Past midterm exam (2017)

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 mutation events would you expect to occur most frequently in non-coding, non-selected regions of the human genome?

- ○ Single nucleotide mutation of A to T
- ○ Single nucleotide mutation of A to G
- ○ Single nucleotide mutation of A to C
- ○ Single nucleotide mutation of C to G
- ○ Single nucleotide insertion of a C between an A and a G (AG becomes ACG)
- ○ Single nucleotide deletion of a C between an A and a G (ACG becomes AG)

Question 2 — 1 pts

Which of the following is a common structural basis for DNA-protein recognition by transcription factors containing the helix-turn-helix domain?

- ○ Complementary base-pairing between nucleotides in transcription factor and nucleotides in binding site

- ○ Phosphorylation of transcription factor by kinase proteins
- ○ Van der Waals interactions between transcription factor and DNA backbone
- ○ Insertion of protein alpha-helix into DNA major groove
- ○ Hydrogen bonding between amino acids and unpaired nucleotides in single-stranded regions

Question 3 1 pts

In a uniform, independent & identically distributed sequence of nucleotides about the length of the HIV genome, roughly how many times would you expect to see the motif ACGACG? Give your answer to one significant figure.

Question 4 2 pts

Flag all correct statements about the Alu element.

- ☐ Copies of Alu comprise about 11% of the human genome
- ☐ Alu operates by a "copy and paste" mechanism, excising its own DNA and then re-integrating
- ☐ Alu is descended from the signal recognition particle RNA
- ☐ Alu has not been an active transposon in the human genome since before the split between New World / Old World monkeys, about 40 million years ago
- ☐ Alus have no associated disease phenotypes currently known

Question 5

2 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]
BED format	[Choose]
Stockholm format	[Choose]
JSON format	[Choose]
XML format	[Choose]

Question 6

1 pts

What is a pseudoknot?

- An RNA secondary structure containing overlapping base-pairs of the form A...B...X...Y where A is paired to B and X is paired to Y
- An RNA secondary structure containing overlapping base-pairs of the form A...X...Y...B where A is paired to B and X is paired to Y
- An RNA secondary structure containing overlapping base-pairs of the form A...X...B...Y where A is paired to B and X is paired to Y
- The unpaired single-stranded region at the end of an RNA stem
- 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 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

What is the mechanism of action of the hammerhead ribozyme?

- Polypeptide bond elongation
- Phosphodiester bond cleavage
- Nucleophilic attack on the alpha carbon of the peptide group
- Hydrogen-bonding between enzyme and substrate
- Denaturing induced by low pH
- Conglomeration of hydrophobic residues

Question 9

1 pts

To the nearest power of 10, how many nucleotides are there in the hammerhead ribozyme?

- 1
- 10
- 100
- 1,000
- 10,000
- 100,000

1,000,000

Question 10

1 pts

Which of the following is a signature of diversifying 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

The Nussinov algorithm for a sequence X can be defined by the recursion

$$N(i,j) = \max \begin{cases} N(i+1, j-1) + \delta_{wc}(X_i, X_j) \\ N(i+1, j) \\ N(i, j+1) \\ \max_k \left(N(i,k) + N(k+1, j)\right) \end{cases}$$

where $N(i,j)$ is the maximum number of complementary basepairs that can be formed by subsequence $X_i \ldots X_j$ and $\delta_{wc}(X_i, X_j)$ is a scoring function that

returns 1 if X_i and X_j are complementary, and 0 if they are not.

For the sequence "gaucua", the (partially filled) Nussinov table is as follows:

i	j	$N(i,j)$	$X_i \ldots X_j$
5	6	1	ua
4	5	0	cu
4	6	?	cua
3	4	0	uc
3	5	0	ucu
3	6	1	ucua
2	3	1	au
2	4	1	auc
2	5	1	aucu
2	6	?	aucua
1	2	0	ga
1	3	1	gau
1	4	?	gauc
1	5	2	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

4 pts

Match each of the RNA folding-related algorithms to its application.

Nussinov algorithm	[Choose]
Zuker algorithm	[Choose]
McCaskill algorithm	[Choose]
Kinfold	[Choose]

Question 13

1 pts

Running RNAfold on a 4kb sequence on my desktop (using the default parameters) takes 7 seconds. For roughly how many seconds would you expect RNAfold to run on an 8kb sequence?

Question 14

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 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 15

1 pts

The Rfam database of RNA domain families has a built-in search tool, Infernal, that builds profiles of RNA sequence alignments and can be used to search RNA sequences. What is the underlying statistical model used by Infernal?

- ( ) Support vector machines
- ( ) Recurrent neural networks

- Hidden Markov models
- Stochastic context-free grammars
- Tree-adjoining grammars
- The partition function

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

$E(x) = 6.72 \times 10^{-21}$ Joules

$E(y) = 9.83 \times 10^{-21}$ 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 40% within intergenic regions, and 50% within gene regions. The proportion of the genome lying inside gene regions is 20%; 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 (haploid) population of size 50, a new mutant arises in the population at time step zero, so that (initially) exactly one of the fifty genotypes in the population has the mutant allele. Assuming that the allele is selectively neutral, what is the probability that it will eventually become fixed in the population by random drift?

Question 22

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?

Question 23

1 pts

The coding scheme invented by Goldman and Birney (2013) for storing information in synthetic DNA molecules works as follows:

First, the (base-2) binary sequence of 0's and 1's to be encoded is converted into a (base-3) ternary sequence of 0's, 1's and 2's. For example, the binary sequence 100110, considered as a number in base 2, corresponds to the number 38 in decimal (base 10); and in ternary (base 3), this number is 1102.

Let the ternary sequence be denoted by $x_1 x_2 x_3 x_4 \ldots$ with (for example)

$x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 2$ for the ternary sequence 1102.

Next, the ternary sequence is converted to a (base-4) quaternary sequence, *in which no digit is ever repeated.* This sequence can be written $y_1 y_2 y_3 y_4 \ldots$ and is generated by setting $y_1 = x_1$ and (for $n > 1$) $y_n = (y_{n-1} + x_n + 1) \mod 4$. So the example ternary sequence 1102 from the paragraph above would be converted to the quaternary sequence 1303.

Finally, the quaternary sequence is converted directly to DNA, using a straightforward mapping (e.g. 0 is A, 1 is C, 2 is G, and 3 is T, so that the quaternary sequence 1303 would be converted to the DNA sequence CTAT). Let this final, DNA sequence be denoted $z_1 z_2 z_3 z_4 \ldots$.

The goal of this scheme is to avoid consecutive identical nucleotides in the DNA sequence ("homopolymer runs"), which can be problematic for sequencing accuracy (for many sequencing technologies, homopolymers tend to be more error-prone when sequenced).

Consider a DNA sequence of 100 nucleotides $(z_1 \ldots z_{100})$ generated to encode some source binary sequence using this technique. You can assume that the source binary sequence is IID and the binary digits are uniformly distributed.

How many bits of Shannon entropy does the 100-nucleotide sequence have *in total* (not per-symbol)? Give your answer to three significant figures.

Question 24 1 pts

Continuing with the Goldman-Birney coding scheme of the previous question, consider picking a random position n from the generated DNA sequence. What is the Shannon entropy, in bits, of the marginal probability distribution $P(z_n)$ for the nucleotide at position n? Give your answer to three significant figures.

Question 25

1 pts

Continuing with the Goldman-Birney coding scheme of the previous question, consider picking a random pair (z_n, z_{n+1}) of consecutive nucleotides from the output DNA sequence. What is the mutual information $M(z_n, z_{n+1})$ in bits? Give your answer to three significant figures.

Question 26

1 pts

Which of the following compression techniques do **not** form a part of the CRAM standard for compressing short reads by alignment to a reference genome?

- [ ] Burrows-Wheeler transform to compress the reference genome
- [ ] Golomb/Elias-Gamma codes to encode distance between reads
- [ ] Huffman codes to encode read quality scores
- [ ] Arithmetic coding to encode distances between mismatches
- [ ] Run-length encoding to encode repeated bases

Question 27

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?

- Using an ideal code for Q to encode a symbol from P is, on average, no better than using an ideal code 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 28

1 pts

The RNA-binding protein database (http://rbpdb.ccbr.utoronto.ca/) lists 416 RNA-binding proteins in the human genome. Suppose, as a hypothetical (that is certainly not true in practice), that all of the following conditions hold:

- each of these sequences recognizes a distinct corresponding RNA binding site,
- the RNA binding site for each of the 416 proteins is the same length (N nucleotides),
- a protein either binds or it doesn't: there is no quantitative degree of binding (more precisely, the binding constant is either zero or infinity),
- at each of the N positions of the binding site, there are exactly two possibilities for what the nucleotide can be, if the protein is to bind,
- no two proteins recognize the same binding site sequence.

What is the minimum integer value of N, if all these conditions hold?

Question 29

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 30

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 31

1 pts

How many different colors have been produced using engineered or naturally-occurring variants of "green fluorescent protein"?

- One: green
- Two: green and yellow
- Three: green, yellow, and red
- More than three

Question 32

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?

Not saved Submit Quiz