

Lezione 1 - Bit parallel

Notazione

Simbolo $T[i]$

Stringa $T[1] \dots T[\ell]$

Sottostringa $T[i:j]$

Prefisso $T[:j]$

Suffisso $T[i:]$

Concatenazione $T_1 T_2$

Problema: Pattern Matching

Input: Ho un testo $T = T[1], \dots, T[n]$ ed un pattern $P = P[1], \dots, P[m]$

Goal: Trovare tutte le occorrenze P in T

Lower Bound: Il minor tempo di calcolo possibile è $\Theta(n+m)$ perché devo almeno leggere T e P

Con un algoritmo banale (uso dei for) otterrei $\Theta(n \cdot m)$. Introduciamo un algoritmo che seppur con tempo $\Theta(n \cdot m)$, è nell'effettivo più veloce in quanto vengono eseguiti pochi calcoli dalla cpu. Costante moltiplicativa minore.

Prima di introdurre l'algoritmo, introduciamo una forma di calcolo definita bit-parallel: eseguiamo 3 operazioni parallelamente su tutti i bit della macchina. Utilizzeremo come operazioni:

- and: $x \wedge y$
- or: $x \vee y$
- xor: $x \oplus y$
- right-shift: $x \gg k$
- left-shift: $k \ll x$

Lezione 1 - Bit parallel

Algoritmo di DomoK:

- Costruiamo una matrice con sulle righe i prefissi P e sulle colonne i caratteri del testo T . Inserisco \sqcup in posizione $M(i,j)$ sse $P[:i] = T[j-i+1:i]$ avendo se la lunghezza del testo fissato il pattern termina con lo stesso carattere:

	A	B	R	A	C	A	D	A	B	R	A
A	$\sqcup A$	O	O	$\sqcup A$	O	$\sqcup A$	O	$\sqcup A$	O	O	$\sqcup A$
AB	O	$\sqcup AB$	O	O	O	O	O	O	$\sqcup AB$	O	O
ABR	O	O	$\sqcup ABR$	O	O	O	O	O	O	$\sqcup ABR$	O

- Nota che per controllare se $M(i+1, j+1)$ matcha, devo solo considerare l'ultimo carattere, in quanto i caratteri precedenti sono stati già controllati da $M(i, j)$.

	A	B	R	A	C	A	D	A	B	R	A
A	$\sqcup A$	O	O	$\sqcup A$	O	$\sqcup A$	O	$\sqcup A$	O	O	$\sqcup A$
AB	O	$\sqcup B$	O	O	O	O	O	O	$\sqcup B$	O	O
ABR	O	O	$\sqcup B$	O	O	O	O	O	O	$\sqcup B$	O

- L'algoritmo che definisce ciò, effettua le seguenti operazioni:

3.1 Right-shift di $C[j-s]$

3.2 Inserisco \sqcup in cima

3.3 Eseguo AND tra $C[j]$ e il valore del right-shift

	A	B			A	B
A	\sqcup	$\sqcup \wedge B=A$			A	\sqcup
AB	O	$\sqcup \wedge B=B$	→	AB	O	\sqcup
ABR	O	$\sqcup \wedge B=R$		ABR	O	O

- Ottimizzo ulteriormente salvandomi le codifiche dell'alfabeto del testo: $U[\sigma]$

- Ottengo in forma compatta la seguente istruzione:

$$C[j] = ((C[j-s] >> s) | (s << C[j-s])) \wedge U(T(j))$$

Lezione 1 - Karp-Rabin

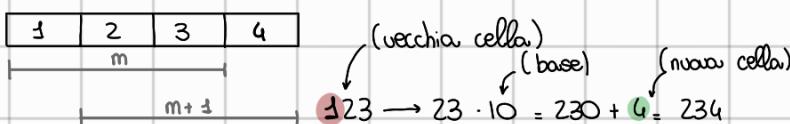
Dopo aver visto DomoKij, affrontiamo un nuovo algoritmo dove accetto come caso pessimo $O(n \cdot m)$ ma sarà raro finirci.

Introduzione all'idea Karp-Rabin: Data una stringa su alfabeto binario (per semplicità), considero la sua fingerprint, ossia il suo valore numerico, e uso una sliding window di ampiezza m su T :

$$\cdot H(S) = \sum_{i=1}^{|S|} 2^{i-1} H(S[i])$$

$$\cdot H(T[i:i+m]) = (H(T[i:i+m-1] \cdot T[i]) / 2 + 2^{m-1} T[i+m])$$

Ovviamente, la sliding window successiva controlla solo l'ultima posizione, in quanto conosce il valore delle posizioni precedenti:



Effettuando solo operazioni algebriche su bit mi assicuro che è una implementazione valida, ma purtroppo non efficiente in quanto con m molto grande il costo delle operazioni non sarebbe unitario ma logaritmico, in quanto proporzionale al numero di bit.

Per ottenere tempi costanti faccio tutte le operazioni mod p primo casuale, ottenendo tutti i numeri $< p$:

· **problema**: ci sono più numeri che mod p sono uguali ($6 \equiv 1 \pmod{5}$ e $11 \equiv 1 \pmod{5}$) → genero falsi positivi.

· **soluzione**: utilizzo K numeri primi: così facendo ho il tempo moltiplicato per K , ma il tasso di errore diminuisce a q^K . Questo è possibile dato che siamo sicuri che non esistano falsi negativi.

Cambio p ogni volta che p genera FP. Se la sequenza genera FP per tutti i K p , allora è FP.

Osservazione: I migliori primi p da scegliere casualmente sono i più grandi vicini alla word size perché generano meno valori uguali.

Classificazione algoritmi probabilistici

· MonteCarlo: sempre veloce, non sempre corretto. Es. Karp-Rabin

· Las Vegas: non sempre veloce, sempre corretto Es. Quicksort con Pivot random.

Oss. Posso passare da MC a LV aggiungendo controlli, non vale viceversa.

Lezione 2 - Suffix-Tree-Array

Dopo aver visto due algoritmi con caso peggiore $O(n \cdot m)$, con il nuovo algoritmo riusciremo ad arrivare al lower bound $O(n+m)$. Cioè nonostante, si dimostrerà che nella pratica non è poi così veloce, e bisognerà "ribassare" la teoria per migliorarlo.

Introduciamo i tries: albero dove gli archi sono etichettati con le lettere, la query sull'albero consiste nel verificare se una parola appartiene all'albero, cioè se esiste il cammino radice-foglia:

· **problema**: non riesco ad identificare le prefissi, in quanto non termino su una foglia.

· **Soluzione**: aggiungo $\$$ alla fine delle parole, dove $\$ \notin \text{Vocabolario}$

Il trie è ottimo in quanto il tempo della query non cambia, dipende solo dal pattern, indipendente dal vocabolario

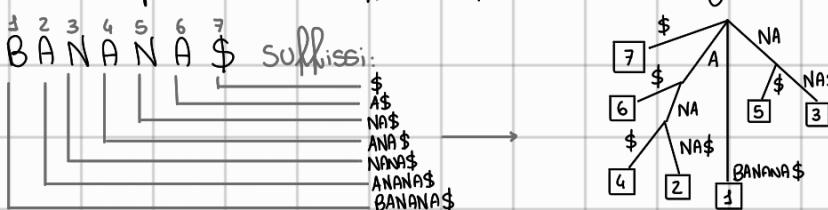
Il legame tra trie e pattern match è che ogni sottostringo è prefisso del suffisso. Se riesco a costruire il trie di tutti i suffissi risolvo il problema: tra i suffissi tali che il prefisso lungo m è uguale al pattern:

· se lo costruisco su un generico dizionario, non ho legami

· se lo costruisco sui suffissi, questi sono legati tra loro

Suffix tree: trie compatto di tutti i suffissi $T\$$: le etichette degli archi uscenti da x hanno iniziali diverse.

Esempio: B A N A N A \\$ suffissi:



Osservazione: affinché il pattern faccia match non è necessario consumare tutte le lettere dell'arco:

se il pattern è NAN matchevi con NANA senza consumare la A.

Definizioni utile:

- path-label(x): concatenazione etichette
- string-depth(x): lunghezza path-label(x)
- pattern-matching(x): visita

Lezione 2 : Suffix-Tree - Array

Problema: $O(n^2)$ di spazio richiesto

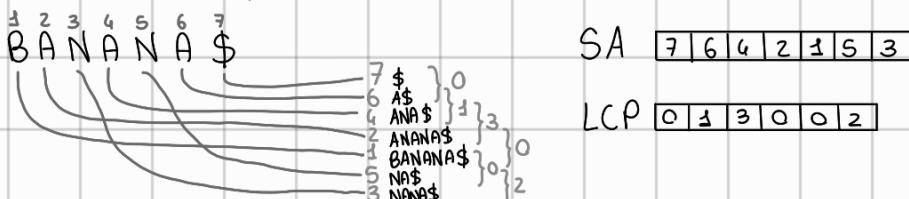
Soluzione: Solviamo gli archi con gli indici.

utilizza 3 puntatori per inizio e uno per length

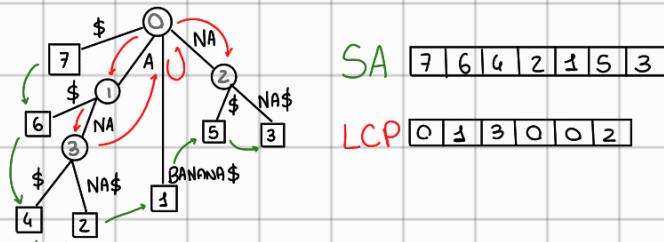
Problema: spazio per carattere 20 n bytes \rightarrow genoma umano in 128 gb e brutta località

Soluzione: suffix-array, bastano 4n bytes \rightarrow genoma umano in 36 gb e buona località

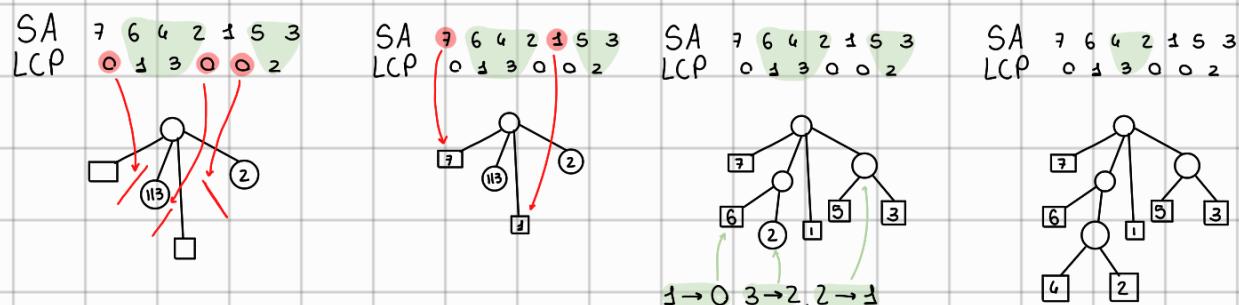
Suffix-array: Array dei suffissi in ordine lessicografico, posizioni iniziali del suffisso memorizzati. Inoltre salvo con $Lcp[i]$: lunghezza prefisso comune tra due suffissi consecutivi $SA[i], SA[i+1]$



Suffix tree \rightarrow Suffix Array: Per costruire il SA partendo da ST mi basta visitare in profondità e ordire l'albero. Per LCP ogni volta che salgo/scendo lungo un arco, tolgo/aggiungo la string-depth. LCP serve per poter attraversare in modo efficiente sia top-down che bottom-up

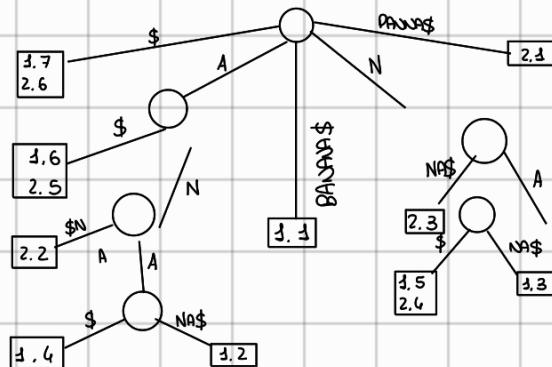


Suffix array (+ LCP) \rightarrow Suffix-tree: Prendiamo LCP: ogni volta che tra 0 so che ho una nuova partizione (prefissi diversi). Altrimenti gli elementi di LCP indicano il numero di nodi mancanti, a cui sottraggo 1. Da eseguire ricorsivamente



Lezione 2 - Suffix-Tree-Array

Suffix-tree generalizzato. Posso prendere ora più stringhe e generalizzare il problema. LCS sul suffix-tree generalizzato. Esempio con BANANA\$ e PANNA\$



Longest Common Subsequence su Suffix-Tree generalizzato: (non pattern matching!)

Solo nelle foglie ho parole che considero (z, z, \dots, n) e le posizioni $(1, \dots, m)$. Come si fa LCS?

Le concateno e distinguo i dollari: BANANA\$, PANNA\$₂ e genero SF. Ora posso avere due

tipi di sottostringhe: quelle che sono suffisso solo a w₂ oppure a entrambe: BANANA\$, PANNA\$₂.

Se un nodo è suffisso di una stringa, lo è anche il nodo padre: etichetto ogni nodo con le stringhe della quale è suffisso in bottom-up (ottimizzo). metto T/F.

Per effettuare LCS leggo l'albero e cerco nodo con le etichette delle stringhe e string-depth maggiore (top-down)

PM SA vs PM ST: tree $O(m + K)$ vs array $O(m \log n + K)$ con $m = |\text{pattern}|$ e $n = |\text{text}|$.

· tree percorre l'albero (lineare)

· array ricerca dicotomica (si aggiunge log)

Riassunto idea PM su GST: solo in ogni nodo un array di K booleani dove K è il numero di parole. Se array[i] = 1, allora il nodo ha almeno una foglia che contiene la i -esima parola

Problema: per ogni nodo abbiamo $K \cdot n$ operazioni e quindi tempo $O(K \cdot n^2)$?

Soluzione: ricordando che ogni nodo ha un solo padre e quindi ogni nodo viene letto 1 volta come padre e 1 come figlio. $O(n \cdot K^2)$ in visita bottom-up.

Lezione 3 - P1 - Suffix-Array

Ideas: Per fare P1 su SA posso fare ricerca dicotomica $O(\log_2 n)$ e dovrei controllare tutto $O(m)$ ottenendo tempo $O(m \log_2 n)$. Vogliamo velocizzare.

Acceleranti: Introduciamo 3 acceleranti per ridurre il tempo (i primi due nell'effettivo, l'ultimo a livello teorico)

Accelerante 1: Avendo un SA e un suo intervallo $SA(L, R)$ di elemento mediano M . Se $S[L]$ e $S[R]$ cominciano con n caratteri, allora i primi n caratteri del loro intervallo sono uguali e posso evitare di confrontarli.

Problema: non ho niente nel LCP che mi dice il numero di caratteri uguali tra L e R .

Accelerante 2: Denoto $l = lcp(L, P)$ e $r = (R, P)$ dove P è il pattern. Mi posso trovare in 3 casi:

Caso $l > r$: Calcolo $m = lcp(L, M)$

1. $l > m$: faccio match sopra m e quindi $r = m \wedge R = M$

2. $l < m$: faccio match + lungo sotto m e quindi $L = M$

3. $l = m$: confronto dal carattere $l+1$

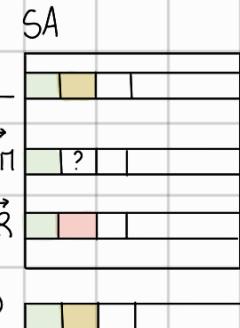
Caso $l < r$: simmetrica a $l > r$

Caso $l = r$: calcolo $m = lcp(L, M)$ e $m' = lcp(M, R)$

$l < m$: faccio match + lungo sotto a M e quindi $L = M$

$l < m'$: caso opposto al precedente: $R = M$

$m = m'$: confronto il carattere successivo



Teoria: ricerca logaritmica e numero costante di operazioni, impiego $O(\log n)$?

Pratica: abbiamo aggiornare l e r $2m$ volte al massimo. Impiego di più di $O(\log n)$

Problema: come faccio ad ottenere i valori necessari per aggiornare l e r ?

Soluzione: se riesco a pre processare il SA per ottenere $lcp(L, M)$ ci impiego $O(m + \log n)$

aggiorno ricerca
↓ ↓
 $O(m + \log n)$

Lezione 3 - Pt1 · Suffix-Array

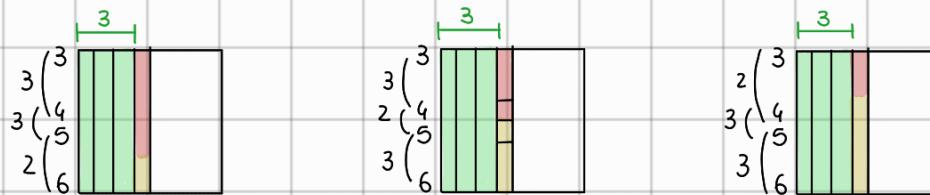
Accelerante 3: Devo considerare il mio array sempre /2.

L
1)
2)
3)
4)
5)
6)
7)
R
8)

Quando calcolo $\text{LCP}(L, R)$ ho due casi:

- se L, R sono adiacenti uso LCP

- altrimenti: considero di saper risolvere i problemi + piccoli (op). se so $\text{LCP}(3, 4)$ e $\text{LCP}(5, 8)$ voglio calcolare $\text{LCP}(3, 8)$ posso usare i casi adiacenti intermedi. nell'esempio $\text{LCP}(4, 5)$. Posso trovarmi in 3 casi:



Posso confermare che $\text{LCP}(5, 8) = \min\{\text{LCP}(3, 4), \text{LCP}(4, 5), \text{LCP}(5, 8)\}$

Conclusioni: Posso processare in tempo lineare $O(n)$ e effettuare la richiesta richiede $O(\log n)$

Sottostringa comune più lunga. Abbiamo visto come trasformare il PM dal GST al SA per migliorare spazio e località. Ora poniamo la stessa trasformazione per il LCS.

Possaggio LCS su ST a LCS su SA. Per poter effettuare il Longest Common Substring sul suffix-array ci basiamo su 3 concetti fondamentali:

- Se nel ST usavamo un root, adesso prendiamo dei sottoorray
- Nel root solleviamo se esistevano nei discendenti entrambe le stringhe, ora basta guardare l'intervallo stesso
- la string depth del ST è simulabile tramite LCP nel ST.

Intervallo buono:

- deve contenere almeno un suffisso di ogni parola
- per ogni parola mi basta esattamente un suffisso.

Intervallo dominante: vogliamo quindi l'ampiezza d'intervallo minima che contenga tutte le parole. Per farlo ci solviamo in un array $\text{last}[i]$ dove memorizzo l'ultimo indice in cui ho visitato la parola i -esima.

Lunghezza LCS: Dato l'intervallo dominante i -esimo, LCS sarà il minimo LCP partendo da j aumentando di 1: $\text{LCS} = \min(\text{lcp}[j, i-1])$. Tempo lineare? No, devo calcolare il minimo lcp e quindi ho due scorrimenti, quindi $O(n^2)$.

Range min query: Data un array A , per trovare il minimo in un range $A[i:j]$ uso una matrice $B[i:h]$ che contiene il minimo di $A[i:i+2^h]$.

Idea: Preprocessiamo i calcoli per ottenere il minimo in modo tale che l'accesso a quest'ultimo sia $\Theta(1)$.

Lezione 4 - rmq

Esempio:

\downarrow	1	2	3	4	5	6	7	8
A.	5	2	4	6	6	1	3	4
—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—

\downarrow	1	2	3	4	5	6	7	8
B.	5	2	4	6	6	1	3	4
—	2	2	4	6	1	1	3	3
—	2	2	3	3	1	—	—	—
—	1	—	—	—	—	—	—	—

$$\log_2 1 = 0 \quad \begin{array}{ccccccccc} 5 & 2 & 4 & 6 & 6 & 6 & 1 & 3 & 4 \\ \nearrow 1 \min & \nearrow 1 \min & & & & & & & \\ 2 & 2 & 4 & 6 & 1 & 1 & 3 & 3 \end{array}$$

$$\log_2 2 = 1 \quad \begin{array}{ccccccccc} 2 & 2 & 3 & 3 & 1 & & & & \\ \nearrow & \nearrow & \nearrow & \nearrow & \nearrow & & & & \\ 2 & 2 & 3 & 3 & 1 & & & & \end{array}$$

$$\log_2 4 = 2 \quad \begin{array}{ccccccccc} 1 & & & & & & & & \\ \nearrow & \\ 2 & 2 & 3 & 3 & 1 & & & & \end{array}$$

$$\log_2 8 = 3 \quad \text{soltuzione ottimale per } A[3:8]$$

Tempi: Conoscendo le due metà. $\min(m_1, m_2)$ è $\Theta(\downarrow)$ $\rightarrow A[i, h] = \min(B[i, h], B[i + \log_2 h, h])$

Implementazione: Dato che se conosco le due metà lavoro in tempo costante, divido l'array in 2 parti: con ciascuna ampiezza pari alla massima potenza di 2 inferiore a h, in modo che possa coprire tutto l'array con minimo spreco, e tra i due prendo il minore.

Formulari: $RMA(v, w) = \min\{B[v, q], B[w - 2^{\lfloor \log_2(w-v+1) \rfloor}, q]\}$

\downarrow 1 2 3 4 5 6 7 8

Esempio: Preso A.

5	2	4	6	6	1	3	4
---	---	---	---	---	---	---	---

 calcolo vari RMA

$$RMA(2, 5): q = \lfloor \log_2(5-2+1) \rfloor = 2, \text{ rmg}(2, 5) = \min\{(2, 2), (2, 2)\}$$

2	4	6	6
---	---	---	---

0	1	2	3
1	5	2	2
2	2	2	2
3	4	4	1
4	6	6	1
5	6	1	
6	1	1	
7	3	3	
8	4		

$$RMA(3, 7): q = \lfloor \log_2(7-3+1) \rfloor = 2, \text{ rmg}(3, 7) = \min\{(2, 2), (4, 2)\}$$

4	6	6	1	3
---	---	---	---	---

0	1	2	3
1	5	2	2
2	2	2	2
3	4	4	1
4	6	6	1
5	6	1	
6	1	1	
7	3	3	
8	4		

Lezione 4 - mq

Osservazione: Il tempo per trovare tutte le occorrenze di LCS nel ST è $O(n+m+K)$ per K occorrenze

Costruzione suffix-array: Voglio tempo lineare. Ho alfabeto Σ con σ simboli, testo T lungo n .

Allora aggrego triple di caratteri ed ottengo alfabeto Σ^3 con σ^3 simboli, testo T lungo $n/3$:

$$T_3 = (T[3], T[2]T[3]) \dots (T[3i+1], T[3i+2]T[3i+3]) \dots$$

$$T_2 = (T[2], T[3]T[4]) \dots (T[3i+2], T[3i+3]T[3i+4]) \dots$$

$$T_1 = (T[1], T[2]T[3]) \dots (T[3i], T[3i+1]T[3i+2]) \dots$$

Ricorsione: Uso un algoritmo ricorsivo basato su radix-sort e merge-sort. La ricorsione è un po' più complicata in quanto, per ottenere linearità raggruppa il testo in triple ed effettua la ricorsione solo su $2/3 n$ del testo, scelto in modo tale che ricaviamo l' $1/3$ mancante.

1. Ricorsione su $T_0 T_3$

2. suffissi $(T_0 T_3) = \text{suffissi}(T_2) \rightarrow T_2[i:] = T[3i+2:]$

3. suffissi T_0 ordinati:

$$T[3i+2]T[3i+3:]$$

$$T_0[i+1:]$$

4. singola passata radix-sort

5. fusione suffissi $(T_0 T_3)$ e T_2

Vantaggio: Il vantaggio rispetto al merge è che i testi sono "accavallati" e posso sfruttare i caratteri in comune.

Lcp: Quando fondiamo, se il carattere di due el. consecutivi sono uguali av Lcp, avremo quello calcolato al passo precedente + 1, altrimenti 0.

L'ezione 5 - Alignment Globale

Programmazione dinamica: Tecnica di programmazione che può essere sfruttata quando il problema è suddivisibile in più sottoproblemi e ha una sottostruttura ottima. Possiamo affermare che la soluzione ottimale ad un problema non cambia se esso diventa sottoproblema di un problema più esteso.

Allineamento: Due sequenze s_1 e s_2 vengono allineate attraverso operazioni di insert (-), ovvero insert e delete. A differenza della distanza di Hamming che misura solo il n° di caratteri diversi, qui non abbiamo il vincolo sull'uguaglianza di lunghezza di testi e possiamo assegnare importanza.

Limitazione: Prese due stringhe s_1 e s_2 e messe in colonna, non posso avere alla stessa colonna 2 (-)

Valore di allineamento: Il valore di allineamento è la somma dei valori delle singole colonne. Viene dato in input, insieme a s_1 e s_2 , la matrice di score, maggiore il coefficiente, maggiore sarà il valore della coppia, ovvero più raro.

Algoritmo di Needleman-Wunsch: L'ultimo componente della mia equazione di ricorrenza è l'ultima colonna. Prendendo ispirazione dalla Edit Distance possiamo trovarci in 3 casi:

$$M[i, j] = \max \begin{cases} M[i, j] + d(s_1[i-1], s_2[j-1]) & |x_i| \\ M[i-1, j] + d(s_1[i-1], -) & |x_i| \\ M[i, j-1] + d(-, s_2[j-1]) & |y_j| \end{cases}$$

Costo/Valore: Se voglio sapere il costo delle operazioni, avrò un problema di minimizzazione, se voglio sapere il valore sarà un problema di massimizzazione.

Condizioni di contorno:

$$M[0, 0] = 0$$

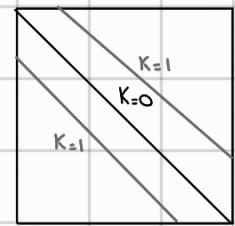
$$M[i, 0] = M[i-1, 0] + d(s_1[i], -)$$

$$M[0, j] = M[0, j-1] + d(-, s_2[j])$$

Lezione 5 - Alignment Globale

Edit Distance: Voglio sapere se la distanza di edit di due sequenze è $= K$. Con l'algoritmo precedente impiego $\Theta(n \cdot m)$. Voglio impiegurci meno.

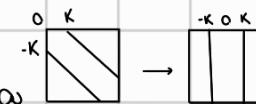
$K=0$ calcolo bande in $\Theta(n)$ in quanto rimango sulla diagonale



$K>1$ calcolo non bande in $\Theta(Kn)$: preso la matrice mi sposto dalla diagonale di max K posizioni, quindi non necessito di tutta la matrice.

Problema: Nell'esempio precedente ho dato per scontato che conosciamo K . Se così non fosse dobbiamo stimare K : partiamo da una banda piccola, così è veloce. Se il percorso è interno allora ho trovato un K buono, altrimenti raddoppio K finché la sol non tocca la banda

Spazio: Per ottenere anche spazio $\Theta(Kn)$ devo traslare la banda



L'ezione 6 - Allineamento Locale e Multiplo

Allineamento locale: date in input due stringhe s_1, s_2 e la matrice di score d , vogliamo individuare le sottostringhe t_1 di s_1 e t_2 di s_2 tali che l'allineamento globale sia massimo.

Algoritmo banale: Calcolo tutte le sottostringhe di s_1 e s_2 e calcolo allineamento globale

Tempo: avrei n^2 sottostringhe per s_1 , m^2 sottostringhe per s_2 e impiego $m \cdot n$ per allineamento globale. $\Theta(n^3 \cdot m^3)$. Vogliamo migliorare

Osservazione: usiamo un problema di massimizzazione del valore e non di minimizzazione del costo in quanto avremmo valore ottimale con caso base $t_1 = t_2 = \epsilon$

Miglioramento: Pensando al problema con i prefissi posso usare la matrice di Needleman-Wunsch dove $M[i, j] = \text{ottimo su } s_1[:i] \text{ e } s_2[:j]$

Tempi migliorato: Ricordando che una sottostringa è prefisso di un suffisso, prendo a coppie di suffissi ed eseguo l'algoritmo sui prefissi. $t_1 = s_1[a:i]$, $t_2 = s_2[b:j]$. Ho $n \cdot m$ suffissi e l'algoritmo è sempre $n \cdot m$, quindi: $\Theta(n^2 \cdot m^2)$

Passaggio globale → locale: $M[i, j] = \text{ottimo fra tutte le sottostringhe che finiscono in } i \text{ e } j$.

Equazioni di ricorrenza: Ora non lavoro più con stringhe come in NW ma con sottostringhe. Dobbiamo elencare tutti i casi per l'ultimo componente, ovvero ultima colonna. Avremo gli stessi casi di NW ma con una novità: l'ultima colonna può essere vuota! (è una sottostringa.)

$$\text{Smith-Waterman: } M[i, j] = \max \begin{cases} M[i-1, j-1] + d(s_1[i], s_2[j]) \\ M[i-1, j] + d(s_1[i], \cdot) \\ M[i, j-1] + d(\cdot, s_2[j]) \\ 0 \end{cases}$$

L'ezione 6 - Allineamento Locale e Multiplo

Calcolo soluzione: Sappiamo quindi calcolare la fine della sottostringa (max nella tabella) ma non dove inizia.

Posizione iniziale: La trovo ricostruendo la soluzione e fermandomi al primo 0 che incontro. So che qualsiasi elemento precedente può solo che abbassare il max in quanto ≤ 0 . Per essere sicuri di trovare almeno uno 0 aggiungo una condizione di controllo

Condizioni di controllo: $M[i,0] = M[j,0] = 0$

Tempo: $O(m \cdot n)$

Osservazione centrale: ogni casella intermedia della matrice non sono risultati temporanei ma valori che riusciamo a risfruttare e ci servono!

Allineamento multiplo: Dobbiamo estendere il problema affinché ragioni su s_1, \dots, s_K stringhe con len diverse e se $K=2$ sia la def. all. glob tra 2 stringhe.

Allineamento: inserimento di indel per ottenere le stringhe estese s_1^*, \dots, s_K^* tutte di uguale lunghezza. Proibisco le colonne con tutti indel ($\forall p \exists S_i^*[p] \neq -$)

Calcolo: Per il calcolo del valore dell'allineamento posso fare la somma delle distanze per ogni coppia $p = \sum_{i,j} d(s_i^*[p], s_j^*[p])$ con controllo che almeno uno dei due sia $\neq -$.

Equazione di ricorrenza: Nell'ultima colonna ho K elementi che possono assumere 2 valori: valore e indel, quindi $2^K - 1$ combinazioni (-1 per caso tutti indel)

Osservazione: Avrei matrice K dimensionale. Se avessi tutte le lunghezze uguali ho una matrice 2^K e tempo $O(2^K \cdot n^K)$. Il problema diventa intrattabile in fretta. Se la lunghezza non fosse fissata diventa NP-complesso. Inoltre il maggior limite è la memoria: con cella=8b, n=1000 e K=4 necessito 3TB di ram

Lezione 7 · Gap

Finora abbiamo considerato gli indel come costo fisso. Ora avranno un costo variabile e dobbiamo cercare di minimizzarlo.

Gap: Sequenza contigua di indel in un allineamento. Spostando il frame di lettura è penalizzante. Un nuovo gap penalizza di più di un prolungamento.

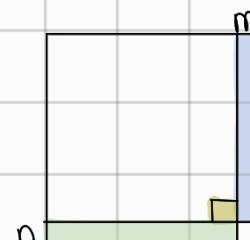
Costo gap: Funzione $P(l)$ che prende in input la lunghezza di un gap e restituisce il costo.

Gap generico - equazione di ricorrenza: Devo elencare tutti i casi possibili per l'ultima colonna. Novità: equazione di ricorrenza con numero di casi variabili.

$$M[i, j] = \max \begin{cases} M[i-1, j-1] + d(s_1[i], s_2[j]) \\ \max_{l>0} M[i, j-l] + P(l) \\ \max_{l>0} M[i-l, j] + P(l) \end{cases}$$

$$M[0, 0] = 0$$

$$M[i, 0] = P(i), M[0, j] = P(j)$$



Tempo $(n \cdot m \cdot (n+m))$. Non riusciamo a migliorare, allora cambiamo la funzione costo.

Gap affine: $P(l)$ è una retta: $P(l) = P_0 + lP_e \rightarrow$ extension
↓
opening

Equazione di ricorrenza: devo separare i casi in cui ho aperto un nuovo gap o sto estendendo.

$$M_1: \text{opt}(S_1[i-1], S_2[j-1]) + d(S_1[i], S_2[j])$$

$$N_1: \text{opt}(S_1[i-1], S_2[j]) + P_0 + P_e$$

$$E_1: \text{opt}(S_1[i-1], S_2[j]) + P_e$$

$$N_2: \text{opt}(S_1[i], S_2[j-1]) + P_0 + P_e$$

$$E_2: \text{opt}(S_1[i], S_2[j-1]) + P_e$$

Lezione 7 - Gap

$$M[i, j] = \max \begin{cases} M[i-1, j-1] + d(s_i, s_j) \\ E_1[i, j], E_2[i, j] \\ N_1[i, j], N_2[i, j] \end{cases}$$

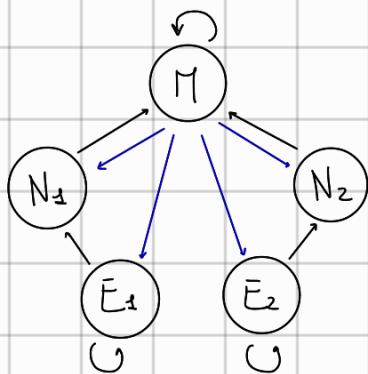
$$E_1[i, j] = \max \begin{cases} E_1[i, j-1] + P_e \\ N_1[i, j-1] + P_e \end{cases} \quad E_2[i, j] = \max \begin{cases} E_2[i-1, j] + P_e \\ N_2[i-1, j] + P_e \end{cases}$$

$$N_1[i, j] = M[i, j-1] + P_o + P_e \quad N_2[i, j] = M[i-1, j] + P_o + P_e$$

Spazio: Ω^{mn} e accesso costante al valore quindi:

Tempo: Ω^{mn})

Sincronizzare: Devo sincronizzare le loro costruzioni.



Scopo allineamento: Trovare nei DB similarità della mia entry e quanto questa sia casuale.

Karlin-Atschul: Quantificazione numerica dell'allineamento trovato.

Punteggio positivo possibile

Punteggio medio negativo

Simboli indipendenti ed equiprobabili

Sequenze infinitamente lunghe

Allineamenti senza gap

Lezione 7 - Gap

Equazione Karlin-Atschul: $E = Kmne^{-\lambda s}$

E : numero allineamenti;

K : costante

n : numero caratteri database

Cresce linearmente rispetto alle dimensioni.

m : lunghezza stringa query

λs : punteggio normalizzato

Decresce esponenzialmente rispetto al valore che mi interessa.

BLAST: Basic Local Alignment Search Tool

value = probabilità che ho trovato allineamenti per caso.

BLAST non può usare Smith-Waterman perché DB troppo grande, preferisce algoritmo euristico.

Ricerca seed: cerco porzioni identiche lunghezza = 3

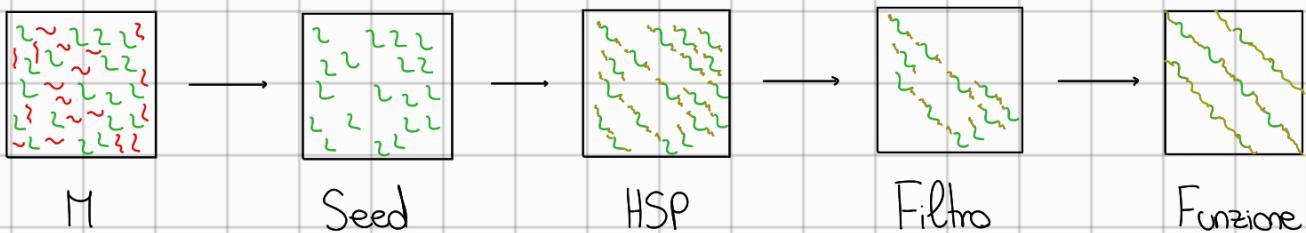
Costruzione HSP: high-scoring segment pair, estensione seed.

Filtro seed: tenuti solo HSP con alta significatività;

Funzione HSP vicine

Smith-Waterman sulle regioni.

In questo modo siamo sicuri di evitare valori inutili.



Lezione 8 - Alberi evolutivi

Evoluzione: effetti visibili in generazioni. Sono casuali: possono portare benefici, malefici o non mutare la capacità riproduttiva della generazione.

Mutazione: esistono diversi tipi.

- inserimento di nuovo gene

- duplicazione singolo nucleotide. Evento più frequente.

- duplicazione intero gene

- inversione di sequenza nucleotidica.

Evento di speciazione: separazione dove abbiamo una biforcazione, una presenta tutte le mutazioni, l'altra nessuna; diventano incompatibili e si creano due specie.



Carattere: caratteristica acquisita una sola volta nell'albero. Sono aspetti morfologici o genetici.

Filogenesi perfetta: ipotesi più restrittiva: un carattere viene acquisito una sola volta.

Input: matrice M con 1 se l'individuo i presenta il carattere j , 0 altrimenti.

Output: albero che spiega M , se esiste

Algoritmo Gusfield: tempo lineare. Usa un radix-sort sulle colonne in ordine decrescente (o in base al numero di 1). così siamo sicuri che una specie che presenta meno caratteri viene prima nell'albero.

Problema: ipotesi troppo restrittiva.

Lezione 8 - Alberi evolutivi

Cambio di stato:

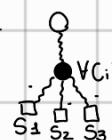
carattere acquisito: lo stato di c passa da 0 a 1

carattere perso: lo stato di c passa da 1 a 0 (backmutation)

Filogenesi persistente: ogni carattere può essere perso una sola volta nell'albero

Dallo: no limite back mutations.

Problema: dallo potrebbe ottenere tutti i caratteri e poi perdere i caratteri $\notin S(c_i)$, quindi non fornisce informazione sull'albero evolutivo.



Massima parsimonia: un albero evolutivo può presentare mutazioni o meno. Vogliamo trovare il minimo numero di mutazioni, in quanto corrisponde all'albero più semplice, e quindi, da un punto di vista evolutivo, il più probabile.

Piccolo prob. parsimonia

Istanza: matrice binaria M con n specie e insieme di m caratteri C.

albero T, le cui foglie corrispondono alla specie di M

$\forall c \in C$ abbiamo un costo w_c fra ogni coppia di stati

Sol. aumm. $\forall c \in C$, una etichettatura λ_c che assegna ad ogni nodo uno degli stati possibili per C

F. obiett.: $\min \sum_{c \in C} \sum_{(x,y) \in E(T)} w_c(\lambda_c(x), \lambda_c(y))$, dove $E(T)$ insieme dei lati di T

Osservazione: ogni carattere può essere gestito separatamente

Lezione 9 - Filogenesi

Algoritmo Sankoff: basato su DP. $P[x, z]$ è la soluzione ottimale del sottoalbero di T che ha radice x , sotto la condizione che x abbia etichetta z .

$$P[x, z] = \begin{cases} 0 & \text{se } x \text{ foglia etichettata con } z \\ \infty & \text{se } x \text{ foglia non etichettata con } z \\ \sum_{s \in F(x)} \min_s \{w(z, s) + P[r, s]\} & \text{altrimenti} \end{cases}$$

dove $F(x)$ insieme figli X in T se x nodo interno

Soluzione ottimale: $\min_s \{P[r, s]\}$ dove r è la radice di T .

Algoritmo Fitch:

limitazioni: caso non pesato: 0 se non cambio, 1 se cambio
albero binario

algoritmo: $S(x)$ insieme stati ottimali per nodo x :

$$S(x) = \begin{cases} \lambda_c(x) & \text{se } x \text{ foglia} \\ S(f_l) \cap S(f_r) & \text{se } S(f_l) \cap S(f_r) \neq \emptyset \\ S(f_l) \cup S(f_r) & \text{se } S(f_l) \cap S(f_r) = \emptyset \end{cases}$$

unificazione: $B(x)$: insieme stati z tc $P[x, z]$ minima. Si dimostra che $B(x) = S(x)$
avendo Fitch-Sankoff. Come estendere Fitch ad albero generico? Prendo il figlio
più presente. Il vero limite è il peso.

Distanza: $d: S \times S \rightarrow R^+$ tc:

riflessiva: $d(x, y) = 0 \longrightarrow x = y$.

$$\forall x, y \in S$$

simmetrica: $d(x, y) = d(y, x)$

$$\forall x, y \in S$$

distanza triangolare: $d(x, y) \leq d(x, z) + d(z, y)$

$$\forall x, y, z \in S$$

Input: matrice M di distanze stimate

Output: albero con le stesse distanze di M , se esiste

Osservazione: a differenza di approcci su caratteri, abbiamo alberi senza radice. Introduco specie esterne chiamate outgroup, così ottengo radice che separa outgroup dal resto.

Lezione 9 - Filogenesi

Osservazione: Altra utilità, gli alberi forniscono un controllo di qualità sui risultati ottenuti.

4° proprietà: $d(x,y), d(x,z), d(y,z)$ raggiunge il massimo in esattamente 2 casi su 3.

Ultrametrica: distanza + 4° proprietà. Allora ottengo sempre un albero con proprietà particolare:
A nodo interno p, tutti i suoi discendenti hanno stessa distanza con p.

Orologio molecolare: posso usare il numero di mutazioni per misurare la distanza tra 2 specie. Non tiene conto della diversità temporale della mutazione. Inoltre lunghezza genoma finito, quindi avrò inversioni. Irrealistico.

Nuovo problema: non voglio solo sapere se esiste o no l'albero T, ma voglio ottenere nel caso non sia possibile, l'albero più vicino all'albero originale.

M matrice $\rightarrow d_M$

T albero $\rightarrow d_T$

Per calcolare la distanza tra ciò che ottengo (albero) e ciò che desidero (matrice) posso usare diverse funzioni obiettivo: scarto quadratico, valore assoluto, il max del valore assoluto. Tendenzialmente si usa la prima in quanto derivata più semplice, ma in ogni caso sono tutte NP-Hard.

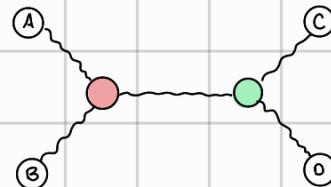
Condizione dei 4 punti: dato albero binario estraggo 4 foglie e il cammino che li collega.

Posso ottenere 3 distanze disgiunte, fissato 1 elemento:

$$d(A,B) + d(C,D) = d(A, \cdot) + d(\cdot, B) + d(C, \cdot) + d(\cdot, D)$$

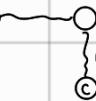
$$d(A,C) + d(B,D) = d(A, \cdot) + d(\cdot, C) + d(B, \cdot) + d(\cdot, D)$$

$$d(A,D) + d(B,C) = d(A, \cdot) + d(\cdot, D) + d(D, \cdot) + d(B, \cdot) + d(\cdot, C)$$



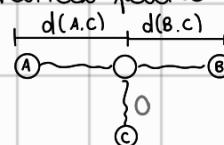
(s) Se la mat. di distanza D è ottenuta da un albero T, allora D soddisfa la cond. 4pt.
nelle 3 distanze avremo uno < degli altri due, e questi uguali. Simile a ultrametrica

Lezione 9 - Filogenesi

Caso degenere: situazione dove $d(A, B) = d(A, C) + d(C, B)$. 

Sbilancio: quanto è vicino ad essere degenere. $s(A, B, C) = d(A, C) + d(B, C) - d(A, B) \geq 0$

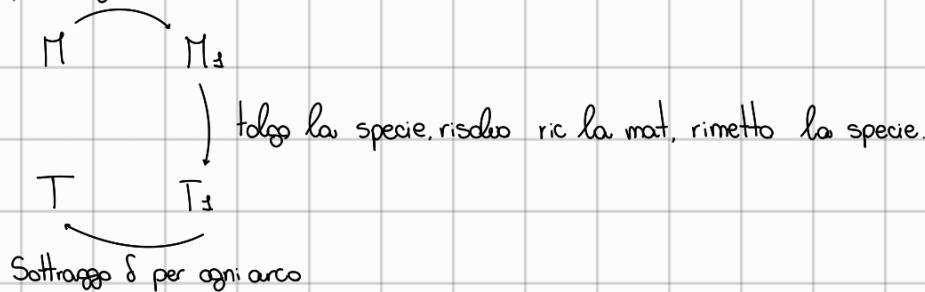
Risulta un caso che non trova mai ma che è in pratica facile da calcolare, in quanto mi permette di calcolare distanze interne



Modifica: modifico M (senza triple degeneri) in maniera controllata ottenendo M_3 (con triple degeneri) così che ottengo un albero T_3 non troppo diverso da T .

Maniera controllata: modifico la triple con sbilancio più vicino a 0. Nel passaggio da T a T_3 avrò gli archi con il costo $-\delta$. Questo causa che ogni distanza avrà -2δ , dove δ è lo sbilancio minimo.

triple degeneri usando la min



Osservazione: tutto ciò ancora non risolve prob. originale.

UPGMA: Unweighted Pair Group with Arithmetic Mean. Fondo i due cluster con minima distanza, ottendone uno. Assegniamo l'altezza ad ogni cluster, che vale la metà della distanza media.

$$D(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i \in C_1} \sum_{j \in C_2} D(i, j)$$

$$h(C) = \frac{1}{2} D(C_1, C_2)$$

La distanza tra C e ognuna delle sue foglie. Di conseguenza produce un ultrametrico, limitato.

Lezione 9 - Filogenesi

Neighbor-joining: Fondi i due cluster che sono vicini fra loro, ma sono anche distante dagli altri.

$D(C_1, C_2)$ uguale a UPGMA

$$v(C) = \frac{1}{n^{\text{cluster}} - 2} \sum_{C_3} D(C, C_3)$$

Fondi cluster con minimo $D(C_1, C_2) - v(C_1) - v(C_2)$ ottenendo C.

Modelli di evoluzione: basati sulla probabilità che in un istante temporale un genoma muta

Jukes-Cantor: equiprobabile.

Kimura 2 parametri, distinguono transizioni da trasversioni

General time reversible, matrice simmetrica

Sono tutti metodi di verosimiglianza. Sono più raffinati ma più costosi della filogenesi su caratteri o distanze. Vengono utilizzati maggiormente in quanto n° specie tendenzialmente piccolo.

Lezione 10 - Grapi di Assemblaggio

Tecnologie: sequenziamento intero genoma umano

PacBio: long read con tasso errore 10% con costi produzione medio/altri

HiSeq: medium / short read con tasso errore 0,1% con costi produzione bassi

Sanger: prima sequenziatore, long read e molto costoso

Oltre al costo del singolo progetto, mai ho l'ammortamento del macchinario.

I macchinari con costi bassi sono costosi.

Ogni strategia richiede la propria tecnologia.

La read è un'estrazione del genoma ma non conosco la provenienza. Devo inserirli per ricostruire. Come fare?

Individuo dove due reads si sovrappongono: se sufficiente lunga, allora prob. erano sovraposte anche nel genoma. Non ne ho la certezza.

Se la sovrapposizione è identica tranne per qualche carattere:

- tasso d'errore

- ipotesi sbagliata: sono regioni diverse.

- organismi diploidi: abbiamo copie cromosomi, non sono perfettamente identici, quindi possono provenire dai due aplotipi.

Assunzione: assumo che l'organismo sia aploide (singolo aploide) e classifichiamo come errore, sostituendo uno nell'altro.

Oss: Il genoma risultante sarà un mix dei due aplotipi

Oss: Il 1° genoma ottenuto era in realtà un mix di 2 persone con contribuzione non uguali, ed erano presenti buchi. Gli ultimi buchi sono stati completati 2 anni fa e dovremmo avere i genomi completi di 300.000 persone.

Lezione 10 - Grafi di Assemblaggio

Assumiamo di non avere errori: voglio trovare la superstringa comune ad un insieme di stringhe. Questo problema (stringhe) è modellabile come grafo.

read = vertice

overlap sufficientemente lungo = archi

L'etichettatura può essere il prefisso da togliere per matchare il suffisso col prefisso dell'altra stringa. Viene chiamato Grafo di Overlap.

Se primo ha overlap con secondo e secondo con terzo, posso avere un overlap più o meno significativo tra primo e terzo. Questo arco tendenzialmente è poco significativo, e quindi dato che vogliamo grafi semplici, rimuoviamo archi transitivi. Ottieniamo lo String Graph che fornisce quindi i path da seguire.

Shortest Substring: più piccolo germa che spiega tutte le mie read.

Istanza: Insieme S di stringhe

Sol. cimm: Superstring T di S. Ogni s_i è sottostringa di T

Funz. abb: $|T|$ germa assemblato, S read

Problema: regioni ripetute

Algoritmo greedy: fondere le due stringhe con massimo overlap. Trova sempre superstring T

È veramente la più corta? no

È sbagliato il modello! L'algoritmo non mi salva.

Mancava informazione nel modello: n° di occorrenze: le parti sovrapposte compaiono più volte rispetto le altre. Esempio: a-long-long-long-time: long compare 3
ng-time compare 2

Se aggiungo quante volte mi aspetto, troppo complicato.

Lezione 10 - Grafi di Assemblaggio

Travelling Salesman Problem:

Istanza: grafo orientato $G = \langle V, A \rangle$ con archi pesati $w: A \rightarrow \mathbb{Q}^+$

Sol ammissibili: Permutazione $\Pi = \langle \pi_1, \dots, \pi_n \rangle$ di V

Funzione obiettivo: $w(\pi_n, \pi_1) + \sum_{i=1}^{n-1} w(\pi_i, \pi_{i+1})$

Algoritmo NP-completo ma è uno dei più semplici.

Similarità: $\exists \text{ read} = \exists \text{ città}$ (vertici)

Differenze: assemblaggio ≠ ciclo (percorso)

lung. stringas ≠ costo ciclo

Prendendo la length è length arco + stringas arrivo. Ho ancora 1 differenza.

Ho un cammino e non un ciclo: inserisco due vertici fintizi, uno di partenza con solo archi uscenti e uno finale con solo entranti + arco per tornare a partenza.

Sembra bello ma in realtà $n \text{ read} > \text{max comp polinomiale TSP}$. Due heuristiche:

Overlap-Layaut-Consensus:

overlap: calcolo overlap e costruisco grafo. Usa SA o PD

layout: fonda e ottengo i contigs. Rimuovo duplicati;

consensus: calcolo nucleotidi.

SBH: non si usa più mai buona idea: se i vertici (stringhe) avessero stessa lunghezza.

Sembra stupido: in che modo semplifica? Dobbiamo cambiare modello

K-mero: sottostringhe lunghe K. Dato un K-mero prendo le due sottostringhe lunghe

K-1, e ne assegno un vertice. Ho arco non in base all'overlap ma in base ai K-meri originali

Lezione 10 - Grafi di Assemblaggio

Grafo di de Bruijn: ho le read. Creo i K-meri. Trovo per ognuno i $2(K-1)$ -meri e tolgo le ripetizioni. Tutti i vertici hanno lunghezza $K-1$, e tutti gli archi hanno overlap - 1. Pochi archi: al max 4 entranti/uscenti (n° caratteri ACGT)

Cambio problema: cerco superstringa dei K-meri. Perdo precisione ma + veloce:
memorizzo con hash

fusione semplice: preso un vertice controllo se esiste uno dei 4 vertici con carattere cambiato (non mi servono archi)

Ma dato che due $(K-1)$ -meri sono collegati da archi, gli archi sono i K-meri.
Toccare tutti i vertici è NP-completo (Hamilton), ma con tutti gli archi è molto più semplice.

Ciclo Euleriano: algoritmo lineare. Il numero di archi entranti deve essere uguale al numero di archi uscenti. Infatti: se grafo连通 e ogni vertice ha l'entrant: l'uscenti allora posso costruire un ciclo euleriano.

Dato che voglio un cammino e non un ciclo mi basta

Grafo Euleriano: Sia $G = \langle V, A \rangle$ grafo orientato. G è semi-euleriano se esistono due vertici s, t t.c. $N_G^-(s) = N_G^+(s) + 1$ e $t \in N_G^-(t) = N_G^+(t) - 1$ mentre per ogni altro vertice w $N_G^-(w) = N_G^+(w)$

Teorema: Sia $G = \langle V, A \rangle$ grafo semi-euleriano e P cammino da t a s . Sia G_1 grafo ottenuto da G togliendo tutti gli archi P , allora G_1 è euleriano.

Teorema: Sia $G = \langle V, A \rangle$ un grafo euleriano e C un ciclo su G . Sia G_2 il grafo ottenuto da G togliendo tutti gli archi di C . Allora G_2 è euleriano.

Lezione 10 - Grafi di Assemblaggio

Riduzione grafi overlap: caso senza errori.

- G grafo overlap con $(a \rightarrow b_1)$ unico arco irriducibile uscente da a e $(a, b_1), \dots, (a, b_n)$ archi uscenti da a. Allora $(b_i \rightarrow b_{i+1})$ con $1 \leq i \leq n-1$ sono archi di G.
- G grafo overlap con $(a \rightarrow b_1)$ unico arco irriducibile uscente da a e $(a, b_1), \dots, (a, b_n)$ archi uscenti da a. Allora $(b_i \rightarrow b_i)$ con $2 \leq i \leq n-1$ sono archi di G.

Dimostrazione: Sia G semi-euleriano. Calcolo un percorso Euleriano:

Parto da $t \in N_G^-(t) = N_G^+(t) - \{t\}$ e vado in un nodo $x: x = s \in N_G^-(s) = N_G^+(s) - \{t\}$:

- 1) si: ho trovato percorso da t a s
- 2) no: attraverso un nuovo arco uscente

Ripeto finché finisco nel vertice t. Sia P_1 cammino da t a s ottenuto.

P_1 ha attraversato tutti gli archi di G?

- 1) si, ho terminato
- 2) no, costruisco ciclo C: soli archi $\notin P_1$. Sicuramente esiste arco perché altrimenti, e sicuramente esiste ciclo perché grafo semi-euleriano. $P_1 + C + P_1$

Siccome è semi-euleriano il processo termina trovando tutti gli archi.

Algoritmo:

- 1) ordinati per lunghezza arco
- 2) marcare "da eliminare" i vertici $b_j, t \in b_i \rightarrow b_j$ con $i < j$
- 3) rimuovere archi che terminano in vertici da eliminare.