Class: Biol/CS123A **Name:** Zayd Hammoudeh

Assignment: #5

Problem: #1 / |

Problem: #2 / |

Problem: #3 / |

Grade: Total / |

Problem #1

Note: The page numbers below are for my studying later on and can be ignored during grading.

Chapter #7:

1.1) When analyzing sequences with evolution in mind, is it the differences between them that we need to quantify?

When analyzing sequences with evolution in mind, it is the differences between them – often summarized as the evolutionary distance or genetic distance – that are of interest and which need to be quantified and scored. (See page 224)

1.2) What is the purpose of the phylogenetic tree representation?

The purpose of the **phylogenetic tree** representation is to summarize the key aspects of a reconstructed evolutionary history. It is a "shorthand method" of representing the history of mutation in an easily understandable way. (See page 225)

1.3) What are species trees, and how are they constructed?

A phylogenetic tree is a diagram that proposes a hypothesis for the reconstructed evolutionary relationships between a set of objects (referred to as **taxa** or **operational taxonomic units** (OTUs)). A species tree is a phylogenetic tree that contains homologous/orthologous sequences from different species; the aim of the tree is to show the relationship between different species. This is in contrast to those trees that attempt to show the relationships between genes and proteins in a large gene family.

Species trees can be constructed from data from other sequences such as morphological features used in traditional taxonomy, presence of certain restriction sites in DNA, or the order of a particular set of genes in the genome. In such cases, the differences may be more qualitative than quantitative. (See page 225)

In a species tree, the different taxa are connected via a set of lines, called **branches** or **edges**. These lines represent the evolutionary relationships between the different species. **External nodes** or **leaves** in the tree are existing species that have not yet evolved into new species or are extinct species that died out without leaving any descendents. External nodes are linked together through **internal nodes**. (See page 225-226)

1.4) Is the evolutionary history of a set of related genes always the same as that of the species from which the genes were selected?

No. The evolutionary history of a set of related genes is often not the same as that of the species from which the genes were selected. (See page 225)

1.5) What is a speciation event? How is it represented in a species tree?

A **speciation event** refers to the point at which the population of an existing species divides into two separate groups that subsequently diverge into separate species. (See definition on page 747). This is represented by an internal node branching to either external nodes or additional internal nodes. The original internal node represents the hypothesized **ancestral state**. (See page 226)

1.6) What does the root represent in a rooted tree?

The **root** in a rooted phylogenetic tree represents the last common ancestor. (See page 227)

1.7) What is the major task of phylogenetic tree reconstruction?

The major task of phylogenetic tree reconstruction is to identify, from the numerous alternatives, the topology that best describes the evolution of the data. (See page 228)

1.8) What are the differences between the following types of phylogenetic trees: cladograms, additive trees, and ultrametric trees?

Cladogram – A rooted tree in which the branch lengths between nodes has no meanings. In this case, only the tree topology is defined. What is more, the ancestors are normally not shown in the tree. Since cladograms give no quantitative information, they are generally not used in sequence analysis.

Additive Tree – Constructed from the same data as a cladogram. However, branch lengths represent a quantitative measure of evolution including being proportional to the number of mutations that have occurred. The units for variation are arbitrary and are instead intended to be proportional to the number of mutations. The evolutionary distance between any two taxa is the sum of the length of the branches connecting them. These trees can be both rooted and unrooted.

Ultrametric Tree – In addition to the properties of an additive tree, all branches in an ultrametric tree has a constant rate of mutation. This is referred to as a **molecular clock** because one can in principle measure the actual times of evolution events from such trees. Ultrametric are always rooted. Present day is at the bottom of the tree and the last common ancestor is at the top of the tree. (See page 228-230)

1.9) What is meant by bootstrap analysis, and how can this analysis be used to construct condensed trees?

Bootstrap analysis is a method of estimating the support present in the sequence data for the topological features of a phylogenetic tree in which many randomized selections of the data are examined to determine their support for each split. (See definition on page 735) When bootstrapping is used, all the observable splits of the tree produced from the original data are listed, and each bootstrap tree produced from sampled data is examined to see if contains the same splits. The percentage of the bootstrap trees that contain each split is either reported in the splits list or displayed on the tree as a number. If a split is not high supported by the bootstraps (e.g. occur in less than 60% of the bootstrap trees), it is sometimes removed. Such trees with removed low percentage splits are called **condensed trees**. (See page 22)

1.10) What are the two conditions that would have made phylogenetic tree reconstruction from a set of homologous sequences considerably easier had they held during sequence evolution?

First, all the sequences evolved at a constant mutation rate for all mutations at all times. Had this condition held, then the number observed differences between any two sequences would have been directly proportional to the time that has elapsed since divergence.

Second, the sequences only diverged a moderate degree such that no position has been subjected to more than one mutation. If his condition held, then once the sequences had been accurately aligned, then all mutational events could have been observed as non-identical bases, and the mutations could have been assumed to be from one base to another. (See page 236)

1.11) Where do most mutations that are retained in DNA come from?

Most mutations that are retained in DNA occurred during DNA replication as a result of uncorrected errors in the replication process. (See page 237)

1.12) What is the difference between synonymous and non-synonymous mutations?

Nucleotide mutations that do not change the encoded amino acid sequence are called **synonymous mutations**. They are generally considered to be neutral since they have no effect. When a nucleotide change alters the encoded amino acid, this change is said to be **non-synonymous**. (See pages 238-239).

1.13) When is it useful to remove the third codon sites from the data before any further analysis?

Nucleotide substitutions at the third codon position are usually synonymous. As such, the accepted mutation rate at those sites is much higher than in the first and second codon positions. This phenomenon is known as **biased mutation pressure**. Hence, if a data

set involves a long evolutionary time scale, then many of the third codon positions may have experienced multiple mutations and show almost random base content. In such cases, it is often useful to remove the third codon sites from the data before performing further analysis. (See page 239)

1.14) What is the key assumption that is made when constructing a phylogenetic tree from a set of sequences?

The key assumption made when constructing a phylogenetic tree from a set of sequences is that all the sequences are derived from a single ancestral sequence (i.e. that they are homologous). (See page 239)

1.15) Explain the process of gene loss. Does gene loss occur solely because of gene duplication?

Initially after gene duplication, there will only be a requirement for one of the genes. Often, instead of the evolving into an alternative function, one of the genes becomes nonfunctional through mutation. This can happen due to the loss of a control region or a modification to the protein sequence renders the protein inactive. Genes that have mutated so as to no longer give rise to protein products are called **pseudogenes**. Usually, pseudogenes continue to accumulate mutations until the pseudogenes are no longer detectable. This process is known as **gene loss**.

Gene loss can occur without gene duplication. (See page 242)

1.16) What is meant by homoplasy?

Sequence similarities that are not due to homology are known as **homoplasy**. Convergent evolution is just one cause of homoplasy. Other causes are parallel evolution and evolutionary reversal. (See page 244)

1.17) What is meant by "horizontal gene transfer" (also known as "lateral gene transfer")? What is it called "horizontal"?

Horizontal gene transfer (HGT), also known as **lateral gene transfer** (LGT), can confuse phylogenetic analysis. It involves the genes from one species being transferred into another species. The term "horizontal" is used to contrast this type of gene transmission with the more typical transmission of genes vertically from parent to offspring. HGT is more prevalent in bacteria and archaea. However, it can happen in eukaryotes through viruses. (See page 246-247)

1.18) What are syntenic regions, and are they easily detected?

A **syntenic region** is a region of a genome that contains a series of genes in a similar order to that found in a region of the genome of a different species. This implies common evolutionary ancestry. (See definition on page 748) However, many changes have been found in the genome at the larger scale than that of individual genes. What is more, chromosomes may have split into smaller fragments that rejoined to make new chromosomes with some sections shuffled or inverted. These types of changes are actually relatively common. The consequent of this is that orthologous genes usually do not occupy equivalent positions even in related species. (See page 248)

1.19) When comparing sequences from two closely related species, which regions will convey useful information for the construction of phylogenetic trees?

When comparing sequences from two closely related species, there will be no information in the highly conserved regions (they will be identical), but the more rapidly changing regions will have useful data. In contrast, when comparing data from two distantly related species, the rapidly changing regions will show nearly uniform dissimilarity; the more conserved regions will have the more useful variation. (See page 249)

1.20) The analysis of which genomic sequence led to the discovery that prokaryotes comprised two quite distinct domains?

The DNA sequence specifying the small ribosomal subunit rRNA (specifically called 16S RNA in prokaryotes) (See page 249)

Problem #2

Part A: Describe the effects of the following two mutations in the BRCA1 gene:

1) **Exon 13: Base 4446 C → T**

By doing a three frame translation on the part exon #13 that is around the mutation and comparing it to the translation of the BRCA1 gene (accession #U14680), you can find the correct codon the mutation. When I did that (from bases 4441 to base 4476), I saw the codon was previously CGA. When the C is mutated to a T, the codon becomes UGA, which is a stop codon. Hence, the gene is truncated prematurely.

2) **Exon 20, 5382 ins C**

This change causes a frame shift mutation. This frame shift mutation causes the last set of codons after base #5382 to be shifted by 1 base. What is more, it causes as slightly premature stop codon as shown below.

Stamdard Amino Acid Sequence After the Insertion:
SQDRKIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELSSFTLGTGVHPIVVVQPDAWTEDNGFHAIGQMCEAPVVTREWVLDSVA
LYQCQELDTYLIPQIPHSHY **STOP**

Mutated Amino Acid Sequence After the Insertion:
LPGQKDLQGARNLLLWALHQHAHRSTGMDGTAVWCFCGEGAFIIHPWHRCPPNCGCAARCLDRGQWLPCNWADV **STOP**

Part B: Describe the effects of the following two mutations in the BRCA1 gene:

1) **3761-3762 del GA**

By doing a three frame translation on this sequence, you can determine the codon location of the mutation. When GA is deleted, it causes a frame shift mutation. From using Expasy and aligning the resulting sequence to the gene translation in Genbank, it is observed that bases 3770 to 3772 (UAG) become a premature stop codon truncating the amino acid sequence.

2) **2616-2617 ins AAGTATCCAT**

By doing a three frame translation on this sequence, you can determine the codon location of this mutation; this analysis shows that base #2616 is the first base in a codon; hence the first two bases in this insertion are the last two bases in its codon. When "AAGTATCCAT" is inserted, the preceding "T" in base location 2616 causes a stop codon when it is combined with the "AA" in the insertion. This leads to premature truncation of the protein.

Problem #3

1. **What is ENIGMA?**

ENIGMA is the Evidence-based Network for the Interpretation of Germline Mutant Alleles consortium.

2. **What is their goal?**

ENIGMA was initiated to evaluate and implement strategies to characterize the clinical significance of the BRCA1 and BRCA2 variants. An example of the work they did in the Splicing Working Group was to report splicing and multifactorial likelihood analysis of 25 BRCA1 and BRCA2 variants from different laboratories.

3. **What does this demonstrate according to the end of the abstract?**

The study demonstrates the added value of collaboration between laboratories, and across disciplines, to collate and interpret information from clinical testing laboratories to consolidate patient management.

4. **What is BIC?**

BIC is the Breast Cancer Information Core database. It contains almost 1,800 distinct gene variants that are reported as having unknown clinical significance, which is more than half of all variants reported in BIC.

5. **Are bioinformatics tools, such as algorithms implemented in web-based programs, that predict splicing effects of nucleotide variants sufficient for predicting splicing? Why?**

While these tools may become standalone diagnostic tools in the future, it is still important at this time to perform splice assays in parallel with the bioinformatic predictions. For example, the tools currently need improved sensitivity and specificity when it comes to predicting the likelihood of disrupted splice sites. This is particularly true for variants that are beyond the highly conserved AG and GU nucleotide pairs. What is more, today's algorithms show poor performance for the prediction of alternatively used mechanisms such as cryptic splice sites.

6. **What is the Splicing Working Group?**

The Splicing Working Group is a part of the ENIGMA Consortium. Specific projects it works on include: identifying optimal standardized protocols and prediction tools for characterizing splicing aberrations as well as assessing the consistency of interpretation of clinical significance of splicing assay results. They also compared splicing aberrations from 25 different BRCA1 and BRCA2 variants to the output of different splicing prediction tools.

7. **Name the four bioinformatics prediction algorithms used in this work?**

Four prediction algorithms they used include: SSF, MaxEntScan, NNSPlice, and GeneSplicer.

8. **What is HGVS?**

HGVS is the Human Genome Variation Society.

9. **What are cryptic sites?**

A **cryptic site** is a splice site defined by a wildtype sequence, but only used when a variant disrupts the native donor or acceptor site.

10. **Use the splice site predictor NNSplice to examine predicted splice sites for the MOG gene (accession number Z48051).**

Tool used: NNSplice predicts donor and acceptor splice sites. The **donor site** in a splice is the 5' end of the intron while the **acceptor site** is the 3' end of the intron. For the minimum donor site score, I used 0.95, and for the minimum acceptor site score, I used 0.97.

Donor Site Predictions:

Number	Start	End	Score	Comments/Validation
1	288	302	0.97	The Genbank record makes no mention of this is an actual splice site. However, to a limited extent, it does not matter since this is before the first coding exon. The Genbank record states the MRNA starts at base #764 which is after this base so this is most likely not a splice site.
2	792	806	0.97	This is not an exon. Exon #1 for MOG begins at base #764. At this high threshold level, the algorithm missed the first splice site.
3	1247	1261	0.99	This is a correct splice site and is confirmed on the Genbank record as the first intron donor site.
4	2450	2464	1.00	This is not a splice site since it is located within intron #1. It could be a cryptic splice site.
5	3615	3629	1.00	This is a correct splice site as confirmed on the Genbank record. It is for the second intron.
6	3836	3850	1.00	This is not a splice site since it is located within intron #2 which goes from base 3622 to base 10105. It could be a cryptic splice site.
7	6577	6591	0.99	This is not a splice site since it is located within intron #2 which goes from base 3622 to base 10105. It could be a cryptic splice site.
8	7732	7746	0.98	This is not a splice site since it is located within intron #2 which goes from base 3622 to base 10105. It could be a cryptic splice site.
9	9825	9839	0.95	This is not a splice site since it is located within intron #2 which goes from base 3622 to base 10105. It could be a cryptic splice site.
10	10113	10127	0.97	This is a correct splice site as confirmed on the Genbank record. It is for the third intron.
11	11611	11625	0.98	This is a correct splice site as confirmed on the Genbank record. It is for the fourth intron.
12	11874	11888	0.99	This is a correct splice site as confirmed on the Genbank record. It is for the fifth intron.
13	12165	12179	0.99	This is not a splice site since it is located within intron #5 which goes from base 11881 to base 14237. It could be a cryptic splice site.
14	14672	14686	0.99	This is a correct splice site as confirmed on the Genbank record. It is for the seventh intron. At this high score level, the algorithm missed the splice site of the sixth intron.
15	16393	16407	0.98	This could be a possible splice site, but since it occurs after the gene's stop codon, it is largely irrelevant. However based off the Genbank record of the mRNA, it is unlikely this is a true splice site.

Acceptor Site Predictions:

Number	Start	End	Score	Comments/Validation
1	681	721	0.98	This is not a true splice site for this gene. Rather the first exon begins at base #764.
2	2240	2280	0.98	This is not a true splice site. It is part of intron #1 which goes from base 1254 to 3273. At this high score, the true acceptor site for intron #1 was filtered out.
3	6780	6820	0.98	This is not a true splice site. It is part of intron #2 which goes from base 3622 to 10105. At this high score, the true acceptor site for intron #2 was filtered out.
4	8724	8764	0.99	This is not a true splice site. It is part of intron #2 which goes from base 3622 to 10105. At this high score, the true acceptor site for intron #2 was filtered out.
5	9934	9974	0.98	This is not a true splice site. It is part of intron #2 which goes from base 3622 to 10105. At this high score, the true acceptor site for

				intron #2 was filtered out.
6	11576	11616	0.99	This is a true splice site and is the acceptor site for intron #3.
7	11709	11749	0.97	This is not a true splice site. It is part of intron #4 which goes from base 11618 to base 11859.
8	11726	11766	0.98	This is not a true splice site. It is part of intron #4 which goes from base 11618 to base 11859.
9	11839	11879	0.97	This is a true splice site and is the acceptor site for intron #4.
10	12135	12175	0.98	This is not a true splice site. It is part of intron #5 which goes from base 11887 to base 14237. At this high score, the true acceptor site for intron #5 was filtered out.
11	12387	12427	0.98	This is not a true splice site. It is part of intron #5 which goes from base 11887 to base 14237. At this high score, the true acceptor site for intron #5 was filtered out.
12	14637	14677	0.99	This is a true splice site and is the acceptor site for intron #6.
13	15044	15084	0.99	This is not a true splice site. It is located within intron #7 which goes from bases 14679 to 15128.
14	15108	15148	0.97	This is a true splice site and is the acceptor site for intron #7.
15	15358	15398	0.98	This is not a true splice site. It is part of exon #8 which goes from base 15129 to base 16323.
16	15802	15842	0.98	This is not a true splice site. It is part of exon #8 which goes from base 15129 to base 16323.
17	16484	16524	0.98	This could be a possible splice site, but since it occurs after the gene's stop codon, it is largely irrelevant. However based off the Genbank record of the mRNA, it is unlikely this is a true splice site since there is not an exon reported to start at this location.