Integrating Biology and Statistics at the freshman level-SYMBIOSIS I

Exploring DNA sequences using probability

Edith Seier & Karl Joplin
seier@etsu.edu; joplin@etsu.edu
East Tennessee State University

HHMI Grant # **52005872**

Symbiosis I is the first of a sequence of three Biology, Mathematics and Statistics integrative courses. Students who pass Symbiosis I get credit for Biology I and Introductory Statistics. Module 5 of Symbiosis I covers the basics of DNA genetics from the biological point of view. The math/statistical component of the module takes DNA as a sequence of the 4 nucleotides C,A,G, T ; questions about sequences are posed and the probability tools to answer them are developed or reviewed. This module prepares the students, at an elementary level, for the future study of bioinformatics. At the same time students put into practice their knowledge of Probability in a relevant context.
From the mathematical point of view the only pre-requisites are the basic understanding of probability, independence and conditional probability, which are covered in Module 4.
The topics covered are: DNA as nucleotide sequences, nucleotide frequency, GC content. Independence and conditional probability in the DNA environment. Transition matrix, graph to represent transition matrices. Probability of a given sequence of nucleotides, repeats of a single nucleotide, length of the repeat, geometric distribution. Palindromes, probability of any palindrome and of specific palindromes, space in between palindromes. Comparing two sequences of nucleotides. Similarities that happen just by chance. Random walks (and their use in testing for similarities).
Part of the teaching material created is displayed below. Microsoft word and free software available in the internet are used in the exploration of sequences.

Data bases, frequencies, GC content

Where can we find DNA data bases?

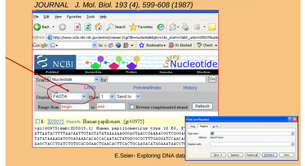
http://www.ncbi.nlm.nih.gov/

In the document 'lookingfor.doc we give complete instructions on how to get either a summary or a sequence of nucleotides from that data base.

E.Seier- Exploring DNA data I 2

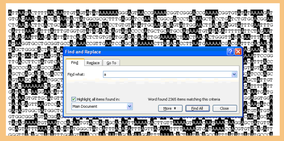
Example : Papilloma virus 18

Sequenced by Cole,S.T. and Danos,O. in 1987. Nucleotide sequence and comparative analysis of the human papillomavirus type 18 genome. JOURNAL J. Mol. Biol. 193 (4), 599-608 (1987)



E.Seier- Exploring DNA data I

How frequent are each one of the nucleotides?

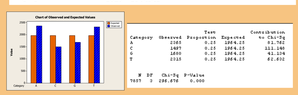


E.Seier- Exploring DNA data I 4

Can the 4 nucleotides considered to b equally frequent?

Table 1 - Frequency of nucleotides in the Human papilloma virus type 18

Row	Nucleotide	Frequency	%
1	G	1580	21.3822
2	C	1497	19.0531
3	A	2265	29.1005
4	T	2315	29.4642



Category	Observed	Proportion Expected	Contribution to Chi-Sq
A	2265	0.25	224.525	221.782
C	1497	0.25	199.23	111.190
G	1850	0.25	199.23	92.106
T	2315	0.25	199.23	81.632

N	DF	Chi-Sq	P-Value
7885	3	288.898	0.000

E.Seier- Exploring DNA data I 5

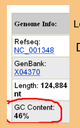
GC content

$$\frac{\#C+\#G}{\#C+\#G+\#T+\#A}*100$$

GC content of HPV 18 is 21.3822+19.0531 ~ 40.4%

Human herpes virus

Explore the GC content of more viruses, you do not need to do the counting, the information is there. Look for (double stranded DNA linear) At http://www.ncbi.nlm.nih.gov/



Look for data base GENOME, look for viruses and display OVERVIEW

Do most viruses have GC content below 50%? Why would be a reason for that?

What is different in the bounding of G with C as opposed to the bounding of T with A?

E.Seier- Exploring DNA data I 6

Dinucleotides

Dinucleotides frequency

- These are the first 100 nucleotides of the BRCA1 gene

CTTTGTGCAACAGTACTTTCCCAGGATCCACA GGAAATACTCAGAGTCCACCTGGACATTTTA CTTATATTCAGTTTCCAAGTGTCAGAGGGGT TCAGGA

- How many times you find the pair CG? How many times you find the pair CA? How many times you find the pair CT? How many times you find the pair CC?

E.Seier- Exploring DNA data I 9

Dinucleotide frequency for the whole BRCA1 gene

Row	Dinucleotide	frequency	P(first)	P(second)	Expected	ratio
1	AG	7706	0.190	0.310	5920.29	1.298899
2	CT	7098	0.295	0.205	6090.81	1.20174
3	CA	7098	0.295	0.190	5648.24	1.25734
4	TG	7997	0.205	0.310	6400.63	1.249411
5	CC	10615	0.295	0.295	8704.98	1.219107
6	GG	11388	0.310	0.310	9678.00	1.176877
7	TT	4828	0.205	0.205	4322.67	1.16908
8	AA	4048	0.190	0.190	3635.92	1.113324
9	GA	5901	0.310	0.190	5522.29	0.994473
10	GC	8935	0.310	0.295	9210.66	0.970077
11	TC	5627	0.205	0.295	6090.92	0.923863
12	AC	4577	0.190	0.295	5645.24	0.807723
13	GT	5020	0.310	0.205	6400.63	0.784306
14	AT	2856	0.190	0.205	3922.57	0.728062
15	TA	2120	0.205	0.190	3922.57	0.540411
16	CG	4154	0.295	0.310	9210.66	0.451000
TOTAL		100718			100718	

Nucleotide frequency:
P(G)=0.31 P(C)=0.295
P(A)=0.19 P(T)=0.205
There are100719 nucleotides in BRCA1 gene sequence. Thus there are 100718 pairs of consecutive letters.
The frequencies of each 'dinucleotide' are displayed in the table
Why the frequencies differ from what we would expect assuming independence?
Portions of the DNA (coding region) are translated into proteins. It is replaced by U and mRNA is 'translated' into aminoacids (which in turn go into proteins)
Thus the 'writing' of aminoacids would put conditions into the pairs of consecutive letters or 'dinucleotides' as opposed as if DNA was just a randomly formed sequence of 4 letters.

E.Seier- Exploring DNA data I 10

Are the consecutive letters independent?

There are100719 nucleotides in that sequence

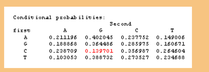
I counted how many times CC,CG,CA,CT, GC,GG,GA,GT,AC,AG,AA,AT, TC,TG,TA or TA happened. The frequencies of each pair are displayed in the table below:

first \ second	A	G	C	T	total
A	4048	7706	4557	2856	19167
G	5901	11388	8935	5020	31244
C	7098	4154	10615	7868	29735
T	2120	7997	5627	4828	20572
				Total 100718	pairs of letters

E.Seier- Exploring DNA data I 11

Conditional probabilities

P(G|C)=Probability of a G given that the previous letter was a C



Conditional probabilities				
First	Second			
	A	G	C	T
A	0.211186	0.402045	0.237752	0.149016
G	0.180868	0.364466	0.265975	0.188671
C	0.238709	0.139701	0.356997	0.264594
T	0.160053	0.388732	0.230527	0.234688

$$P(G|C)=\frac{P(C\ first\ and\ G\ next)}{P(C)}=\frac{\frac{10071}{29735}}{\frac{100718}{100718}}=\frac{41.54}{29735}=0.1397$$

- Overall the proportion of G is 31%, so if we pick one location at random the probability that it is a G is 0.31 , but if we know that the previous one is a C, the probability is only 0.13

E.Seier- Exploring DNA data I 12

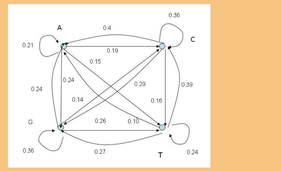
Transition probability matrix

(rounded values)

	A	G	C	T
A	0.21	0.40	0.24	0.15
G	0.19	0.36	0.29	0.16
C	0.24	0.14	0.36	0.26
T	0.10	0.39	0.27	0.24

E.Seier- Exploring DNA data I 13

Graph to represent the transition probabilities (BRCA1 gene)

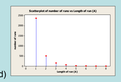


E.Seier- Exploring DNA data I 14

Repeats

One letter repeats

How many times in the BRCA1 gene we find A, AA, AAA, AAAA,...? (We used microsoft Word) We even found one run of 24 As



	f1		f2		f3		f4
TAT	394	TAAT	144	TAAAT	37	TAAAAT	20
TAC	515	TAAG	115	TAAAG	32	TAAAAG	3
TAG	664	TAAC	88	TAAAC	26	TAAAAC	4
CAT	950	CAAT	138	CAAAT	46	CAAAAT	6
CAC	1757	CAAG	377	CAAAG	102	CAAAAG	13
CAG	3228	CAAC	231	CAAAC	46	CAAAAC	15
GAT	807	GAAT	164	GAAAT	43	GAAAAT	10
GAC	1386	GAAG	406	GAAAG	85	GAAAAG	22
GAG	2459	GAAC	255	GAAAC	64	GAAAAC	12
Total	12160		1918		481		105

Calculating the probability of one nucleotide repeats under different circumstances

Example 1:
Assume the successive letters are independent and that all the letters are equally likely
How much is the probability of AAAAA ?
$P(AAAAA)=(0.25)^5$
But we are not yet discussing which is the 6^{th} letter, if it is a A again it will not be a repeat of length 5, so we need to consider the probability of AAAAAnoA also written as $AAAAA\bar{A}$
So the probability of a repeat of length 5 would be
$P(AAAAA\bar{A})=(0.25)^5 0.75$

Example 2:
Now calculate the probability of $AAAA\bar{A}$ assuming that the consecutive letters are **independent** but that 10% of the nucleotides in that section of DNA are A.
$P(AAAA\bar{A})=(0.1)^4 0.81$

The geometric distribution

Think of a Bernoulli trial, : success and failure.

Example: Success: A Failure:not A (either G,C or T)

BRCA1 gene, P(A)=0.19 , thus the probability of 'failure' is 1- P(A)=0.81.

Variable : X number of successes before the first failure

Possible outcomes	Value of the variable	Probability
$A\bar{A}$	1	$0.19^{*}0.81$
$AA\bar{A}$	2	$0.19^{*}0.19^{*}0.81=(0.19)^2 0.81$
$AAA\bar{A}$	3	$0.19^{*}0.19^{*}0.19^{*}0.81=(0.19)^3 0.81$
$AAAA\bar{A}$	4	$(0.19)^4 0.81$

Repeating words- challenge : how to do it fast

Example of 'internal repeat' (repeated 'word' within the same sequence).

TGCTTTTAGGCACATTTTAGTTGTTTTACTTAAGCTAATTGCATATCTTAGGCACATTTGGCTTAGGCACATTT

Karlin et.al(1983) Converting words into numbers

Let A=0, C=1, G=2 and T=3 , we will call V the 'value' of the letter
Imagine a word of length L, the number assigned to the word will be:

$$1+\sum_{i=0}^{L-1} V_i 4^{L-i}$$

Where V_i is the 'value of the letter' in the ith position
For example for the10 letter word
TTAGGCACAT
3302210103
Note the value of each letter is multiplied by 4^{L-i} where it is the length of the word, in the case L=10 and i is the position of the letter in the word.

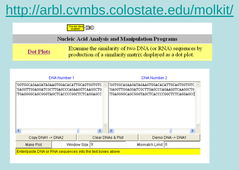
i	1	2	3	4	5	6	7	8	9	10
$L-i$	9	8	7	6	5	4	3	2	1	0
	4^9	4^8	4^7	4^6	4^5	4^4	4^3	4^2	4^1	4^0
V_i	3	3	0	2	2	1	0	1	0	3
$V_i 4^{L-i}$	$3*4^9$	$3*4^8$	0	$2*4^6$	$2*4^5$	$1*4^4$	0	$1*4^2$	0	3

The sum of the last row is 993551 and we need to add 1 to that value according to the formula, thus the numerical representation of the word TTAGGCACAT is 993556.

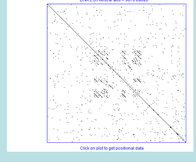
First the search can be done for shorter words; then the search for longer words can be done at the location where the shorter words were found to be repeated. For example if we are to find that the word TTAGGCACAT is repeated in the sequence, the sequence also has been TTAGGCACA repeated

Free software to do dotplots (Colorado State Univ.)

http://arbl.cvmbs.colostate.edu/molkit/



Dotplot to locate internal repeats



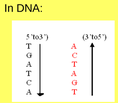
- Only plots long repeats
- *There are tests (of the permutation test type) to check if the repeats we see can happen just by chance or if we could consider that there are more long repeats than expected.*

Palindromes

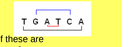
Palindromes

In plain language, a word or phrase that reads the same in both directions: ANNA, ABBA

In DNA:



- the complementary strand, reads the same *(in its own direction)*
- To check if a sequence is a palindrome



- Which of these are palindromes?
- GATC
- AGGCCT
- AAGGCCTT

Module 5 section 7- E. Seier 2

How many different 4 letter palindromes can you think of?

	1	2	3	4
1	A	A	T	T
2	A	C	G	T
3	A	G	C	T
4	A	T	A	T
5	C	A	T	G
6	C	C	G	G
7	C	G	C	G
8	C	T	A	G
9	G	A	T	C
10	G	C	G	C
11	G	G	C	C
12	G	T	A	C
13	T	A	T	A
14	T	C	G	A
15	T	G	C	A
16	T	T	A	A

In the first two boxes any of the 4 letters can go
4*4 =16 possibilities
We don't have a choice for boxes 3 & 4 if we want it to be a palindrome.
How many possible '4 letter words' are there? 4*4*4*4=256
How many possible '4 letter palindromes' are there?'4*4=16
We pick a 4 letter word, what is the probability that it is a palindrome?

Module 5 section 7- E. Seier 3

Looking for palindromes in a sequence

First 100 nucleotides of Staphilococus Aureus JH1

ATTAAAATTCTCGTATTAGCTCATTGATTATCTAGTCATAATTCAAGCAACTACTACAATATAACAA AATCCTTTTTATAACGCAAGTTCATTTTATACT

In Microsoft Word we can use EDIT>FIND to look for palindromes

	1	2	3	4	FREQ
1	A	A	T	T	2
2	A	C	G	T	0
3	A	G	C	T	1
4	A	T	A	T	1
5	C	A	T	G	0
6	C	C	G	G	0
7	C	G	C	G	0
8	C	T	A	G	1
9	G	A	T	C	0
10	G	C	G	C	0
11	G	G	C	C	0
12	G	T	A	C	0
13	T	A	T	A	3
14	T	C	G	A	0
15	T	G	C	A	0
16	T	T	A	A	1

In a sequence of 100 nucleotides, how many '4 letter words' can you define? 100-3=97

We found that 9 were palindromes.

The most frequent one was TATA

Module 5 section 7- E. Seier 4

Why do palindromes have a biological interest?

One of the reasons for the interest in palindromes is that enzymes that are used to cut DNA, and are called, 'restriction enzymes' have palindromes as their targets. Those restriction enzymes are commonly called '4-cutters' or '6-cutters' depending if their targets are palindromes of length 4 or 6. To read more about 'cutters' you can check http://www.scq.ubc.ca/?p=249

Some of the most known restriction enzymes and the sequence they recognize:

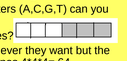
Restriction enzyme	Sequence they recognize
Sau3A	GATC
	CTAG
EcoRI ('ekko r 1')	GAATTC
	CTTAAG
BglII ('bagel 2')	AGATCT
	TCTAGA
HindIII	AAGCTT
	TTCGAA

(EcoR1 is obtained from the bacteria Escherichia coli

Module 5 section 7- E. Seier 6

Palindromes of length 6

- How many possible words of 6 letters (A,C,G,T) can you think of ? 4*4*4*4*4=46= 4096
- How many of them are palindromes?



The first 3 letters are free to be what ever they want but the last 3 are determined by the first ones 4*4*4= 64

- Assuming equal probabilities for each one of the nucleotides , and assuming independence, the probability that a 6 letter word is a palindrome is 64/4096=1/64
- In a sequence that is N letters long, what is the expected number of 6 letter palindromes under the assumption of equal probability and independence?

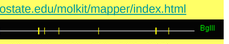
(N-5)/64

Module 5 section 7- E. Seier 5

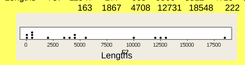
Breast Cancer Gene BRCA1-BX251925 (100719bp) and BglII (targets AGATCT)

How long would the segments be if we use BglII to cut BRCA1-BX251925? Use CSU software:

http://arbl.cvmbs.colostate.edu/molkit/mapper/index.html



Lengths	737	11947	277	669	9900	3512	4729	5601	12878
		163	1867	4708	12731	18548	222	3980	



N	Mean	Minimum	Median	Maximum
16	5779	163	4344	18548

Module 5 section 7- E. Seier 11