

Genome 540 Discussion

Jan. 20th 2026



Office Hours Today!

3-4 pm

Foege 3rd Floor

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

<https://www.dnabaser.com/articles/IUPAC%20ambiguity%20codes.html>

HW2 Overview

Longest exact sequence match
from HW1 is ~200 bp.

How surprising is that?

Let's find out: Generate some
simulated sequences, and run
HW1 on those to compare.

HW2 Overview

First part: Write a program to

- Read a fasta file
(to use for specifying models)
- Generate sequences
(using said specified models)
- Save generated sequences as fasta files.

Second part:

- Run HW1 on your simulated fastas.

HW2 Overview

Generating sequences:

- Equal frequency
- Matched nucleotide frequency
(order-0 markov)
- Matched dinucleotide frequency
(order-1 markov)
- Matched dinucleotide frequency x10

HW2 Overview

Equal frequency assumption

$$P(A) = P(C) = P(T) = P(G) = 1/4$$

HW2 Overview

Order-0 Markov:

reference sequence: ACTGATGTGCAA

Counts:

A: 4

C: 2

G: 3

T: 3

divide by

length of

sequence

→

Frequencies:

A: 1/3

C: 1/6

G: 1/4

T: 1/4

$P(A) = 0.33$

$P(C) = 0.16$

$P(G) = 0.25$

$P(T) = 0.25$

HW2 Overview

Order-1 Markov:

Lets do an example!!

Reference sequence:

ACTTGATGGAATGGTACAATGAGCGATC

HW2 Overview

Reference sequence:

ACTTGATGGAATGGTACAATGAGCGATC

May need to transpose to match the homework template :)

		First nuc			
Second nuc		A	C	T	G
	A	2	1	4	1
	C	2	0	1	1
	T	1	1	2	4
	G	4	1	1	1

		Given nuc			
P(next nuc)		A	C	T	G
	A	0.22	0.33	0.5	0.14
	C	0.22	0	0.12	0.14
	T	0.11	0.33	0.25	0.57
	G	0.44	0.33	0.12	0.14

HW2 Overview

Run HW1 on each of your simulated fasta files.

seq1: 10MB region of human genome,
from HW1

seq2: The simulated sequence

HW2 Overview

How do you expect each of the models will alter the histogram? The length of the longest sequence match?

HW2 Overview

Please!!

Run diff (or some other comparison tool) to make sure your formatting matches that of the template!

Makes things easier for me :)

```
hex ~/G/0/g/h/example/  diff hw1_template.txt hw1_test.txt
2,5c2,5
< Name: Conor Camplisson
< Email: concamp@uw.edu
< Language: C++
< Runtime: 8.08 sec
---
> Name: Karl Young
> Email: youngkar@uw.edu
> Language: Rust
> Runtime: xx sec
158c158
< Position: 338240
---
> Position: 82469
162c162
< Position: 82469
---
> Position: 338240
171,175d170
<
< int main() {
<     do_analysis();
<     return 0;
< }
```