Distribuzioni e Teoria dell'informazione

Algoritmi per l'intelligenza artificiale

Vincenzo Bonnici
Corso di Laurea Magistrale in Scienze Informatiche
Dipartimento di Scienze Matematiche, Fisiche e Informatiche
Universitá degli Studi di Parma

2025-2026

DNA e organizzazione

Tipicamente, la maggior parte del **genoma** di una cellula è memorizzata tramite il **DNA** (RNA).
Sappiamo giá che esistono vari **livelli di organizzazione** di questa macromolecola.
Se consideriamo il DNA come una sequenza di caratteri, ignorando la sua **natura tridimensionale**, allora alcuni **elementi strutturali** a livello di informazione sono:

- regioni trascrizionali, ovvero i geni
- regioni di regolazione, promotori inibitori e co-fattori
- hotpost di ricombinazione
- regioni dedicate alla topologia strutturale
- regioni legate a fattori evolutivi senza esporre una diretta caratterizzazione funzionale, che rimangono le meno studiate e le meno conosciute

Una infinitá di possibili genomi

Il DNA puó essere visto come un nastro su cui l'informazione genomica è memorizzata.
Il testo di tale nastro é scritto nel linguaggio nucleotidico, che é composto da 4 simboli (A, C, G, T).

Se fissiamo una lunghezza n, il numero totale di stringhe diverse aventi lunghezza n che possiamo formare é 4^n. Quindi, ad esempio, per una stringa lunga 100, possiamo avere 4^{100} possibili stringhe, che é un numero maggiore di 10^{60}.

Se pensiamo ai genomi reali, essi sono lunghi da qualche migliaio a miliardi di caratteri, il ché implicata una infinitá di possibili stringhe che possono essere **esplorate** dalla evoluzione.

Paradosso evolutivo

Si suppone che i primi proto-genomi siano state delle molecole di RNA, e l'adozione del DNA sia successiva ad esse.

In entrambi i casi, la macromolecola (RNA o DNA che sia) non é solo un supporto di memorizzazione della informazione genomica, in quanto essa é parte attiva del sistema cellula. Ovvero, esso implementa un sistema di calcolo con cui la cellula organizza i proprio processi interni e reagisce agli stimoli esterni.

Paradosso evolutivo

Siamo portati a pensare al DNA semplicemente come il nastro su cui é memorizzata l'informazione genomica. Tuttavia, esso é parte fondamentale della intelligenza computazionale della cellula.

La cellula é quindi un sistema di calcolo, il cui attore principale é il DNA. Esso possiede le informazioni per codificare gli altri elementi della cellula. tuttavia, il DNA ha bisogno che gli altri elementi siano già presenti nella cellula per poter esprimere tale informazione, e soprattutto che tali elementi siano **compatibili** con esso.

Nasce da quí un paradosso: é nato prima il genoma (RNA/DNA) o la cellula (membrana e il suo contenuto escluso il DNA/RNA)?

DNA e trasmissione della informazione

Un'altra importante funzione del DNA è la trasmissione della informazione in esso contenuta alla progènie.
In tale prospettiva, il DNA, ovvero il genoma da esso rappresentato, riassume la storia evolutiva di una specie.

Forze *strutturali* del genoma

Il genoma di una specie deve compiere anche un'altra funzione, ovvero quella di adattarsi ai cambiamenti tramite l'evoluzione.
Esso deve quindi esplorare le infinite possibilitá di cambiamento.

Per questo motivo, un genoma é soggetto a due forze distinte:

- strette regole organizzative che permettono di rappresentare l'informazione utile ad esprimere le sue funzionalità primarie
- adattabilitá alla nuove situazioni, che si manifesta come l'espressione di un processo di mutazione (spesso) randomico

Darwin

Nella teoria Darwiniana tre fattori sono fondamentali alla evoluzione:

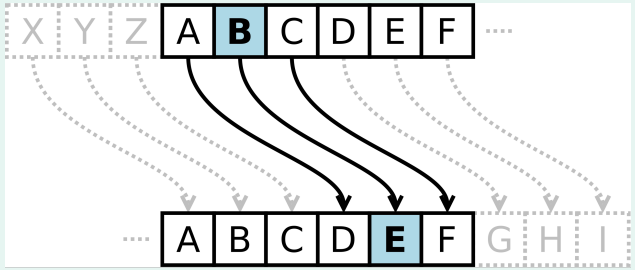
- ereditarietá
- variabilitá
- selezione naturale

La variabilitá, ovvero il processo di esplorazione, avviene principalmente durante la procreazione tramite fenomeni di ricombinazione, ma anche di alterazione casuale e errori nella copia del DNA.
Un aspetto cruciale della teoria di Darwin é che vi é un flusso bidirezionale dell'informazione tra specie e singolo individuo.

Basi della teoria dell'informazione

Il codice di Giulio Cesare

Nel settimo secolo A.C. Giulio Cesare usava crittografare i suoi messaggi con un codice prodotto da un cifrario a sostituzione, anche detto a scorrimento.



Rottura di Al-Kindi

Il codice venne forzato dal matematico Al-Kindi che si accorse che dato un testo scritto in un determinato linguaggio, se il teso é abbastanza lungo da poter trarne una informazione sufficiente, allora ogni lettera ha una specifica frequenza che si differenzia da tutte le altre lettere. Quindi é possibile rompere un cifrario a scorrimento semplicemente guardando in modo inverso le frequenze dei caratteri.
Questa applicazione puó essere considerato uno dei primi esempi di teoria dell'informazione.

Curve di Zipf

Il matematico statistico Zipf, autore della cosídetta legge di Zipf, introdusse il concetto di curva di Zipf che é stato applicato allo studio dei linguaggi naturali e non solo.

In tale curva vengono rappresentati gli elementi di una distribuzione, ordinandoli prima per la loro frequenza, ovvero calcolando il loro rango.

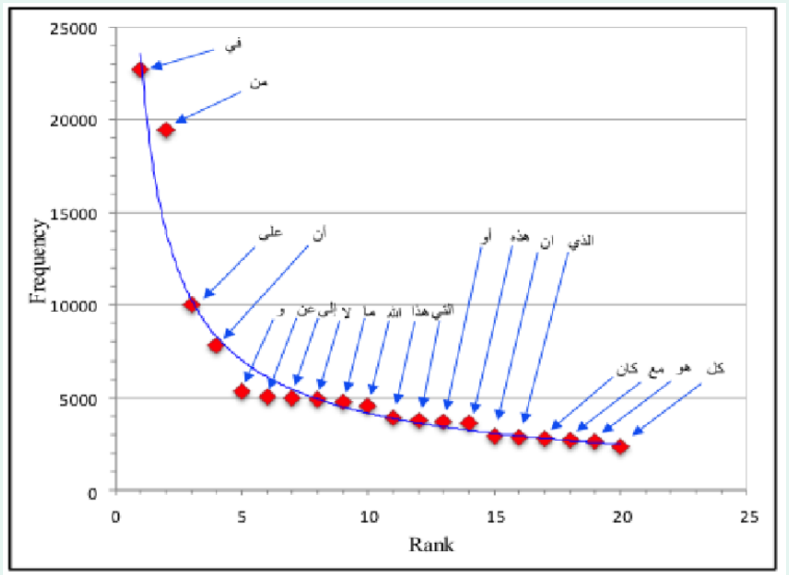


Figure: Curva di Zipf per alcune parole del linguaggio arabo.

Variabili e distribuzioni

Data una **variabile** X, indichiamo con $\hat{X}$ il suo
oointervallo di variabilitá, ovvero l'insieme dei valori che X puó assumere.

Una **distribuzione discreta** di n occorrenze e k oggetti ci informa su come le n occorrenze si distribuiscono sui k oggetti designati.
Una distribuzione é detta di **probabilitá** se la somma dei valori da essa assegnata é uguale ad 1, ovvero essa rappresenta delle probabilitá piú che delle molteplicitá.

Basi della teoria dell'informazione

Informazione

L'**informazione** associata ad un possibile elemento $x \in \hat{X}$ che avviene con probabilità $p(x)$ é data da

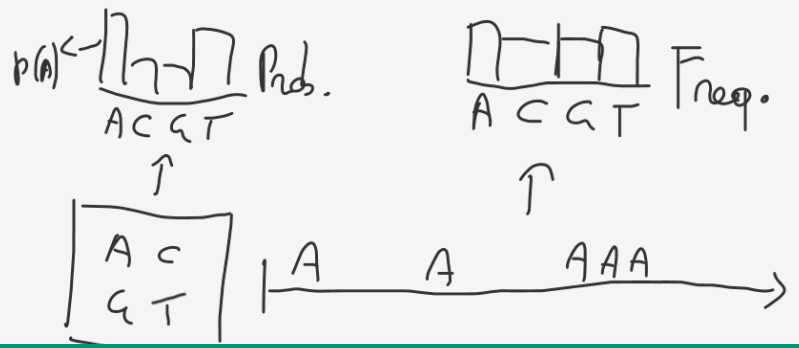
$$I(x \in \hat{X}) = log_2(\frac{1}{p(x)}) = -log_2(p(x))$$

Sorgete informativa

Una **sorgente informativa** é la coppia (X, p), dove X é una variabile e p é una funzione di probabilitá assegnata a tutti i suoi possibili valori.

Probabilitá vs frequenza

Bisogna sempre ricordarsi che probabilitá é un concetto che si esprime *a priori*, mentre la frequenza é un concetto che si esprime *a posteriori*



Entropia

L'entropia di una sorgente informativa si definisce come

$$H(X, p) = -\sum_{x \in \hat{X}} p(x) log_2(p(x))$$

essa quantifica l'informazione media della sorgente ed é, quindi, una media pesata della informazione. Il peso é dato dalla probabilitá stessa di ogni possibile valore/elemento.

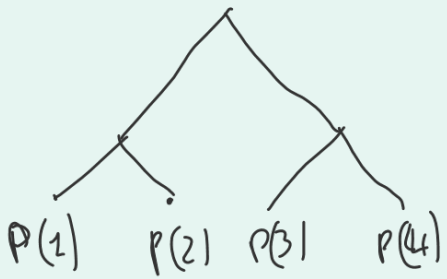
Prop. 1: proprietá di equipartizione dell'entropia

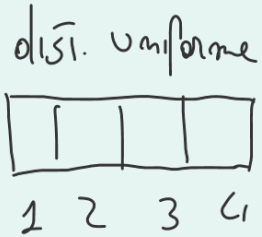
L'entropia raggiunge il suo valore massimo quando tutti gli elementi sono equiprobabili, ovvero quando p é uniforme.

$$\forall x \in \hat{X} \rightarrow p(x \in \hat{X}) = \frac{1}{n} \text{ per } |\hat{X}| = n$$

Un esempio concreto: alberi di decisione

Il giocatore A deve indovinare un numero che il giocatore B ha pensato.





$$-\sum_{i=1}^{4} p(x_i) log_2(x_i) = -\sum_{i=1}^{4} \frac{1}{4} log_2(\frac{1}{4}) = -\sum_{i=1}^{4} (\frac{1}{4} log_2(1)) - (\frac{1}{4} log_2(4))$$

$$= -\sum_{i=1}^{4} -(\frac{1}{4} log_2(4)) = log_2(4)$$

Proposizione 1.1

Se la base del logaritmo é n, allora l'entropia massima é 1.

$$H(X,p) = -\sum_{i=1\ldots n} p(x_i) log_n(p(x_i)) = -\sum \frac{1}{n} log_n \frac{1}{n}$$

$$= \frac{-n(log_n \frac{1}{n})}{n} = -(log_n \frac{1}{n}) = -(log_n(1) - log_n(n)) = log_n(n) = 1$$

□

Stringhe

Stringa

Dato un **alfabeto** Γ, una **stringa** α è una sequenza contigua di caratteri $a_1 a_2 \dots a_n$ tale che $a_i \in \Gamma$.
La **lunghezza** della stringa viene indicata con $|\alpha| = n$.

Indichiamo con λ la **stringa vuota**.

Sottostringhe

Indichiamo con $\alpha[i]$, con $1 \leq i \leq |\alpha|$, il carattere in posizione i-esima della stringa α.

Indichiamo con $\alpha[i, j]$, con $1 \leq i \leq j \leq |\alpha|$, la **sottostringa** di lunghezza $j - i + 1$ di α dalla posizione i alle posizione j entrambe incluse.

$\alpha[1, i]$, con $1 \leq i \leq |\alpha|$, é un **prefisso** di α.
$\alpha[i, |\alpha|]$, con $1 \leq i \leq |\alpha|$, é un **suffisso** di α.

Fattori

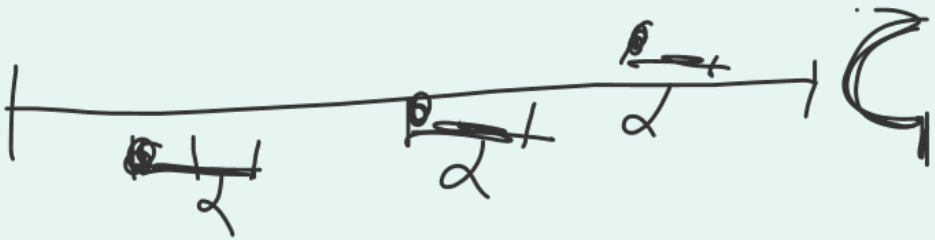
Dato un genoma $G \in \Gamma*$, l'insieme dai **fattori** (tutte le possibili sottostringhe) di G é dato da

$$D(G) = \{G[i,j] : 1 \leq i \leq j \leq |G|\}$$

Occorrenze

Data una parola $\alpha \in D(G)$, l'**insieme delle posizioni** in cui α occorre in G é dato da:

$$pos_G(\alpha) = \{i : G[i,j] = \alpha\}$$



La **molteplicitá** di α in G é data dal numero di occorrenze di α in G:

$$mult_G(\alpha) = |pos_G(\alpha)|$$

Genomi come stringhe

Occorrenze

- un **hapax** é una parola che occorre una solo volta, $\alpha : mult_G(\alpha) = 1$
- un **repeat** é una parola che occorre almeno due colte, $\alpha : mult_G(\alpha) > 1$
- un repeat α é **massimale** se $\forall x \in \Gamma \Rightarrow mult_G(\alpha x) = 1$
- un hapax $\alpha = (a_1, \ldots, a_{n-1}, a_n)$ è **minimale** se $mult_G((a_1, \ldots, a_{n-1})) > 1$
- un **nullomero** é una parola che non occorre in G, $\alpha : mult_G(\alpha) = 0 \Rightarrow \alpha \notin D(G)$

Elongazioni

Data una parola α, una **elongazione** di α é una parola αx con $x \in \Gamma$.

- dato un genoma G, data una parola $\alpha \in D(G)$, una elongazione di α in G é data data $\alpha x : \alpha x \in D(g)$
- un **memer** α é una parola tale che $\forall x \in \Gamma \rightarrow \alpha x \in D(G)$.

Genomi come stringhe

K-meri

Data una lunghezza di parola k, i fattori di G lunghi k sono anche detti k-mer.
L'insieme dei k-meri di G, per un determinato k é dato da

$$D_k(G) = \{\alpha \in D(G) : |\alpha| = k\}$$

Applicazioni pratiche

Conoscere l'insieme dei fattori o dei k-meri di un genoma é di essenziale importanza nella bioinformatica per molteplici applicazioni, tra cui ricostruibilitá e analisi dei genomi tramite NGS.

Distribuzioni

Una **distributione** $\phi : A \mapsto B$ é una funzione che mappa gli elementi dell'insieme A (dominio) agli elementi dell'insieme B (codomino). L'elemento caratterizzante di una funzione che rappresenta una distribuzione é il fatto che essa determina come gli elementi di A si distribuiscono secondo gli elementi di B.

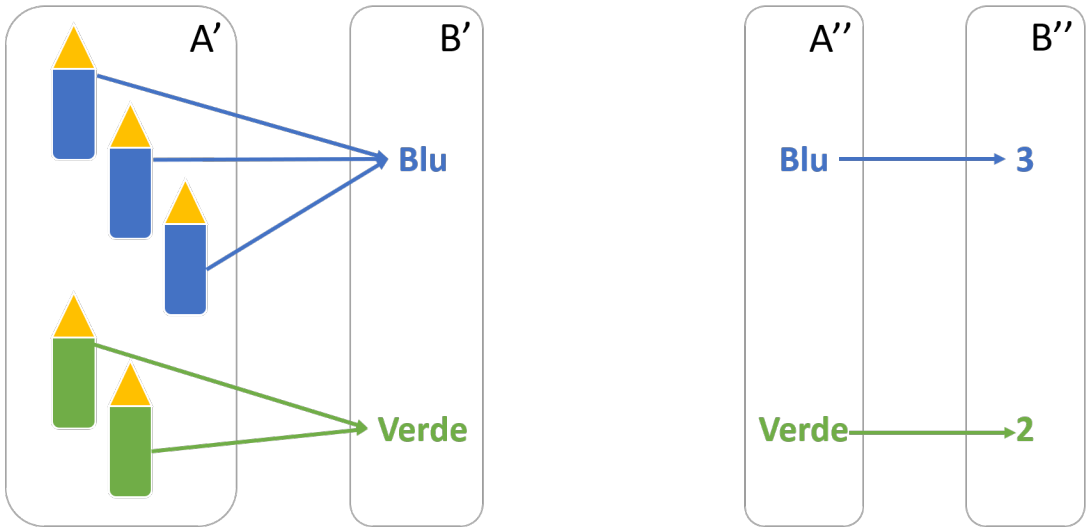
Distribuzione discreta

Una distribuzione é detta **discreta** se il suo dominio é un insieme discreto.

Distribuzione di molteplicitá

Una distribuzione é detta **di molteplicitá** se il suo codominio é l'insieme dei numeri naturali e se essa rappresenta una misura di molteplicitá, ovvero una quantitá.

Data una distribuzione originaria ϕ' avente dominio A' (le matite) ed codominio B (i colori), una distribuzione di molteplicitá ϕ'' mappa gli elementi di B in un insieme di molteplicitá A'' t.c. $A'' \subseteq \mathbb{N}$ e $\sum_{x \in B} \phi''(x) = |A'|$.



Distribuzione di frequenza

Una distribuzione é detta **di frequenza** se il suo codominio é l'insieme dei numeri reali $\mathbb{R}$ e se $\sum_{x \in A} \phi(A) = 1$.

Data una distribuzione di molteplicitá $\phi_M : A \mapsto B$, una distribuzione di frequenza puó essere ottenuta trasformando le molteplicitá, ovvero gli elementi di B, in frequenze. Si ottiene cosí una distribuzione $\phi_F : A \mapsto B'$ t.c. $\phi_F(x \in A) = \frac{\phi_M(x)}{\sum_{y \in A} \phi_M(y)}$.

Distribuzione di probabilitá

Una distribuzione di frequenza é detta **di probabilitá** se gli elementi del suo codominio rappresentano delle probabilitá.

Distribuzioni genomiche

Una distribuzione ci aiuta quindi a rappresentare in modo quantitativo la distribuzione dei possibili stati di un determinato fenomeno.
Essa é quindi un determinato punto di vista del fenomeno in esame e ne cattura solo un ben determinata proprietá quantitativa.
Punti di vista diversi ci aiutami a catturare e a studiare proprietá diverse dello stesso fenomeno.
Motivo per cui é possibile definire diversi tipi di distribuzione, in base al tipo di elementi del dominio e del codominio, che ci aiutano a catturare determinate proprietá, nel caso particolare delle stringhe di DNA.

Raggruppiamo quindi le distribuzioni in base al tipo di elemento del loro dominio.

Distribuzioni il cui dominio sono i k-meri

La **distribuzione della moltiplicitá di parola** informa della molteplicitá di ogni k-meri relativamente ad un genoma G preso in esame

$$\alpha \in \Gamma^k \mapsto mult_G(\alpha)$$

alternativamente

$$\alpha \in D_k(G) \mapsto mult_G(\alpha)$$

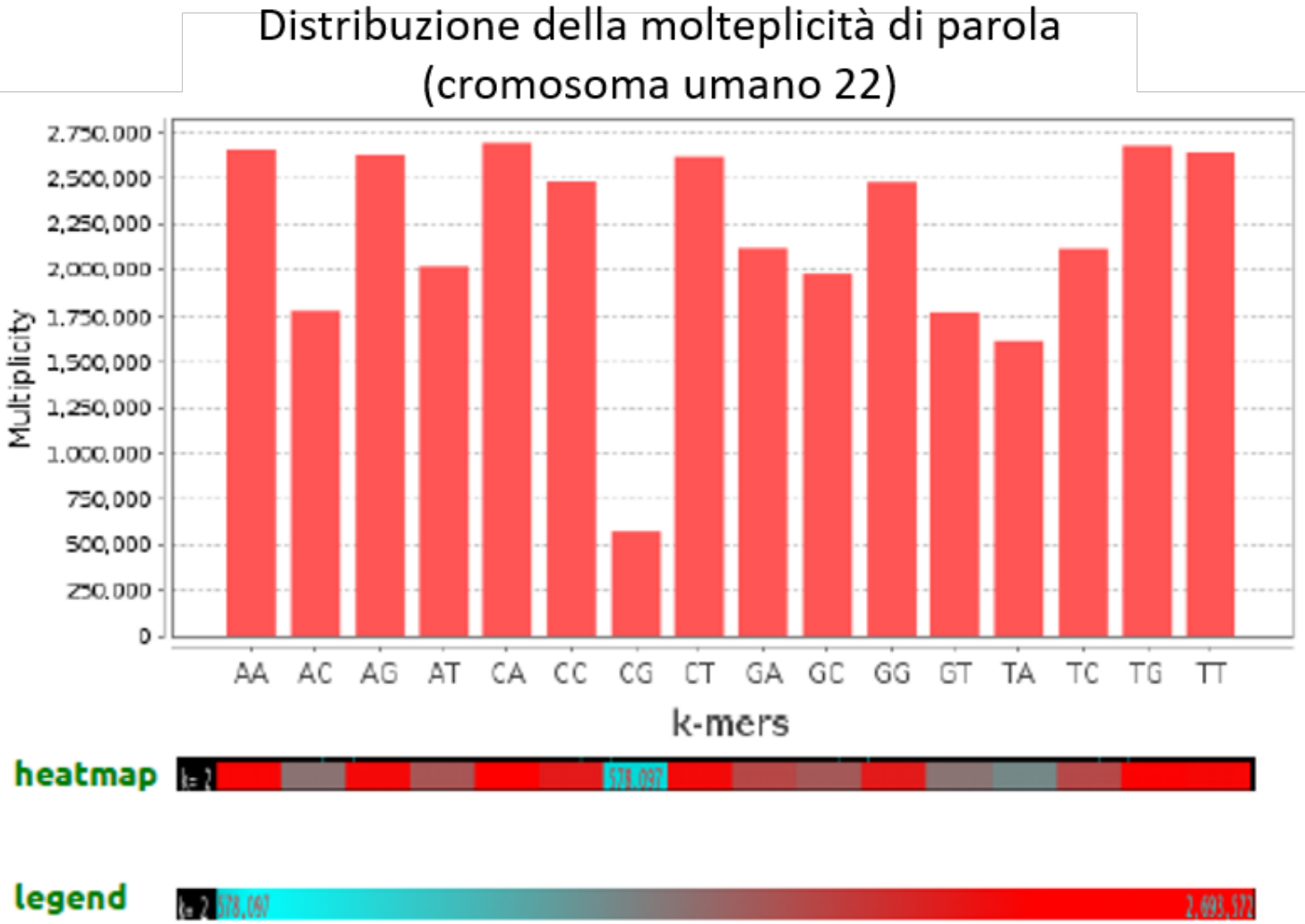
oppure

$$\alpha \in D(G) \mapsto mult_G(\alpha)$$

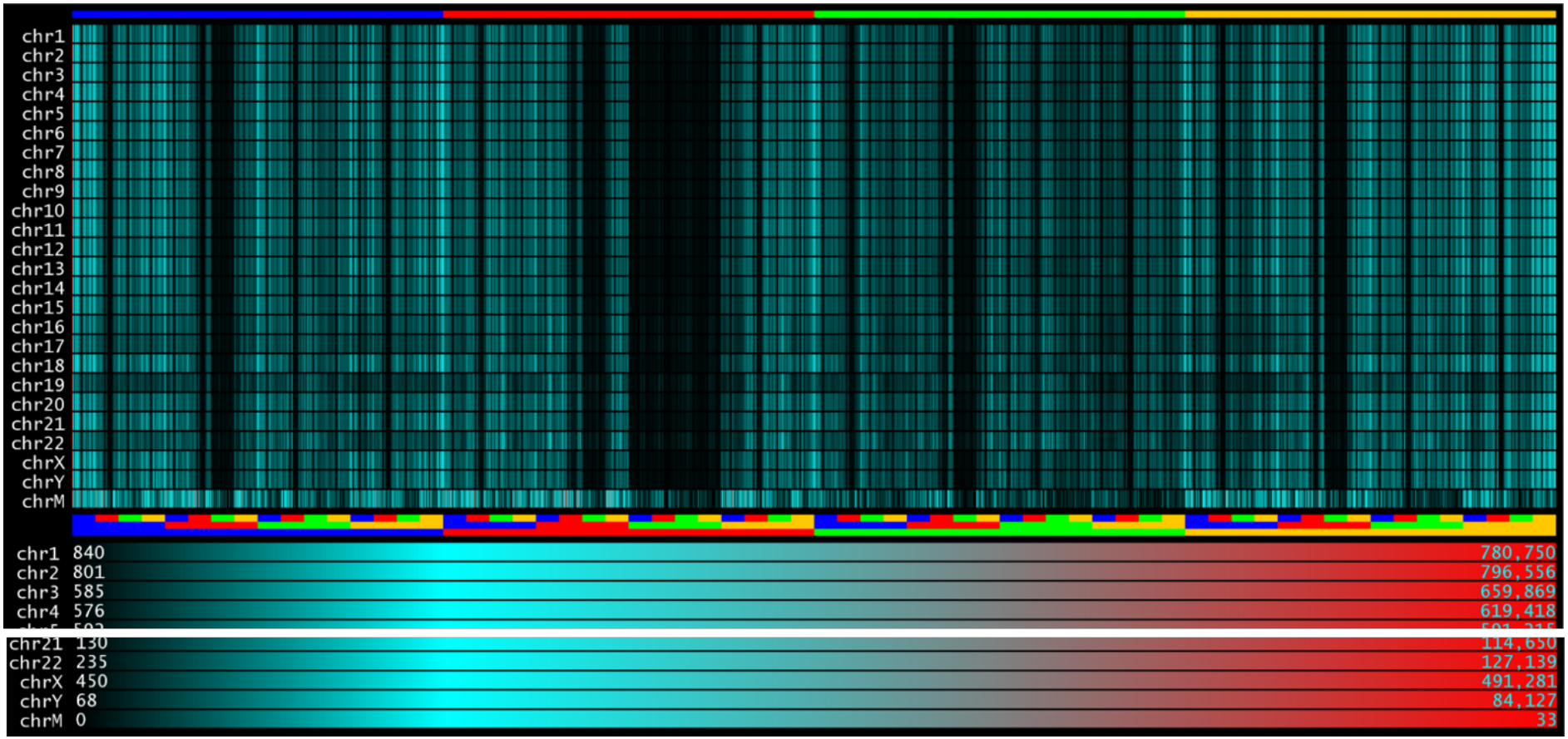
Analogamente la **distribuzione della frequenza di parola** informa della frequenza dei k-meri presenti in un genoma

$$\alpha \in D_k(G) \mapsto \frac{mult_G(\alpha)}{|G| - k + 1}$$

Distribuzioni genomiche

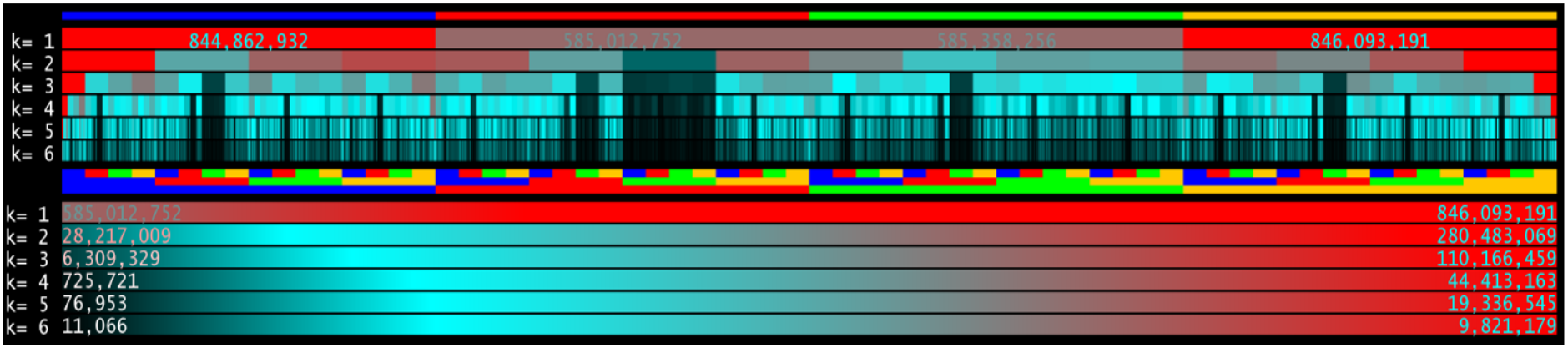


Distribuzioni genomiche

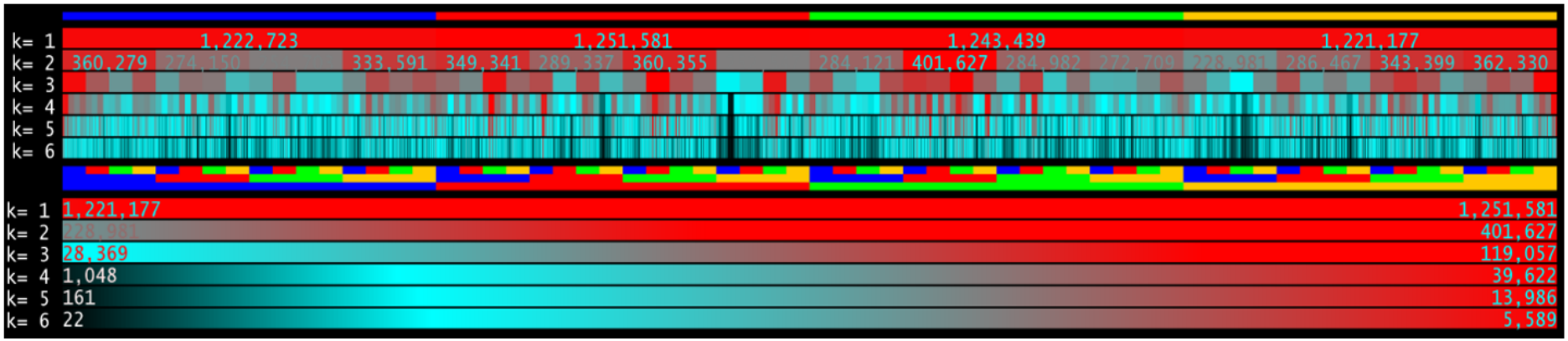


Distribuzioni genomiche

Homo sapiens



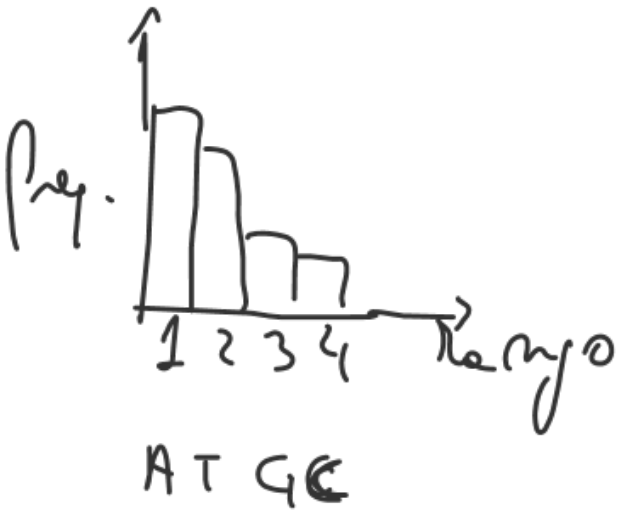
Escherichia coli



Distribuzioni il cui dominio é il rango

La **curva di Zipf** informa della molteplicitá di ogni rango, relativamente ad un genoma G preso in esame. I ranghi sono ottenuto ordinando i k-meri in base alla loro molteplicitá.

$$i = rank(\alpha) : \alpha \in D_k(G) \mapsto mult_G(\alpha)$$



Distribuzioni il cui dominio é la molteplicitá

Dato un valore di molteplicitá m_i, la **distribuzione di co-molteplictiá** informa circa il numero di parole aventi tale molteplicitá nello specifico genoma preso in considerazione.

Sia $M(G) = \{m_i : \exists \alpha \in D(G), mult_G(\alpha) = m_i\}$, allora

$$m_i \in M(G) \mapsto |\{\alpha : \alpha \in D(G), mult_G(\alpha) = m_i\}|$$

. Analogamente, sia $M_k(G) = \{m_i | \exists \alpha \in D_k(G), mult_G(\alpha) = m_i\}$, allora

$$m_i \in M_k(G) \mapsto |\{\alpha : \alpha \in D_k(G), mult_G(\alpha) = m_i\}|$$

.

Distribuzioni il cui dominio é la lunghezza di parola

La **distribuzione della lunghezza di parola** informa del numero di k-meri al variare di k

$$k \mapsto |D_k(G)|$$

La **distribuzione degli hapax** informa del numero di hapax al variare di k

$$k \mapsto |H_k(G)|$$

dove $H_k(G) = \{\alpha \in D_k(G) : mult_G(\alpha) = 1\}$

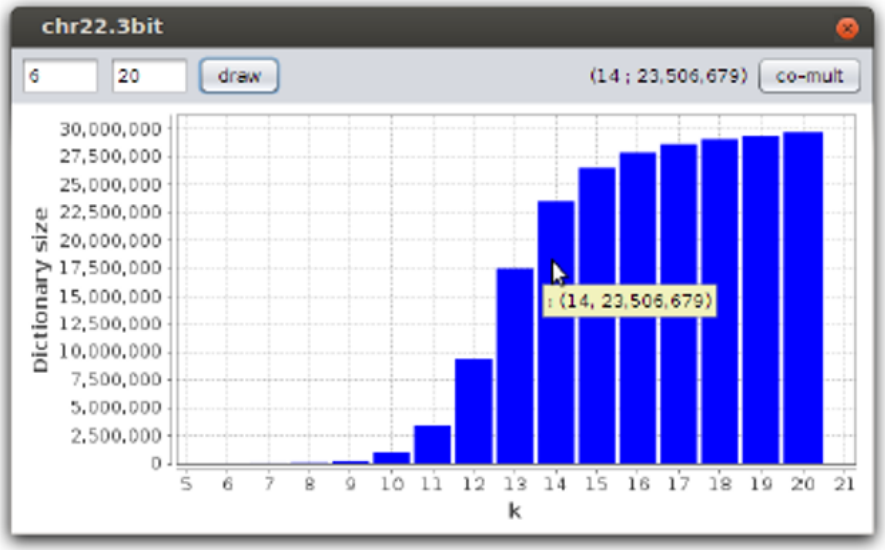
La **distribuzione dei repeat** informa del numero di repeat al variare di k

$$k \mapsto |R_k(G)|$$

dove $R_k(G) = \{\alpha \in D_k(G) : mult_G(\alpha) > 1\}$

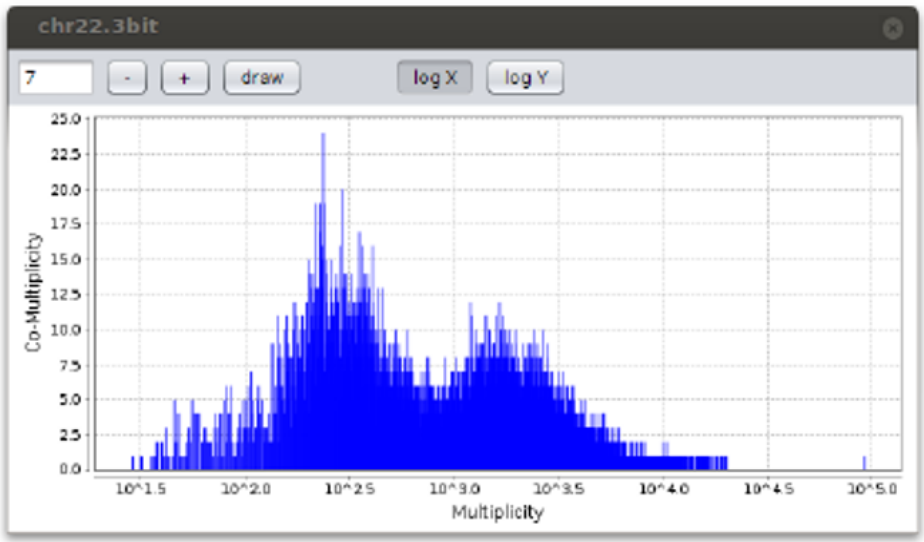
Distribuzioni genomiche

Distribuzione della lunghezza di parola



Chromosome 22 of
Homo sapiens

Distribuzione di co-molteplicità



Indici statistici delle distribuzioni

Il **momento** di origine m e ordine k di una variabile casuale X é definito come il valore atteso della k-esima potenza dei valori di X.

Di seguito si assume X variabile aleatoria con dominio in valori reali $x_1, x_2, \ldots, x_n$ rispettivamente con probabilitá $p_1, p_2, \ldots, p_n$.

1° momento = media (o valore atteso)

$$E[X] = \sum_i p_i x_i$$

Indici statistici delle distribuzioni

2° momento

$$E[X^2] = \sum_i p_i x_i^2$$

Varianza = deviazione dalla media del 2° momento

$$var(X) = E[(x - \mu)^2] = \sum_i p_i (x_i - \mu)^2$$

dove μ é la media.

Varianza = deviazione dalla media del 2° momento

Inoltre (utile per calcolare in modo iterativa la varianza)

$$var(X) = E[X^2] - (E[X])^2$$

,infatti

$$\sum_i p_i(x_i - \mu)^2 = \sum_i p_i x_i^2 + \sum_i p_i \mu^2 - 2\mu \sum_i p_i x_i$$

$$= E[X^2] + \mu^2 - 2\mu^2 = E[X^2] - (E[x])^2$$

Deviazione standard

La deviazione standard, ovvero di quanto in media mi discosto dalla media, é definita come:

$$sd(X) = \sqrt{var(X)}$$

Coefficiente di variazione

É definito come

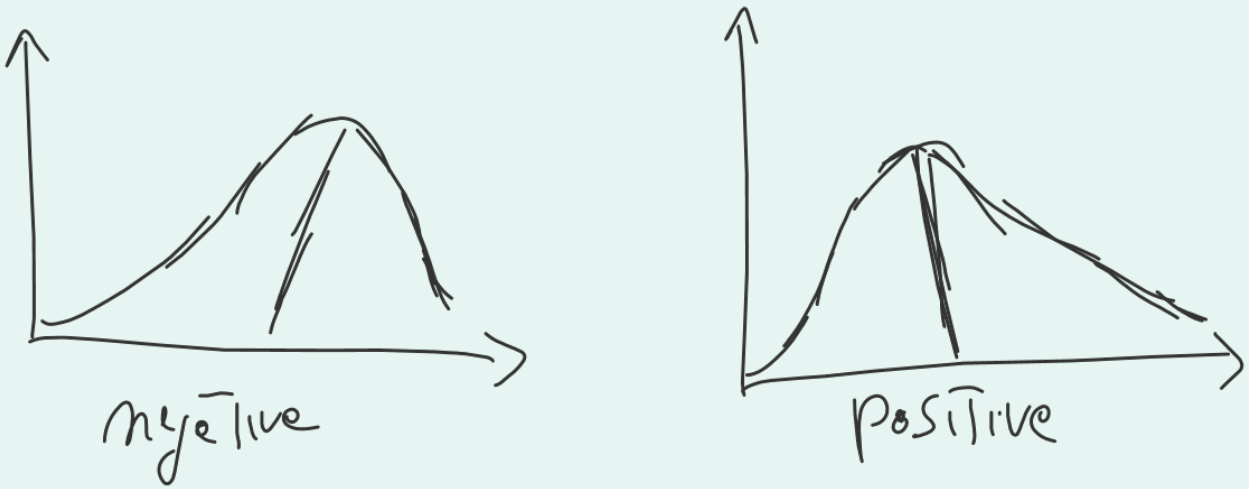
$$\frac{sd(X)}{|\mu|}$$

3° momento = skewness

É una misura di asimmetria definita come

$$\gamma_1 = E[(\frac{x - \mu}{\sigma})^3]$$

dove μ é la media e σ é la deviazione standard.



Indici statistici delle distribuzioni

3° momento = skewness

Inoltre,

$$\gamma_1 = \frac{E[X^3] - 3\mu E[X^2] + 3\mu^2 E[X] - \mu^3}{\sigma^3}$$

$$= \frac{E[X^3] - 3\mu(E[X^2] - \mu E[x]) - \mu^3}{\sigma^3}$$

$$= \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3} = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{\frac{3}{2}}}$$

4° momento = indice di Kurtosis

É un indice di allungamento, o appiattimento, della coda della distribuzione.

$$Kurt[X] = E[(\frac{X-\mu}{\sigma})^4] = \frac{E[(X-\mu)^4]}{(E[X-\mu)^3])^2}$$

Indici di dispersione e di diversitá

Un **indice di dispersione** riassume la misura con la quale i valori di una distribuzione statistica quantitativa sono distanti da un valore centrale (media o mediana).

Se la variabile statistica in esame é di tipo nominale, esso allora prende il nome di **indice di diversitá**.

Una variabile é detta **nominale** quando il carattere (statistico) che stiamo studiando assume stati discreti non ordinabili.

Indici di dispersione e di diversitá

Indice di Simpson

L'indice di Simposon é un misura quantitativa che riflette quanti tipi (specie) diverso ci sono in un dataset (comunitá).

Tale indice é una rappresentazione statistica della biodiversitá in tre aspetti principali:

- richezza (richness): il numero di specie diverse
- equitabilitá: grado di omogeneità col quale gli individui sono distribuiti nelle varie specie che compongono una comunità
- dominanza: il grado con cui una specie é dominante verso le altre specie facenti parte della comunitá

Indice di Simpson

Esso é definito come

$$\lambda = \sum_{i=1}^{R} p_i^2$$

dove R é la ricchezza, ovvero il numero di specie distinte nella comunitá, e p_i é l'abbondanza della i-esima specie. Esso rappresenta la probaiblitá che due entitá prese a caso dal dataset siano dello stesso tipo (specie).

Indice di Shannon (entropia di Shannon)

L'indice di Shannon, detto anche comunemente entropia, é un indice di diversitá basato sulla media geometrica pesata della abbondanza proporzionale delle specie di una comunitá.

$$H = -\sum_{i=1}^{R} p_i ln(p_i) = -\sum_{i=1}^{R} ln(p_i^{p_i})$$

$$= -(ln(p_1^{p_1}) + ln(p_2^{p_2}) + \dots ln(p_R^{p_R}))$$

$$= -ln(p_1^{p_1} p_2^{p_2} \dots p_R^{p_R}) = ln(\frac{1}{p_1^{p_1} p_2^{p_2} \dots p_R^{p_R}}) = ln(\frac{1}{\prod_{i=1}^{R} p_i^{p_i}})$$

Numeri di Hill

I numeri di Hill sono una generalizzazione delle misure di diversitá.
Dato un ordine q, la diversitá é misurata prendendo in considerazione il reciproco della media generalizzata M_{q-1} della abbondanza proporzionale dei tipi nel dataset:

$$ {}^{q}D = \frac{1}{M_{q-1}} = \frac{1}{\sqrt[q-1]{\sum_{i=1}^{R} p_i p_i^{q-1}}} = (\sum_{i=1}^{R} p_i^q)^{\frac{1}{1-q}} $$

Numeri di Hill

Per $q = 0$ é in pratica la media della abbondanze.
Per $q = 1$ non é definita, tuttavia si puó calcolare il limite per q che tende a 1 e si ottiene

$$^1D = \frac{1}{\prod_{i=1}^{R} p_i^{p_i}} = e^{-\sum_{i=1}^{R} p_i ln(p_i)}$$

L'entropia di Rényi é una generalizzazione della entropia di Sannon, per $q \neq 1$

$$^qH = \frac{1}{1-q} ln(\sum_{i=1}^{R} p_i^q) = ln(\frac{1}{\sqrt[q-1]{\sum_{i=1}^{R} p_i p_i^{q-1}}}) = ln(^qD)$$

Per $q = 2$ equivale all'inverso dell'indice di Simpson

$$^2D = \frac{1}{\lambda}$$

Confrontare distribuzioni

Riassumere una distribuzione in un solo valore numero, un indice, puó essere spesso svantaggioso, soprattutto se lo scopo é confrontare distribuzioni tra loro.

A tale fine, possiamo utilizzare misure di

- similaritá/dissimilaritá tra insiemi
- vicinanza/prossimitá per spazi vettoriali
- correlazione tra variabili
- divergenza tra distribuzioni di probabilitá.

Confrontare distribuzioni

Similiartá di Jaccard (per insiemi)

Ci dice quanto due insiemi siano simili, ovvero quanto hanno in comune. Essa puó essere ad esempio definita sui dizionari dei k-meri di due genomi G_1 e G_2:

$$J_k(G_1, G_2) = \frac{|D_k(G_1) \cap D_k(G_2)|}{|D_k(G_1) \cup D_k(G_2)|}$$

Similiartá di Jaccard generalizzata (per multiinsiemi)

Ci dice quanto due multiinsiemi siano simili, ovvero considera anche le molteplitiá degli elementi. Sia $D_k(G_1, G_2) = D_k(G_1) \cup D_k(G_2)$, allora

$$J'_k(G_1, G_2) = \frac{\sum_{\alpha \in D_k(G_1, G_2)} \min(mult_{G_1}(\alpha), mult_{G_2}(\alpha))}{\sum_{\alpha \in D_k(G_1, G_2)} \max(mult_{G_1}(\alpha), mult_{G_2}(\alpha))}$$

Si puó facilmente dimostrare che entrambe sono definite a valori in $[0 \dots 1]$.

Confrontare distribuzioni

Distanze

Dato un genoma G, possiamo immaginarlo come in vettore in uno spazio n dimensionale. Dove le dimensioni sono i suoi k-meri.

Se peró si vogliono confrontare due genomi, le dimensioni da prendere in considerazione devo essere le medesime.
Motivo per cui, come dimensioni possiamo usare Γ^k oppure $D_k(G_1) \cup D_k(G_2)$.
Per k relativamente piccoli, Γ^k ci dá la possibilitá di mettere in relazione i confronti di piú genomi!

Distanza del coseno

Misura il coseno dell'angolo formato da due vettori in uno spazio n dimensionale, $A, B \in \mathbb{R}^n$:

$$cos\_sim(A, B) = \frac{< A|B >}{< A|A >< B|B >}$$

$$= \frac{\sum_{i=1}^{n} A[i]B[i]}{\sqrt{\sum_{i=1}^{n} A[i]^2}\sqrt{\sum_{i=1}^{n} B[i]^2}}$$

dove $< | >$ é i prodotto vettoriale e $A[i]$ é il valore del vettore A nella dimensione i-esima.

Si puó dimostrare che se lo spazio é binario, ovvero $\{0, 1\}^n$ allora tale misura equivale al coefficiente di Tanimoto

Distanza di Minkowski

É una metrica in uno spazio vettoriale normato (ovvero uno spazio su cui é definita una norma).
Dato $k \in \mathbb{N}^+$ e due vettori $A, B \in \mathbb{R}^n$, é definita come:

$$d_k(A,B) = (\sum_{i=1}^{n} |A[i] - B[i]|^k)^{\frac{1}{k}} = \sqrt[k]{\sum_{i=1}^{n} |A[i] - B[i]|^k}$$

É la generalizzazione della distanza Euclidea, per $k = 2$, e della distanza di Manhattan, per $k = 1$.

Confrontare distribuzioni

Distanza di Hamming

Conta il numero di posizioni in cui due vettori divergono (non hanno gli stessi valori).
É una distanza di Mahanattan per vettori booleani.
É definita come:

$$h(A, B) = \sum_{i=1}^{n} |A[i] - B[i]|$$

Puó essere normalizzata dividendo gli addenti per il valore massimo.

Correlazione di Pearson

Ci informa del livello di concordanza tra due variabili reali, ovvero il grado di concordanza del loro cambiamento.

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

dove $cov(X, Y)$ é la **covarianza** tra i due vettori X e Y, definita come la somma dei prodotti delle deviazioni rispetto alla media, ovvero:

$$cov(X, Y) = E[(X - E[X])(Y - E[Y])] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

. Se le due variabili sono **concordanti**, quindi **correlate**, la covarianza é > 0. Se le due variabili sono **disconcordanti**, quindi **anti-correlate**, la covarianza é < 0.
σ_X e σ_Y sono le deviazioni standard, che sono sempre > 0.

Confrontare distribuzioni

Correlazione di Spearman

Molto simile alla correlazione di Pearson, ma i valori sono confrontati tramite il rango. x_i non é quindi il valore i-esimo della variabile X ma il valore per cui i é l'i-esimo rango.

Correlazione di Kendall

Conta il numero di inversioni di rango tra coppie contigue tra le due variabili.

$$r = \frac{\#\text{ coppie concordanti} - \#\text{ coppie discordanti}}{\frac{n(n-1)}{2} = \#\text{coppie totali}}$$

Divergenza di Kullback-Leibler

Date due distribuzioni di probabilitá, P e Q, **definite sullo stesso dominio** D, la divergenza di Kullback-Leibler di P da Q ci informa sul guadagno di informazione nell'utilizzare P piuttosto che Q.

$$KL(P||Q) = \sum_{x \in D} P(x) log(\frac{P(x)}{Q(x)})$$

Non é simmetrica, ovvero $KL(P||Q) \neq KL(Q||P)$, e in linea generale non ha un massimo.

Un modo per renderla simmetrica é utilizzare la misura $\frac{KL(P||Q)+KL(Q||P)}{2}$. Tuttavia...

Divergenza di Jensen-Shannon

É una nozione di divergenza simmetrica basata sulla distribuzione media $A = \frac{P+Q}{2}$ e non sulla media delle divergenze

$$JSD(P, Q) = \frac{KL(P||A) + KL(Q||A)}{2}$$

Ha un valore massimo di 1 se la base del logaritmo é 2, altrimenti non ha un limite superiore.

Confrontare distribuzioni

f-divergenze

Le divergenze viste precedentemente fanno parte di una famiglia di misure chiamate f-divergenze, il cui scopo é quello di rappresentare la divergenza come media dell'odds ratio (rapporti di probabilitá) ponderata da una funzione f.

Odds ratio

Con il termine inglese **odds** si intende il rapporto tra la probabilità p di un evento e la probabilità che tale evento non accada, cioè la probabilità $1 - p$, ovvero $\frac{p}{1-p}$.
L'**odds ratio** è definito come l'odds della malattia tra soggetti esposti, diviso l'odds della malattia tra soggetti non esposti, al fine di definire la correlazione tra un fattore di rischio e una malattia:

$$OR = \frac{\frac{P(malattia|esposti)}{1-P(malattia|esposti)}}{\frac{P(malattia|non\_esposti)}{1-P(malattia|non\_esposti)}}$$

Divergenza di Hellinger

É una importante f-divergenza definita come:

$$HE(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{x\in D}(\sqrt{P(x)} - \sqrt{Q(x)})^2}$$

Puó anche essere scritta come

$$HE^2(P,Q) = 1 - \sum_{x\in D}\sqrt{P(x)Q(x)}$$

Importante proprietá é che essa é a valori in $[0 \ldots 1]$.

Informazione

In concetto di sorgente informativa (X, p) é stato originariamente pensato per l'analisi matematica dei processi di comunicazione in modo da esprimere la natura probabilistica dell'informazione.

L'**informazione** é una funzione inversa della probabilitá. É la controparte a posteriori di una incertezza a priori rapresentata dalla probabilitá, misurando il guadagno di conoscenza una volta che un evento é accaduto. Quindi, un evento piú raro é anche piú informativo.

Informazione

Dato un evento E con probabilitá $P(E)$, l'informazione é definita come

$$I(E) = log(\frac{1}{P(E)}) = -log(P(E))$$

Informazione di eventi congiunti

Il *log* garantisce l'**additivitá** per un evento (E, E') congiunto ed indipendente (ovvero $P(E, E') = P(E) \cdot P(E')$) t.c. $I((E, E')) = I(E) + I(E')$.

Informazione

Data una sorgente informativa (X, p), l'informazione di tale sorgente é definita come

$$I((X,p)) = H(X,p) = \sum_{a \in \hat{X}} p(a) I(p(a)) = -\sum_{a \in \hat{X}} p(a) log(p(a))$$

ovvero, essa é la media pesata delle informazioni dei singoli eventi scaturiti della sorgente informativa.

Abbiamo gá visto la proporietá di equipartizione , ovvero la prop. per cui H é massima se tutti gli eventi della sorgente informativa sono equiprobabili.

Dalla entropia finisca a quella informazionale

Il concetto di entropia era giá stato utilizzato prima di Shannon. Il termine venne coniato da R. Clausius, un fisico tedesco che volle studiare l'entropia dei sistemi termodinamici. Entropia dal greco "en-tropos", verso interno.

Il secondo principio della termodinamica (per sistemi isolati) afferma che la variazione di entropia termodinamica S in un sistema é sempre maggiore o uguale a 0

$$\Delta S \geq 0$$

Dalla entropia finisca a quella informazionale

Nello studio dei gas perfetti, Boltzmann fu il primo a formulare l'entropia nella forma

$$H = \sum_{i=1}^{m} n_i log_2(n_i)$$

dove n_i indica il numero di molecole di un gas appartenenti ad una determinata classe di velocitá i. Tale formula é la rappresentazione microscopica della entropia di Clausius.

Dalla entropia finisca a quella informazionale

L'entropia di Clausius puó essere definita come

$$S = k \cdot log_e(w)$$

dove k é la costante di Blotzmann e w é il numero di micro-stati distinguibili associati al macro-stato del sistema.

Dato V^n il numero di arrangiamenti possibili di n molecole in V celle di volume, w é il numero di modi distinti per dividere n particelle in m classi di velocitá, quindi

$$w = \frac{n!}{n_1!n_2!\ldots n_m!}$$

da cui

$$S = k \cdot ln \frac{n!}{n_1! n_2! \dots n_m!}$$

per l'approssimazione di Stirling ($ln(n!) \simeq n \cdot ln(n)$)

$$S = k \cdot n \cdot ln(n) - k(n_1 ln(n_1) + \dots + n_m ln(n_m))$$

possiamo sostituire n_i come $n \cdot p_i$, ovvero la probabilitá p_i moltiplicata per il totale delle particelle n

$$S = k \cdot n \cdot ln(n) - k((n \cdot p_1) ln(n \cdot p_1) + \dots + (n \cdot p_m) ln(n \cdot p_m)))$$

$$S = k \cdot n \cdot ln(n) - k \cdot n \cdot ln(p_1 + \dots + p_m) - k \cdot n(p_1 ln(p_1) + \dots + p_m ln(p_m))$$

essendo $(p_1 + \dots + p_m) = 1$

$$S = -k \cdot n(p_1 ln(p_1) + \dots + p_m ln(p_m)) = -k \cdot n \sum_{i=1}^{n} p_i ln(p_i)$$

Principio

L'entropia di una sorgente informativa (X, p) é proporzionale al logaritmo del numero w di sorgenti informative distinte che forniscono gli stessi valori di X con la stessa distribuzione di probabilitá p.

Ovvero, l'entropia di una sorgente é proporzionale al numero di sorgenti aventi la stessa entropia.

Da cui, l'entropia di una sorgente é proporzionale al numero di variabili stocastiche aventi la stessa distribuzione di probabilitá.

Principio circolare dell'entropia

$$H(X, p_x) = c \cdot log_2|\mathbb{X}|$$

per una qualche costante c, e t.c. $\mathbb{X} = \{(Y, p_y) : \hat{X} = \hat{Y}, p_x = p_y\}$

Il principio circolare ci dice che l'entropia é determinata da p_X ma, allo stesso tempo, corrisponde al numero i modi con cui ci puó realizzare p_x partendo dalla variabile aleatoria X.
Abbiamo quindi due spazi:

- lo spazio interno, dato dall'insieme degli eventi della sorgente
- lo spazio esterno, dato dalla classe delle sorgenti aventi la stessa distribuzione di probabilitá

Distribuzione di probabilitá congiunta

Siano X e Y due variabili aleatorie, e sia $p(x, y)$ la loro **probabilitá congiunta** dei due eventi $x \in \hat{X}$ e $y \in \hat{Y}$.
La **distribuzione di probabilitá congiunta** di X e Y é data da

$$p_{X,Y} = (p(x, y)|x \in \hat{X}, y \in \hat{Y})$$

ed essa é definita sul prodotto cartesiano $\hat{X} \times \hat{Y}$.

Entropia congiunta

Data la sorgente informativa $(X \times Y, p_{X,Y})$, definiamo **entropia congiunta** come l'entropia di tale sorgente informativa

$$H((X \times Y, p_{X,Y})) = - \sum_{x \in \hat{X}, y \in \hat{Y}} p(x, y) log_2(p(x, y))$$

Indichiamo come $p_{(x,y)}$ la probabilitá congiunta $(x, y) \mapsto p(x, y)$ degli eventi congiunti $x \in \hat{X}$ e $y \in \hat{Y}$.
Indichiamo con $p_x \cdot p_y$ il prodotto delle probabilitá semplici, $(x, y) \mapsto p(x)p(y)$.
Allora:

$$I(X, Y) = KL(p_{(x,y)}, p_x \times p_y) = \sum_{x \in \hat{X}, y \in \hat{Y}} p(x, y) log(\frac{p(x, y)}{p(x)p(y)}) =$$

$$= \sum_{x \in \hat{X}, y \in \hat{Y}} p(x, y) log(\frac{p(x, y)}{p(x)}) - \sum_{x \in \hat{X}, y \in \hat{Y}} p(x, y) log(p(y))$$

Informazione mutua

Sia p una probabilitá definita su $X \times Y$ (prodotto cartesiano), e siano p_x e p_y le probailitá associate a X e Y.
L'**informazione mutua** $I(X, Y)$ é data dalla divergenza entropica delle distribuzioni p_x e p_y.

$$I(X,Y) = KL(p_{(x,y)}, p_x \times p_y) = \sum_{x \in \hat{X}, y \in \hat{Y}} p(x,y) log(\frac{p(x,y)}{p(x)p(y)}) =$$

$$= \sum_{x \in \hat{X}, y \in \hat{Y}} p(x,y) log(\frac{p(x,y)}{p(x)}) - \sum_{x \in \hat{X}, y \in \hat{Y}} p(x,y) log(p(y))$$

il teorema di Bayes ci dice che $\frac{p(x,y)}{p(x)} = p(y|x)$, inoltre $\sum_{x \in \hat{X}} p(x) = 1$, da cui

$$= \sum_{x \in \hat{X}, y \in \hat{Y}} p(x,y) log(p(y|x)) - \sum_{y \in \hat{Y}} p(y) log(p(y)) = H(Y) - H(Y|X)$$

ovvero la media della inf. di Y meno la media della inf. di Y dato X.

Informazione mutua

Si puó dimostrare che l'informazione mutua puó essere espressa in modo equivalente come:

$$
\begin{aligned}
I(X, Y) &= H(Y) - H(X|Y) \\
&= H(X) - H(Y|X) \\
&= H(X) + H(Y) - H(X, Y) \\
&= H(X, Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$

Cross-entropia

Data una variabile aleatoria X e due distribuzioni di probabilità discrete, P e Q, definite su di essa, la cross-entropia è definita come:

$$H(P||Q) = -E_P[log_2(Q)] = -\sum_{x \in \hat{X}} P(x) log_2(Q(x)) = H(P) + KL(P||Q)$$

misura il numero medio di bit necessari per identificare un evento estratto da $\hat{X}$ nel caso sia utilizzato uno schema ottimizzato per una distribuzione di probabilità Q piuttosto che per la distribuzione P.