Gaps are commonly observed in a sequence alignment. Briefly describe why gaps are important in sequence alignment, and what biological phenomena could have caused this observation in an alignment?

Gap allows for maximisation of similarity of sequences, i.e. adjusting the residue positions using gap allow for more identical residues to match up. This is caused by insertions and/or deletions (sometimes this is referred to as indels).

Why is sequence conservation (or homology) commonly observed at the protein level? Give TWO (2) reasons.

Multiple codons code for the same amino acid (codon degeneracy issue), thus DNA sequences do not adequately capture the conservation of gene function exhibited at the protein level.

Protein sequences also have a larger alphabet size (i.e. 20 amino acids) compared to DNA (4 bases), so the information content of protein sequences is greater.

Name one limitation of the progressive multiple sequence alignment approach and suggest an alternative approach to address this limitation. (2 marks)

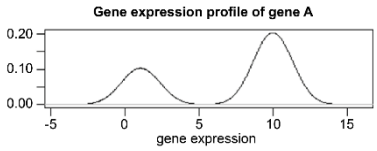
Limitation: Errors in the early steps of progressive MSA are fixed and propagated throughout the process.

Alternative: refine the earlier alignments iteratively and see if better alignments (i.e. alignments with higher scores) can be achieved (i.e. the iterative progressive MSA approach).

Describe one reason why RNA-sequencing may be more advantageous than microarray technology for capturing gene expression data. (3 marks)

RNA-seq covers entire transcriptome without relying on a reference; more information can be extracted e.g. splicing, SNPs, non-coding RNAs; greater sensitivity in estimating expression than microarrays.

The plot below shows the gene-expression distribution for Gene A in a population of 1000 cells, as observed in a single cell RNA-sequencing experiment.



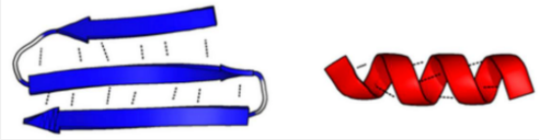
There are two sub-populations within the 1000 cells that express different levels of Gene A.

Name the open source project that contains comprehensive bioinformatics software in the R language.
Bioconductor

Name two (2) computational approaches for gene finding.

Intrinsic (or ab initio), Homology-based (based on sequence similarity)

Name four (4) specific sequence features of eukaryotic genes that can be incorporated into a Hidden Markov model for predicting protein-coding genes. (4 marks) (motifs of) splice site, Poly-A site, (sequence characteristics of) promoter region, codon usage bias, presence of ORFs (and any other plausible features).



A. The schematic structure on the left (in the figure above) is a transmembrane domain, consisting of three alpha strands. True or false? (1 mark)

B. The schematic structure on the right is an alpha helix. True or false? (1 mark)

C. The dotted lines in both structures represent hydrogen bonds. True or false? (1 mark)

F. Both proline and glycine tend to support the formation of beta strands; therefore their absence cause kinks in beta strands. True or false? (1 mark)

A: _____; B: _____; C: _____; D: _____.

Selected Answer: [None Given]

Correct Answer: A: False; B: True; C: True; D: False.
Alternative answer: True: BC, False: AD

Which of the following properties of a phylogenetic tree are an indication of insufficient evidence to resolve a tree during phylogenetic inference? Multifurcating branch points.

Gene Ontology is organised in three structured, species-independent ontologies. Name these three ontologies
Biological Process (2/3), Molecular Function (2/3), Cellular Component (2/3).

Sequence Analysis

What is homology? Orthology? Paralogy?

The study of a common evolutionary origin. Same origin, different species. Same origin, same species.

What does a phylogenetic tree represent? What does "unrooted", "ultra-metric", "additive" mean for a tree?

Evolutionary relationships; descendent, time/distance. An ultra-metric tree has a root, and a constant rate of mutation along all branches; additive trees indicate progress of divergence of species additively.

What is an evolutionary model? Describes how sequences change often relative time/rate of evolution

How do we infer a phylogenetic tree?

Sequence similarity gives alignment; Evolutionary models explain differences; Measuring distances; pairing, clustering, inferring ancestor states.

What are "sequence motifs"?

Short (often linear) sequence snippets, statistically enriched, carry function/structural characteristics; not necessarily strongly conserved.

How can we represent motifs and how does searching for them work?

Consensus; reg. expression; PWM. Matching by window over sequence; prob. Score.

What is a sequence logo and why is it informative?

Visualisation. Displays level of agreement/level of unpredictability or information entropy in each "column" by

"height of stack", shows striking/dominant feature/s.

What method finds "labels" of nodes in a tree?

Maximum parsimony; ML and other prob models.

Why is sequence conservation (or homology) commonly observed at the protein level? Give two reasons.

Codon degeneracy issue – codon is triplet – multiple codons may code for the amino acid

4 nucleotides (A, C, G, T) vs. 20 amino acids (so there are greater diversity and they are much more informa content)

"Two protein sequences that share 99% identity are 99% homologous, and 1% non-homologous." Do you agree?

No. homology is a statement, 2 seqs are either homologous or not, there's no extent of homology.

Provide one reason why log-adds scores are used in most substitution matrices.

We can sum them up, instead of dealing with multiplication of probability.

Name a key difference between a global alignment and a local alignment.

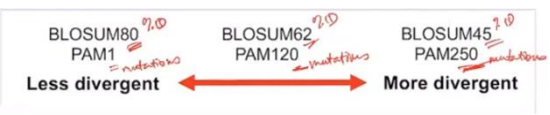
Global: comparing whole sequences | Needleman-Wunsch | less prone to give us false homology.

Local: localised similar regions | Smith-Waterman | more prone to give us false homology.

PAM and BLOSUM are two commonly used scoring matrices for amino acid substitutions. Name one key difference between PAM and BLOSUM.

PAM, the smaller the unit, the more different the sequences are, the number is number of mutations.
BLOSUM, the higher number suggest the higher similarity, so this is because this number is percent identity.

The smaller the unit in PAM, the more similar sequences are,
whereas the larger the value in BLOSUM more similar sequences are.



Marcy would like to align two protein seqs that share less than 25% identity. She decided to use PAM1 as the scoring matrix. Do you think the PAM matrix she has chosen is appropriate in this case? Briefly explain why.

No. Maybe PAM250. Based on highly similar sequences is, this based on sequences that share 85% identity or more, so given that Marcy wants to compare sequences they share less than 25% identity, then PAM1 would not be appropriate.

Gaps are commonly observed in a sequence alignment. Briefly describe why gaps are important in sequence alignment, and what biological phenomena could have caused this observation in an alignment.

In sequence alignment, we want to maximize similarity between or among the sequences that we're comparing, so gap would help us do that, so that the similar regions will be aligned together.

From the biological perspective, it reflects the insertions and deletions that would have occurred during the evolutionary history of the sequences.

Sequence alignment

Alice would like to compare two DNA sequences, seq1 (ACCATCGGA) and seq2 (ATCCATGGGA). She came up with two pairwise alignments, alignment *X* and alignment *Y* (shown below). She would like to determine which of these two alignments is more optimal using the associated scoring scheme (below).

```
seq1 A-CCATCGGA
seq2 ATCCATGGGA
alignment X

seq1 ACC-ATCGGA
seq2 ATCCATGGGA
alignment Y
```

scoring scheme

```
match     =  3
mismatch  =  1
gap open  = -3
gap extend = -1
```

Alice has determined that alignment *Y* is more optimal than alignment *X*. Do you agree with Alice? Briefly explain why and show your calculations. (2 marks)

NO.

alignment X: 8(3) + 1 + (-3) = 22 ✓✓

alignment Y: 7(3) + 2(1) + (-3) = 20

Sequence alignment

Consider the multiple sequence alignment below of six protein sequences. Conserved aligned positions are in boldface (asterisk underneath); positions that consist of amino acids of similar physicochemical properties are noted with a colon. Dr Stephens has informed you that this protein family should also have a conserved DFG region near the end of the sequences.

```
seq1 DRHSDNINVKKTGQLFHIIDFG
seq2 DRHSDNIMIRESGQLFHIIDFG
seq3 DRHNSNIMVKDDGQLFHIDAFG
seq4 DRHNDNIMITETGNLFHID-FG
seq5 DRHNDNLMVTKGGRLFHID-FG
seq6 DLKPENLLIDQQGYIQVID-FG
     *   :*: :   *       **
```

Based on the given information, do you think this alignment can be further improved? Briefly explain your answer and suggest a solution. (2 marks)

we don't need to have a column of gap, because you know gap is a penalty will just reduce the score.

Dynamic programming

Below is the dynamic programming matrix (with the associated scoring scheme) that is used to find the optimal Needleman-Wunsch alignment between two sequences, *X* (CATG) and *Y* (CAG). Use this matrix to answer A and B.

Scoring scheme
match = 2
mismatch = 0
gap = -2

X→ Y↓	gap	C	A	T	G
gap	0	-2	-4	-6	-8
C	-2	2	0	-2	-4
A	-4	0	?	2	0
G	-6	-2	2	4	4

A. Calculate the score and traceback direction for the box marked with a question mark. Show your calculations. (2 marks)

4? max ← (-2) + 0 = -2
↑ (-2) + 0 = -2
↖ 2 + 2 = 4 ✓

B. Based on the completed matrix, write out the optimal pairwise alignment of sequences *X* and *Y*. (2 marks)

Seq X CATG
Seq Y CA-G

Name one key limitation of progressive multiple sequence alignment. Errors get propagated.

Name two key approaches adopted in Clustal to establish the hierarchical order for which sequences are progressively aligned. UPGMA + Neighbour joining

Most multiple sequence alignment (MSA) approaches are based on progressive and iterative methods. Name two other MSA approaches. - Consensus(T-COFFEE) – HMM

Analysis of high-throughput sequences

You are given a *de novo* genome assembly that consists of ten (10) contigs. The length of each contig, ordered from the longest to the shortest, is summarised in the table below, together with the cumulative sum of these contig lengths. Based on this table, determine the N75 length of the assembly. Show your calculations. (2 marks)

Contig	Length (bases)	Cumulative sum
contig1	512,654	512,654
contig2	502,310	1,014,964
contig3	335,021	1,349,985
contig4	325,321	1,675,306
contig5	298,654	1,973,960
contig6	278,645	2,252,605
contig7	198,756	2,451,361
contig8	188,868	2,640,229
contig9	102,100	2,742,329
contig10	99,865	2,842,194

total assembled bases

N50 50% total assembled bases = 2,842,194 × 1/2 = 1,421,097

2,842,194 × 3/4 = ...

The 4 longest contigs make up ≥ 50% of the bases

∴ N50 length = length of the 4th longest contig = 325,321

N75 ⇒ 75%

Analysis of high-throughput sequences

In a genome sequencing project, 350,230,200 new sequence reads were generated. Of these reads, 343,123,286 mapped to the human reference genome. Based on this information, discuss the reasons why the new reads are likely from a human genome and not a microbial or plant genome. (1 mark)

343,123,286 / 350,230,200 × 100 = 98 % ⇒ likely human

hypothetically: 343,123 / 350,230,200 × 100 = ? →

Name a type of a continuous representation of a motif. Briefly explain one advantage of representing a motif using a continuous approach. Sensible to frequency of character; Enable matches to be ranked

Consider the following alignment of six proteins. Shaded positions are known to be functional, and the alignment illustrates that they are also conserved.

```
p110β:      SYVLGIG----------DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT
p110δ:      TYVLGIG----------DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT
p110α:      TFILGIG----------DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT
p110γ:      TFVLGIG----------DRHNDNIMITETGNLFHIDFGHILGNYKSFLGINKERVPFVLT
p110_dicti: TYVLGIG----------DRHNDNLMVTKGGRLFHIDFGHFLGNYKKKFGFKRERAPFVFT
cAMP-kinase: QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA
```

A. **Outline two different representations of the conserved positions, which can be used to search other sequences.** Consensus sequence, position specific similarity matrices

B. **Discuss relative advantages and disadvantages of the chosen representations for assigning function to novel sequences.** *Consensus sequence*, limitation: we can only have one resident one particular letter representing each position so we're very limited in what we can capture. *position specific similarity matrices* which are a little bit more intricate and have a little more functionality, as we can look at the variability within a particular position like different options for a particular position, as well as a numerical estimate for those.

Q1a – 2016 Exam

Consensus sequences and regular expressions

ChIP-seq is an assay that allows the experimenter to identify pieces of DNA to which a particular protein is bound. It is often used for determining DNA sequence patterns for binding transcription factors. Suppose you decided to work with a transcription factor YFP2 and ran a ChIP-Seq experiment which showed that YFP2 bound to an 8-mer. From the counts of nucleotides in matching sequences, you calculate the probability matrix for each matching position independently, given below:

Determine the consensus sequence from the probability matrix. Determine the consensus sequence for the alternate strand.

	1	2	3	4	5	6	7	8
A	0.30	0.10	0.25	0.09	0.15	0.30	0.20	0.50
C	0.05	0.10	0.60	0.01	0.40	0.01	0.20	0.15
G	0.45	0.10	0.05	0.20	0.40	0.19	0.30	0.15
T	0.20	0.70	0.10	0.70	0.05	0.60	0.30	0.20

Option 1 (5C, 7G)	Option 2 (5C, 7T)	Option 3 (5G, 7G)	Option 4 (5G, 7T)
GTCTCTGA	GTCTCTTA	GTCTGTGA	GTCTGTTA
CAGAGACT	CAGAGAAT	CAGACACT	CAGACAAT

Determine the consensus sequence.

```
G A G G T A A A C
T C C G T A A G T
C A G G T T G G A
A C A G T C A G T
T A G G T C A T T
T A G G T A C T G
A T G G T A A C T
C A G G T A T A C
T G T G T G A G T
A A G G T A A G T

T A G G T A A G T
```

Question 6 from 2015 Final Exam

You have discovered four DNA binding sites, with 8 nucleotides each, for a transcription factor TF1.

A. Specify the "consensus" sequence motif of the four sites.
B. Identify the most specific "regular expression" of the four sites that matches all of them.
C. Determine how many different DNA sequences that can be matched by the regular expression you correctly specified in (B).

```
1: ACGATAAC
2: ACGATACC
3: AAGATAAC
4: ACAATAAC
```

A. Consensus sequence	B. Regular expression	C. Potential DNA matches
ACGATAAC	. Matches any character [xyz] Matches a single character in brackets {xyz} Matches a single character not in brackets A[CA][GA]ATA[AC]C	8 matches A[CA][GA]ATA[AC]C = ACGATAAC A[CA][GA]ATA[AC]C = ACGATACC A[CA][GA]ATA[AC]C = ACAATAAC A[CA][GA]ATA[AC]C = ACAATACC A[CA][GA]ATA[AC]C = AAGATAAC A[CA][GA]ATA[AC]C = AAGATACC A[CA][GA]ATA[AC]C = AAAATAAC A[CA][GA]ATA[AC]C = AAAATACC

A position weight matrix for a motif is shown below. Based on the information provided, identify the position in sequence X where the motif is most likely to occur, and the score of this motif match.

C.

	1	2	3	4	5
A	1.825	-7.869	-7.869	2.097	-7.869
C	-8.072	-8.072	-8.072	-8.072	1.895
G	-0.632	-8.065	1.830	-8.065	-8.065
T	-7.856	2.110	-2.243	-7.856	-7.856

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

$$P(x|M) = \sum_{i=1}^{K} w_{i,x}$$

	Sequence	Calculation	Score
1	CGTGA	(-8.072) + (-8.065) + (-2.243) + (-8.065) + (-7.869)	-34.314
2	GTGAC	(-0.632) + (2.110) + (1.830) + (2.097) + (1.895)	7.3
3	TGACG	(-7.856) + (-8.065) + (-7.869) +(-8.072) + (-8.065)	-39.927
4	GACGT	(-0.632) + (-7.869) + (-8.072) + (-8.065) + (-7.856)	-32.494
5	ACGTT	(1.825) + (-8.072) + (1.830) + (-7.856) + (-7.856)	-20.129
6	CGTTA	(-8.072) + (-8.065) + (-2.243) + (-7.856) + (-7.869)	-34.105
7	GTTAT	(-0.632) + (2.110) + (-2.243) + (2.097) + (-7.856)	-6.524
8	TTATG	(-7.856) + (2.110) + (-7.869) + (-7.856) + (-8.065)	-29.536
9	TATGA	(-7.856) + (-7.869) + (-2.243) + (-8.065) + (-7.869)	-33.902
10	ATGAC	(1.825) + (2.110) + (1.830) + (2.097) + (1.895)	9.757
11	TGACG	(-7.856) + (-8.065) + (-7.869) + (-8.072) + (-8.065)	-39.927

Position 10 is the most likely start

17

Q1ab – 2013 Exam

Calculating Position-Specific Scoring Matrices

What steps are needed to create PWM1 (no pseudo counts added) and PWM2 (pseudo counts added)?

	PWM1	PWM2
Step 1	Get the counts for each nucleotide at each position (PFM)	Get the counts for each nucleotide at each position (PFM)
Step 2	NO pseudo counts added	YES pseudo counts added
Step 3	Calculate the probabilities for each nucleotide at each position	Calculate the probabilities for each nucleotide at each position
Step 4	Correct nucleotide frequencies using uniform background distribution (PPM1)	Correct nucleotide frequencies using uniform background distribution (PPM2)
Step 5	Calculate log_2 weights (PWM1)	Calculate log_2 weights (PWM2)

```
AAGATA
AGGATA
CGGATT
TGAATA
AAGATA
AGGATA
CGGATT
TGAATA
```

```
A: 0.25
C: 0.25
G: 0.25
T: 0.25
```

PFM :

	1	2	3	4	5	6
A	4	2	2	8	0	6
C	2	0	0	0	0	0
G	0	6	6	0	0	0
T	2	0	0	0	8	2

⇒

PFM + pseudos :

	1	2	3	4	5	6
A	5	3	3	9	1	7
C	3	1	1	1	1	1
G	1	7	7	1	1	1
T	3	1	1	1	9	3

```
AAGATA
AGGATA
CGGATT
TGAATA
AAGATA
AGGATA
CGGATT
TGAATA
```

PPM1 :

	1	2	3	4	5	6
A	0.5	0.25	0.25	1	0	0.75
C	0.25	0	0	0	0	0
G	0	0.75	0.75	0	0	0
T	0.25	0	0	0	1	0.25

PPM2 :

	1	2	3	4	5	6
A	0.42	0.25	0.25	0.75	0.08	0.58
C	0.25	0.08	0.08	0.08	0.08	0.08
G	0.08	0.58	0.58	0.08	0.08	0.08
T	0.25	0.08	0.08	0.08	0.75	0.25

```
A: 0.25
C: 0.25
G: 0.25
T: 0.25
```

PFM ←	nucleotide counts (PFM)	→ PFM
NO ←	pseudo counts?	→ YES
PFM to PPM1 ←	probabilities	→ (PFM + pseudos) to PPM2
PPM1 / 0.25 ←	background correction	→ PPM2 / 0.25
PPM1 to PWM1 ←	weights	→ PPM2 to PWM2

PPM1 / 0.25 :

	1	2	3	4	5	6
A	2	1	1	4	0	3
C	1	0	0	0	0	0
G	0	3	3	0	0	0
T	1	0	0	0	4	1

PPM2 / 0.25 :

	1	2	3	4	5	6
A	1.67	1	1	3	0.33	2.33
C	1	0.33	0.33	0.33	0.33	0.33
G	0.33	2.33	2.33	0.33	0.33	0.33
T	1	0.33	0.33	0.33	3	1

PPM1 / 0.25 :

	1	2	3	4	5	6
A	2	1	1	4	0	3
C	1	0	0	0	0	0
G	0	3	3	0	0	0
T	1	0	0	0	4	1

PPM2 / 0.25 :

	1	2	3	4	5	6
A	1.67	1	1	3	0.33	2.33
C	1	0.33	0.33	0.33	0.33	0.33
G	0.33	2.33	2.33	0.33	0.33	0.33
T	1	0.33	0.33	0.33	3	1

PWM1

	1	2	3	4	5	6
A	1	0	0	2	0	1.58
C	0	0	0	0	0	0
G	0	1.58	1.58	0	0	0
T	0	0	0	0	2	0

PWM2

	1	2	3	4	5	6
A	0.74	0	0	1.58	-1.60	1.22
C	0.00	-1.60	-1.60	-1.60	-1.60	-1.60
G	-1.60	1.22	1.22	-1.60	-1.60	-1.60
T	024	-1.60	-1.60	-1.60	1.58	0

Note: the final step is doing a log base 2 on PPM

Determine a position *frequency* matrix (PFM), representing the counts of nucleotides; convert this to a position *probability* matrix (PPM); show both matrices in your response.

	1	2	3	4	5	6	7	8	9
A	3	6	1	0	0	6	7	2	1
C	2	2	1	0	0	2	1	1	2
G	1	1	7	10	0	1	1	5	1
T	4	1	1	0	10	1	1	2	6

PFM:

	1	2	3	4	5	6	7	8	9
A	4	7	2	1	1	7	8	3	2
C	3	3	2	1	1	3	2	2	3
G	2	2	8	11	1	2	2	6	2
T	5	2	2	1	11	2	2	3	7

PPM:

	1	2	3	4	5	6	7	8	9
A	0.29	0.50	0.14	0.07	0.07	0.50	0.57	0.21	0.14
C	0.21	0.21	0.14	0.07	0.07	0.21	0.14	0.14	0.21
G	0.14	0.14	0.57	0.79	0.07	0.14	0.14	0.43	0.14
T	0.36	0.14	0.14	0.07	0.79	0.14	0.14	0.21	0.50

G A G G T A A A C
T C C G T A A G T
C A G G T T G G A
A C A G T C A G T
T A G G T C A T T
T A G G T A C T G
A T G G T A A C T
C A G G T A T A C
T G T G T G A G T
A A G G T A A G T

T A G G T A A G T

26

Use the probability matrix (from Q4b) to determine the *maximum probability of a match* anywhere in the DNA promoter sequence TGAGGTAAACA.

PPM:

	1	2	3	4	5	6	7	8	9
A	0.29	0.50	0.14	0.07	0.07	0.50	0.57	0.21	0.14
C	0.21	0.21	0.14	0.07	0.07	0.21	0.14	0.14	0.21
G	0.14	0.14	0.57	0.79	0.07	0.14	0.14	0.43	0.14
T	0.36	0.14	0.14	0.07	0.79	0.14	0.14	0.21	0.50

	Sequence	Calculation	Score
1	TGAGGTAAA	0.36 * 0.14 * 0.14 * 0.79 * 0.07 * 0.14 * 0.57 * 0.21 * 0.14	0.00000092
2	GAGGTAAAC	0.14 * 0.50 * 0.57 * 0.79 * 0.79 * 0.50 * 0.57 * 0.21 * 0.21	0.00031
3	AGGTAAACA	0.29 * 0.14 * 0.57 * 0.07 * 0.07 * 0.50 * 0.57 * 0.14 * 0.14	0.00000063

The maximum probability of match is 0.00031.

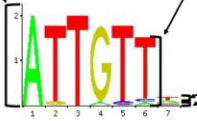
A sequence logo for the transcription factor Sox5 is below. Describe in words or in mathematical terms how a sequence logo is determined; in particular, what determines the height of all and individual letters for each position.

Entropy is represented by the total height of a column in *bits*.

$$H_i = -\sum f_{i,a} \log_2 f_{i,a}$$

The total height of characters in a column is the **Information Content.**

$$I_i = H_{max} - H_i$$



The height of an individual character is the **frequency of the residue.**

$$n_{i,a}$$

An example logo for so-called signal peptides is provided below. What is the maximum height a letter can take when the 20-letter amino acid alphabet is used for the analysed sequences? It is sufficient that you provide the expression that will give the answer value.

The maximum entropy, or when each character is equally likely, is the scenario of the highest uncertainty and determines the height of a column.

$$H_i = -\sum f_{i,a} \log_2 f_{i,a}$$

For nucleotides, it was:

$f_{i,a} = \frac{1}{4}$ for all $a \in \{A, C, G, T\}$

$$H_{max} = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)$$

OR

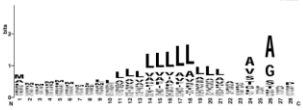
$$H_{max} = -\sum_{i=1}^{4} \frac{1}{4}\log_2 \frac{1}{4}$$

$$\boldsymbol{H_{max} = 2\ bits}$$

For proteins it would be:

$$H_{max} = -\sum_{i=1}^{20} \frac{1}{20}\log_2 \frac{1}{20}$$

$$\boldsymbol{H_{max} = 4.32\ bits}$$



Gene Expression Analysis

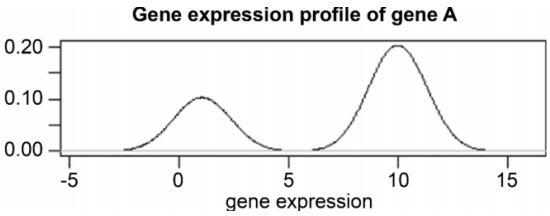
A. **Name two technology platforms for generating high throughput gene expression data.**
RNA-Seq and Microarray

B. **What is the name of the largest repository that houses publicly available gene expression data?**
GEO (database within NCBI)

C. **For a given list of genes, a common approach is to investigate whether a specific pathway is enriched in this list. What is the statistical test used for over-representation analysis to determine whether the enrichment of a pathway is significant?** Fisher's exact test or hypergeometric test

D. **Two genes are said to be differentially expressed when their gene expression profiles are different between two groups. What is the statistical test often used to determine whether two genes are differentially expressed?** T-test or limma

The plot below shows the gene-expression distribution for Gene A in a population of 1000 cells, as observed in a single cell RNA-sequencing experiment.



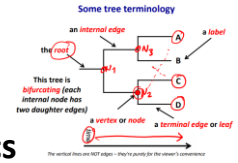
Based on the information contained in this plot, identify the statement that is TRUE.

I. **Gene A is alternatively spliced into two different isoforms.**

II. → ***There are two sub-populations within the 1000 cells that express different levels of Gene A.***

III. **Gene A is differentially expressed.**

IV. **The probe-set for Gene A shows non-specific binding.**



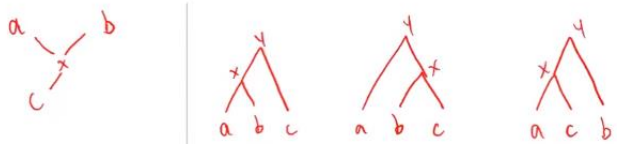
Phylogenetics

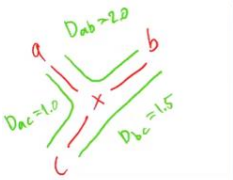
Question 3. Phylogenetics. **Total: 10 marks**

You are provided with sequences for three genes, a, b and c.

A. How many structurally distinct **unrooted**, bifurcating trees can be constructed with a, b and c? (1 mark)

B. How many structurally distinct **rooted**, bifurcating trees can be constructed with a, b and c? (1 mark)





$D_{ax} = D_{ab} - D_{bx} = 2.0 - D_{bx}$

$D_{bx} = D_{bc} - D_{cx} = 1.5 - D_{cx}$

$D_{cx} = D_{ac} - D_{ax} = 1.0 - D_{ax}$

$D_{ax} = 2.0 - (1.5 - (1.0 - D_{ax}))$

$= 2.0 - 1.5 + 1.0 - D_{ax}$

$2D_{ax} = 2.0 - 1.5 + 1.0 = 1.5$

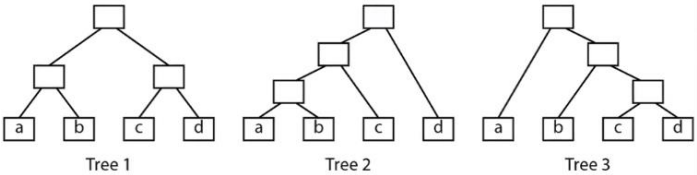
$D_{ax} = 0.75$

C. Draw an unrooted, bifurcating tree for a, b and c, with evolutionary distances assigned to all branches. The *additive* distances are as follows: 2.0 between a and b, 1.0 between a and c, and 1.5 between b and c. (3 marks)

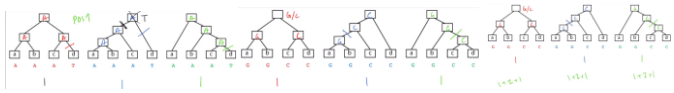
Poll: What is the distance between the "linking node" *x* and *c*? A: 0.25, B: 0.50, C: 0.75, D: 1.00, E: 1.25

D. Which tree below (1-3) assumes the *greatest* evolutionary change, i.e. has the poorest parsimony score? Explain your answer. You should assume that all mutations are weighted equally. (3 marks)

The trees represent potential phylogenetic relationships between four tri-nucleotide sequences a, b, c and d:



Sequences: a=ACG b=AAG c=ACC d=TTC



E. Which phylogenetic tree construction algorithm is described by the following steps? Complete the missing step 4. (2 marks)

1. Compute all pairwise distances (*n*-by-*n* matrix).
2. Find 2 leaves in tree that are close to each other, but far from other leaves. (closest)
3. Determine/invent parent, distances for this parent and collapse and remove pair of leaves from step 2; $n = n - 1$.
4. _____.

Poll:
A. Maximum parsimony
B. Maximum likelihood
C. UPGMA
D. Neighbor joining

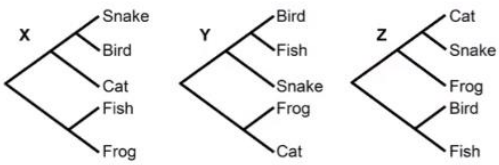
Question 5. Phylogenetics. **Total: 10 marks**

A. Briefly explain the advantage of using a Poisson-corrected distance and a Gamma-corrected distance in phylogenetics. (2 marks)

Poisson distance correction
accounts for multiple mutations at site – definitively happens over longer timeframes

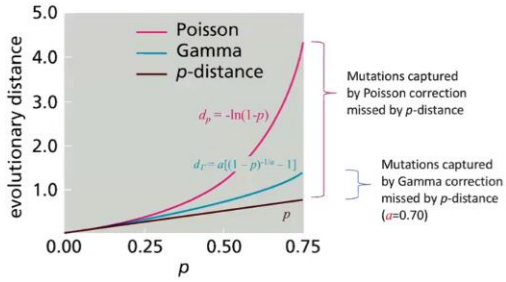
Gamma distance correction
accounts for site-specific rates – believed to occur when selective pressure varies across sites

B. Three phylogenetic trees (X, Y and Z) below were reconstructed based on the five sequences shown on the right.



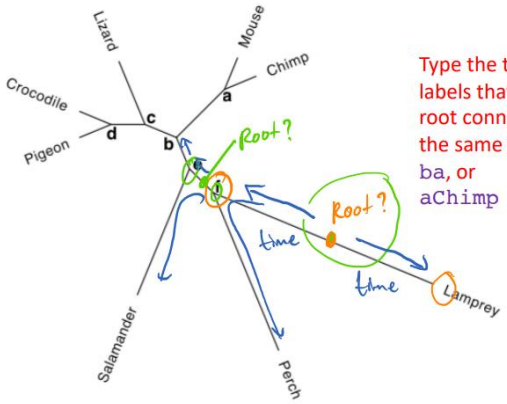
Fish	ATAC
Snake	TTTA
Frog	TTTC
Bird	TTAC
Cat	TTTA

Based on this information, which of these trees is the most parsimonious? Show your calculations and steps you took to identify the most parsimonious tree. (3 marks)



P8: Exam question from 2019

The unrooted phylogenetic tree below was inferred from eight orthologous sequences representing different species. Each internal branch point is labelled and the leaves are labelled with the species names.

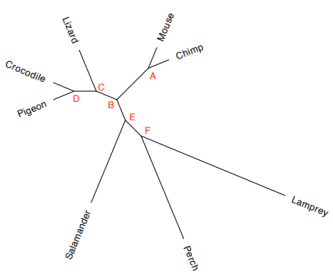


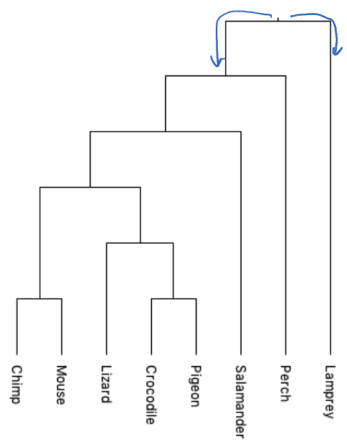
Type the two node labels that the root connects on the same row, e.g. ba, or aChimp

Based on this tree, answer the following questions.

A. To root the tree with lamprey as an outgroup, on which branch should the root be placed? A branch could be from a to b, or from a to Chimp, etc. (2 marks)

B. Draw the rooted tree in A. (2 marks)





To use molecular data to reconstruct evolutionary history requires making a number of reasonable assumptions. Which of the following statements are INCORRECT?

The molecular sequences used in phylogenetic construction are homologous.
The molecular sequences used in phylogenetic construction share a common origin.
Parent branch splits into two or more daughter branches at any given point.

Genome Analysis

Example of past year exam question

Obs = TGCT

S1 = EEII
S2 = IEEI

(P|S1) = (0.2*0.5)*(0.3*0.5)*(0.5)*(0.2*0.6)*(0.3*0.6) = 0.000162

(P|S2) = (0.3*0.6) *(0.4)*(0.3*0.5)*(0.3*0.5)*(0.5)*(0.3*0.6) = 0.000145

Name two computational approaches for gene finding.

Identity search, Similarity search- Homology based, Ab initio approaches

Name three specific sequence features of eukaryotic genes that can be incorporated into a Hidden Markov model for predicting protein-coding genes. Exons and introns, Regulatory sequences (promoter and enhancer), Splice donor and accepter sides, 5’ Cap and 3’ PolyA tail

You are designing a Hidden Markov model for the identification of protein-coding genes in eukaryotic genomes. Which specific sequence characteristics of protein-coding genes could be included in the model and how would you train the model?

Biological databases

Name two tasks that you can perform using a database of biological sequence

To identify homologous sequence; primer design; find molecular function.

You have been given the amino acid sequence of protein known as BraC. You have been asked to find as much information as possible about BraC.

Describe what types of biological databases that are available for your research and the types of information they will provide.

NCBI -> Run BLAST -> shared similarity to other known protein sequence in the database (homologous sequence) -> based on the hormone sentences if they're similar, it's likely to infer the same function

UniProtKB -> Run BLAST

PDB -> based on sequence of morality, you would get protein structure information

Annotation database -> protein function, PTM, localization

Literature (PubMed) -> what studies have been done on this protein

Ontology

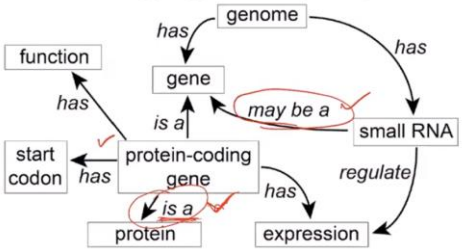
Briefly explain "ontology" and name TWO reasons why ontology is important for organising biological data. Standardised vocabulary that defines concepts/terms and their relations or constraints in a domain (knowledge / field / discipline).

- Share common understanding
- Reuse and recycle knowledge
- Make assumptions explicit (no ambiguity)

Ontology

The figure below depicts a diagram of ontology relationships. Classes are denoted by boxes and the relationships between them by arrows. A triple is an ontology instance that consists of the subject, predicate and object.

In this figure, identify TWO (2) instances (triples) that are inappropriate or incorrect. For each instance, briefly justify your answer. (2 marks)



FastA

You are given a nucleotide sequence *X*. The following hash table in the FastA algorithm describes the *Ci* index for each 3-tuple, and the position at which the 3-tuple is first found in sequence *X*. A position 0 indicates that the 3-tuple is not found.

The first 17 rows are shown below; all other 3-tuples are not found in *X*. The 3-tuple for each index is shown below as a guidance.

Based on this information, write out sequence *X*.

AAACAATA

AACAATA

AAAAACACACAAAATATA

Seq X: AAACAATA

Hash table — key → value

3-tuple	Index (C_i)	First-found position in *X*
AAA →	0	1
AAC →	1	2
AAG →	2	0
AAT →	3	5
ACA →	4	3
ACC →	5	0
ACG →	6	0
ACT →	7	0
AGA →	8	0
AGC →	9	0
AGG →	10	0
AGT →	11	0
ATA →	12	6
ATC →	13	0
ATG →	14	0
ATT →	15	0
CAA →	16	4

A BLAST search result with an E-value of 0.001 must be significant. Regardless of which database we use in the search. Do you agree with the statement above? Please justify.

Disagree. Database size matters -> it affect E-value

BLAST

Two sequences shown below (Seq1 and Seq2) are of equal length at 190 bp. If searched using BLASTN against the NCBI NR nucleotide database, which of the two sequences is more likely to have significant hits? Briefly explain your answer. (2 marks)

Seq1: low-complexity Seq

```
AAAAATTTTTAAAAATTTTTAAAAATTTTTAAAAATTTTTAAAAATTTTTAAAAATTTTTAAAA
ATTTTTAAAAATTTTTAAAAATTTTTAAAAATTTTTAAAAATTTTTTTTTAAATTTTTTATTTT
ATTTATATATTTTTTTTTTTTTTAAATTATATATTATTTTTTTTTTTTTTTTTTTTTATTA
```

Seq2:

```
TACGAAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTTGGTAAG
ATGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCCATAACTGCCTGACTAGAGTACGG
TAGAGGGTGGTGGAATTTCCTGTGTAGCGGTGAAATGCGTAGATATAGGAAGGAACACCAGT
```

The key thing here is brings us back to the low complexity masking with blast, obviously you can see sequence one only has A and T, so this is a low complexity sequence. And even if you find any hits it would not be biologically meaningful can really know for sure how meaningful that particular match that would be. Whereas, in sequence two, it seems like a normal DNA sequence to me, and that if we find a hit in a database, it's more biological meaningful.

BLAST has a higher specificity that the FastA algorithm when searching for homologous sequences in a large database, because the masking of low-complexity regions in BLAST reduces potential false positives. Do you agree? Justify.

Masking of low complexity, the region is to reduce false positive. Sequences just being identical by chance, we want to avoid that, which is why we mask all the low complexity regions generally and loss.

Name two examples of biological data types. Sequences, graphs

Name two characteristics/factors that would affect how a user use a database. Type of data, Availability

Two key challenges in the maintenance of biological databases are the shareable **and** interoperable **of these data among diverse laboratories and computer systems or platforms.**

Name two main functions of biological databases. Make biological data available to scientists, Make biological data available in computer-readable form

Name two main applications of biological databases in bioinformatics. Identification of biological entities, Inference of function

What is an ontology? What is a main characteristic of an ontology? Ontology defines (specifies) the concepts, relationships, and other distinctions that are relevant for modelling a domain. It takes the form of the definitions of representational vocabulary (classes, relations etc.), which provide meanings for the vocabulary and formal constraints on its coherent use.

Why do we need ontology? Give two reasons. It used to share common understanding of the structure of information among people or software agents, to enable reuse of domain knowledge, and to make domain assumptions explicit.

What is basis for constructing an ontology triple? Subject, predict, object

Name two examples of ontology databases. Reactome, KEGG

Protein bioinformatics

What is protein secondary structure and how can we predict it? Why is it useful?

Local structural classes (3- or 8-class; alpha, beta, coil). By statistics collected over "window". Chou-Fasman. Starting point for prediction of tertiary and quaternary structure. Insights into biological function of protein. Facilitate alignment for homology modelling of distantly related proteins.

How can we find protein-protein interactions, by experiment and by bioinformatics? Binding sites?

By homology, Conservation (maybe spread in sequence but come together in structure); Many same-charged residues (electro-static interaction); Hydrophobic patch (unusual at surface; interaction by hydrophobic forces).

What is a biological network? Exemplify what nodes and edges may mean.

Flexible, joint representation of multiple components (nodes) in a system, with topology reflecting their relationships (edges); can be interaction, pathway, or causal/regulatory relationships.

Final exam 2018: Protein bioinformatics

The Chou-Fasman propensity (*P*) values for each amino acid to form an α-helix or a β-strand, respectively is shown in the table (right). A larger value denotes higher propensity.

Based on your understanding of the formation of protein secondary structure and the propensity table, predict the secondary structure class (**α-helix**, **β-strand** or **coil**) at the highlighted position of the following amino acid sequences. You do not need to use Chou-Fasman's algorithm, but similar principles should apply in determining your answers.

(a) His – Lys – Glu – Ile – Cys – Leu – Pro – Ile – Val – Phe – Lys – Asp

...

(d) Justify the predictions for (a)-(c) by explaining what the table is based on and the strategy with which predictions were made. Calculations are not required.

Amino acid	α-helix		β-strand	
	Designation	*P*	Designation	*P*
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74
Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	B	0.57	B	0.55
Gln	f	1.11	f	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	B	0.69	F	1.4

(a)

His – Lys – Glu – Ile – Cys – Leu – Pro – Ile – Val – Phe – Lys – Asp

1.00 1.16 1.51 1.08 0.70 1.21 0.57 1.08 1.06 1.13 1.16 1.01

0.87 0.74 0.37 1.60 1.19 1.30 0.55 1.60 1.70 1.38 0.74 0.54

(b)

Arg – Pro – Met – Ala – Lys – Thr – Gln – Ala – Phe – Cys – Gly

0.98 0.57 1.45 1.42 1.16 0.83 1.11 1.42 1.13 0.70 0.61

0.93 0.55 1.05 0.83 0.74 1.19 1.10 0.83 1.38 1.19 0.75

(c)

Pro – Gly – Cys – His – Pro – Ser – Tyr – Ala

0.57 0.61 0.70 1.00 0.57 0.77 0.69 1.42

0.55 0.75 1.19 0.87 0.55 0.75 1.40 0.83

The correct answers were (a) *beta*, (b) *alpha*, (c) *coil*

What *sensitivity* to the class *alpha helix* do I get by predicting *alpha helix* for all three questions?

(a) observing not α predicting α FP
(b) observing α predicting α TP
(c) observing not α predicting α FP

- Sensitivity $\frac{tp}{tp+fn}$
- Specificity $\frac{tn}{tn+fp}$

Two decimal places multiplied by 100

e.g. 100 50 33 0

correct 100 50 33 0

What *specificity* to the class *alpha helix* do I get by predicting *alpha helix* for all three questions?

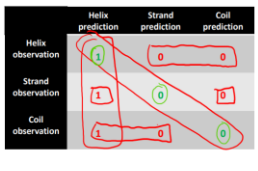
(a) observing not α predicting α FP
(b) observing α predicting α TP
(c) observing not α predicting α FP

- Sensitivity $\frac{tp}{tp+fn}$
- Specificity $\frac{tn}{tn+fp}$ $\frac{0}{0+2}$

multiply by 100

e.g. 100 50 33 0

What *accuracy* (Q_3) do I get by predicting *alpha helix* for all three questions?



$$Q_k = \frac{\sum_{j=1}^{k} tp(j)}{\sum_{j=1}^{k} tp(j) + fn(j)} \cdot 100$$

0.33 $\frac{1}{3}$

TRUE:

Pathway databases are never static, and entities are regularly updated.

Molecular clock is an assumption where the rate of evolutionary change of any specified protein is approximately constant over time and over different lineages.

Gene Ontology describes the attributes of genes and gene products. It does not represent protein structures, gene regulatory networks and biological pathways

Ontology in biology is a systematic, unambiguous description of specific biological attributes.

Sequence ontology describes features and attributes of biological sequences, e.g. binding sites, and exons.

BioPax describes attributes of biological pathways, not packaging of biological materials

FastA adopts the hashing-and-chaining algorithm to identify k-mers for seeding an alignment.

FastA algorithm uses only exact matches

Significance of the expect (E) value in BLAST is dependent on the size of the database

The BLAST algorithm searches for high-scoring segment pairs that are statistically significant

Low-complexity regions are usually masked in BLAST searches

Short DNA or amino acid sequences can carry functionally significant meaning– E.g. nuclear localisation signals and transcription factor binding sites

They are challenging to discover due to: – Random locations in genome/protein sequence – Degeneracy

Motifs can be represented both: – Discretely (e.g. consensus sequence or regular expression) – Continuously (e.g. position probability matrix)

Aligning large sequences fails to pick up short sequence motifs with functional and structural features, e.g. localization signals, binding sites

Motif representations can be derived from motif members (sharing features)

– Manually, by alignment, informed exploration and/or by precise experimental methods

– Via databases storing profiles/motifs as regular expressions, PWMs and profile HMMs

– Using discovery methods like Gibbs and MEME (do not require prior alignment)

A “logo” visualises a sequence pattern

PWM scoring can be used to find motifs