AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO201: Intelligence of Biological Systems - 1

Probabilities of Patterns in a String

Dr. Manjusha Nair M
Amrita School of Computing, Amritapuri

Email : manjushanair@am.amrita.edu
Contact No: 9447745519

The probability that there exist a 9-mer appearing three or more times in a randomly generated DNA string of length 500 is approximately 1/1300.

How????

Contents

- Possible k-mers of length k = 3 ?
- What is the probability that some k-mer appears *t* times in a text?
- What is Pr(4,2,"01",1) ?
- What is Pr(4,2,"11",1) ?
- What is Pr(4,2,"01",2) ?
- What is Pr(4,2,"11",2) ?
- Overlapping Words paradox
- Binomial Coefficient
- What is Pr(*500, 4, 9, 3*)?

Possible k-mers of length k = 3 ?

- Alphabets : A, T, G, C [The no of symbols : 4]
- Possible 3- mers are:
  - AAA AAT AAC AAG ATA ATT ATC ATG ACA ACT ACC ACG AGA AGT AGC AGG.....
- Number of possible combinations at k=3 is
  - $4^3 = 64$
- Generally, Number of possible combinations for any k-mer, is
  - 4^k

What is the probability that some k-mer appears ***t*** times in a text?

- **Notation**
- Pr(*N, A, pattern, t*):
  - Probability that k-mer *pattern* appears *t* times in a text with length *N* and alphabet *A*.
- **What is Pr(4,2,"01",1) ?**
- For eg., What is the probability that pattern "01" appears in a binary string (A = 2) of length 4?

What is Pr(4,2,"01",1) ?

- **What is Pr(4,2,"01",1) ?**
  - What is the probability that pattern "01" appears one times in a binary string (A = 2) of length 4?
- 0000 0001 0010 0011 0100 0101 0110 0111 1000 1001 1010 1011 1100 1101 1110 1111
- *01 is a substring of 11 of these 4-mers.*
- Number of possible 4-mers : $2^4 = 16$
- Each 4-mer can be generated with probability *1/16*
- Probability that pattern "01" appears in a binary string is : 11/16

What is Pr(4,2,"11",1) ?

- What is the probability that pattern "11" appears one times in a binary string (A = 2) of length 4?
- 0000 0001 0010 0011 0100 0101 0110 0111 1000 1001 1010 1011 1100 1101 1110 1111
- *11 is a substring of 8 of these 4-mers*
- Number of possible 4-mers : $2^4 = 16$
- Each 4-mer can be generated with probability *1/16*
- Probability that pattern "11" appears in a binary string is : 8/16 = ½

"11" is less likely than "01"

What is Pr(4,2,"01",2) ?

- **What is Pr(4,2,"01",2) ?**
  - What is the probability that pattern "01" appears **twice** in a random binary string (A = 2) of length 4?
- 0000 0001 0010 0011 0100 0101 0110 0111 1000 1001 1010 1011 1100 1101 1110 1111
- *0101* is the only binary string containing *01* twice.
- Number of possible 4-mers : $2^4 = 16$
- Each 4-mer can be generated with probability *1/16*
- Probability that pattern "01" appears twice or more in a binary string is : 1/16

What is Pr(4,2,"11",2) ?

- **What is Pr(4,2,"11",2) ?**
  - What is the probability that pattern "11" appears **twice** in a random binary string (A = 2) of length 4?
- 0000 0001 0010 0011 0100 0101 0110 0111 1000 1001 1010 1011 1100 1101 1110 1111
- *0111*, 1110 and 1111 have at least two occurrences of *11*.
- Number of possible 4-mers : $2^4 = 16$
- Each 4-mer can be generated with probability *1/16*
- Probability that pattern "11" appears twice or more in a binary string is : 3/16

Results so far

- What is Pr(4,2,"01",1) ?
  - ANS: 11/16
- What is Pr(4,2,"11",1) ?
  - ANS: 8/16 = ½
- What is Pr(4,2,"01",2) ?
  - ANS: 1/16
- What is Pr(4,2,"11",2) ?
  - ANS:3/16

Different K mers have different probabilities of occurring multiple times

Overlapping Words paradox

- 1110 – two overlapping occurrences of “11”
- 1111 - three overlapping occurrences of “11”
- “01” can never occure more than twice in a binary 4-mer
- Overlapping words paradox makes computing Pr(*N, A, pattern, t*), a complex problem.
  - Because the probability depends on the choice of the ‘k-mer’.
- So we approximate the computation of Pr(*N, A, pattern, t*), rather than making it exact.

What is our approximation?

- We assume that the k-mer pattern is not overlapping.
- We approximater Pr(*N, A, pattern, t*) as:
  - Count the number of ways to insert t instances of a k-mer pattern into a fixed string of length n = N-t. k

Binomial Coefficient

The Binomial Coefficient $\binom{m}{k}$ represents the number of of ways to choose $\boldsymbol{k}$ out of $\boldsymbol{m}$ objects and is equal to

$$\frac{m!}{k!\,(m-k!)}$$

There are 20 students in a class. They need to pick four students to serve as the student representatives for an upcoming trip. How many ways are there to choose the representatives?

$$\binom{20}{4} \qquad \frac{n!}{k!(n-k)!} = \frac{20!}{4!(20-4)!} = \frac{20!}{24 \times 16!} = \frac{116280}{24} = 4845$$

Binomial Coefficient

- Pr(*N, A, pattern, t*):
  - Probability that k-mer *pattern* appears *t* times in a text with length *N* and alphabet *A*.
- Consider a string ***Text*** of length ***N*** with ***t*** occurrences of k-mer.
- Here, ***Text*** is a sequence of n = ***N- t. k*** symbols interrupted by ***t*** insertions of the k-mer.
- Text = **ATAT**GAGCGC**AT**GCCT
- Pattern = AT, N= 16, t=3, k=2
- n = N – t.k = 16-3. 2 = 16-6 = 10

Binomial Coefficient

- Pr(*N, A, pattern, t*):
  - We want to count the number of ways to intersect ***t*** instances of k-mer ***Pattern*** into a fixed text of length ***n*** = ***N-t.k***
- Consider again the problem of embedding two occurrences of "01" into "222" (n = 3)

0101222	0120122	0122012	0122201	2010122
X X XXX	X XX XX	X XXX X	X XXXX	XX X XX
2012012	2012201	2201012	2201201	2220101
XX XX X	XX XXX	XXX X X	XXX XX	XXXX X

- We will therefore have n+t occurrences of "X", from which we must select t for the placements of Pattern,giving a total of $\binom{n+t}{t}$

Binomial Coefficient

- Pr(*N, A, pattern, t*):
  - We want to count the number of ways to intersect ***t*** instances of k-mer ***Pattern*** into a fixed text of length ***n = N-t.k***,
  - Total number of ways to choose t out of n+t is $\binom{n+t}{t}$
- We then multiply this by the number of strings of length ***n*** , in which we can insert ***t*** instances of ***Pattern*** to have approximate total of $\binom{n+t}{t} \cdot A^n$
- Dividing by the number of strings of length N, we have our desired approximation,
- Pr(*N, A, pattern, t*)≈ $$\frac{\binom{n+t}{t} . A^n}{A^N}$$

Binomial Coefficient

- Pr(*30, 4, ACTAT, 3*)?
  - What is the probability that the 5-mer *ACTAT* occurs at least 3 times in a random DNA string of length N = 30?

n = N- t. k = 30-3. 5 = 15

$$\Pr(30, 4, ACTAT, 3) = \frac{\binom{n+t}{t} \cdot A^n}{A^N} = \frac{\binom{15+3}{3} \cdot 4^{15}}{4^{30}} \approx 7.599\ x\ 10^{-7}$$

What is Pr(*500, 4, 9, 3*)?

- Pr(*500, 4, 9, 3*)?

$$\text{Let } p = \Pr(N, A, Pattern, t) \approx \frac{\binom{n+t}{t} A^n}{A^N}$$

- Pr(*500, 4, 9, 3*) ≈ 1/1300

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO201: Intelligence of Biological Systems - 1