizione della stringa aminoacidica. Sono caratteri ambigui la 'X' (qualunque aminoacido possibile), la B (in caso di ambiguità tra 'N' e 'D', aminoacidi simili dal punto di vista chimico-fisico), la 'Z' (in caso di ambiguità tra 'Q' o 'E') e 'J' (ambiguità tra I ed L). I caratteri 'O' ed 'U' indicano rispettivamente la pirrolisina, un particolare tipo di lisina, usualmente indicata con K, e la selenocisteina, un particolare tipo di cisteina, indicata con C).

2.1.2 Proteomi batterici

I proteomi batterici sono stati prelevati solo dal già menzionato RefSeq, perché non presenti in altre collezioni di dati ripuliti da errori e ridondanze.

Un proteoma batterico è in genere un ordine di grandezza più piccolo di quello di eucarioti superiori, quali l'uomo e il topo. Per potere confrontare i risultati ottenuti dall'analisi statistica tra le diverse specie, è necessario che i campioni statistici studiati, di fatto i proteomi, siano confrontabili per quanto riguarda la lunghezza L. Per poter fare ciò, si sono assemblati diversi proteomi batterici al fine di ottenere oggetti di dimensioni adatte al confronto statistico.

2.2 Analisi statistica dei k-peptidi

Dopo la scelta del set dei dati si è passati all'analisi statistica dei peptidi, così come spiegato nel capitolo 1.

Tutti i codici per il conteggio dei k-peptidi ($N^{obs}(X_1, X_2, \cdots, X_k)$) e per il calcolo dei valori attesi ($N^{exp}(X_1, X_2, \cdots, X_k)$) sono stati scritti in linguaggio C++. L'elaborazione dei dati e la creazione dei grafici è stata effettuata mediante il programma Matlab.

L'analisi statistica è stata fatta per k-peptidi con k=2, 3, 4. Non è possibile spingersi oltre con le attuali dimensioni dei proteomi. Infatti, ricordando che tutte le possibili parole lunghe k che si possono formare con un alfabeto di 20 aminoacidi sono 20^k, si ha che tutti i possibili pentapeptidi sono $20^5 = 3.2 \cdot 10^6$. Osservando che il numero di k-parole secondo la (1.1) è dello stesso ordine della

Tabella 2.1: Caratteristiche dei proteomi analizzati. Notazione: L = numero totale di aminoacidi, N_P = numero di proteine, L_P = lunghezza media delle proteine. Per l'uomo e per il topo è stato utilizzato un unico proteoma reperibile presso il *Consensus Coding Sequence* database (CCDS) [17]; per i batteri l'analisi è stata effettuata su un set di proteomi combinati reperibili dal *Reference Sequence* database (RefSeq) [15].

	L	N_P	L_P	# proteomi combinati
Homo Sapiens (RefSeq)	16455036	34180	481.42	1
Mus Musculus (RefSeq)	22447725	46902	478.61	1
Homo Sapiens (IPI)	28653368	68043	421.11	1
Mus Musculus (IPI)	25507684	53997	472.39	1
Homo Sapiens (CCDS)	9429284	17751	531.20	1
Mus Musculus (CCDS)	6480301	13374	484.54	1
Proteobacteria (RefSeq)	10767179	34188	314.94	12
Archeabacteria (RefSeq)	9471267	31417	308.46	15

lunghezza L del proteoma[4] e che le dimensioni medie di un proteoma sono di $10^6 \div 10^7$ aminoacidi, si capisce come il campione sia appena sufficiente affinché ogni pentapeptide sia rappresentato una volta.

2.2.1 Composizione aminoacidica

Nel caso $k = 1$ ci si riduce allo studio dei singoli aminoacidi. Poiché non esiste un valore osservato di ordine inferiore, le $p^{obs}(X_i)$ non possono essere confrontate con un valore atteso, come per i k-peptidi con $k \geq 2$ (vedi capitolo 1). Tuttavia esiste una distribuzione con cui poter operare un confronto. Jordan et al. [18] hanno infatti mostrato che la disomogeneità di frequenza tra gli aminoacidi deriva in parte dalla degenerazione del codice genetico: la frequenza di ogni aminoacido è

[4] Ciò avviene se $N_P \ll L$, ipotesi vera per qualunque proteoma reale.

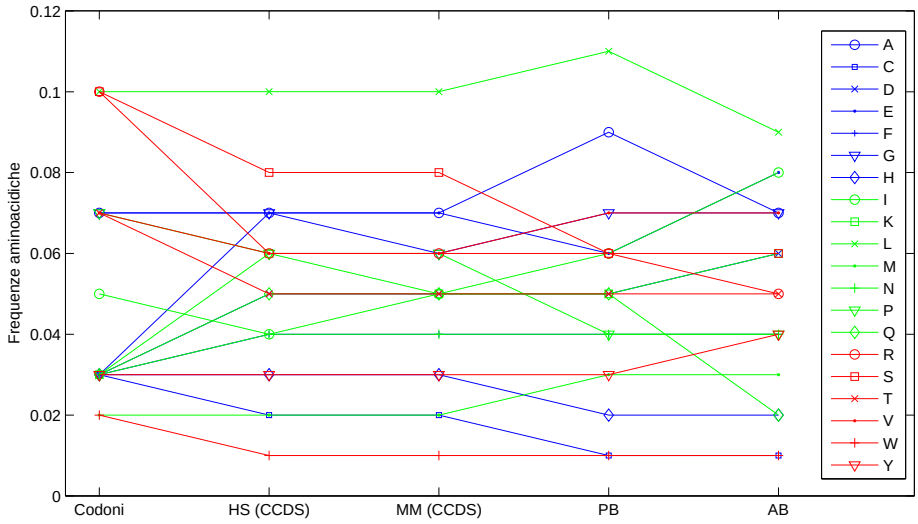


Figura 2.1: Distribuzioni aminoacidiche per le diverse specie. Per l'uomo e il topo sono presenti i dati del solo CCDS. Per il confronto di frequenze in set di dati diversi ma appartenenti alla stessa specie vedi figura 2.4.

proporzionale al numero di codoni da cui è codificato[5]. Esiste quindi una sorta di valore atteso, basato questa volta sulle regole applicate dal DNA per codificare gli aminoacidi. In figura 2.1 confrontiamo le distribuzioni aminoacidiche ottenute per i nostri proteomi con la frequenza legata al codice genetico.

In figura 2.1 per l'uomo e il topo sono presenti soltanto i dati relativi al proteoma CCDS. Infatti, come mostreremo meglio in seguito (vedi paragrafo 2.3), i risultati forniti dalle tre collezioni RefSeq, IPI e CCDS sono sostanzialmente equivalenti. Per questo motivo in seguito, così come fatto adesso, mostreremo soltanto le informazioni sui proteomi umano e murino relativi al CCDS.

[5]I 20 aminoacidi sono codificati nel DNA da triplette di nucleotidi. Utilizzando gruppi di tre lettere di un alfabeto a 4 simboli, come quello nucleotidico, si possono formare 64 parole ($= 4^3$); queste triplette vengono anche dette codoni. Essendo il numero dei possibili codoni maggiore del numero di aminoacidi, si può intuire come la corrispondenza non sia biunivoca, ma piuttosto tale da associare più triplette alla stessa lettera. Il fenomeno va sotto il nome di *degenerazione del codice genetico*.

2.2.2 Confronto tra valori osservati e valori attesi

Per $k \geq 2$ si è potuto operare un confronto tra modello teorico e dati. Si sono conteggiate le occorrenze di ciascun k-peptide facendo scorrere sul proteoma una finestra di lunghezza k che non si sovrapponesse tra due proteine; le parole contenenti uno dei caratteri ambigui elencati in 2.1.1 sono state opportunamente tralasciate[6]. Dopo aver computato per ogni k-peptide anche il relativo valore atteso, si è potuto operare un confronto tra teoria e dati studiando il grado di correlazione (lineare) sussistente tra i valori p^{obs} e p^{exp} nonché la distribuzione del loro rapporto (equivalente anche alla distribuzione di N^{obs}/N^{exp}) per ogni k e per ogni organismo. In tabella 2.3 sono riportati il coefficiente di Pearson per la correlazione lineare (r) e il valore medio con relativa deviazione standard delle distribuzioni del rapporto p^{obs}/p^{exp}.

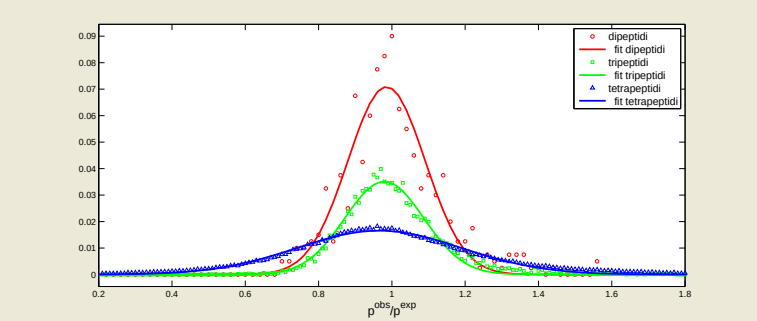
Ciascuna di queste ultime distribuzioni è ben fittata da una curva gaussiana

Si può osservare come per ciascuna distribuzione si abbia $< p^{obs}/p^{exp} >= 1$ con una deviazione standard alla prima cifra decimale. Ciò mostra la validità dell'approssimazione di correlazione a $k-1$ corpi discussa nel primo capitolo per prevedere i valori di occorrenza dei peptidi.

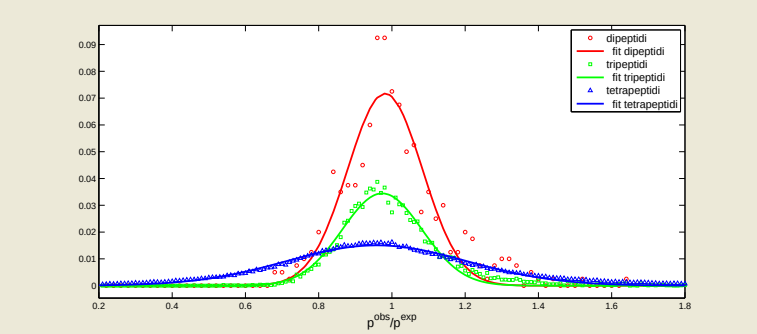
2.3 Equivalenza dei database RefSeq, IPI e CCDS

Per uomo e topo l'analisi in k-peptidi è stata eseguita su informazioni provenienti da tre diverse banche dati: il RefSeq, l'IPI e il CCDS. Tuttavia, così come mostra

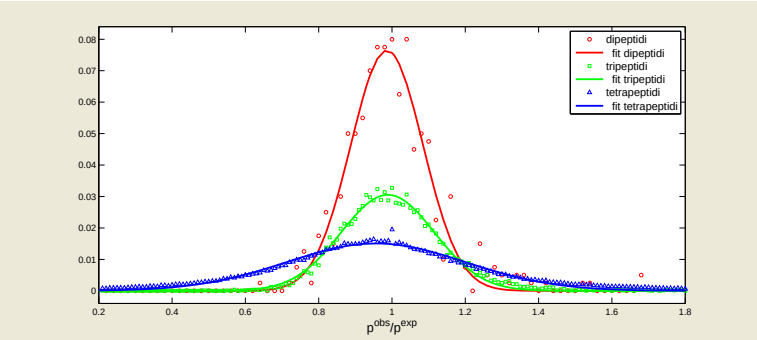
[6]Se così non fosse, l'alfabeto aminoacidico non sarebbe costituito da 20 simboli e la nostra analisi ne risulterebbe inficiata.



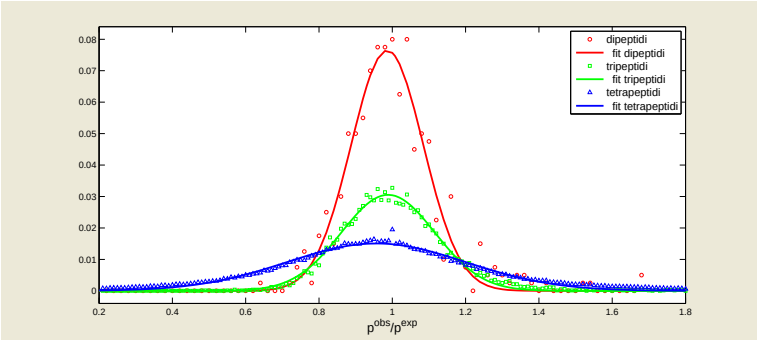
(a) Uomo (CCDS)



(b) Topo (CCDS)



(c) Archibatteri (RefSeq)



(d) Protobatteri (RefSeq)

Figura 2.2: Curve di distribuzione del rapporto p^{obs}/p^{exp} per $k = 2, 3, 4$ in diverse specie.

Tabella 2.2: Media e deviazione del rapporto tra valori osservati e valori attesi dei k-peptidi con k=2, 3, 4 per uomo e topo nei tre database RefSeq, IPI e CCDS.

	Dipeptidi p^{obs}/p^{exp}	**Tripeptidi** p^{obs}/p^{exp}	**Tetrapeptidi** p^{obs}/p^{exp}
HS (RefSeq)	$1,00 \pm 0,14$	$1,00 \pm 0,14$	$1,00 \pm 0,13$
MM (RefSeq)	$1,00 \pm 0,14$	$1,00 \pm 0,14$	$1,00 \pm 0,14$
HS (IPI)	$1,01 \pm 0,17$	$1,01 \pm 0,17$	$1,01 \pm 0,16$
MM (IPI)	$1,01 \pm 0,18$	$1,01 \pm 0,16$	$1,01 \pm 0,17$
HS (CCDS)	1.00 ± 0.13	1.01 ± 0.17	1.01 ± 0.35
MM (CCDS)	1.00 ± 0.14	1.01 ± 0.17	1.01 ± 0.38

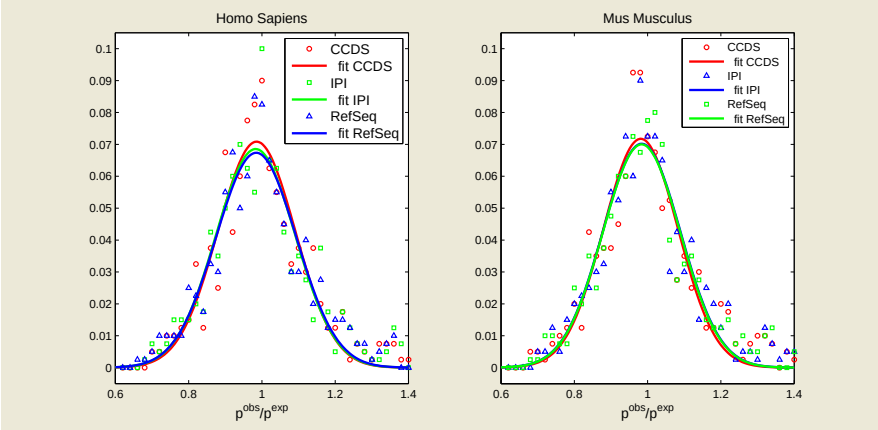
il confronto delle distribuzioni aminoacidiche di figura 2.4 e i dati relativi alla distribuzione di p^{obs}/p^{exp} riportati in tabella 2.2 e in figura 2.3, i risultati che si ottengono sono praticamente equivalenti.

Pertanto di qui in avanti si proseguirà l'analisi del proteoma umano e murino soltanto sui dati relativi ad un solo database. In particolare, si è preferito utilizzare il CCDS perché in esso risulta completamente assente il carattere ambiguo 'X', talvolta presente nei proteomi delle altre due banche dati.

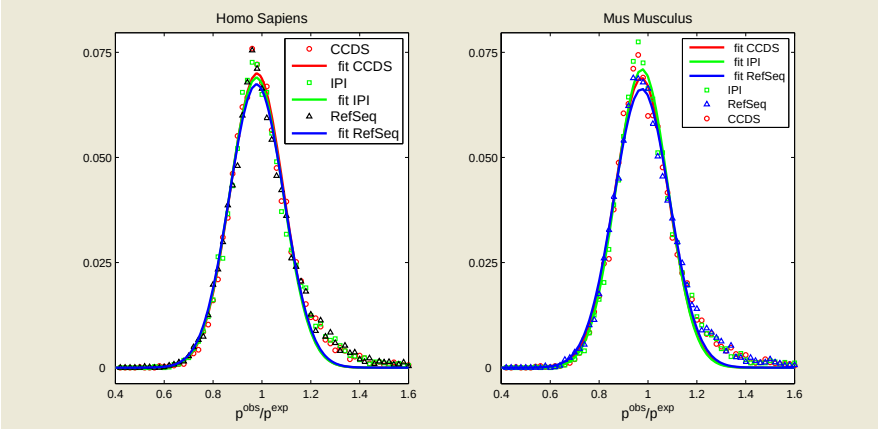
2.4 Definizione di motivi e nascita dei dizionari

2.4.1 Peptidi sovrarappresentati, sottorappresentati e nullomeri

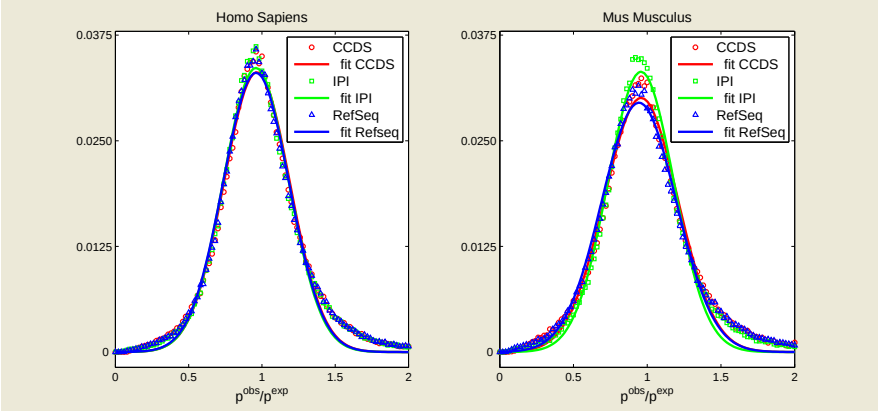
Nel paragrafo 2.2.2 si è discusso della relazione che intercorre tra osservato e atteso nelle specie in esame. In particolare, si è visto come la distribuzione di p^{obs}/p^{exp} abbia un andamento gaussiano. In questo modo nasce in maniera



(a) Dipeptidi

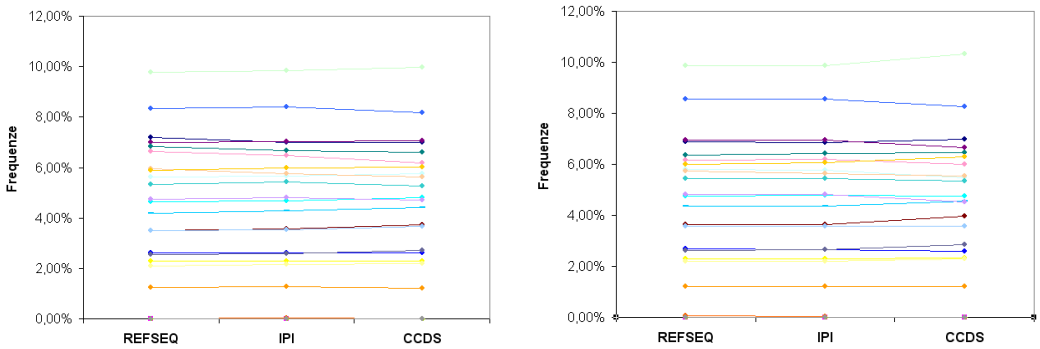


(b) Tripeptidi



(c) Tetrapeptidi

Figura 2.3: Confronto tra le distribuzioni del rapporto p^{obs}/p^{exp} per i tre database RefSeq, IPI e CCDS in uomo e topo. Come si può osservare, i risultati ottenuti mediante i tre database sono del tutto equivalenti.



(a) Homo Sapiens (b) Mus Musculus

Figura 2.4: Confronto tra le frequenze aminoacidiche dei tre diversi proteomi (RefSeq, IPI e CCDS) per (a) uomo e per (b) topo. Come si può vedere, ogni aminoacido ha una frequenza pressoché costante in ciascun proteoma.

naturale la definizione di *motivo*, cioé di parola che occorre molte più volte rispetto a quanto previsto dal modello teorico: un k-peptide $X_1, X_2, \cdots, X_k$ si definisce statisticamente significativo se si ha:

$$\frac{p^{obs}(X_1, X_2, \cdots, X_k)}{p^{exp}(X_1, X_2, \cdots, X_k)} \geq \left\langle \frac{p^{obs}}{p^{exp}} \right\rangle + 2\sigma. \tag{2.1}$$

σ è la deviazione standard, per cui si ha $\sigma^2 = \left\langle \left(\frac{p^{obs}}{p^{exp}} \right)^2 \right\rangle - \left\langle \frac{p^{obs}}{p^{exp}} \right\rangle^2$. I peptidi per cui vale la (2.1) sono in un certo senso sovrarappresentati all'interno del proteoma. Denotiamo l'insieme dei k-motivi di un proteoma con il simbolo Z_k^{over}.

Si è parimenti interessati a quei peptidi che deviano particolarmente dall'atteso, ma perché sottoespressi. In maniera simmetrica a (2.1), un k-peptide $X_1, X_2, \cdots, X_k$ è considerato sottorappresentato se vale la condizione:

$$\frac{p^{obs}(X_1, X_2, \cdots, X_k)}{p^{exp}(X_1, X_2, \cdots, X_k)} \leq \left\langle \frac{p^{obs}}{p^{exp}} \right\rangle - 2\sigma \tag{2.2}$$

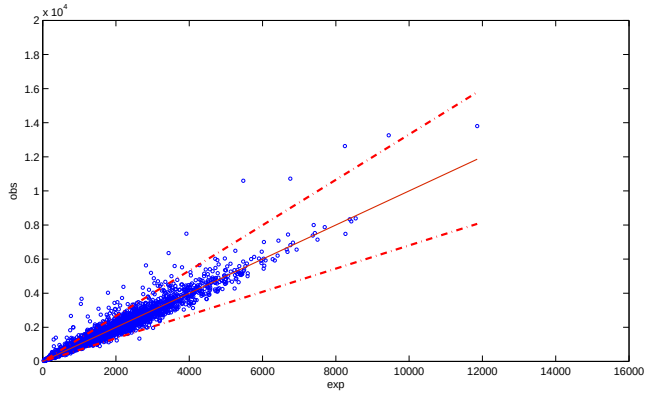
dove σ ha sempre il significato di deviazione standard. L'insieme di questi peptidi è indicato con Z_k^{under}.

Infine esistono anche peptidi che, pur avendo una probabilità p^{exp} non nulla di occorrere, non sono rappresentati. Questi peptidi vengono talvolta chiamati *nullomeri* [19]; ne indichiamo l'insieme con Z_k^{null}.

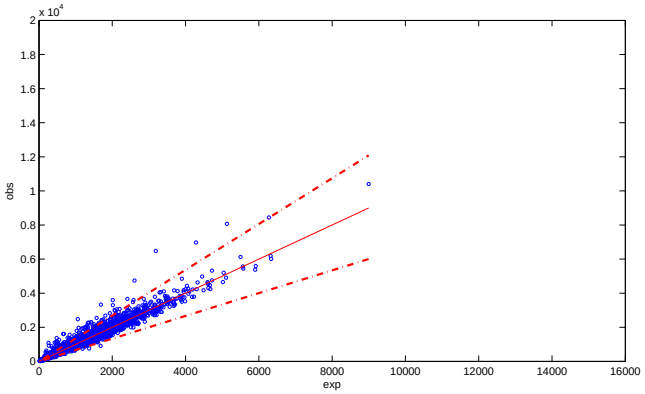
Gli insiemi Z_k^{over}, Z_k^{under} e Z_k^{null} vengono anche chiamati *dizionari*, seguendo l'analogia linguistica già richiamata nell'introduzione e nel capitolo 1. In tabella 2.3 è riportato il numero degli elementi di ciascun dizionario. È immediato notare come per qualunque k il numero dei peptidi sovrarappresentati sia di gran lunga maggiore di quello dei sottorappresentati, nonostante le definizioni (2.1) e (2.2) siano simmetriche. Questo risultato può essere facilmente visualizzato rappresentando ciascuna parola con un punto in un grafico che abbia N^{exp} sulle ascisse e N^{obs} sulle ordinate (vedi figura 2.5). Su un grafico di questo tipo è possibile tracciare un angolo (rette in rosso tratteggiate nell'immagine) all'esterno del quale i punti rappresenteranno parole sovrarappresentate (parte superiore) o sottorappresentate (parte inferiore). Infatti, la tangente dell'angolo formato dalla retta superiore con l'asse delle ascisse è proprio uguale a $\left\langle \frac{p^{obs}}{p^{exp}} \right\rangle + 2\sigma$ mentre vale $\left\langle \frac{p^{obs}}{p^{exp}} \right\rangle - 2\sigma$ per la retta sottostante. La retta centrale rappresenta invece la bisettrice del quadrante; i punti con rapporto $N^{obs}/N^{exp} = 1$ vi si dispongono sopra.

Tabella 2.3: Caratteristiche dei k-peptidi, con $k = 2,\ 3,\ 4$, in ciascun organismo. Notazione: r = coefficiente di Pearson tra distribuzione attesa e distribuzione osservata, $p^{(obs)}/p^{(exp)}$ = rapporto medio tra distribuzione di probabilità osservata e distribuzione di probabilità attesa, Z_k^{over} = dizionario di parole sovrarappresentate, Z_k^{under} = dizionario di parole sottorappresentate, Z_k^{null} = dizionario di parole attese ma mai osservate.

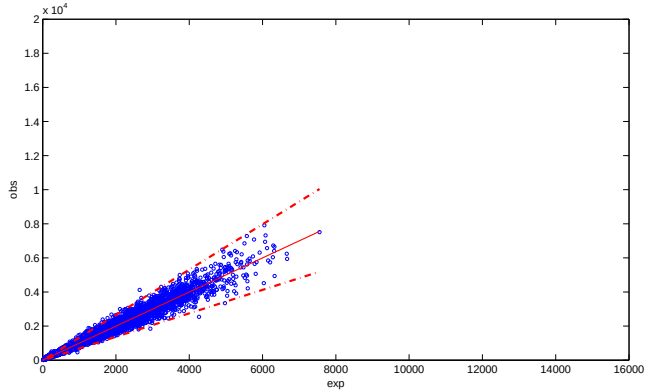
	Homo Sapiens CCDS				
	r	$p^{(obs)}/p^{(exp)}$	Z_k^{over}	Z_k^{under}	Z_k^{null}
Dipeptidi	0.9698	1.0026 ± 0.1350	17	6	0
Tripeptidi	0.9754	1.0061 ± 0.1628	219	22	0
Tetrapeptidi	0.9461	1.0113 ± 0.3520	4265	1094	562
	Mus Musculus CCDS				
	r	$p^{(obs)}/p^{(exp)}$	Z_k^{over}	Z_k^{under}	Z_k^{null}
Dipeptidi	0.9719	1.0018 ± 0.1370	18	5	0
Tripeptidi	0.9780	1.0053 ± 0.1689	246	20	0
Tetrapeptidi	0.9427	1.0139 ± 0.3846	4581	732	996
	Proteobacteria RefSeq				
	r	$p^{(obs)}/p^{(exp)}$	Z_k^{over}	Z_k^{under}	Z_k^{null}
Dipeptidi	0.9850	1.0011 ± 0.1514	13	8	0
Tripeptidi	0.9886	1.0032 ± 0.1422	240	104	0
Tetrapeptidi	0.9733	1.0175 ± 0.4614	3513	20	1433
	Archeabacteria RefSeq				
	r	$p^{(obs)}/p^{(exp)}$	Z_k^{over}	Z_k^{under}	Z_k^{null}
Dipeptidi	0.9845	1.0026 ± 0.1329	16	3	0
Tripeptidi	0.9844	1.0082 ± 0.1604	198	46	0
Tetrapeptidi	0.9720	1.0257 ± 0.5308	2809	0	2134



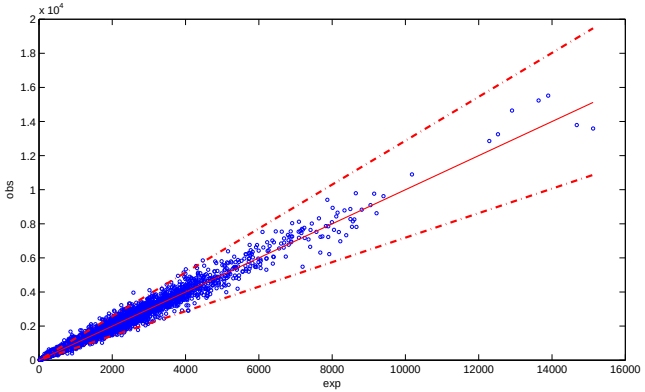
(a) Uomo



(b) Topo



(c) Archibatteri



(d) Protobatteri

Figura 2.5: Confronto del valore atteso (N^{exp}) rispetto al valore osservato (N^{obs}) di tripeptidi in specie diverse. I punti al di fuori dell'angolo individuato dalle linee tratteggiate rappresentano le parole statisticamente significative (parte superiore esterna all'angolo) e le parole sottorappresentate (parte inferiore esterna all'angolo). Se un tripeptide è rappresentato quanto atteso si collocherà sulla bisettrice. I grafici per $k = 2$ e $k = 4$ sono riportati in appendice B.

2.5 Complessità di sequenza

Non tutti i motivi anche se significativi racchiudono lo stesso contenuto di informazione. Ad esempio, gli omopolimeri (segmenti peptidici formati da un solo tipo di aminoacido) risultano particolarmente frequenti perché prodotti da processi di duplicazione di interi tratti di codice genetico. Essi non trasportano quindi un'informazione biologica dello stesso calibro di quella di un peptide formato da aminoacidi tra loro tutti diversi.

Esistono importanti algoritmi [20], integrati nella maggior parte dei database che mascherano tratti di sequenze nucleotidiche e peptidiche a basso contenuto informativo, detti anche segmenti a bassa complessità[7], al fine di favorire l'emergere dell'informazione biologica.

Abbiamo classificato le parole dei dizionari in due soli gruppi di complessità: consideriamo peptidi a bassa complessità tutte quelle parole che contengono due lettere consecutive uguali; tutti gli altri saranno peptidi ad alta complessità.

In tabella 2.4 riportiamo per ogni dizionario il numero di questi peptidi, distinguendo gli omopolimeri tra i peptidi a bassa complessità.

[7] La maggior parte di questi algoritmi utilizzano la misura di informazione data dall'entropia di Shannon [21].

Tabella 2.4: Caratteristiche dei dizionari Z_k^{over} in ciascun organismo. Gli *omopolimeri* sono peptidi che consistono della ripetizione di una sola lettera (ad esempio: $AAAA$), i peptidi *a bassa complessità* nella nostra definizione contengono almeno due lettere consecutive uguali (ad esempio: $AACD$ oppure $AEED$). Tutti gli altri peptidi rientrano nel gruppo di parole *ad alta complessità*.

	Homo Sapiens CCDS		
	Omopolimeri	Bassa complessità	Alta complessità
Dipeptidi	11	-	6
Tripeptidi	13	7	199
Tetrapeptidi	11	589	3665

	Mus Musculus CCDS		
	Omopolimeri	Bassa complessità	Alta complessità
Dipeptidi	11	-	7
Tripeptidi	13	8	225
Tetrapeptidi	11	691	3879

	Proteobacteria RefSeq		
	Omopolimeri	Bassa complessità	Alta complessità
Dipeptidi	3	-	10
Tripeptidi	3	16	221
Tetrapeptidi	2	511	3000

	Archeabacteria RefSeq		
	Omopolimeri	Bassa complessità	Alta complessità
Dipeptidi	5	-	11
Tripeptidi	4	19	175
Tetrapeptidi	2	428	2379

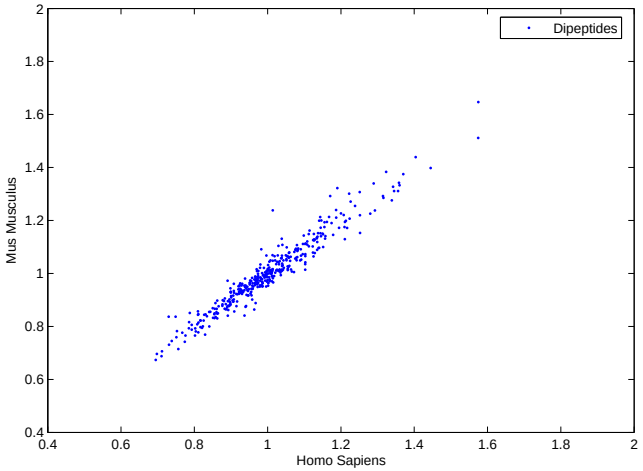
2.6 Specie a confronto: il diverso utilizzo delle parole

Uno degli scopi della nostra analisi statistica mira a cogliere la differenza nell'utilizzo delle parole e dei motivi lungo il corso dell'evoluzione. Per questa ragione si sono selezionati proteomi di organismi tanto lontani lungo la scala evolutiva.

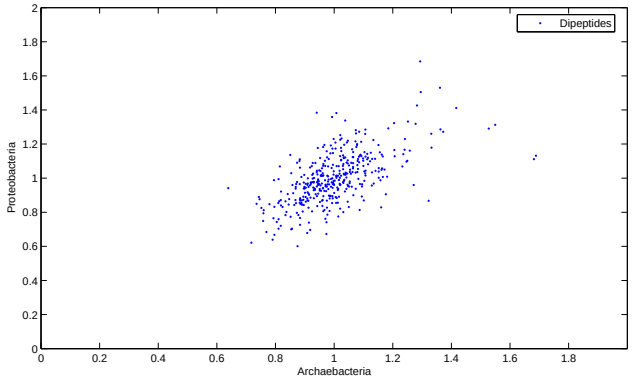
2.6.1 I k-peptidi: correlazione tra specie

Prescindendo dal valore di significatività statistica, è interessante sottolineare il diverso impiego delle parole aminoacidiche negli organismi studiati. Nelle figure 2.6, 2.7 e 2.8 sono messi in relazione i valori p^{obs}/p^{exp} di proteomi diversi. Ogni possibile peptide è rappresentato da un punto i cui valori di ascissa e ordinata corrispondono a valori di p^{obs}/p^{exp} in specie diverse.

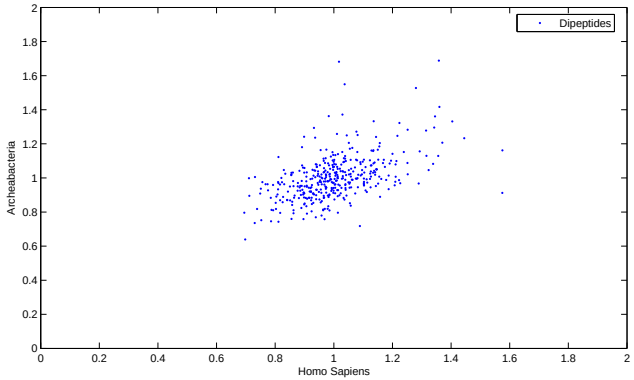
È possibile osservare come la correlazione lineare più forte, che va comunque diminuendo all'aumentare di k, si abbia paragonando i proteomi di uomo e topo. In tutti gli altri casi, anche quando si confrontano tra di loro i due diversi tipi di batterio, la correlazione è appena una parvenza che sbiadisce all'aumentare della lunghezza delle parole. In questo modo emerge che, fissata una parola, nei due proteomi se ne fa un impiego molto diverso (rapporto p^{obs}/p^{exp} molto diverso tra i due organismi). Se ne deduce pertanto che la pressione evolutiva non si manifesta a livello di composizione aminoacidica, ma nella formazione del legame peptidico.



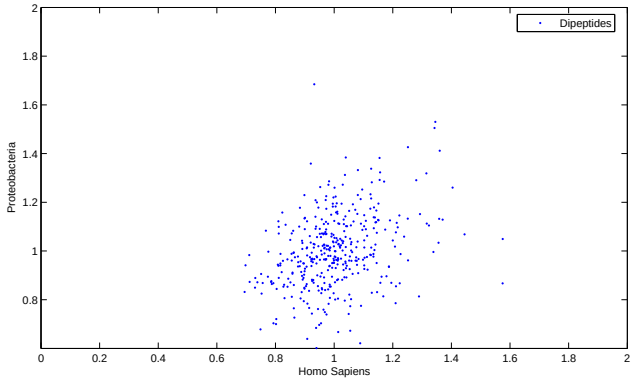
(a) Uomo - Topo



(b) Archibatteri - Protobatteri

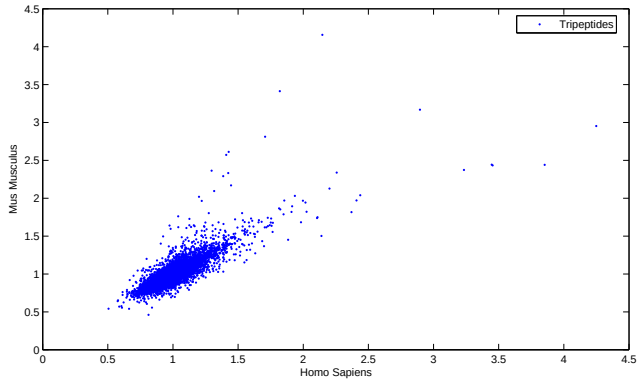


(c) Uomo - Archibatteri

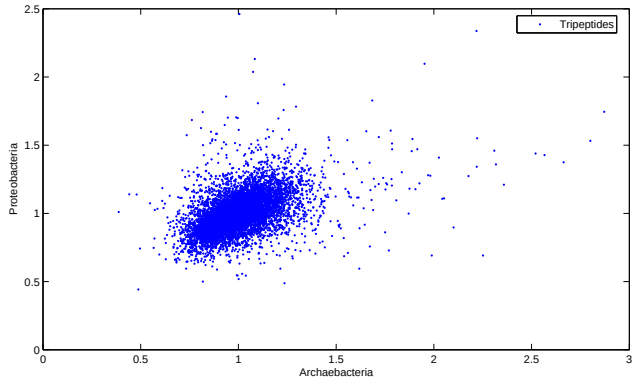


(d) Uomo - Protobatteri

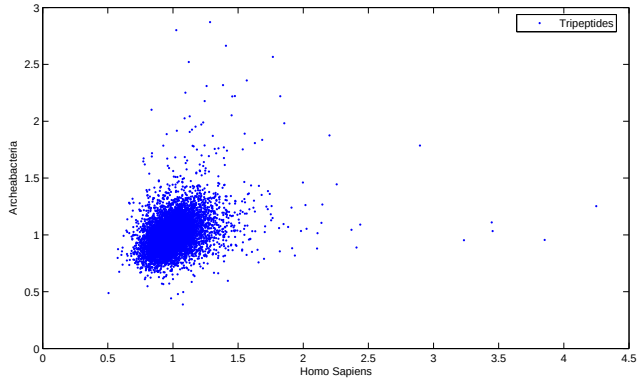
Figura 2.6: Grafici relativi al confronto del rapporto p^{obs}/p^{exp} per dipeptidi in specie diverse. I valori sugli assi indicano il rapporto p^{obs}/p^{exp}. Se un dipeptide possiede lo stesso rapporto in entrambe le specie, il punto che lo rappresenta si collocherà sulla bisettrice. Riportiamo i coefficienti di Pearson r: (a) $r = 0.9677$, (b) $r = 0.6148$, (c) $r = 0.5159$, (d) $r = 0.3455$.



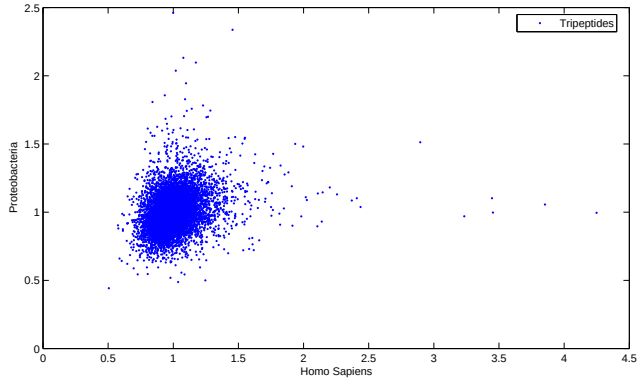
(a) Uomo - Topo



(b) Archibatteri - Protobatteri

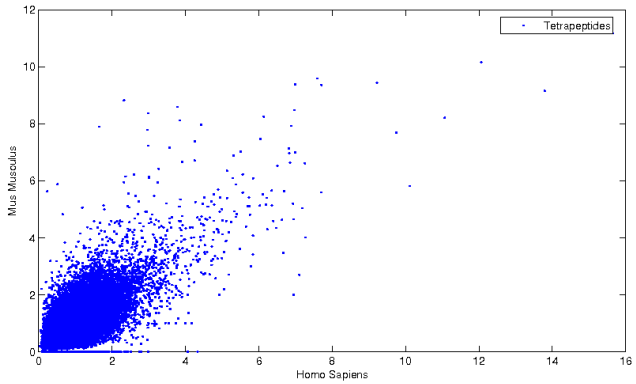


(c) Uomo - Archibatteri

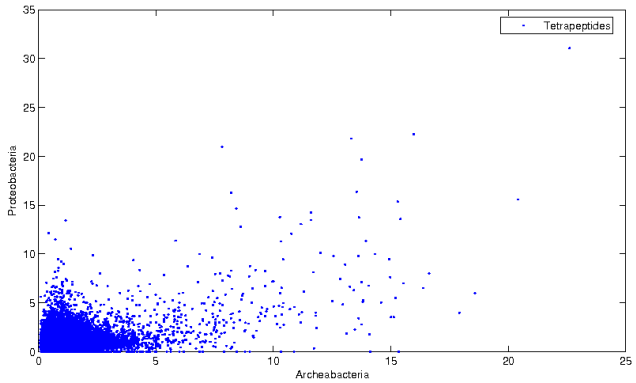


(d) Uomo - Protobatteri

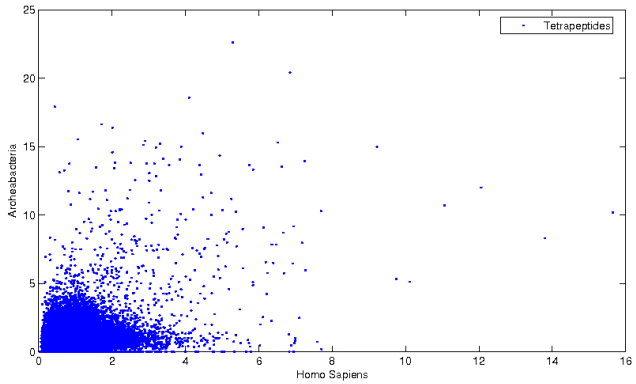
Figura 2.7: Grafici relativi al confronto del rapporto p^{obs}/p^{exp} per tripeptidi in specie diverse. I valori sugli assi indicano il rapporto p^{obs}/p^{exp} in ciascuna specie. Se un tripeptide possiede lo stesso rapporto in entrambe le specie, il punto che lo rappresenta si collocherà sulla bisettrice. Riportiamo i coefficienti di Pearson r: (a) $r = 0.8347$, (b) $r = 0.4707$, (c) $r = 0.3397$, (d) $r = 0.2641$.



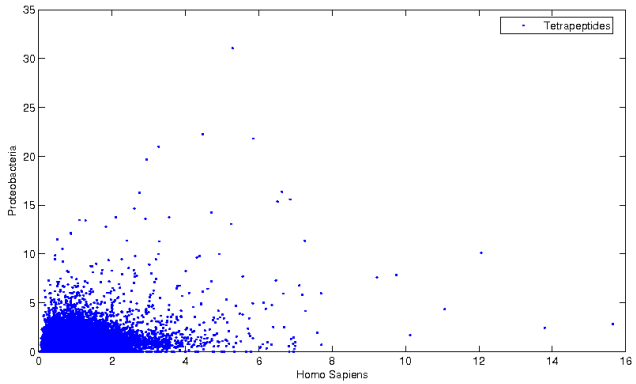
(a) Uomo - Topo



(b) Archibatteri - Protobatteri



(c) Uomo - Archibatteri



(d) Uomo - Protobatteri

Figura 2.8: Grafici relativi al confronto del rapporto p^{obs}/p^{exp} per tetrapeptidi in specie diverse. I valori sugli assi indicano il rapporto p^{obs}/p^{exp}. Se un tetrapeptide possiede lo stesso rapporto in entrambe le specie, il punto che lo rappresenta si collocherà sulla bisettrice. Riportiamo i coefficienti di Pearson r: (a) $r = 0.6708$, (b) $r = 0.3301$, (c) $r = 0.2221$, (d) $r = 0.1587$.

2.6.2 I k-motivi: copertura del proteoma

Dai valori riportati in tabella 2.3, rappresentati nel grafico di figura 2.9, si possono osservare delle lievi fluttuazioni della frazione di k-motivi tra un proteoma e l'altro. Esse non sono però tali da poter giustificare la netta differenza tra eucarioti e batteri che si osserva analizzando il rapporto:

$$\frac{\sum_{(X_1,X_2,\cdots,X_k)\in Z_k^{over}} N^{obs}(X_1, X_2, \cdots, X_k)}{L-(k-1)N_P} \tag{2.3}$$

ossia la frazione di parole contenute nel proteoma che sono motivi. Il rapporto (2.3) può essere denominato *copertura del proteoma*, poiché indica "quanto proteoma" è occupato da motivi. La differenza in copertura tra eucarioti e batteri è mostrata in figura (2.10): nonostante entrambi posseggano pressocché lo stesso numero di motivi all'interno del proprio dizionario, uomo e topo ne fanno un impiego ben maggiore.

Si potrebbe pensare che il risultato derivi da come i motivi vengono selezionati. Tuttavia, se si procede fissando la stessa percentuale di parole significative per ciascuna specie (questo equivale, una volta ordinate le parole per significatività statistica, a prenderne lo stesso numero), il risultato non cambia (vedi figura 2.11).

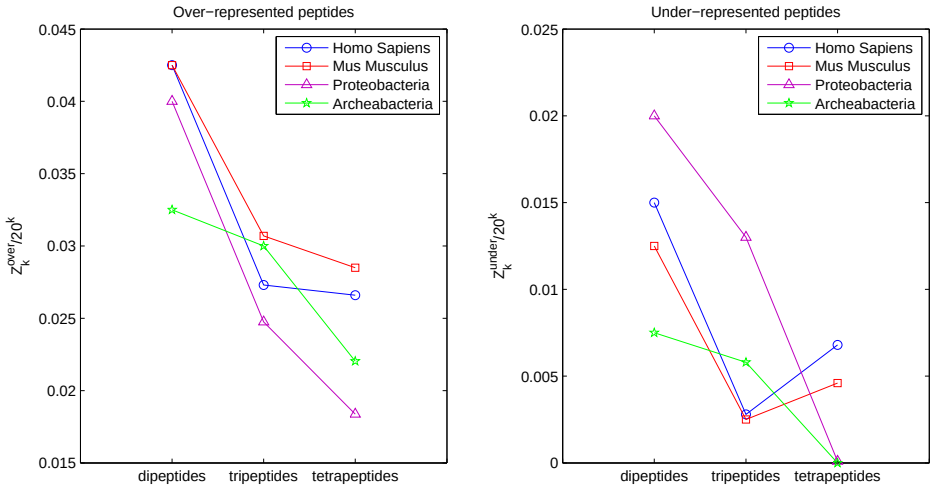


Figura 2.9: (a) Frazione di parole significative rispetto al numero di parole possibili ($Z_k^{over}/20^k$); (b) frazione di parole sottoespresse rispetto al numero di parole possibili ($Z_k^{under}/20^k$).

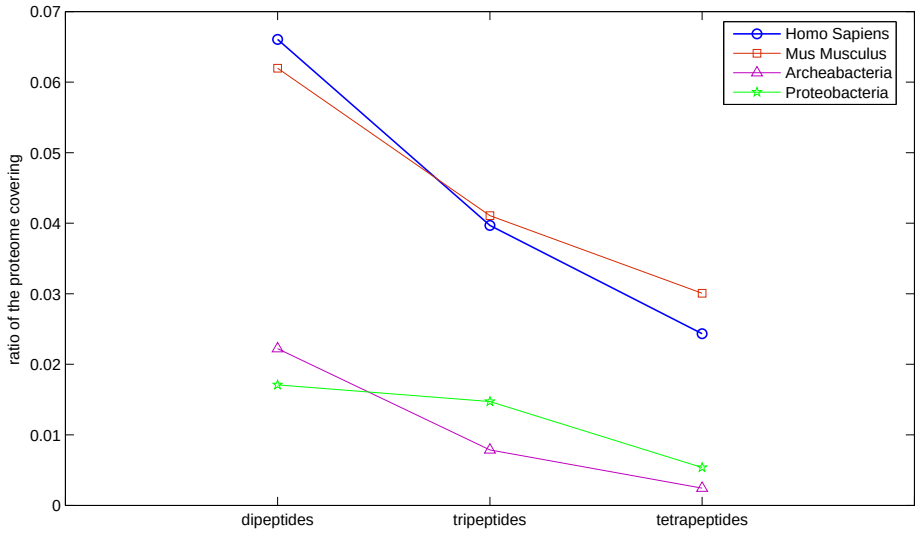
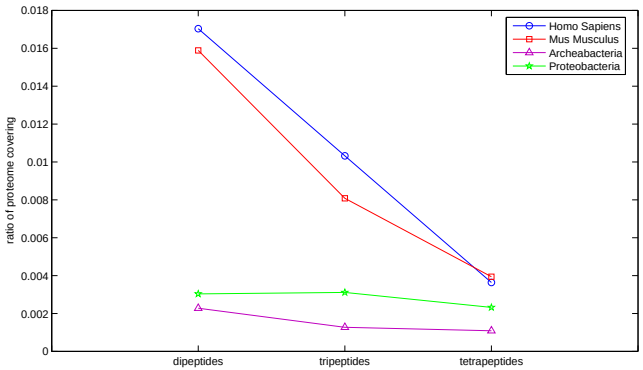
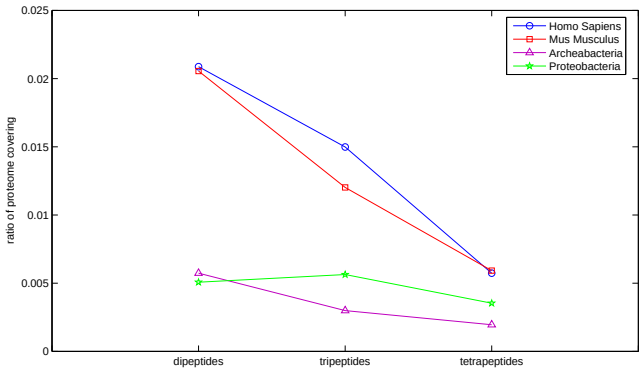


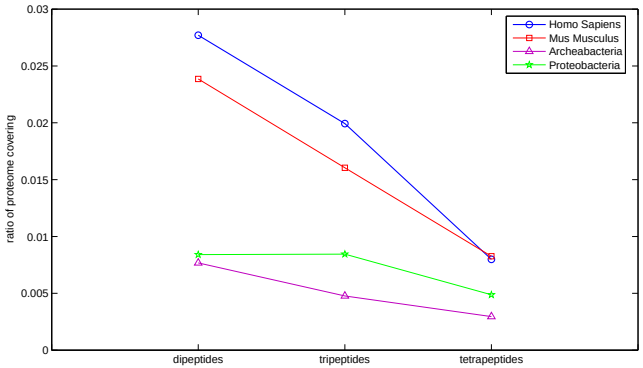
Figura 2.10: Frazione di proteoma ricoperto da motivi (vedi la (2.3)).



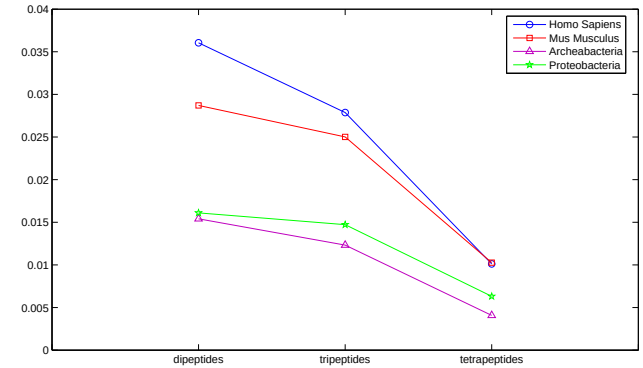
(a) 1%



(b) 1.5%



(c) 2%



(d) 2.5%

Figura 2.11: Copertura del proteoma avendo fissato la stessa percentuale di parole significative per ogni specie e per ogni k.

Capitolo 3

Grafi di motivi

If there is any area in which network thinking could trigger a revolution, I believe that biology is it.

A.L. Barabàsi

François Jacob nel 1977 affermava [22]:

> *Nature is a tinkerer and not an inventor.*

Jacob intendeva sottolineare il meccanismo con cui la selezione naturale e la pressione evolutiva agiscono: la natura non crea nuove strutture, ma tende a *ri*organizzare, *ri*combinare, *ri*ciclare strutture già esistenti. L'affermazione è particolarmente calzante se applicata all'ambito genetico, tra l'altro ambito di ricerca del suo autore. Infatti, a livello genetico l'espletamento di una nuova funzione biologica raramente coincide con la creazione di nuovo materiale genetico, ma piuttosto con il riarrangiamento di quello già presente.

Nel nostro caso i motivi non sono semplici oggetti a sé stanti, ma vengono correlati e combinati aggiungendo così nuovo contenuto informativo a quello presente se considerati come unità individuali. Siamo interessati ad analizzare proprio queste interazioni che condurranno alla formazione di strutture peptidiche di notevole interesse nello studio delle funzioni svolte da proteine: i domini funzionali.

3.1 Cos'è un dominio?

All'interno di una proteina, un *dominio* è una parte della sua sequenza che ha una struttura propria, ripiegandosi indipendentemente dal resto della catena peptidica. [31, 32]. La lunghezza di un dominio può andare dai 25 ai 500 aminoacidi, con una lunghezza media di circa 100. In natura si osservano spesso domini diversi assemblarsi per formare multidomini.

La maggior parte dei domini non è specifici di una proteina o di una famiglia di proteine, ma è presente in una grande varietà delle proteine di un organismo. Spesso ad un dominio è associata una particolare funzione biologica. I domini possono essere considerati delle vere e proprie unità genetiche mobili, utilizzate dalla natura per generare nuove sequenze [31, 32].

Proprio per via di queste caratteristiche, molti dei segmenti peptidici contenuti in un dominio risultano sovraespressi nel senso discusso nel capitolo 2 e prendono parte ai dizionari Z_k^{over}. Un dominio può quindi essere immaginato come una combinazione significativa (la nozione di combinazione significativa sarà discussa nel paragrafo 3.2.2) di motivi aminoacidici. Un modo quindi per localizzare

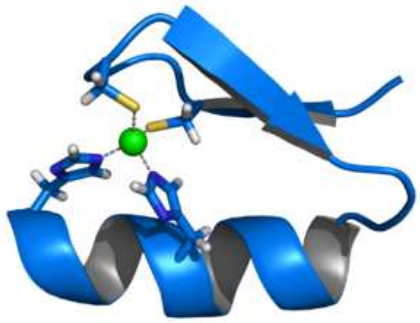


Figura 3.1: Un esempio di dominio. Si tratta dello *zinc finger*, uno tra i più diffusi essendo in grado di legarsi al DNA. I domini zinc finger sono in genere formati da circa trenta residui aminoacidici e sono così chiamati per via della presenza di un atomo di zinco. Sono caratterizzati da una α-elica sulla cui faccia si trovano residui di istidina (H) e da un foglietto β contenente due residui di cisteina (C). Sia l'istidina che la cisteina vanno ad interagire con l'atomo di zinco, dando stabilità alla struttura. Molto spesso i domini zinc finger possono essere collegati l'uno all'altro grazie ad interazioni testa coda.

domini proteici conosciuti o individuarne di nuovi è la ricerca di co-occorrenze di parole sovraespresse.

Partendo da questa ipotesi di lavoro, vogliamo studiare le correlazioni tra motivi nel nostro set di dati proteomici. A tal fine, stimeremo innanzitutto la probabilità che due parole cooccorrano in una finestra di l aminoacidi così da poter operare un confronto con le cooccorrenze osservate e individuare quali di esse siano significative. I risultati saranno poi studiati e analizzati facendo ricorso alla teoria dei grafi [7] e ad algoritmi di clustering [28, 29].

3.1.1 Probabilità di cooccorrenza di motivi

Utilizzando la stessa notazione introdotta nel capitolo 1, denotiamo con P un proteoma costituito da N_P proteine, per un totale di L aminoacidi. Indichiamo

con L_P la lunghezza media di una proteina; si ha ovviamente:

$$L_P = \frac{L}{N_P}.$$

Indichiamo con X e Y due parole di lunghezza k e con P_x e P_y la loro frequenza nel proteoma. Si vuole stimare il numero di volte che all'interno di una finestra lunga l aminoacidi la parola X sia seguita[1] dalla parola Y. La finestra non si deve sovrapporre tra due proteine.

Per fare ciò, stimiamo innanzitutto la probabilità che ciò accada in una sola proteina lunga L_P. Il calcolo può essere svolto in due fasi:

- Per ogni posizione della parola X dal primo aminoacido al $(L_p - l - 1)$-esimo, vi sono l possibili posizioni per la parola Y, costituite dai primi l aminaocidi successivi a X. Tenendo in conto la frequenza delle parole X e Y, il numero di cooccorrenze atteso è dato da: $(L_P - l - 1)lP_xP_y$.

- Quando la parola X è situata al $(L_p - l + i)$-esimo aminoacido, con $i = 1, \ldots, l$, per via della lunghezza finita della sequenza, vi sono solo $l - i$ possibili posizioni per la parola Y. In questo caso, il numero atteso di occorrenza è: $[(l-1) + (l-2) + \cdots + 1]\, P_xP_y = \frac{l(l-1)}{2} P_xP_y$

Questi due termini vanno sommati per ottenere il numero atteso totale per una proteina di lunghezza L_P. Infine per ottenere $N^{exp}(Y|X)_l$, ossia il numero atteso di volte che Y segua X in una finestra di lunghezza l in tutto il proteoma, è

[1] L'ordine delle due parole è importante e viene tenuto in conto nel nostro formalismo.

necessario moltiplicare per il numero totale di proteine. Si ha

$$\begin{aligned} N^{exp}(Y|X)_l &= \left[(L_P - l - 1)lP_xP_y + \frac{l(l-1)}{2}P_xP_y\right]N_P = \\ &= \left(L_P - \frac{l}{2} - \frac{3}{2}\right)lP_xP_yN_P. \end{aligned} \tag{3.1}$$

Questa quantità può essere confrontata con i conteggi $N^{obs}(Y|X)_l$, estratti dal proteoma.

3.2 Grafi di motivi

3.2.1 Cos'è un grafo?

Per poter rappresentare un sistema di interazioni come quelle intercorrenti tra motivi, torna molto utile la teoria dei grafi [23, 24]. Un *grafo* o *network* è un insieme di elementi detti *nodi* o *vertici* collegati fra loro da *archi* o *links*. Più formalmente, si dice grafo una coppia ordinata $\mathscr{G} = (\mathscr{N}, \mathscr{L})$ di insiemi, con $\mathscr{N}$ insieme dei nodi ed $\mathscr{L}$ insieme degli archi, tali che gli elementi di $\mathscr{L}$ siano coppie di elementi di $\mathscr{N}$, ovvero $\mathscr{L} \subseteq \mathscr{N}^2$. Talvolta in un link tra due nodi si definisce un verso. In tal caso il grafo si dirà *diretto* e gli elementi di $\mathscr{L}$ saranno coppie ordinate di elementi $\mathscr{N}$. Si può anche presentare la necessità di distinguere i link tra di loro, assegnando a ciascuno un peso. Un grafo di questo tipo si dice *pesato* [32, 26, 27]

Un grafo può essere facilmente rappresentato in forma di matrice. Nel caso di un grafo non pesato (pesato), si può infatti definire una matrice A, detta *matrice*

di adiacenza (*matrice dei pesi*), i cui elementi a_{ij} sono uguali a 1 (al peso del link) se esiste il link tra i nodi i e j, e sono zero altrimenti.

Si osservi che secondo la definizione di $\mathscr{G}$ data, in $\mathscr{L}$ possono essere presenti coppie costituite dallo stesso elemento di $\mathscr{N}$ o più coppie identiche fra loro. Ciò corrisponde alla presenza nel grafo di link che collegano un nodo con sé stesso (i cosiddetti *self-loop*) e di archi multipli tra due nodi. Nella costruzione del grafo di cui si parlerà nel paragrafo successivo queste eventualità verranno escluse [25].

3.2.2 Costruzione del grafo

Il grafo che si vuole costruire deve rappresentare le correlazioni tra motivi del proteoma e deve consentire di individuare i patterns di motivi che originano domini funzionali. Nei grafi che consideriamo, ogni nodo rappresenta uno dei motivi ad alta complessità di un certo dizionario Z_k^{over}; viene tracciato un arco diretto dal nodo che rappresenta la parola X al nodo che rappresenta la parola Y se è statisticamente significativo il numero di cooccorrenze delle due parole in una prefissata finestra di lunghezza l. Più rigorosamente il link è tracciato se vale la condizione:

$$\frac{N^{obs}(Y|X)_l}{N^{exp}(Y|X)_l} \geq \left\langle \frac{N^{obs}(Y|X)_l}{N^{exp}(Y|X)_l} \right\rangle + 2\sigma \tag{3.2}$$

dove σ rappresenta la deviazione standard della distribuzione $N^{obs}(Y|X)_l/N^{exp}(Y|X)_l$ (vedi analogia con 2.4.1). Il valore $N^{obs}(Y|X)_l/N^{exp}(Y|X)_l$ costituisce anche il peso del link. Il grafo che ne emerge è quindi un grafo diretto e pesato.

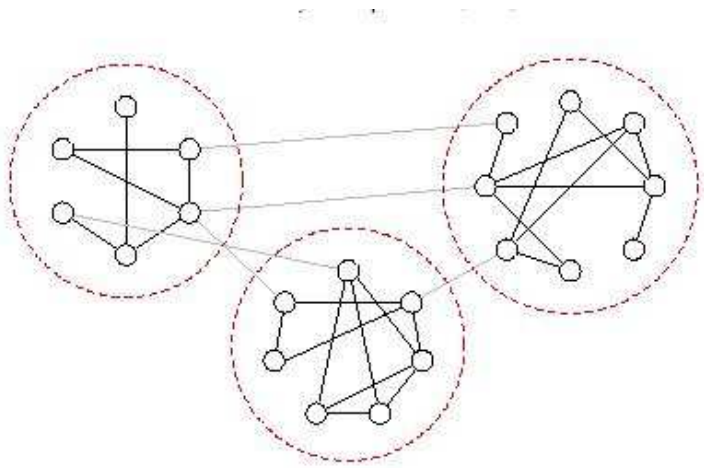


Figura 3.2: In un grafo le comunità possono essere definite come gruppi di nodi per i quali all'interno del gruppo vi è un'alta densità di archi. Nel caso mostrato in figura vi sono tre comunità, contrassegnati dai cerchi tratteggiati.

3.3 Presenza di comunità nella struttura di un grafo

Il concetto di comunità (o cluster) e la sua prima formalizzazione matematica sono stati proposti nell'ambito delle scienze sociali [33]. Dato un grafo $\mathscr{G} = (\mathscr{N}, \mathscr{L})$, una *comunità* (o *cluster*) è un sottografo $\mathscr{G}' = (\mathscr{N}', \mathscr{L}') \subseteq \mathscr{G}$ i cui nodi sono tra loro "fortemente coesi". Il concetto di coesione strutturale dei nodi $\mathscr{G}$ può essere quantificata in diversi modi; pertanto non esiste un'unica definizione formale di comunità [7].

In questo lavoro non ci interessa entrare nei dettagli sulle possibili definizioni alternative di comunità, né sui numerosi algoritmi per determinare le comunità presenti in un dato grafo. Intenderemo semplicemente per comunità un gruppo di nodi tale che la densità di archi "all'interno del gruppo" sia più alta della densità

di archi "tra gruppi" di nodi. Il concetto può essere ulteriormente espresso nella maniera seguente: una comunità è un gruppo di nodi tale che un *random walker* che inizia il suo cammino su uno di questi, rimarrà prevalentemente a muoversi sui nodi del gruppo. Questa seconda definizione è legata al metodo pratico con cui determineremo la presenza di comunità nei nostri grafi, metodo che sarà discusso nel paragrafo successivo [29].

La determinazione di clusters all'interno del nostro grafo ci permetterà di trovare quei gruppi di motivi che più frequentemente cooccorrono. Mostreremo come questi gruppi diano identità ai domini funzionali proteici.

3.3.1 Markov Clustering Algorithm

Il Markov Clustering (MCL) [28, 29] è un algoritmo basato su una procedura di bootstrapping applicata ad una matrice stocastica (o matrice di Markov) derivata dalla matrice di adiacenza del grafo. L'algoritmo è stato scelto perché uno dei pochi adatto ad essere usato per grafi diretti e pesati, nonché per la facilità nell'implementazione.

L'algoritmo

Come già introdotto nel paragrafo 3.3, i cluster di un grafo sono caratterizzati dalla presenza di un alto numero di archi tra i nodi che ne fanno parte. Per questo motivo ci si aspetta che per coppie di nodi appartenenti allo stesso cluster il valore del cammino più lungo sia relativamente basso; al contrario, per nodi facenti parte di cluster diversi esso assume valori relativamente alti. Ciò implica

che sia molto improbabile passare da un cluster all'altro mediante un random walk.

Per chiarire meglio quanto appena affermato, chiariamo il concetto di random walk su un generico grafo pesato. Supponiamo che il random walker si trovi, ad un dato istante, su un nodo i. Il nodo j su cui egli si troverà all'istante successivo viene scelto tra tutti i primi vicini di i, con una probabilità proporzionale al peso del link che si trova tra i e j. Così facendo, è possibile costruire una matrice di transizione M di dimensione $N \times N$, i cui elementi sono:

$$M_{ij} = \frac{a_{ij}}{\sum_m a_{im}}. \tag{3.3}$$

Si può facilmente verificare come M sia una matrice stocastica. Una matrice si dice *stocastica* (oppure *di Markov*) se soddisfa i seguenti requisiti:

- è quadrata;
- tutti i suoi elementi sono dei reali appartenenti all'intervallo $[0, 1]$;
- la somma di tutti gli elementi di una qualsiasi colonna vale 1.

L'algoritmo è basato su una procedura di bootstrapping: viene calcolata la probabilità dei random walk all'interno del grafo usando in maniera iterativa due particolari set di operatori. I due operatori sono chiamati operatori di *espansione* e *inflazione* e vanno applicati a matrici stocastiche. L'operatore di espansione applicato ad una matrice coincide con farne il quadrato secondo il normale prodotto riga per colonna. L'operatore di inflazione agisce su una matrice facendone la po-

tenza di Hadamard seguita da uno step di scaling, in modo tale che la matrice risultante sia ancora una matrice stocastica.

Formalmente: sia $M \in R^{k \times k}$ una matrice stocastica e $r > 0$ un numero reale; l'operatore di inflazione $\Gamma_r : R^{k \times k} \to R^{k \times k}$ agisce su M nella maniera seguente:

$$(\Gamma_r M)_{pq} = \frac{(M_{pq})^r}{\sum_{i=1}^{k} (M_{iq})^r}. \tag{3.4}$$

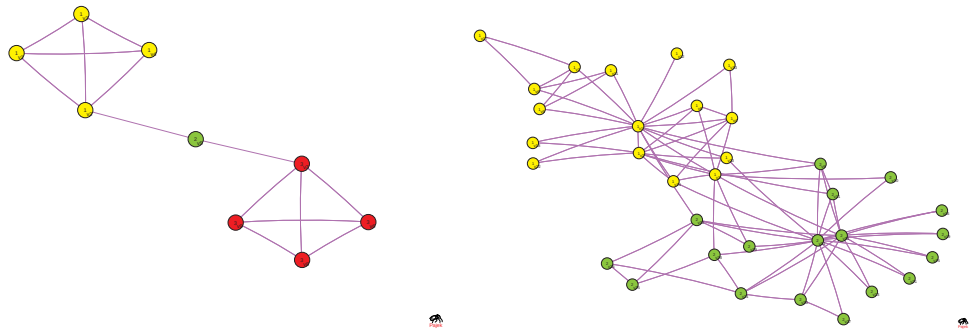
L'algoritmo consiste nell'applicazione successiva dell'operatore di inflazione alternato all'operatore di espansione. Il metodo converge ad una matrice doppiamente idempotente: essa non varia più sotto l'azione dei due operatori. La matrice ottenuta restituisce un grafo sconnesso in cui ogni componente contiene i nodi appartenenti allo stesso cluster nel grafo di partenza.

L'operatore di inflazione (3.4) dipende da un parametro r, detto *granularità*. L'incremento di questo parametro aumenta la "forza" dell'operatore di inflazione con l'effetto di aumentare il numero di cluster nella matrice di convergenza.

Implementazione e test dell'algoritmo

L'algoritmo è stato implementato in ambiente Matlab.

Per testare il corretto funzionamento dell'algoritmo, l'MCL è stato utilizzato su grafi la cui suddivisione in cluster fosse banale o già studiata. In particolare riportiamo i risultati di due test: il primo (figura 3.3 (a)) è un piccolo network di 9 nodi la cui suddivisione in comunità è più che intuitiva; il secondo (figura 3.3 (b)) è un network sociale di 34 nodi che rappresenta i rapporti di amicizia all'interno di un karate club [30]. I risultati dell'algoritmo sono per il primo network gli



(a) (b)

Figura 3.3: Test per il funzionamento dell'algoritmo MCL. I diversi colori dei vertici indicano appartenenza a cluster diversi; il valore del parametro r è stato settato a 2. (a) Piccolo network test formato da nove nodi la cui divisione in comunità è più che intuitiva. La suddivisione in cluster effettuata dall'algoritmo coincide con quella intuitiva. (b) Network sociale dei rapporti di amicizia in un karate club, studiato da Zachary. I risultati dell'MCL sono identici a quelli trovati in [30].

stessi della nozione intuitiva di comunità, per il secondo identici a quelli trovati dal sociologo W. Zachary.

3.4 Alcuni risultati

Il metodo per la costruzione e lo studio di un grafo di motivi descritto nei paragrafi precedenti è stato applicato ad uno dei proteomi del set introdotto nel capitolo 2. In particolare, si è scelto di studiare le correlazioni dei motivi ad alta complessità del dizionario Z_k^{over} nel proteoma umano (database CCDS).

Il network risultante consiste di 199 nodi, come si può verificare osservando la tabella 2.3. Fissando per la finestra di cooccorrenza $l = 36$, valore pari alla lunghezza della proteina più piccola, si tracciano 992 archi. Si osserva la presenza

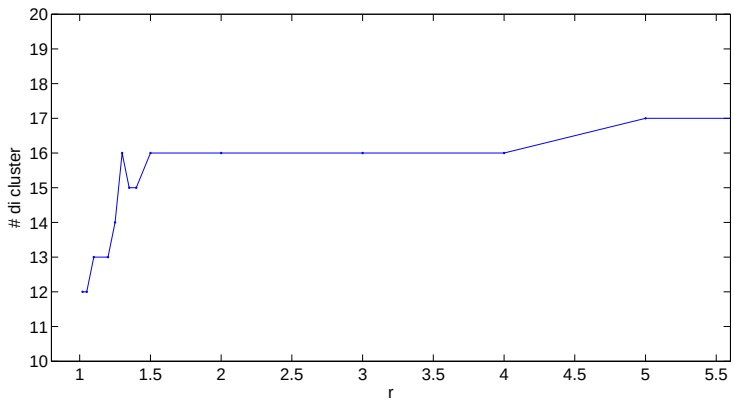


Figura 3.4: Numero di cluster in funzione del parametro di granularità r per il grafo studiato. Si osserva un lungo plateau per $r \in [1.5, 4]$

di nodi isolati (76), ovvero non collegati a nessun altro nodo.

3.4.1 Il parametro r

La determinazione dei cluster all'interno del grafo è in parte legata alla scelta del valore del parametro di granularità r, presente nell'implementazione dell'algoritmo MCL. L'ideatore dell'algoritmo, Stijn Van Dongen [28, 34], suggerisce di settare il valore all'interno dell'intervallo $[1.1, 5]$. Per avere una maggiore consapevolezza nella scelta del valore di r, abbiamo studiato al suo variare la variazione del numero di cluster. I risultati sono mostrati in figura 3.4.

Come si può osservare, esiste una largo plateau per l'intervallo $[1.5, 4]$ che fissa il valore del numero di cluster a 16. Risulta quindi ininfluente la scelta di un valore in tale intervallo. Nelle nostre elaborazioni successive è stato fissato $r = 2$.

3.4.2 Come si trovano i domini?

Una volta individuati i cluster mediante l'MCL, è necessario dare loro un'identità "biologica". Come già anticipato, si vuole verificare se i motivi appartenenti allo stesso cluster occorrono più frequentemente nella stessa classe di domini. In tal modo è possibile far corrispondere ad ogni cluster uno o più domini funzionali.

Al momento, l'unico espediente per questo tipo di analisi è una ricerca in database di domini[2] dando come chiave di volta in volta le coppie di motivi interessate. L'ovvio problema di questo tipo di analisi è legato al fatto che può essere svolta solo manualmente.

La rappresentazione del network nonché la classificazione in domini effettuata sono riportate in figura 3.5.

[2]Uno dei più curati nonché quello da noi utilizzato è il prosite [35].

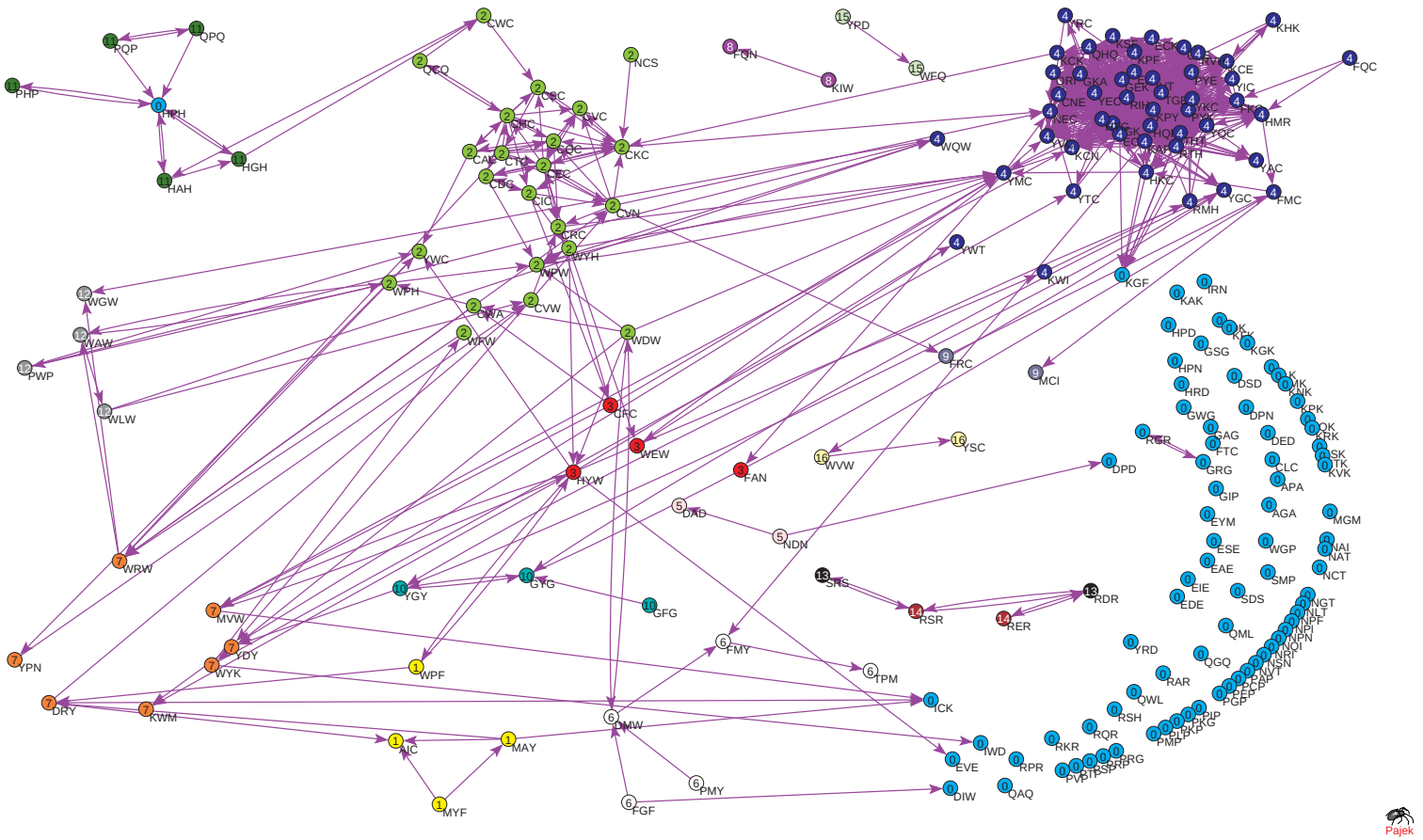


Figura 3.5: Grafo di motivi in cui i nodi appartenenti allo stesso cluster sono stati evidenziati con numeri e colori diversi. È stato utilizzato l'algoritmo MCL con $r = 2$.Caratterizzazione dei cluster: **1** (giallo) - olfactory receptor, **2** (verde chiaro) - cystein-rich tandem repeats, **3** (red) - non identificato, **4** (blu) - zinc finger, **5** (rosa) - cadherin, **6** (bianco) - olfactory receptor e serine-threonine kitinase, **7** (arancione) - non identificato, **8** (viola) - homeobox protein, **9** (grigio chiaro) - non identificato, **10** (azzurro) - keratin and nuclear ribonucleoproteins, **11** (verde scuro) - prevalentemente forkhead boxprotein, **12** (grigio) - tratti ricchi di triptofano, **13** (nero) - diversi domini proteici, di cui la maggior parte si trovano in RNA binding proteins, **14** (bordeaux) - diversi domini proteici, di cui la maggior parte si trovano in RNA binding proteins, **15** (verde acqua) - homeobox proteins (di tipo diverso da quelle del gruppo 8), **16** (giallo chiaro) - non identificato.

Conclusioni e sviluppi futuri

A conclusion is the place

where you got tired of thinking.

Martin Henry Fischer

L'analisi linguistica di proteoma e genoma è un argomento di ricerca tra i più caldi, perché solo recentemente si sono resi disponibili sequenziamenti completi. Vi è una miniera di informazioni biologiche ancora da interpretare, tra le quali capire come le caratteristiche genotipiche si riflettano sul fenotipo, originando differenze così spiccate nel corso dell'evoluzione.

In questo lavoro sono due i risultati più importanti ottenuti. Il primo riguarda il diverso utilizzo di parole e di motivi esistente tra i batteri e gli eucarioti superiori analizzati, ossia uomo e topo (vedi paragrafo 2.6). In particolare i dati sulla copertura del proteoma sono in accordo con l'alto numero di sequenze ripetute riscontrate nei proteomi degli eucarioti e dovute a fenomeni di duplicazione genica [36]. Il secondo importante punto del lavoro, del tutto nuovo in letteratura, è la definizione di network di motivi che ha fornito un metodo globale[3] per analizzare e mettere in relazione contemporaneamente più domini proteici. In

[3]Il metodo è globale in quanto permette di effettuare un'analisi di tutto il proteoma e non di un solo gruppo di sequenze.

particolare il metodo del network di motivi potrà risultare particolarmente utile per multidomini o domini sovrapposti in alcune particolari famiglie di proteine.

Il metodo ha quindi grosse potenzialità che vanno sviluppate escogitando un modo per rendere automatica l'individuazione dei domini dopo aver indentificato i cluster. In questo modo sarà possibile studiare applicazioni in cui i motivi del grafo siano tetrapeptidi, analisi ancora proibitiva per via dell'elevato numero di nodi (3365 nel caso dell'uomo).

Appendice A

Processi stocastici e catene di Markov

Le catene di Markov sono una particolare classe di processi stocastici discreti sia nello spazio che nel tempo[12].

Sia X una variabile casuale discreta i cui possibili stati spaziali appartengano ad un alfabeto $S = s_1, s_2, \cdots, s_N$ e la cui funzione di probabilità sia $p(x) \equiv Prob\,[X = x]\,,\ x \in S$. In altre parole, se x è una delle possibili realizzazioni delle variabile casuale X, allora la funzione $p(x)$ è la probabilità che la variabile casuale X assuma il valore x.

Un processo stocastico è una sequenza di variabili casuali. Indichiamo con $X^n = X_1, X_2, \cdots, X_n$ una sequenza di n variabili casuali per $n = 1, 2, \cdots$, eventualmente tra di loro legati da una dipendenza arbitraria. Un processo stocastico

è caratterizzato dalla funzione di probabilità congiunta:

$$p(x_1, x_2, \cdots, x_n) \equiv Prob\left[(X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n)\right] \tag{A.1}$$

dove $x_1, x_2, \cdots, x_n = x^n \in S^n$. In una forma più compatta si può anche scrivere $p(x^n) \equiv Prob\,(X^n = x^n)$. Un processo stocastico si può configurare come un sistema dinamico, vale a dire come un sistema il cui stato cambia nel tempo:

$$X_1 \to X_2 \to \cdots \to X_n$$

Il pedice delle variabili X rappresenta un indice temporale; l'espressione X_t sta ad indicare lo stato del sistema al tempo t, con $t = 1, 2, \cdots, n$. Ad ogni istante t il sistema può assumere uno degli stati dello spazio S. Una realizzazione del processo stocastico è una data sequenza del tipo $x_1, x_2, \cdots, x_n$, e rappresenta una possibile traiettoria di durata temporale n. In altre parole, il sistema dinamico si troverà nello stao x_1 all'istante $t = 1$, passerà poi allo stato x_2 all'istante $t = 2$ e così via fino all'istante $t = n$. In teoria, si possono avere N^n sequenze diverse di lunghezza n. In generale, però non tutte sono possibili, mentre alcune hanno una probabilità di occorrenza maggiore di altre. Tutto ciò è racchiuso nella funzione di probabilità congiunta (A.1).

Assieme alla funzione di probabilità congiunta, è utile definire la funzione di probabilità condizionale:

$$p\left(x_{n+1}|x_1, \cdots, x_n\right) = \frac{p\left(x_1, x_2, \cdots, x_{n+1}\right)}{p\left(x_1, x_2, \cdots, x_n\right)} \tag{A.2}$$

La probabilità condizionale $p(x_{n+1}|x_1, \cdots, x_n)$ rappresenta la probabilità che il sistema passi allo stato x_{n+1} al tempo $t = n + 1$ dopo aver percorso la sequenza $x_1, \cdots, x_n$.

Un processo stocastico discreto X^n si dice processio markoviano o catena di Markov se la funziona di probabilità congiunta ha la forma:

$$p(x_{n+1}|x_1, \cdots, x_n) = p(x_{n+1}|x_n) \tag{A.3}$$

per $\forall n$ e per $x_1, x_2, \cdots, x_{n+1} \in S$. Ciò implica che la distribuzione di probabilità congiunta si possa fattorizzare come segue:

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2|x_1) \cdots p(x_n|x_{n+1}) \tag{A.4}$$

Le catene di Markov per cui le probabilità condizionali $p(x_t|x_{t-1})$ non dipendono dal tempo (dall'indice temporale t) ma solo dagli stati spaziali si dicono catene di Markov stazionarie. Solitamente viene utilizzata la seguente notazione compatta:

$$p(X_t = x'|X_{t-1} = x) = \pi(x'|x) \tag{A.5}$$

La probabilità totale di una transizione dallo stato X_{t-1} allo stato X_t è ovviamente uguale a 1:

$$\sum_{x'} \pi(x'|x) = 1 \tag{A.6}$$

Le probabilità di transizione $\pi(x'|x)$ possono essere rappresentate in forma di matrice. Se si suppone di avere una catena di Markov con k stati, ossia $x \in S$ con $|S| = k$, le probabilità di transizione si scrivono nella forma di una matrice

$k \times k$:

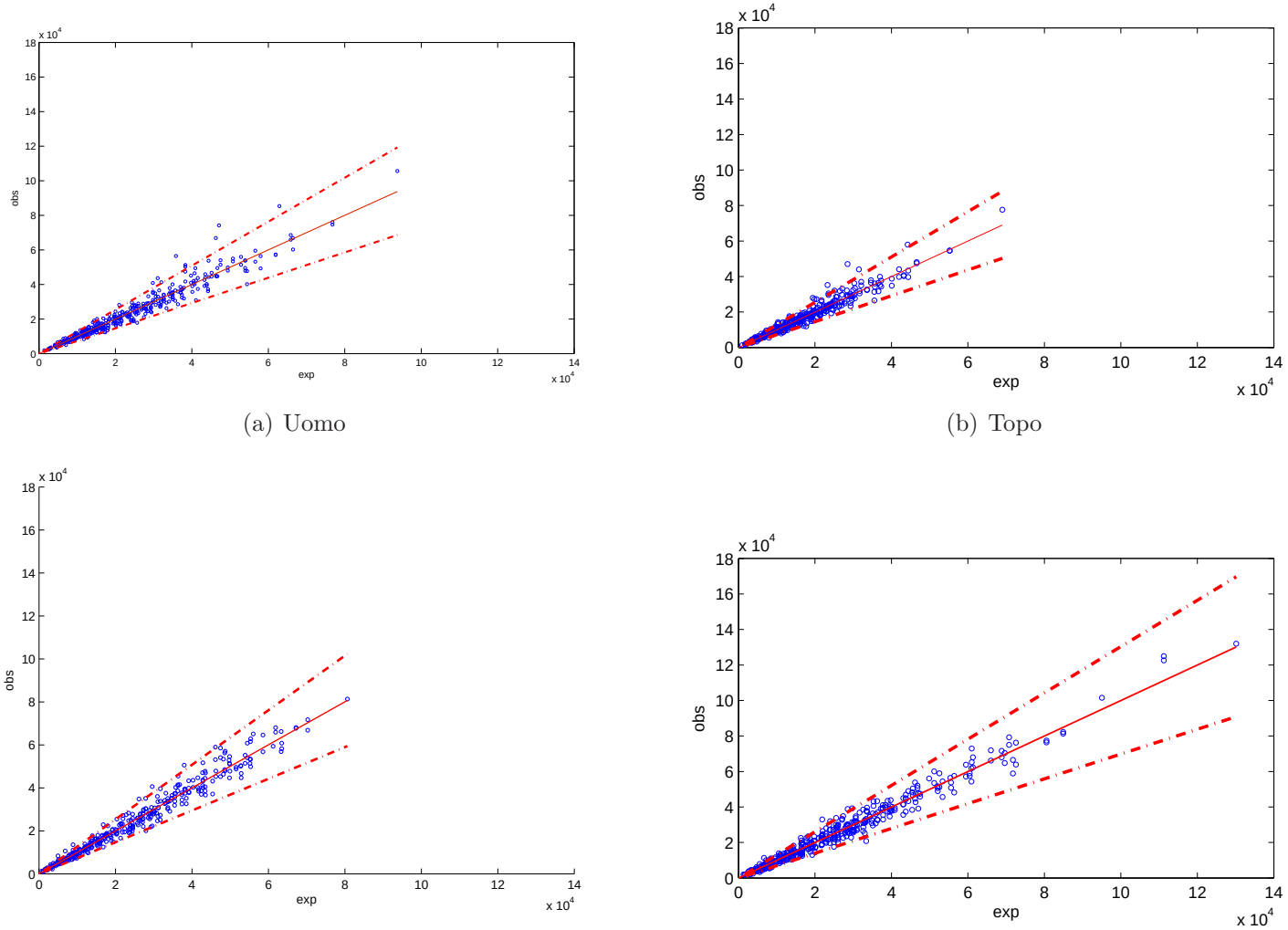
$$\Pi = \begin{pmatrix} \pi(1|1) & \pi(1|2) & \cdots & \pi(1|k) \\ \pi(2|1) & \pi(2|2) & \cdots & \pi(2|k) \\ \vdots & & \ddots & \vdots \\ \pi(k|1) & \pi(k|2) & \cdots & \pi(k|k) \end{pmatrix} \tag{A.7}$$

Si noti che l'elemento $\Pi_{ij} = \pi(i|j)$ rappresenta la probabilità di transizione dallo stato j allo stato i, e non il viceversa.

Appendice B

Ulteriori risultati e grafici

Si riportano per completezza in questa appendice tutti quei risultati e grafici che non sono stati inseriti nei capitoli precedenti a vantaggio della fluidità del testo.



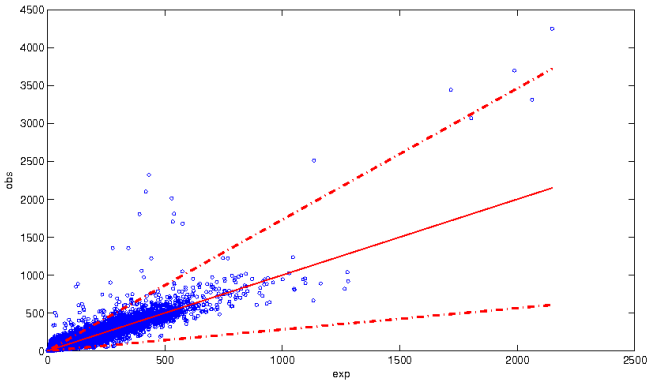
(a) Uomo

(b) Topo

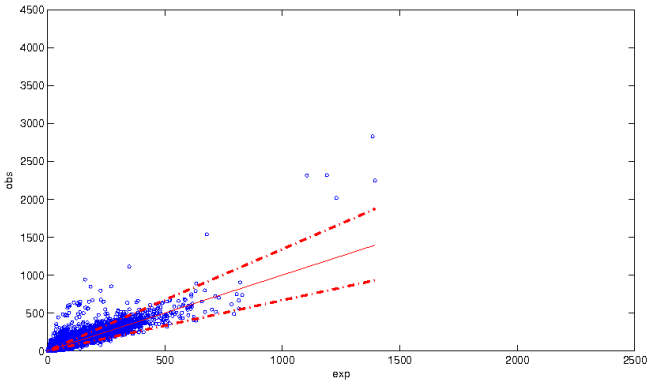
(c) Archibatteri

(d) Protobatteri

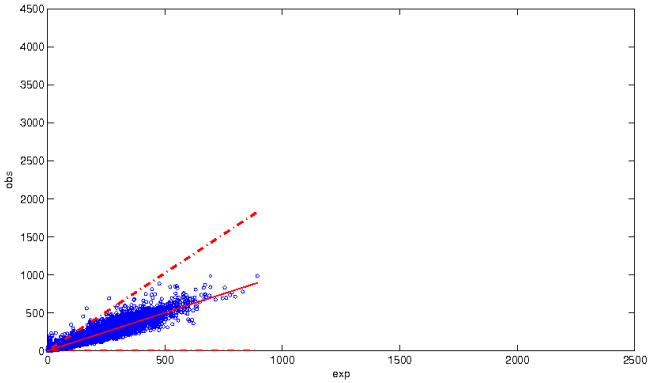
Figura B.1: Confronto del valore atteso (N^{exp}) rispetto al valore osservato (N^{obs}) per i dipeptidi in specie diverse. I punti al di fuori del angolo individuato dalle linee tratteggiate rappresentano le parole statisticamente significative (valore osservato è maggiore di due volte lo scarto quadratico medio) Se un tetrapeptide è rappresentato tanto quanto atteso si collocherà sulla bisettrice.



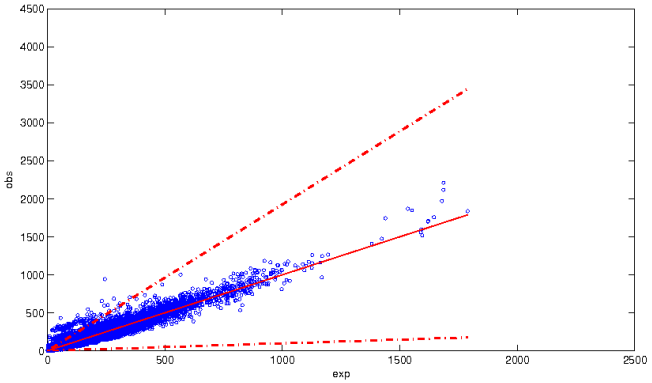
(a) Uomo



(b) Topo

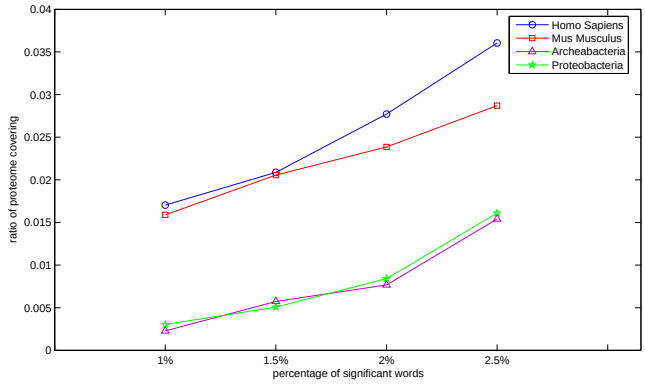


(c) Archibatteri

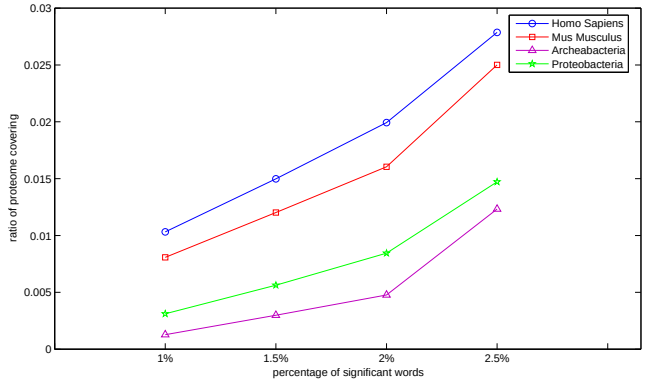


(d) Protobatteri

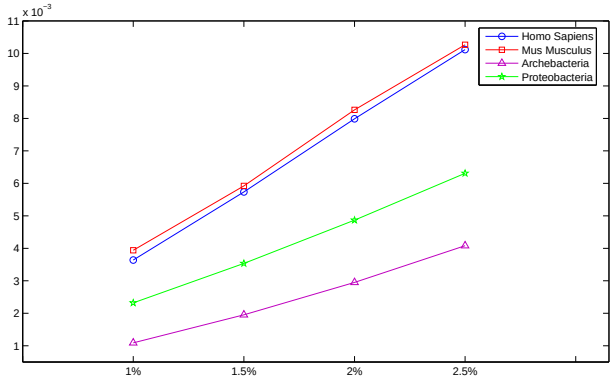
Figura B.2: Confronto del valore atteso (N^{exp}) rispetto al valore osservato (N^{obs}) per i tetrapeptidi in specie diverse. I punti al di fuori dell'angolo individuato dalle linee tratteggiate rappresentano le parole statisticamente significative (valore osservato è maggiore di due volte lo scarto quadratico medio) Se un tetrapeptide è rappresentato tanto quanto atteso si collocherà sulla bisettrice.



(a) dipeptidi



(b) tripeptidi



(c) tetrapeptidi

Figura B.3: Copertura del proteoma avendo fissato la stessa percentuale di parole significative per ogni specie e per ogni k.

Ringraziamenti

La genesi del lavoro qui presentato affonda le sue radici durante il mio primo anno di studi, quando seguivo alla SSC il corso di 'Ordine e disordine della materia' tenuto dal Prof. Vito Latora. In quel periodo ho scoperto come discipline tanto diverse come la fisica e la biologia possano interagire nell'ambito dello studio dei sistemi complessi. Successivamente il mio interesse per la biofisica e la biologia computazionale è stato ulteriormente approfondito frequentando il corso di 'Genomica e proteomica umana' tenuto sempre presso la SSC dal Prof. Daniele Condorelli. È stato proprio allora che ho capito che avrei voluto svolgere il lavoro di ricerca per la mia tesi in questo campo 'poco convenzionale'. L'opportunità mi è stata data proprio dal Prof. Latora e dal Prof. Condorelli, con cui ormai lavoro da circa due anni: a loro va la mia massima riconoscenza non solo per aver seguito questo lavoro di tesi con pazienza e manifestando la massima disponibilità, ma anche per avermi guidato e iniziato al mondo della ricerca.

Un particolare riconoscimento va al Prof. Andrea Giansanti e al Dott. Vittorio Rosato per i continui confronti e le interessanti discussioni fatte assieme; da loro è partita l'idea che ha portato alla nascita del progetto di ricerca presentato in questo lavoro. Un sentito ringraziamento va anche al Prof. Gioavnni Piccitto,

che in qualità di tutor nei primi quattro anni del mio corso di studi, mi ha sempre dato preziosi e utili consigli, e al Dott. Yamir Moreno per avermi ospitato a Zaragoza, presso il BIFI, dove ho potuto sfruttare particolare risorse di calcolo per completare questo lavoro, e per avermi fatto crescere nella conoscenza della teoria dei network complessi.

Enorme riconoscenza va poi al mio "team di informatici, ingegneri informatici e info-fisici", per avermi soccorso in tutti quei momenti in cui i computer sembrano assumere volontà propria... Grazie quindi a Salvo Scellato (The Tarro), a Sandro Meloni (Pigiamino), a Paolo Giarrusso (Blaisorblade), Antonio Mio (l'infofisico), a Filippo Bannò (Oppilif) e a Daniele Marletta (Shaoran).

A Federico e Alessandra, i miei più cari amici del periodo universitario, presenze su cui ho sempre potuto contare, va un grazie speciale per tutte le risate e l'allegria che mi hanno regalato, sia nei momenti lieti che in quelli difficili.

Il grazie più dolce va a te, Salvo, per essere sempre stato al mio fianco durante l'ultimo e più pesante anno di studi, per aver sempre mostrato interesse e curiosità nei confronti del mio lavoro, per avermi aiutato a stare bene ed essere serena e, non per ultimo, per avere pazientemente ascoltato decine di volte le presentazioni powerpoint per i talk :).

Il contributo maggiore, tuttavia, perché questo iter fosse portato a compimento è di tutti coloro che hanno sempre creduto in me e che per me hanno sempre tifato, perché senza 'supporters' è impossibile andare avanti... Quindi grazie Mamma, Claudia, Emma e Nonnino. E grazie anche a te Papà, che pur non avendo potuto assistere ai miei successi accademici, pur non avendo potuto seguirmi nella strada che mi ha condotto fin qui, mi hai regalato l'amore e la

passione per la logica, per la scienza e per la precisione.

Bibliografia

[1] Kennedy, D. and Norman, C., *What don't we know?*, Science **309** (2005) 78–102.

[2] V. Rosato, N. Pucello, G. Giuliano, *Evidence for cystein clustering in thermophilic proteomes*, Trends in Genetics, **18**:278-281 (2002).

[3] I. Rigoutsos, A. Floratos, *Combinatorial pattern discovery in biological sequences: the TIREISIAS algorithm*, Bioinformatics, **14**:55-67 (1998).

[4] Brendel, V., Beckmann, J.S., and Trifonov, E.N., *Linguistics of Nucleotide Sequences: Morphology and Comparison of Vocabularies*, Journal of Biomolecular Structure & Dynamics **4** (1986) 11-21.

[5] Berezovsky, I.N., Kirzhner, A., Kirzhner, V.M., Rosenfeld, V.R. and Trifonov, E.N., *Protein Sequences Yield a Proteomic Code*, Journal of Biomolecular Structure & Dynamics **21** (2003) 317-325 .

[6] Pietrokovski, S., *Comparing nucleotide and protein sequences by linguistic methods*, Journal of Biotechnology **35** (1994) 257-272.

[7] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D. U., *Complex networks: structure and dynamics*, Physics Reports **424** (2006) 175-308.

[8] Ferraro, L., Giansanti, A., Giuliano, G. and Rosato, V., *Co-expression of statistically over-represented peptides in proteomes: a key to phylogeny?*, arXiv:q-bio.MN/0410011 v2 at http://lanl.arXiv.org .

[9] Otaki, J. M., Ienaka, S., Gotoh, T. and Yamamoto, H., *Availability of short amino acid sequences in proteins*, Protein Sci. **14** (2005) 617-625.

[10] Giansanti, A., Bocchieri, M., Rosato, V. and Musumeci, S., *A functional homology between the chitinase from Plasmodium falciparum and human chitotriosidase explains self-sustaining of malaria*, Parasitol. Res. DOI 10.1007/s00436-007-0529-4.

[11] Qi, J., Wang, B. and Hao, B., *Whole proteome Prokaryote Phylogeny Without Sequence Alignment: a K-string Composition Approach*, J. Mol. Evol. **58** (2004) 1-11 .

[12] Cox, D.R. and Miller, H.D., *The theory of stocastic processes*, Chapman&Hall/CRC.

[13] http://www.ncbi.nlm.nih.gov/

[14] http://www.ebi.ac.uk/

[15] http://www.ncbi.nlm.nih.gov/RefSeq/

[16] http://www.ebi.ac.uk/IPI/

[17] http://www.ncbi.nlm.nih.gov/CCDS/

[18] I.K. Jordan et al., *A universal trend of amino acid gain and loss in protein evolution*, Nature, **433**:633-638 (2005).

[19] Hampikian, G. and Andersen, T., *Absent sequences: nullomers and primes*, Pacific Symposium on Biocomputing **12**:355-366 (2007).

[20] John C. Wootton and Scott Federhen, *Statistics of local complexity in amino acid sequences and sequences databases*, Computers Chem., vol.17 **2**:149-163 (1993).

[21] Romero P., Obradovic Z., Li X., Garner E.C., Brown, C.J. and Dunker, A.K., *Sequence complexity of disordered protein*, Proteins, **41**:38-48 (2001).

[22] Jacob, F., *Evolution and Tinkering*, Science, **196** 4295 (1977).

[23] Barabàsi, A.L. and Oltvai, Z., *Network biology: understanding the cell's a functional organization*, Nature Review Genetics, **5**:101–113 (2004).

[24] Albert, R., *Scale-free networks in cell biology*, Journal of Cell Science, **118**:4947–4957 (2005).

[25] Bollobàs, B., *Modern Graph Theory*, Graduate Texts in Mathematics, Springer, New York (1998).

[26] West, D. B., *Introduction to Grraph Theory*, Prentice-Hall, Englewood Cliffs, NJ (1955).

[27] Wasserman, K. Faust, *Social Networks Analysis*, Cambridge University Press, Cambridge (1994).

[28] Van Dongen, S., *Graph Clustering by Flow Simulation*, PhD thesis, University of Utrecht, May 2000.

[29] Enriht, A. J., Van Dongen, S. and Ouzounis C. A., *An efficient algorithm for large-scale detection of protein families*, Nucleic Acids Research, **30**:1575-1584 (2002).

[30] W.W. Zachary, *An information flow model for conflict. and fission in small groups*, J. Anthropol. Res. **33**:452 (1977).

[31] Valle, G., Helmer Citterich, M., Attimonelli, M. e Pesole, M. ,*Introduzione alla bioinformatica*, Zanichelli, 2003.

[32] www.wikipedia.org

[33] Wasserman, S. and Faust, K., *Social Networks Analysis*, Cambridge University Press, Cambridge (1994).

[34] http://micans.org/mcl/man/

[35] http://www.expasy.ch/prosite/

[36] Marcotte, E.M., Pellegrini, M., Yeates T. O. and Eisenberg, D., *A Census of Protein Repeats*, J. Mol. Biol., **293**:151-160 (1998).