Analisi di algoritmi per il Motif Finding

Tommaso Papini
tommaso.papini1@stud.unifi.it

Gabriele Bani
gabriele.bani@stud.unifi.it

UNIVERSITÀ
DEGLI STUDI
FIRENZE

11 Dicembre 2015

Un po' di background

DNA:

- ✓ sequenza di nucleotidi
- ✓ 4 tipi di nucleotide: A, T, C, G
- ✓ *l-mer*: sottosequenza di DNA di lunghezza *l*

Motifs

In biologia può essere necessario ricavare certe sequenze di DNA "nascoste"

- ✓ pattern di nucleotidi ripetuti (l-mer)
- ✓ utili a capire determinati comportamenti biologici
  - sequenze di attivazione di geni specifici

Il problema del Motif Finding

Il problema del Motif Finding consiste nel ricavare un set di t **l-mer** da un insieme di t sequenze di DNA.

Input

- ✓ *DNA*: matrice di nucleotidi $t \times n$
  - t sequenze di DNA
  - ognuna di lunghezza n
- ✓ l: lunghezza del motif cercato

Output

- ✓ $s = (s_1, s_2, \ldots, s_t)$: lista di t posizioni iniziali di l-mer il più simili tra loro

Un primo esempio

CGGGGCTATGCAACTGGGTCGTCACATTCCCCTTTCGATA
TTTGAGGGTGCCCAATAAATGCAACTCCAAAGCGGACAAA
GGATGCAACTGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGATGCAACTCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCATGCAACTTC
CTGCTGTACAACTGAGATCATGCTGCATGCAACTTTCAAC
TACATGATCTTTTGATGCAACTTGGATGAGGGAATGATGC

Un primo esempio

CGGGGCTATGCAACTGGGTCGTCACATTCCCCTTTCGATA
TTTGAGGGTGCCCAATAAATGCAACTCCAAAGCGGACAAA
GGATGCAACTGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGATGCAACTCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCATGCAACTTC
CTGCTGTACAACTGAGATCATGCTGCATGCAACTTTCAAC
TACATGATCTTTTTGATGCAACTTGGATGAGGGAATGATGC

Mutazioni random

CGGGGCTATcCAgCTGGGTCGTCACATTCCCCTTTCGATA
TTTGAGGGTGCCCAATAAggGCAACTCCAAAGCGGACAAA
GGATGgAtCTGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGAaGCAACcCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCtTGgAACTTC
CTGCTGTACAACTGAGATCATGCTGCATGCcAtTTTCAAC
TACATGATCTTTTGATGgcACTTGGATGAGGGAATGATGC

Come trovare l'l-mer più simile tra tutti?

Allineamento

CGGGGCTATcCAgCTGGGTCGTCACATTCCCCTTTCGATA

TTTGAGGGTGCCCAATAAggGCAACTCCAAAGCGGACAAA

GGATGgAtCTGATGCCGTTTGACGACCTAAATCAACGGCC

AAGGAaGCAACcCCAGGAGCGCCTTTGCTGGTTCTACCTG

AATTTTCTAAAAAGATTATAATGTCGGTCCtTGgAACTTC

CTGCTGTACAACTGAGATCATGCTGCATGCcAtTTTCAAC

TACATGATCTTTTGATGgcACTTGGATGAGGGAATGATGC

Profilo e Consenso

Allineamento		A	T	C	C	A	G	C	T
		G	G	G	C	A	A	C	T
		A	T	G	G	A	T	C	T
		A	A	G	C	A	A	C	C
		T	T	G	G	A	A	C	T
		A	T	G	C	C	A	T	T
		A	T	G	G	C	A	C	T
Profilo	**A**	5	1	0	0	5	5	0	0
	T	1	5	0	0	0	1	1	6
	G	1	1	6	3	0	1	0	0
	C	0	0	1	4	2	0	6	1
Consenso		A	T	G	C	A	A	C	T

Score

Come definire la "bontà" di un set di l-mer?

Funzione score

Si definisce una funzione score sul vettore $s = (s_1, s_2, \ldots, s_t)$ di posizioni iniziali:

$$Score(s, DNA) = \sum_{j=1}^{l} M_{P(s)}(j)$$

dove

- ✓ $P(s)$: matrice profilo su s
- ✓ $M_{P(s)}(j)$: elemento massimo nella colonna j-esima di $P(s)$

Si cerca il set di posizioni iniziali s che massimizzi $Score(s, DNA)$!

Score: l'esempio di prima

Allineamento		A	T	C	C	A	G	C	T
		G	G	G	C	A	A	C	T
		A	T	G	G	A	T	C	T
		A	A	G	C	A	A	C	C
		T	T	G	G	A	A	C	T
		A	T	G	C	C	A	T	T
		A	T	G	G	C	A	C	T
Profilo	**A**	5	1	0	0	5	5	0	0
	T	1	5	0	0	0	1	1	6
	G	1	1	6	3	0	1	0	0
	C	0	0	1	4	2	0	6	1
Consenso		A	T	G	C	A	A	C	T

$$Score(s, DNA) = 5 + 5 + 6 + 4 + 5 + 5 + 6 + 6 = 42$$

Score

Quanto può valere lo score?

$$Score(s, DNA) = \begin{cases} l \cdot t, & \text{nel caso migliore} \\ \frac{l \cdot t}{4}, & \text{nel caso peggiore} \end{cases}$$

- ✓ lt corrisponde al caso in cui tutti gli l-mer sono identici
- ✓ $\frac{lt}{4}$ corrisponde al caso in cui gli l-mer siano diversi in tutte le posizioni

Algoritmi brute force

Algoritmi greedy

Algoritmi randomizzati

Conclusioni

Domande? Grazie!

Fine.

Domande? *Grazie!*

Domande? Grazie!