

Riconoscimento e recupero dell'informazione per bioinformatica

Clustering: similarità

Manuele Bicego

Corso di Laurea in Bioinformatica

Dipartimento di Informatica - Università di Verona

Sommario

- ⇒ Definizioni preliminari
- ⇒ Similarità tra punti
- ⇒ Similarità tra sequenze
 - ⇒ dynamic time warping
- ⇒ Similarità tra insiemi
- ⇒ Un esempio biologico: il BLAST

Definizioni

⇒ Coefficiente di similarità:

⇒ indica la “forza” della relazione tra due oggetti

⇒ maggiore è la somiglianza tra questi oggetti, più alto è il coefficiente di similarità

⇒ Dissimilarità (distanza):

⇒ concetto simile ma che misura le differenze tra due oggetti

⇒ In generale si può parlare di “misure di prossimità”

\mathcal{X} Dominio del problema , $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$

$$proximity(\mathbf{x}_i, \mathbf{x}_j) = f : \mathcal{X}^2 \rightarrow \mathbb{R}$$

Definizioni

⇒ Concetto di “metrica” (già vista nella lezione sulla rappresentazione): misura di prossimità con particolari caratteristiche

⇒ Definizione: (dissimilarità/distanza metrica):

⇒ misura di dissimilarità che soddisfa alle seguenti proprietà:

1. Positivity: $d_{ij} \geq 0$
2. Reflexivity: $d_{ii} = 0$
3. Definiteness: $d_{ij} = 0$ objects i and j are identical
4. Symmetry: $d_{ij} = d_{ji}$
5. Triangle inequality: $d_{ij} < d_{ik} + d_{kj}$

Altra rappresentazione

⇒ Matrice di prossimità:

⇒ matrice che descrive i valori della funzione per tutte le possibili coppie

$$\begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

⇒ simmetrica / asimmetrica, dissimilarità / similarità, ...

Una distanza asimmetrica

- ⇒ Esempio: distanza tra sequenze di DNA
- ⇒ Distanza= inverso dell'alignment score basato sulla matrice di sostituzione



- ⇒ Distanza simmetrica se la matrice di sostituzione è simmetrica (“A → T” è uguale a “T → A”)

	A	T	C	G
A	-	α	α	α
T	α	-	α	α
C	α	α	-	α
G	α	α	α	-

Jukes-Cantor

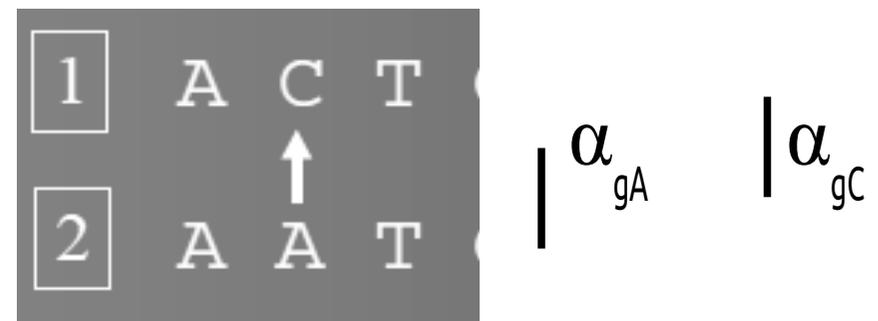
Una distanza asimmetrica

⇒ Se la matrice di sostituzione non è simmetrica, allora non è la stessa cosa allineare la sequenza 1 su 2 o allineare la sequenza 2 su 1

⇒ Esempio: Distanza Tajima-Nei per sequenze di DNA

⇒ Pesa in modo diverso le sostituzioni tenendo conto della frequenza che i nucleotidi hanno all'interno delle sequenze

	A	T	C	G
A	-	α_{GT}	α_{GC}	α_{GG}
T	α_{GA}	-	α_{GC}	α_{GG}
C	α_{GA}	α_{GT}	-	α_{GG}
G	α_{GA}	α_{GT}	α_{GC}	-



Commenti

- ⇒ La scelta della misura di similarità/distanza è cruciale e influenza enormemente l'uscita del clustering

- ⇒ Informazione a priori:
 - ⇒ contesto applicativo
 - ⇒ tipo di pattern (vettore, sequenza, dati mancanti)
 - ⇒ dimensionalità del pattern
 - ⇒ scala
 - ⇒ cardinalità dell'insieme
 - ⇒ requisiti (velocità vs precisione): e.g. retrieval by content
 - ⇒ (esperienza del ricercatore)

Trasformazione

⇒ Similarità e dissimilarità misurano la stessa quantità da due punti di vista differenti

⇒ Trasformazione: $d(\mathbf{x}_i, \mathbf{x}_j)$ una distanza

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{a}{d(\mathbf{x}_i, \mathbf{x}_j)} \quad \text{con } a > 0$$

$$s(\mathbf{x}_i, \mathbf{x}_j) = d_{max} - d(\mathbf{x}_i, \mathbf{x}_j) \quad \text{con } d_{max} = \max_{i,j} d(\mathbf{x}_i, \mathbf{x}_j)$$

$$s(\mathbf{x}_i, \mathbf{x}_j) = \ln(d_{max} + k - d) \quad \text{con } k > 0$$

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{kd}{1+d} \quad \text{con } k > 0$$

Misure per pattern vettoriali

- ⇒ Campo molto investigato: esistono molte misure diverse!
 - ⇒ vettori numerici
 - ⇒ vettori categorici
 - ⇒ vettori binari

- ⇒ Distanze tra vettori numerici
 - ⇒ distanza euclidea
 - ⇒ distanza di Manhattan
 - ⇒ distanza Maximum
 - ⇒ distanza di Mahalanobis
 - ⇒ distanza di Minkowski
 - ⇒ misura coseno (similarità)

Vettori numerici

Nozioni preliminari: vettori $\mathbf{x} = [x_1 \dots x_d]$, $\mathbf{y} = [y_1 \dots y_d]$

⇒ distanza euclidea

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^d (x_j - y_j)^2} = [(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T]^{1/2}$$

⇒ Molto utilizzata

⇒ distanza di Manhattan (city block distance)

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^d |x_j - y_j|$$

⇒ Tutti i percorsi più brevi hanno la stessa lunghezza

⇒ Utilizzata nei circuiti dove i fili possono andare solo orizzontalmente o verticalmente

Vettori numerici

⇒ Maximum distance (distanza “sup”)

$$d(\mathbf{x}, \mathbf{y}) = \max_{1 \leq j \leq d} |x_j - y_j|$$

⇒ Distanza di Mahalanobis

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})\Sigma(\mathbf{x} - \mathbf{y})^T}$$

⇒ Scalamento degli assi

⇒ Pro: invariante alle rotazioni/traslazioni/trasformazioni affini

⇒ Contro: calcolo della matrice di covarianza

Vettori numerici

⇒ Distanza di Minkowsky

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^d (x_j - y_j)^p \right)^{1/p}$$

⇒ Generalizzazione della distanza euclidea ($p=2$) e di quella di manhattan ($p=1$)

⇒ Similarità coseno

$$d(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

⇒ Similarità (non distanza)

⇒ Tiene conto della lunghezza dei vettori

Misure per dati categorici

⇒ Dati discreti (exe DNA)

⇒ simple matching dissimilarity measure

$$\delta(x, y) = \begin{cases} 0 & \text{se } x = y \\ 1 & \text{se } x \neq y \end{cases}$$

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^d \delta(x_j, y_j)$$

Misure per dati binari

Dati binari: 0 o 1

⇒ Distanza di Hamming: numero di posizioni dove i due vettori differiscono

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^d (x_j - y_j)^2$$

⇒ Esempio: la distanza di Hamming tra 1011101 e 1001001 è 2

Misure per dati binari

Similarità di Jaccard: misura del grado di overlap fra 2 insiemi A e B

⇒ L'intersezione di A e B divisa per l'unione di A e B

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

⇒ La distanza di Jaccard si ottiene facilmente facendo

$$JD(A, B) = 1 - J(A, B)$$

Misure per dati binari

⇒ Da questa misura si può ricavare la versione per dati binari:

⇒ Date 2 stringhe binarie A e B:

$$JD(A, B)_{bin} = 1 - \frac{M_{11} + M_{00}}{M_{11} + M_{01} + M_{00}} = \frac{M_{10} + M_{01}}{M_{11} + M_{01} + M_{00}}$$

⇒ M_{11} n. di simboli dove sia A che B valgono 1.

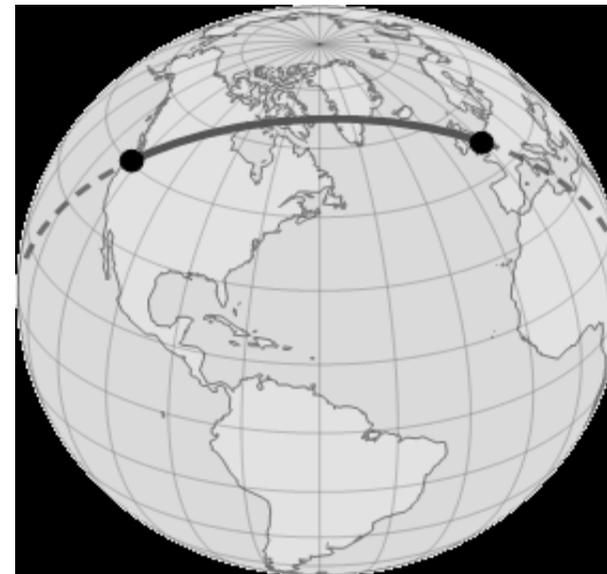
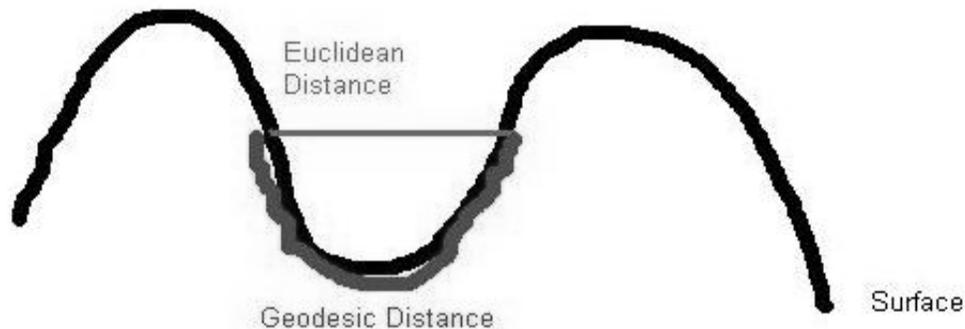
⇒ M_{00} n. di simboli dove sia A che B valgono 0.

⇒ M_{01} n. di simboli dove A vale 0 e B vale 1.

⇒ M_{10} n. di simboli dove A vale 1 e B vale 0.

⇒ Molto utilizzata in bioinformatica

Distanza geodesica



Tiene conto del “supporto”: il percorso che effettivamente occorre fare per raggiungere B partendo da A

- Misura più realistica, quindi più accurata
- Molto più difficile da misurare, occorre conoscere il supporto

Misure per pattern sequenziali

⇒ Devono gestire:

⇒ l'ordinalità dei dati

⇒ il fatto che ci possono essere sequenze di lunghezza diversa

⇒ Classi di misure

⇒ misure per dati vettoriali (se le sequenze sono della stessa lunghezza)

⇒ misure basate sul concetto di MSC (Massima sottosequenza comune)

⇒ misure basate su modelli probabilistici

⇒ misure basate su il concetto di editing (Edit Distance)

Misure per pattern sequenziali

⇒ misure per dati vettoriali (se le sequenze sono della stessa lunghezza)

⇒ Esempio: distanza euclidea

⇒ Problemi:

⇒ non tengono conto della sequenzialità dei dati

⇒ non possono gestire sequenze a lunghezza diversa

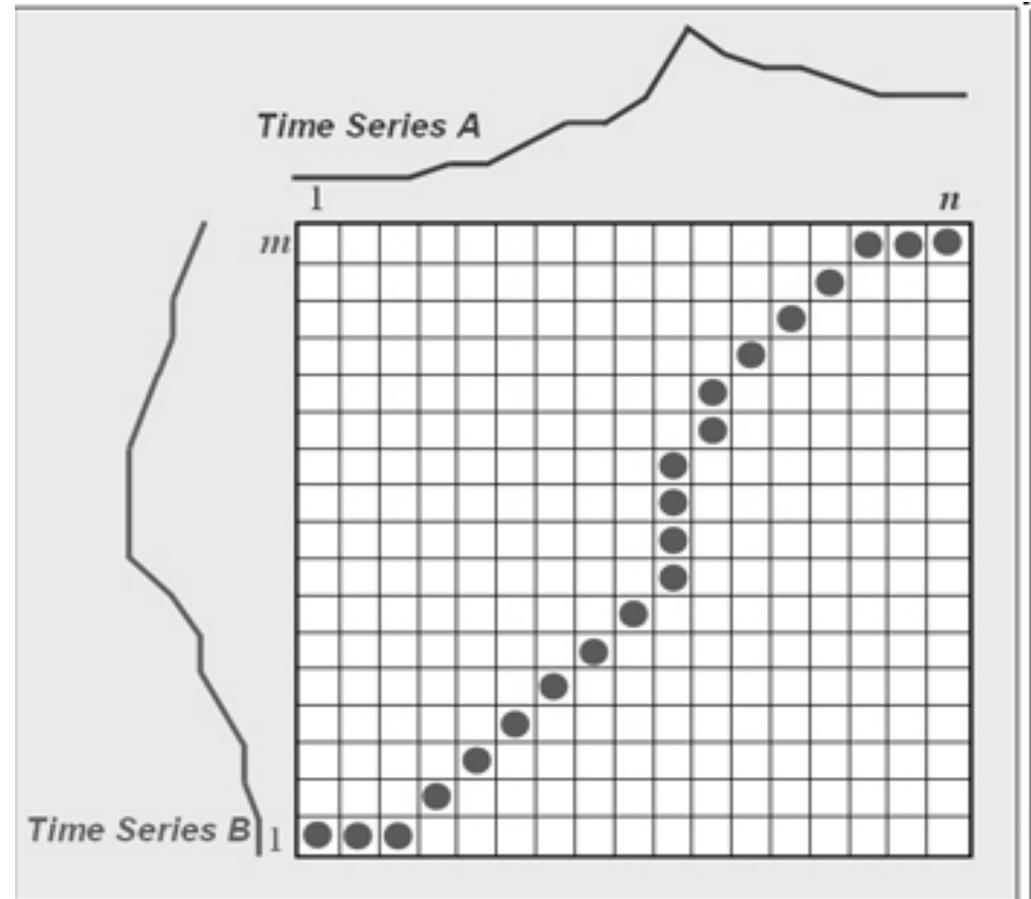
Misure per pattern sequenziali

Una misura più adatta a pattern sequenziali: Dynamic Time Warping

- ⇒ Dynamic time warping: metodo per gestire “accelerazioni” e “decelerazioni” della sequenza
 - ⇒ usato in speech processing
 - ⇒ GOAL: “allineare” due sequenze “deformando” (warping) l’asse temporale fino a quando non si trova un match ottimale.
- ⇒ IDEE:
 - ⇒ estendere le sequenze ripetendo i simboli
 - ⇒ calcolare la distanza tra le sequenze estese

Misure per pattern sequenziali

- ⇒ IDEA: warping path:
 - ⇒ come allineare diversi punti delle due sequenze
 - ⇒ ogni punto rappresenta un matching tra due punti delle due sequenze
 - ⇒ ci possono essere più "match" per lo stesso punto



Dynamic time warping (dettagli)

Date due sequenze

$$\Rightarrow X = x_1, x_2, \dots, x_r$$

$$\Rightarrow Y = y_1, y_2, \dots, y_s \quad (\text{in generale "r" diverso da "s"})$$

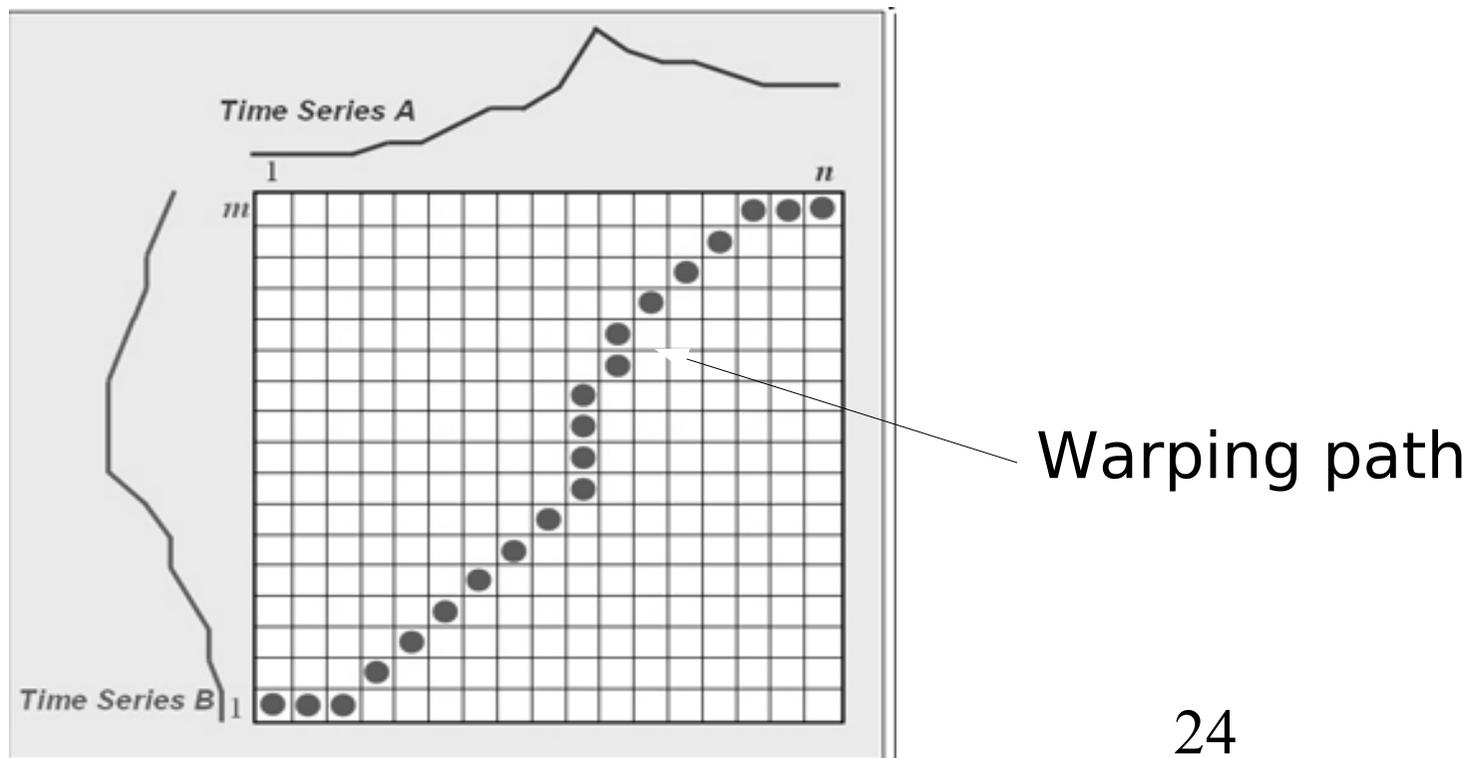
Si costruisce una matrice M di dimensione $(r \times s)$

M_{ij} = distanza tra i punti x_i e y_j (tipicamente la distanza euclidea)

M_{ij} rappresenta l'"allineamento" tra i punti x_i e y_j

Dynamic time warping (dettagli)

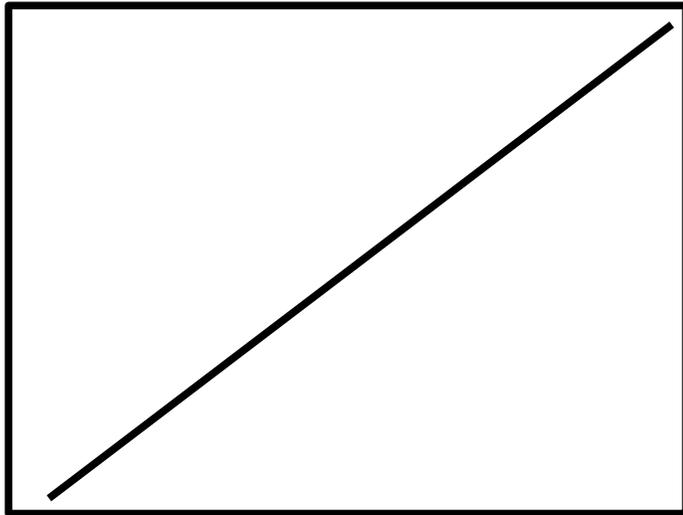
- ⇒ L'idea è di costruire un allineamento tra le due sequenze utilizzando la matrice M
- ⇒ un possibile allineamento tra le due sequenze è rappresentato da un percorso nella matrice (warping path)



Dynamic time warping (dettagli)

k è la lunghezza del warping path

$$\max \{ r, s \} \leq k \leq r + s - 1$$



Situazione
migliore



Situazione
peggiore₂₅

Dynamic time warping (dettagli)

⇒ Proprietà di un warping path

Monotonicità $w_h = (i, j), w_{h-1} = (i', j') \Rightarrow i \geq i', j \geq j'$

ci si muove sempre in alto o verso destra

Continuità $w_h = (i, j), w_{h-1} = (i', j') \Rightarrow i \leq i' + 1, j \leq j' + 1$

si viaggia sempre su celle adiacenti, non ci sono “salti”

Condizioni al contorno $w_1 = (1, 1), w_k = (r, w)$

si parte in basso a sx e si finisce in alto a dx

Dynamic time warping (dettagli)

Nota: ci sono molti percorsi che soddisfano le proprietà del warping path

Il percorso ottimale è quello che minimizza il costo di warping $DTW(X, Y)$

$$DTW(X, Y) = \min_k \sum_{h=1}^k M_{w_h} \quad \begin{array}{l} w_h = (i_h, j_h) \\ M_{w_h} = M_{i_h, j_h} \end{array}$$
$$= \min_k \sum_{h=1}^k d(x_{i_h}, y_{j_h})$$

Misure per pattern sequenziali

- ⇒ Misure basate sul concetto di MSC (Massima sottosequenza comune)
- ⇒ MSC without scaling:
 - ⇒ IDEA: due sequenze sono simili se hanno una sottosequenza in comune molto grande
 - ⇒ Approcci:
 - ⇒ caso discreto: classico problema della massima sottosequenza comune (vedi corso di algoritmi)
 - ⇒ caso continuo: occorre definire quando due valori “corrispondono” (match)
 - ⇒ ESEMPIO: x_i matches y_j se $|x_i - y_j| < \delta$
- ⇒ Tipica soluzione con Programmazione Dinamica

Misure per pattern sequenziali

⇒ MCS con Local Scaling

⇒ IDEA: due sequenze sono simili se hanno un numero “sufficientemente elevato” di coppie di sottosequenze simili

⇒ due sottosequenze sono simili se la prima può essere scalata e traslata appropriatamente in modo da assomigliare all'altra

⇒ SOLUZIONI: R-tree, percorso più lungo in un grafo aciclico

Misure per pattern sequenziali

- ⇒ misure basate su modelli probabilistici:
 - ⇒ modellare la sequenza con modelli probabilistici
 - ⇒ estrarre la similarità sfruttando i modelli
- ⇒ ESEMPIO: similarità tra sequenze con Hidden Markov Models
 - ⇒ Costruisco un modello per la sequenza 1
 - ⇒ La similarità tra la sequenza 2 e la sequenza 1 si misura vedendo quando bene la sequenza 2 è spiegata dal modello della sequenza 1 (si misura la likelihood)

Misure per pattern sequenziali

- ⇒ Edit distance: misura quante “modifiche” occorre effettuare su una sequenza per ottenerne un'altra
 - ⇒ Sostituzioni, inserzioni, cancellazioni
- ⇒ In altre parole, il “costo” che si deve pagare per convertire un pattern nell'altro
- ⇒ ESEMPIO: i simboli sono lettere, i pattern sono parole di un testo scritto.
- ⇒ Possibili “editing”:
 - ⇒ sostituzioni: “pattern” → “pastern”
 - ⇒ inserzioni: “pattern” → “patterns”
 - ⇒ cancellazioni: “pattern” → “pttern”

Misure per pattern sequenziali

⇒ Edit distance

$$D(A, B) = \min_j [C(j) + I(j) + R(j)]$$

⇒ j varia tra tutte le possibili variazioni necessarie per ottenere B da A

⇒ Soluzione algoritmica: programmazione dinamica (dettagli nel cap 8.2.2 del Theodoridis)

⇒ Molto utilizzato per automatic editing e text retrieval (trovare il best match tra un pattern e un database di patterns)

⇒ Utilizzabile per trovare la distanza tra due sequenze di geni

Misure per insiemi

⇒ Obiettivo: trovare una misura di prossimità per insiemi.

⇒ Dati non ordinati

⇒ cardinalità diversa

⇒ Altro punto di vista:

⇒ misure di similarità tra clusters (i clusters sono insiemi)

⇒ utilizzabile negli algoritmi gerarchici (si vedrà meglio in seguito)

⇒ **ESEMPI**

⇒ nearest neighbor distance: distanza tra i punti più vicini

⇒ farthest neighbor distance: distanza tra i punti più lontani

⇒ average neighbor distance: distanza media tra tutti i punti

⇒ mean based distance: distanza tra i “rappresentanti” dei cluster

⇒ Media

⇒ Medoide

⇒ Vettore più centrale

Un esempio biologico: BLAST

⇒ Basic Local Alignment Search Tool

- ⇒ algoritmo per confrontare sequenze biologiche (nucleotidiche o aminoacidiche)
- ⇒ confronta una sequenza di test con un database di sequenze, ritornando le più simili
- ⇒ uno degli algoritmi più famosi di bioinformatica
 - ⇒ affronta un problema molto importante
 - ⇒ è computazionalmente efficiente -- la ricerca effettuata con algoritmi di programmazione dinamica è assolutamente inefficiente, vista la mole di dati presente oggi -- BLAST è 50 volte più veloce

Un esempio biologico: BLAST

- ⇒ IDEA: cerca di allineare due sequenze, lo score di allineamento rappresenta la misura della bontà del match

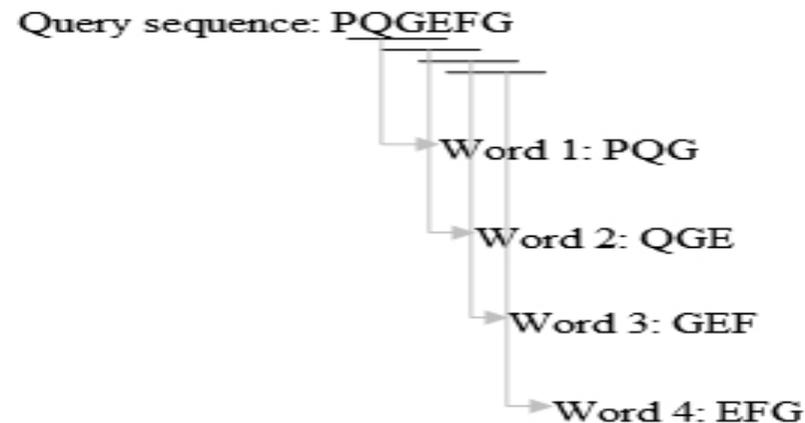
- ⇒ Assunzioni / Idee per velocizzare l'approccio
 - ⇒ non cercare l'allineamento "ottimale"
 - ⇒ non effettuare la ricerca in tutto lo spazio delle sequenze
 - ⇒ utilizzare una serie di euristiche per velocizzare l'approccio

- ⇒ Input dell'algoritmo:
 - ⇒ sequenza query (sequenza sconosciuta)
 - ⇒ sequenza target (o database)

Un esempio biologico: BLAST

PASSI dell'ALGORITMO

1. Rimuovere le regioni di bassa complessità della sequenza query
 - ⇒ regioni della sequenza con ripetizioni di pochi tipi di simbolo
 - ⇒ possono confondere il programma nello trovare regioni significative
2. Creare una lista delle “word” di K lettere della sequenza query



Un esempio biologico: BLAST

3. cercare, in tutte le sequenze del database, tutte le word di lunghezza K che hanno un buon match con le word della sequenza query
 - ⇒ buon match = score di allineamento sopra una certa soglia
 - ⇒ utilizzo della “substitution matrix” per calcolare lo score
 - ⇒ lo score considera l’allineamento senza gap
 - ⇒ ogni word trovata si chiama “hit” (o “hotspot”)
 - ⇒ allineamento senza gap è molto veloce: possibilità di memorizzare una volta per tutte le posizioni delle word in tutto il database
4. utilizzare ogni “hit” come “seme” per allargare la regione di similarità
 - ⇒ cercare di estendere la coppia di similarità a dx e a sx fino a quando lo score di similarità non diminuisce
 - ⇒ il risultato si chiama HSP (High Scoring segment pair)

Un esempio biologico: BLAST

5. visualizzare tutti gli HSP con uno score sufficientemente alto
 - ⇒ vengono listati in ordine di score
6. fornire un'analisi statistica degli score risultanti: l'E-value
 - ⇒ misura il numero di hit che si potrebbero vedere “per caso”, in un database di sequenze casuali
 - ⇒ dipende dalla dimensionalità del database e dalla lunghezza della sequenza di query
 - ⇒ la significatività statistica è proporzionale al valore di tale indice (valori attorno allo zero supportano fortemente i risultati)

Un esempio biologico: BLAST

Note:

- ⇒ Eventualmente si può gestire anche la presenza di più HSP in una stessa sequenza del database
- ⇒ si può utilizzare on line:
<http://blast.ncbi.nlm.nih.gov/Blast.cgi>
- ⇒ utilizzatissimo per il buon compromesso tra accuratezza e velocità (negli anni sono state presentate molte varianti)
 - ⇒ l'articolo dove viene presentato è il più citato degli anni 90

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). "Basic local alignment search tool". *J Mol Biol* **215** (3): 403-410