Past midterm exam (2018)

Started: Apr 7 at 6:47am

Quiz Instructions

Answer as many questions as you can in the time allowed.

Question 1

1 pts

Which of the following genomic variants would you expect to observe with the highest rate *per site* in the human genome? ("Rate per site" can be defined as the frequency with which the mutation is observed, divided by the number of sites where it could have occurred. So if there are 1,000 sites in the genome where mutation X could possibly occur, but in a given individual only 12 of those sites actually exhibit mutation X, then X's rate per site is 0.012.)

○ Single nucleotide substitution

○ Multiple substitutions at a run of consecutive nucleotides

○ Deletion of an individual nucleotide

○ Duplication of an entire gene

○ Microsatellite expansion or contraction

○ Translocation of a chromosome

Question 2

1 pts

Which of the following properties of the EcoRI restriction enzyme most plausibly explains why its binding sites are palindromic?

○ The protein contains the PD..D/EXK motif

○ The enzyme cleaves a phosphodiester bond

○ The enzyme structure contains a four-helix bundle

○ The protein structure is a homodimer and each monomeric unit binds DNA

○ On cutting the DNA, the enzyme leaves a 3' overhang

Question 3

1 pts

In a uniform, independent & identically distributed sequence of nucleotides the length of the *E.coli* genome, roughly how many times would you expect to see the motif AGCATGCT? Give your answer to two significant figures.

Question 4

2 pts

Flag all correct statements about transposable elements (TEs).

□ TEs constitute at least 10% of the human genome

□ Like viruses, TEs can be classified by whether they use DNA or RNA for their genetic material

□ TE insertion only ever occurs at one spot in the genome, and causes no directly observable phenotypes

□ There are examples of TEs depositing binding sites in promoter regions upstream of genes, which can affect host gene expression

□ TEs have been used as tools for genetic engineering and gene therapy

Question 5

3 pts

Match each file format to the best description of the type of data it primarily models.

FASTA format	[Choose]
New Hampshire format	[Choose]
GFF format	[Choose]
BigBed format	[Choose]
CRAM format	[Choose]
BigWig format	[Choose]
WIG format	[Choose]

Question 6

1 pts

Select all true statements about pseudoknots.

□ A pseudoknot is generally more stable than a stem-loop with the same number of base-pairs

□ Transfer RNAs (tRNA) structures contain pseudoknots

□ A pseudoknot contains base-pairs in the arrangement A...X...B...Y where A is paired to B and X is paired to Y

□ One hypothesized role for pseudoknots is to trigger programmed ribosomal frameshifts e.g. in virus genomes

□ A pseudoknot is a junction between three or more helical stems in an RNA structure

Question 7

1 pts

Starting with an RNA sequence that is completely unfolded, the formation of the first base-pair is often energetically disfavored. Why is this?

○ Because there is an entropy cost for an otherwise unconstrained polymer to loop back on itself

○ Because base-pair hydrogen bonding itself is energetically unfavorable: it is only the stacking of adjacent base-pairs that stabilizes the structure

○ Because the first base-pair that is formed is random, and may not be a Watson-Crick basepair

○ Because of van der Waals interactions and steric constraints between adjacent basepairs

○ Because favorable stem formation requires a very specific loop sequence, such as a tetraloop or triloop

Question 8

1 pts

Which of the following chemical reactions or interactions is not known to be part of the repertoire of ribozymes?

○ Polypeptide bond elongation

○ Phosphodiester bond cleavage

○ Redox reactions

○ Recognition of small-molecule ligands

○ Hydrogen bond-mediated pairing between complementary nucleotides

Question 9

1 pts

To the nearest power of 10, how many nucleotides are there in the genome of the DNA herpesvirus that causes infectious mononucleosis?

○ 1

○ 10

○ 100

○ 1,000

○ 10,000

○ 100,000

○ 1,000,000

Question 10

1 pts

Which of the following is a signature of purifying selection?

○ GC content over 50%

○ GC content under 50%

○ GC content exactly 50%

○ Ka/Ks > 1

○ Ka/Ks < 1

○ Ka/Ks = 1

○ Free energy of folding is positive

○ Free energy of folding is negative

Question 11

4 pts

Consider the following dynamic programming recursion for a score designed to measure the RNA folding potential of a sequence X:

$$N(i,j) = \max \begin{cases} N(i+1, j-1) + \delta(X_i, X_j) & \text{if } j > i+2 \\ \max_{k=i}^{j-1} \left(N(i,k) + N(k+1,j)\right) & \text{if } j > i \\ 0 \end{cases}$$

where $\delta(a, b)$ is the following scoring function

a	b	$\delta(a,b)$
A	U	2
C	G	3
G	C	3
U	A	2
G	U	1
U	G	1
All other pairs of *(a,b)* values	All other pairs of *(a,b)* values	0

This recursion can be viewed as a modified version of the Nussinov algorithm. For the sequence "GAUCUA", the (partially filled) table is as follows:

i	j	$N(i,j)$	$X_i \ldots X_j$
5	6	0	UA
4	5	0	CU
4	6	?	CUA
3	4	0	UC
3	5	0	UCU
3	6	2	UCUA
2	3	0	AU
2	4	0	AUC
2	5	2	AUCU
2	6	?	AUCUA
1	2	0	GA
1	3	0	GAU
1	4	?	GAUC
1	5	3	GAUCU
1	6	?	GAUCUA

Fill in the missing values of the table.

N(4,6)	[Choose] ▴▾
N(2,6)	[Choose] ▴▾
N(1,4)	[Choose] ▴▾
N(1,6)	[Choose] ▴▾

Question 12

3 pts

How does the algorithm in the previous question (let's call it Algorithm X) differ from the Nussinov algorithm? Select all correct statements that apply.

- □ Algorithm X can find pseudoknotted structures
- □ Algorithm X enforces a minimum separation between paired bases
- □ Algorithm X allows wobble base-pairs
- □ Algorithm X does not allow bulges in stems
- □ Algorithm X scores base-pairs more highly if they form more hydrogen bonds

Question 13

1 pts

Folding a 10kb sequence using an open-source implementation of Zuker's algorithm takes about 20 seconds on my laptop. Roughly how many seconds should I expect it to take to fold a 100kb sequence?

Question 14

1 pts

Review the proposed RNA folding exercise of the previous question. Aside from the running time, what do you think is likely to be the biggest practical obstacle to running this analysis, as described?

○ You can't predict big RNA structures using Zuker's algorithm

○ Zuker's algorithm neglects basepair stacking, which is important

○ Zuker's algorithm only counts the number of basepairs, not their free energies

○ A typical laptop may not have enough memory for the task described

○ Zuker's algorithm is not implemented in any open source software packages

Question 15

1 pts

Which of the following functions can **not** be asymptotically bounded from above using a big-O notation bound of the form $\mathcal{O}(x^n)$, for large enough x at some value of n?

1. $\mathrm{a(x) = (x+3)(x-5)}$
2. $b(x) = \sum_{i=1}^{20} \sum_{j=i}^{20} \sum_{k=j}^{20} \sum_{l=k}^{20} \sum_{u=i}^{j} \sum_{v=i}^{j} x^3$
3. $c(x) = x \log(x) + x^4$
4. $d(x) = x \exp(3x)$

5. $f(x) = 1/x$
6. $g(x) = x^2 \sin(x)$
7. $h(x) = \frac{(x+5)(x-2)}{x+3}$

□ a(x)

□ b(x)

□ c(x)

□ d(x)

□ f(x)

□ g(x)

□ h(x)

Question 16

1 pts

A Metropolis-Hastings sampler, started in a particular state x, proposes a move from state x to state y. The sampler's proposal distribution is symmetric (that is, if the sampler were to be started in state y, it would propose the reverse move $y \rightarrow x$ with the same frequency that it proposes the forward move $x \rightarrow y$ when it's started in state x).

The sampler is designed to sample an energy landscape defined by $E(x)$, spending more time in states with lower energy. More precisely, the sampler is designed such that the number of samples at a particular state u should (asymptotically, in the limit of running the sampler for a large number of iterations) be proportional to $\exp(-E(u)/k_B T)$, where $k_B = 1.38064852 \times 10^{-23}$ Joules/Kelvin is Boltzmann's constant and T is the temperature (in Kelvin). The sampler accomplishes this by probabilistically rejecting some proportion of energy-increasing moves $u \rightarrow v$, accepting only a fraction $A(u, v)$ of such moves (energy-decreasing moves are always accepted).

The values of the energy function for the two states x and y are

$E(x) = 4.85 \times 10^{-21}$ Joules

$E(y) = 6.24 \times 10^{-21}$ Joules

What, to three significant figures, is the probability that the proposed move $x \rightarrow y$ is accepted by the sampler if the temperature is 293 Kelvin?

Question 17

1 pts

A genome has a GC content of 60% within intergenic regions, and 50% within gene regions. The proportion of the genome lying inside gene regions is 30%; the rest is intergenic.

A position in the genome is randomly sampled; the nucleotide at that position is a G. What, to three significant figures, is the posterior probability that the sampled position was in a gene region?

Question 18

1 pts

In the probabilistic interpretation of the k-means algorithm, what is the underlying probability distribution that explains the observed data, and what is its relationship to the clustering algorithm?

○ A mixture of binomial distributions, with one mixture component per cluster

○ A mixture of Poisson distributions, with one mixture component per cluster

○ A mixture of Gaussian distributions, with one mixture component per cluster

○ A mixture of binomial distributions, with one mixture component per datapoint

○ A mixture of Poisson distributions, with one mixture component per datapoint

○ A mixture of Gaussian distributions, with one mixture component per datapoint

Question 19

1 pts

Which probability distribution is most appropriate for estimating the statistical significance of a sequence alignment score?

○ Gaussian distribution

○ Binomial distribution

○ Extreme value distribution

○ Exponential distribution

○ Gamma distribution

Question 20

1 pts

I roll a fair six-sided die, and then flip a fair coin as many times as the number on the die. (For example, if the die roll comes up 3, then I flip the coin 3 times.)

What is the expected number of times that the coin comes up heads?

Question 21

1 pts

In a Wright-Fisher model with mutation and a diploid population of size N=25, totalling 2N=50 chromosomes, a new mutant arises in the population at time step zero, so that (initially, at time t=1) exactly one of the fifty chromosomes has the mutant allele. Assuming that the allele is selectively neutral, and under the usual perfect mixing assumptions of the Wright-Fisher model, what is the mean number of chromosomes that will have the mutant allele at time t=1, i.e. after one generation has elapsed?

Question 22

1 pts

Continuing the previous question about the Wright-Fisher model, what is the **variance** of the number of chromosomes with the mutant allele at time t=1?

Question 23

1 pts

A FASTA file contains a 100-kilobase DNA sequence (named "test") which has GC content of 70%, has the same nucleotide composition as its reverse complement, and can be regarded as an IID sequence. The file is run through an efficient general-purpose compression utility. Roughly how many **bytes** in length would you expect the compressed file to be? Give your answer to three significant figures.

Question 24

3 pts

There is considerable ongoing interest in *DNA storage*, the process of encoding binary files molecularly as DNA sequences. This question concerns an encoding scheme for DNA storage.

Specifically, consider the following algorithm for encoding a binary sequence $b_1 b_2 b_3 \ldots b_L$as a sequence of DNA nucleotides $n_1 n_2 n_3 \ldots n_L$:

- Define p_i to be the "previous" nucleotide, as follows:
  - If $i = 1$ then $p_i \;=\; A$
  - If $i > 1$ then $p_i \;=\; n_{i-1}$
- For $i = 1$ to L:
  - The nucleotide n_i should be chosen so as to satisfy **all three of** the following conditions:
    - n_i is not equal to p_i;
    - n_i is not equal to the Watson-Crick complement of p_i;
    - n_i is related to p_i by a transition substitution if $b_i = 0$, and by a transversion substitution if $b_i = 1$.

The binary input sequence ($b_1 b_2 b_3 \ldots b_L$) may be regarded as a random variable, in which case the encoded DNA sequence ($n_1 n_2 n_3 \ldots n_L$) is also a random variable. Assume that the input is a uniformly-distributed IID sequence of random bits.

Select all true statements about this code.

□ The Shannon entropy of the encoded DNA sequence ($n_1 \ldots n_L$) is L bits.

□ The encoded binary sequence can always be accurately and uniquely recovered (decoded) even if there are single-nucleotide duplication errors when sequencing the DNA (e.g. an "A" is sometimes read by the decoder as "AA"), as long as there are no other kinds of error.

□ The encoded binary sequence can be accurately and uniquely decoded even if there are multi-nucleotide deletion errors when sequencing the DNA (e.g. "GACT" is misread as "GT"), as long as there are no other kinds of error.

□ The encoded binary sequence can be accurately and uniquely decoded even if there are occasional single-nucleotide transition errors when sequencing the DNA (e.g. "GACT" is misread as "GAGT"), as long as there are no other kinds of error.

□ The code can be said to be "ideal", in that it compresses the maximum amount of information into a sequence of L nucleotides

□ Given the encoded DNA sequence and its reverse complement, it will not (in general) be possible to tell which is the real encoded sequence and which is the reverse-complement. In other words, the encoded binary sequence can NOT be accurately and uniquely decoded if the decoder does not know which DNA strand it is on, even in the absence of sequencing errors.

□ The marginal distribution of any individual nucleotide n_i is a uniform distribution.

□ The encoded sequence $n_1 n_2 n_3 \ldots n_L$ is an IID sequence.

□ In the absence of sequencing errors, and assuming that the length L, strand orientation, and start- & end-points of the DNA sequence $n_1 n_2 n_3 \ldots n_L$ are known correctly, the input binary sequence $b_1 b_2 b_3 \ldots b_L$ can always be recovered uniquely, without ambiguity.

Question 25

1 pts

The so-called "2bit" file format is a compressed bioinformatics file format for storing DNA sequence data. Its complete technical specification is unnecessary to answer this question (but may be found here (http://genome.ucsc.edu/FAQ/FAQformat.html#format7) if you need clarification).

Briefly, the format describes some metadata (such as the name and length of the sequence), followed by the DNA sequence itself which uses 2 bits per nucleotide. The file can contain multiple such records (so the first sequence can be followed by another metadata block and then another sequence, and so on).

Select all the conditions under which the file size of randomly-generated DNA sequence encoded using this format would asymptotically approach the Shannon limit for compression of the encoded DNA.

□ The file must be encoded as a text flatfile, not as binary

□ The encoded DNA is long compared to the metadata

□ The file must contain only one DNA sequence

□ The encoded DNA is an IID sequence

□ There are no repeated bases in the encoded DNA

□ The encoded DNA is uniformly distributed

Question 26

1 pts

Which of the following is a correct compression-oriented interpretation of Gibbs' Inequality, $D(P||Q) \geq 0$ where $D(P||Q)$ is the relative entropy of two probability distributions P and Q?

○ To encode a symbol that was sampled from distribution P, a code that is ideal for Q takes, on average, at least as many bits as a code that is ideal for P.

○ The average number of bits used by an ideal code to compress a symbol sampled from P is the Shannon entropy of Q.

○ The maximum possible value of the Shannon entropy for P is the log of the number of outcomes in Q.

○ The random variables modeled by P and Q are independent.

○ It is possible to transmit a signal error-free on a noisy channel, if sufficient redundancy is introduced.

Question 27

1 pts

What is the "central dogma of molecular biology"?

○ Information flows from RNA to proteins, never from proteins to RNA

○ Sequence determines structure, which determines function

○ Information wants to be free

○ Amino acids can be modeled as hydrophobic or hydrophilic

○ Every rule in biology has its exceptions

Question 28

1 pts

What is the mechanism of the nucleic acid logic circuit described in the following paper:

Science. 2006 Dec 8;314(5805):1585-8.
Enzyme-free nucleic acid logic circuits.

Seelig G, Soloveichik D, Zhang DY, Winfree E.

○ RNA auto-cleavage using an edited hammerhead ribozyme

○ Allosteric unfolding and refolding of nucleic acid complexes

○ Transcriptional control using engineered transcription factors and promoters

○ Transcriptional control using engineered terminator stem-loops

○ Translational control using an engineered ribosome binding site

Question 29

1 pts

Given a uniform distribution over nucleotides, what is the probability that two independently sampled nucleotides will form a canonical Watson-Crick base-pair?

Question 30

1 pts

Given a uniform distribution over nucleotides, what is the mutual information of two independently sampled nucleotides?

Quiz saved at 9:48pm Submit Quiz