**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).

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

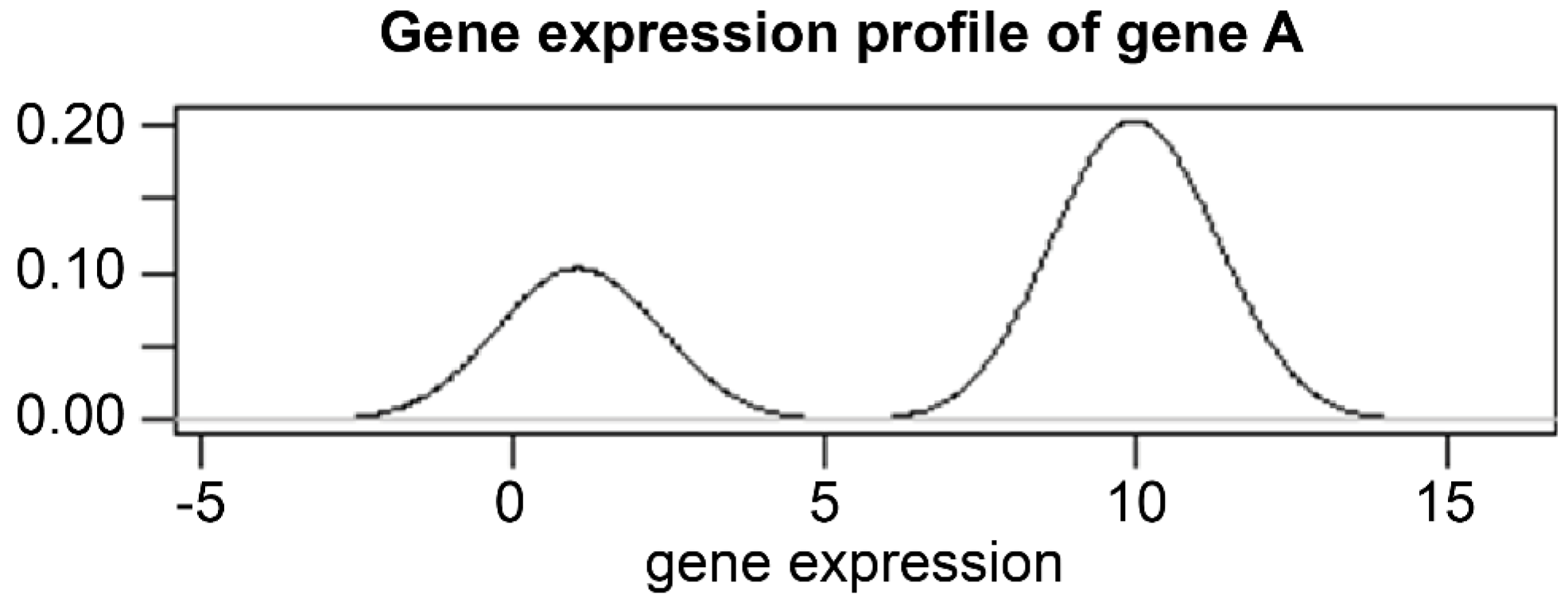
Multiple codons code for the same amino acid (codon degeneracy issue), thus DNA sequences do not adequately capture the conservation of gene function exhibited at the protein level.

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.

**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).

**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).

**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

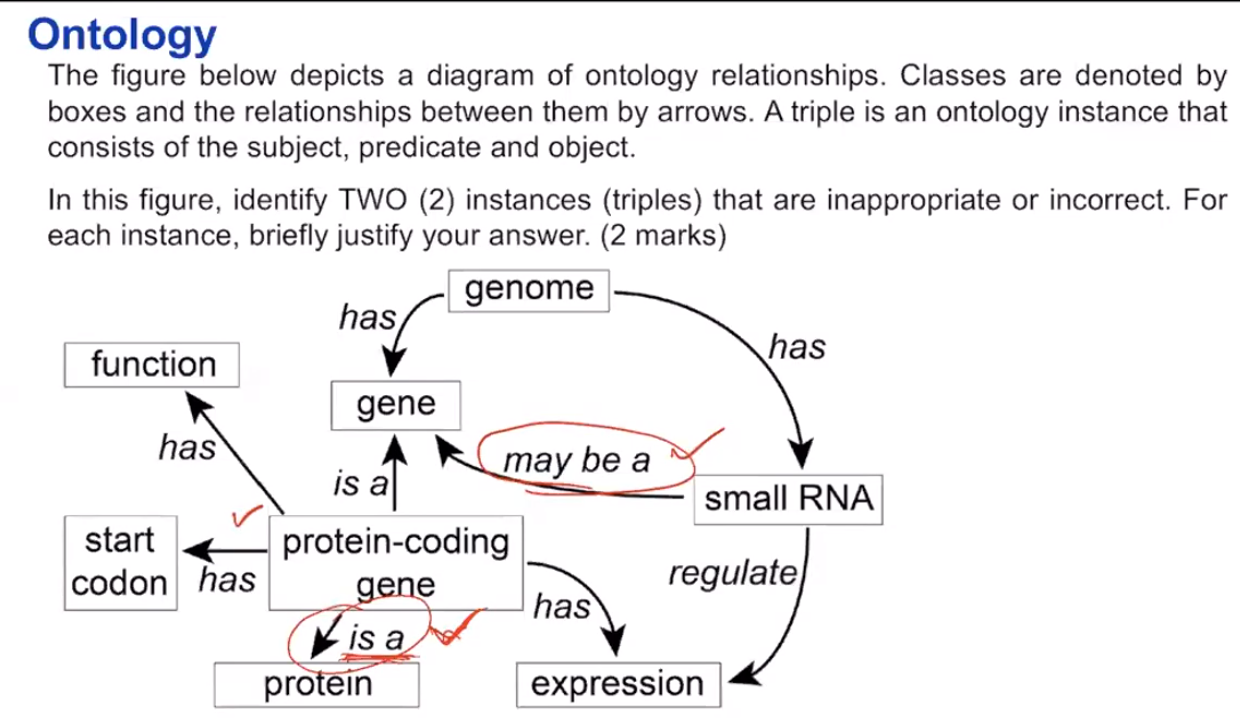
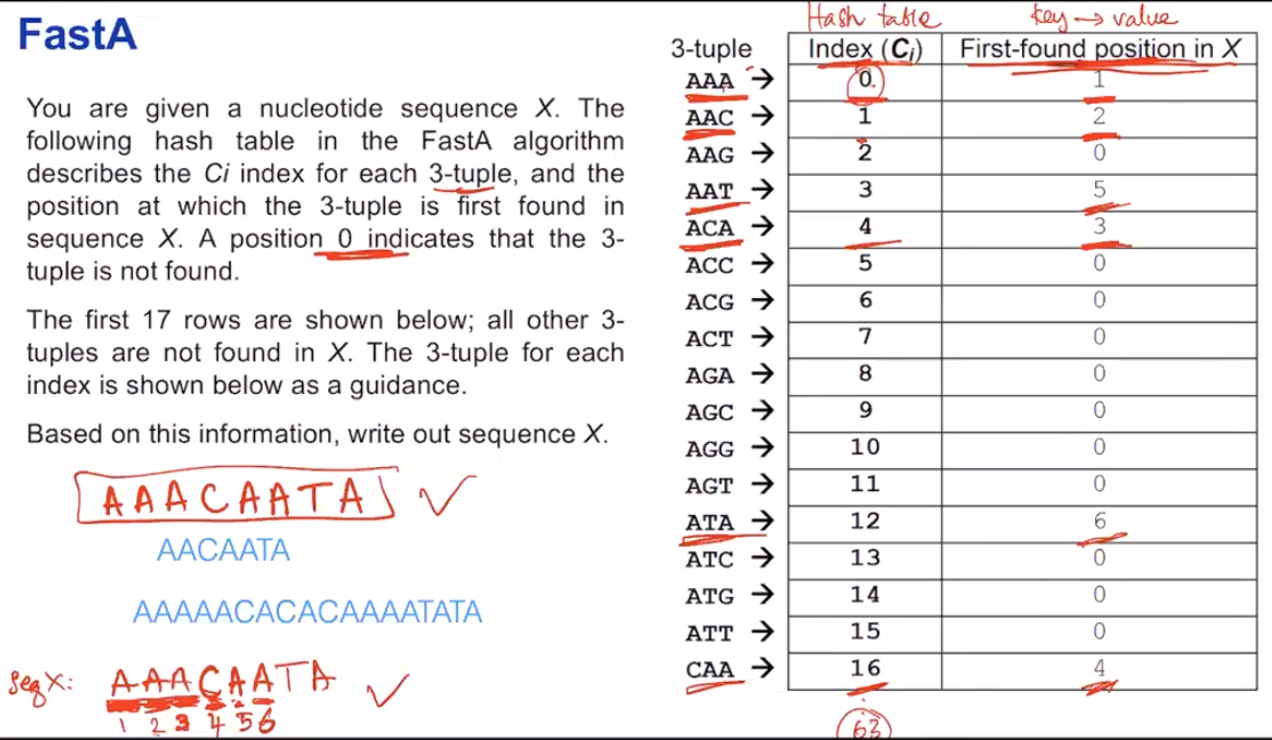
Literature (PubMed) -> what studies have been done on this protein

**Ontology**

**Briefly explain “ontology” and name TWO reasons why ontology is important for organising biological data.**

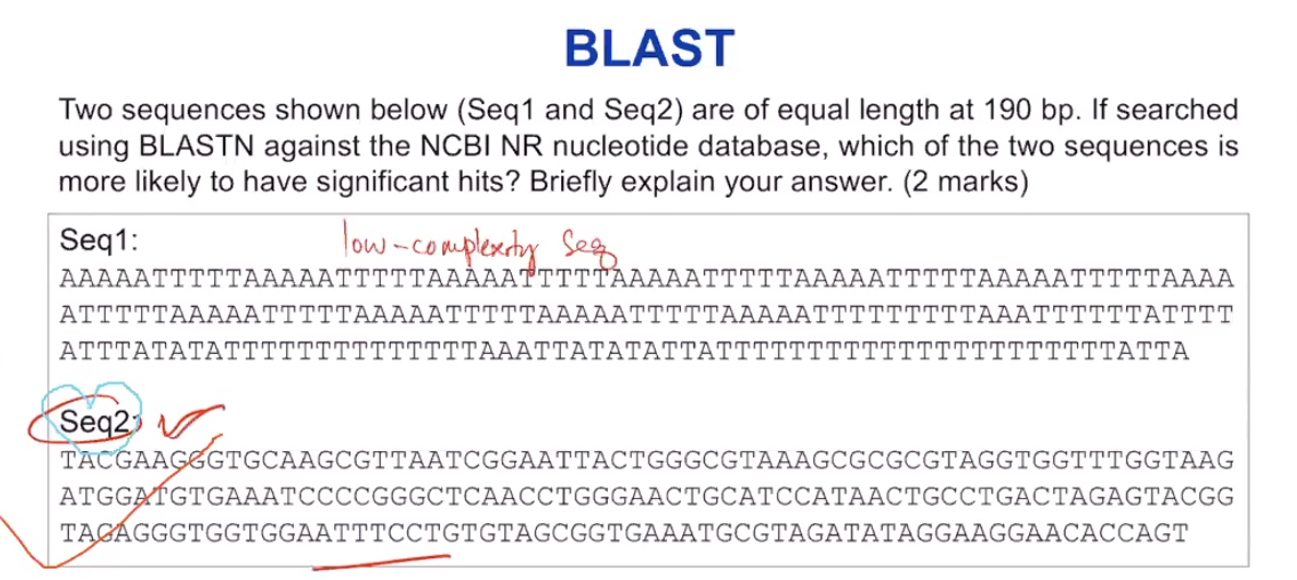
Standardised vocabulary that defines concepts/terms and their relations or constraints in a domain (knowledge / field / discipline).

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

**** ****

**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



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.

**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