Question 1. Sequence alignment **(Total 6 marks)**

A. Global alignment is often used to determine homology between genes. For what type of problem is local alignment more appropriate? (1 mark)

Local alignment is appropriate when evolutionarily very distant genes where functional domains are of interest, rather than overall homology (e.g. zinc-finger or homeo domains for DNA binding). Or, for piece-by-piece whole-genome alignment, where regions have been subject to major translocations, inversions and duplications.

B. What do you need to know and how would you use sequence alignment methods to determine whether two genes are “orthologs”, or not? Similarly, whether they are “paralogs” or not? (2 marks)

Homology is an evolutionary relationship that either exists or does not—two genes are homologous if they derive from the same ancestral gene. An ortholog is a homolog associated with a speciation event. A paralog is a homolog that arose through a duplication event. So: you need to know if the sequences are sourced from different species.

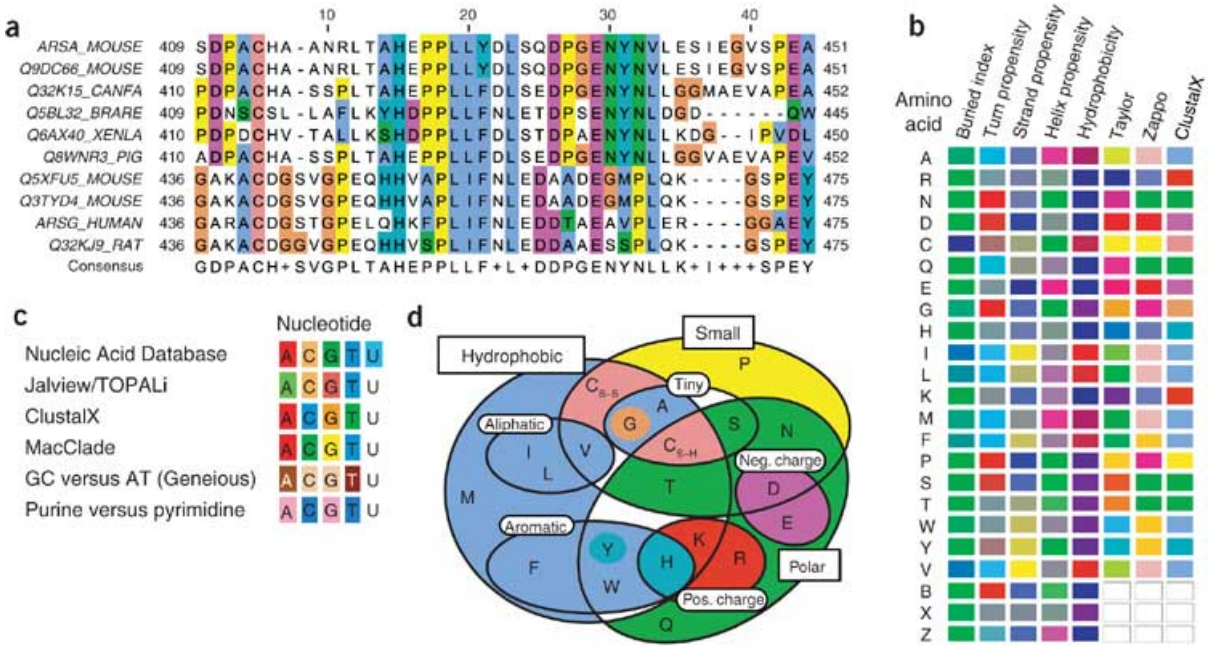
Similarity is a measure of the quality of alignment between two sequences. High similarity is supporting evidence for homology. So: we need to perform alignment between the sequences to ascertain level of similarity.

Similar sequences may be orthologs or paralogs. We cannot separate between orthologs and paralogs by sequence alignment alone

C. Below a multiple alignment is given. Explain why a single “multiple” alignment can be more informative than multiple “pairwise” alignments. (2 marks)

With multiple sequences, diversity is captured and represented, with highly conserved regions supported by greater number of sequences, resulting in more significant scores when aligned.

D. Use the example below to describe the layout and colours, and provide a brief interpretation of the sequences in the alignment. (2 marks)



Colours mostly represent physicochemical properties, which are interesting to visualize if conserved. Conservation is seen by the agreement of letters (representing amino acids for the proteins above) and the agreement of colours. As an example, hydrophobic residues are (in many schemes) coloured blue/green/cyan, to be detected around column 20 where many residues are strongly conserved in the above alignment. The conserved alignment may indicate shared function or structural properties, essential to the proliferation of the sequence product.

Question 1 (alternate). Sequence alignment **(Total 6 marks)**

The optimal alignment between the strings

X = ACTGA and

Y = TACGA

can be determined by using the following scoring system, building the alignment score matrix below.

The score for a "matched symbol" is +2, a "mismatch" is -1 and a "gap" is -2. You should use this scoring system when answering the following questions.

A. What is the optimal global alignment between *X* and *Y*? What is the optimal global alignment score? (4 marks)

	Y=	**T**	**A**	**C**	**G**	**A**
X=	0					
A						
C						
T						
G						
A						

B. What is the optimal local alignment between *X* and *Y*? What is the optimal local alignment score? (2 marks)

	Y=	**T**	**A**	**C**	**G**	**A**
X=	0					
A						
C						
T						
G						
A						

	Y=	T	A	C	G	A
X=	0	-2 (GX)	-4 (GX)	-6 (GX)	-8 (GX)	-10 (GX)
A	-2 (GY)	-1 (M)	0 (M)	-2 (GX)	-4 (GX)	-6 (GX)
C	-4 (GY)	-3 (M/GY)	-2 (M/GY)	2 (M)	0 (GX)	-2 (GX)
T	-6 (GY)	-2 (M)	-4 (*)	0 (GY)	1 (M)	-1 (M/GX)
G	-8 (GY)	-4 (GY)	-3 (M)	-2 (GY)	2 (M)	0 (M/GX)
A	-10 (GY)	-6 (GY)	-2 (M)	-4 (*)	0 (GY)	4 (M)

M = Match [diagonal]
GY = Gap in X [up]
GX = Gap in Y [left]

X= –ACTGA
Y= TAC–GA
Score= 4

	Y=	T	A	C	G	A
X=	0	0	0	0	0	0
A	0	0	2 (M)	0	0	0
C	0	0	0	4 (M)	2 (GX)	0
T	0	2 (M)	0	2 (GY)	3 (M)	1 (M/GX)
G	0	0	1 (M)	0	4 (M)	2 (M/GX)
A	0	0	2 (M)	0	2 (GY)	6 (M)

X= ACTGA
Y= AC–GA
Score= 6

Question 2. Sequence database searching **(Total 3 marks)**

A heuristic is an approximation or 'rule of thumb' pertaining to a particular problem; heuristic algorithms are algorithms that admit sub-optimal (i.e., approximate) solutions to a problem, typically for performance reasons.

BLAST and FastA are database search algorithms, both of which use heuristics. Give a high-level outline of the common algorithmic approach used by both BLAST and FastA, identifying the main aspect of these algorithms that makes them heuristic, and why this heuristic may result in sub-optimal solutions.

BLAST uses finite state machine/finite state automata to identify matches of k-mers between a query sequence and a subject sequence. The k-mer matches are then used to seed a pairwise alignment, extending from both ends, resulting in high-scoring pairs (HSPs). BLAST allows for non-exact matches.

FastA uses hashing-and-chaining to document the positions at which the k-tuples are found on the query sequence. Perfect matches between two sequences can then be identified, followed by local alignment. FastA only handles exact matches.

Question 2 (alternate). Sequence database searching **(Total 4 marks)**

BLAST and FastA are both database search algorithms that use fast, sub-optimal alignment of sequences to speed up searches. Both algorithms start by scanning a query sequence in order to find sub-sequence matches.

A. What common assumption about high-scoring alignments underlies the BLAST and FastA scanning approaches? (2 marks)

Sequences that share a high extent of identity are assumed to be homologous.

B. Outline the major difference between the BLAST and FastA algorithms for *k*-mer/*k*-tuple matching. (2 marks)

BLAST uses finite state machine/finite state automata / allows for non-exact matches.

FastA uses hashing-and-chaining / only handles exact matches.

Question 3. Sequence motifs **(Total 4 marks)**

Below is an alignment of eight nucleotide sequences.

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

A. Explain how you construct a log-odds position weight matrix from this data, and a set of sequences from the same region of DNA as background.
(2 marks)

```
Background A: 22/48 = 0.458; C: 2/48 = 0.04; G: 12/48 = 0.25; T: 12/48
= 0.29
Col 1:
A: 4/8 / (22/48) = 4*48 / 8*22 = 1.1                    log_2(1) ~= 0
C: 2/8 / ( 2/48) = 2*48 / 8*2 = 6                       log_2(6) ~= 2.5
G: 0/8 / (12/48) = 0                                    log_2(0) = -Inf
T: 2/8 / (12/48) = (2*48)/(8*12) ~=0.9                  log_2(0.9) ~=0
Col 4:
A: 8/8 / (22/48) = (8*48) / (8*22) = 2.2                log_2(2.2) ~= 1
```

B. Present rough numeric estimates if you were to assume a uniform nucleotide background. Tip: $\log_2(1)=0$ $\log_2(2)=1$ $\log_2(4)=2$ $\log_2(1/2)=-1$ $\log_2(1/4)=-2$
(2 marks)

Question 3 (alternate). Sequence motifs **(Total 6 marks)**

The following matrix contains the counts of particular nucleotides at binding positions 1 to 6 for the *Arnt* transcription factor.

```
A  [   4  19   0   0   0   0 ]
C  [  16   0  20   0   0   0 ]
G  [   0   1   0  20   0  20 ]
T  [   0   0   0   0  20   0 ]
```

By normalising each column to add to 1, the matrix becomes a position probability matrix (PPM). The PPM can be used to find binding sites of *Arnt*, in so-called promoter regions of genes.

A. Specify the consensus sequence (with the most dominant nucleotides at each position). (1 mark)

CACGTG

B. Describe the basic steps by which new binding sites can be found using a PPM. (2 marks)

The strategy is to evaluate the probability of the data given a PPM *q*. Putative binding sites are those which exceed a threshold value. The position is identified by scanning over the sequence, scoring each position *i*, by reference to PPM positions *u*. S(i)=q_1,R(i) * q_2,R(i+1) * q_3,R(i+2) * ... * q_m,R(i+m-1) where *R(i)* is the residue index at sequence position *i*

C. Determine the highest scoring position in the following sequence. (1 mark)

i. C C C A C G T A A C G T G G

Position 8: AACGTG

D. So-called pseudo counts are commonly used to compensate for the lack of experimental data. As an example, we can add a single count to each cell. Does this make a difference to the outcome of the question above (b)? Explain your answer and why this modification may alleviate issues of data sparseness. (2 marks)

```
A [ 5  20  1   1   1   1 ]          A [.21 .83 .04 .04 .04 .04]
C [17  12  1   1   1   1 ]          C [.71 .04 .88 .04 .04 .04]
G [ 1   2  1  21   1  21 ]          G [.04 .08 .04 .88 .04 .88]
T [ 1   1  1   1  21   1 ]          T [.04 .04 .04 .04 .88 .04]
```

Position 3 **C A C G T A** .71 x .83 x .88 x .88 x .88 x .04 = 0.016

Position 8 **A A C G T G** .21 x .83 x .88 x .88 x .88 x .88 = 0.105

No substantial difference in the example. The pseudo counts ensures that no combination is "impossible" (allowing sequences with near perfect matches but one orthogonal position to

receive a positive score).

Question 4. Databases and ontology **(Total 3 marks)**

Examine the following pieces of the Gene Ontology, and answer the following questions:

A. Identify the part of the ontology term definition that specifies where the term is located in the GO conceptual hierarchy

Cellular component

B. From which of the three GO hierarchies does this term come?

```
id: GO:0005739
name: mitochondrion
namespace: cellular_component (B)
def: "A semiautonomous, self replicating organelle that occurs in varying
numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic
cells. It is notably the site of tissue respiration." [ISBN:0198506732]
subset: goslim_candida
subset: goslim_generic
subset: goslim_pir
subset: goslim_plant
subset: goslim_yeast
synonym: "mitochondria" EXACT []
xref: NIF_Subcellular:sao1860313010
xref: Wikipedia:Mitochondrion
is_a: GO:0043231 ! intracellular membrane-bounded organelle
is_a: GO:0044444 ! cytoplasmic part
(A)
```

Question 4 (alternate). Databases and ontology **(Total 3 marks)**

A. What is UniProt?

Public resource of protein sequence and functional information

B. What is the main database for Biological Macromolecular Structures?

Protein Data Bank (PDB)

C. What are the three ontologies in the Gene Ontology?

Biological Process, Molecular Function, Cellular Component

Question 5. Gene and protein expression **(Total 3 marks)**

A. What kind of molecules do expression microarrays measure? (1 mark)
RNA

B. Expression microarray data is known to be "noisy". Describe two ways of reducing this problem: one based on the experimental design and one based on post-processing the probe intensities. (2 marks)

(i) Two-colour procedure where two samples are labelled with different fluorophores and hybridized together on a single array—making direct comparisons possible.

(ii) By normalization, standardizing each value relative to the average intensity and deviation of each probe on the array. Log-transforming to make distributions "normal".

Question 5 (alternate). Gene and protein expression **(Total 4 marks)**

A. You have successfully completed your two-channel microarray experiment and have obtained Red and Green intensity values for each feature/probe on the array.
   i. Describe how you would initially transform the data so that increases and decreases in gene expression are treated equally. (1 mark)
   ii. What method would you use to normalise the intensities of the two dyes, as well as to eliminate other systematic differences due to unequal experimental conditions. (1 mark)

B. You wish to normalise data values between arrays. Provide details of a method for normalising data between arrays. (2 marks)

Question 6. Protein structure **(Total 6 marks)**

A. What is the primary, secondary, tertiary and quaternary structure of a protein? (2 marks)

Sequence, local conformation, 3D structure, inter-protein interaction complex

B. List, and in a sentence describe, the intrinsic and environmental properties that determine the secondary structure of a protein. (2 marks)

Temperature, pH, ligands, environment -> aqueous or membrane?

C. Briefly describe a method or principle that can be used to predict secondary structure from an amino acid sequence. (2 marks)

Chou-Fasman (Determines the probability of finding a given amino acid combination in a particular secondary structure. Uses scores over a window of amino acids. Rules, e.g. 4/6 amino acids need to have a propensity score exceeding 100, are used. Propensity scores are derived from large sets of proteins to form rules for certain secondary structures.)

Question 7. Genome analysis (Total 4 marks)

What are the unique bioinformatics challenges to finding ORFs in eukaryotic genes versus prokaryotic genes? Explain your answer.

Question 8. Phylogeny (Total 3 marks)

Phylogenetic trees are calculated by applying mathematical models to infer evolutionary relationships between organisms, based on a set of characters that describe their differences. The most common characters are nucleotide or protein MSAs, but morphological information has also been used. There are four main categories of phylogenetic reconstruction methods:

(1) distance matrix based methods, e.g. neighbor joining
(2) probabilistic methods including maximum likelihood
(3) maximum parsimony methods

Pair these categories with X, Y and Z in the statements below, like so:

1. → Y [write X, Y or Z]
2. → Z
3. → X

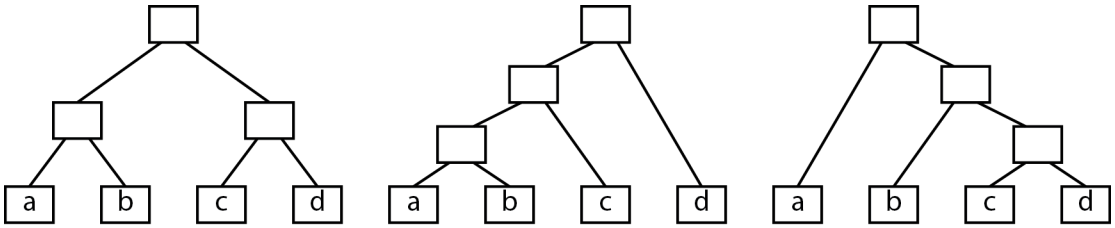
X is the principle of choosing simpler hypotheses in preference to those requiring a more complex explanation. X approaches create trees using the minimum number of ancestors needed to explain the observed characters.

Y estimates the mean evolutionary time (measured as the mean number of changes per site) since two species diverged from their most recent common ancestor. However, because they reduce the estimate of most recent common ancestor to a single value, information on character evolution is lost.

Z searches a set of tree and evolutionary models to find the ones most likely to generate the observed characters. Some variants of Z offer more flexibility, as they allow optimization of all aspects of a tree (model, topology, branch length). But this comes at a cost: they require computationally expensive techniques.

Question 8 (alternate). Phylogeny (3 marks)

Consider the following phylogenetic trees.



Tree 1 Tree 2 Tree 3

True or false:

(A) All trees are ultrametric True
(B) Only one tree is bifurcating False
(C) All trees are un-rooted False
(D) Tree 2 can be a gene tree for a gene X, tree 3 can be a gene tree for gene Y, when both X and Y are from the same species True
(E) Maximum parsimony can identify the ancestral sequences from those at the leaves True
(F) In determining the score for a tree, maximum parsimony can NOT account for multiple mutations at the same site True

END OF EXAMINATION