

Solutions I. Genes/DNA

1. Ribose has hydroxyl (-OH) group at the 2' position, deoxyribose does not (only has -H)
2. A & G
3. A-T, C-G, G-C, T-A
4. ACGGTGATC → 5' GATCACCGT 3'
5. 60%
6. Answer a, 5' to 3'
7. Answer b, guanine

8. 5' ATGGTTTACTTGAA 3'
 3' TACCAAATGAACTT 5' = non-coding strand

 5' AUGGUUUUACUUGAA 3' = mRNA transcript
 Met-Val-Leu-Leu-Glu = 3-letter abbreviations
 MVLLE = 1-letter abbreviations

9.
 - a. Arabidopsis thaliana, 80 amino acids, AT1G65484
 - b. MGLKM...PRTGS
10.
 - a. Under "protein coding genes" you see 2 different protein-coding transcripts. Exons are blue, coding sequences are yellow, introns are the thin black lines. Each transcript has 2 exons and 1 intron.
 - b. Yes, GU at the start of the intron and AG at the end. The bases at the start of the second exon vary
 - c. Download the data (see screenshots) and generate a new text file where the CDSs have been concatenated.

Primary Data

Name	AT1G65484.1
Type	mRNA
Description	transmembrane protein
Position	Chr1:24347750..24348619 (+ strand)
Length	870 bp

Attributes

Computational_description
unknown protein; FUNCTIONS IN: molecular_function unknown; INVOLVED IN: biological_process unknown; LOCATED IN: endomembrane system; Has 30201 Blast hits to 17322 proteins in 780 species: Archae - 12; Bacteria - 1396; Metazoa - 17338; Fungi - 3422; Plants - 5037; Viruses - 0; Other Eukaryotes - 2996 (source: NCBI BLINK).

Conf_class 3
Conf_rating ****

Dbxref gene:4515100870 UniProt:B3H4Y2

Id AT1G65484.1
Seq_id Chr1
Source Araport11

Region sequence

FASTA

```
>Chr1 Chr1:24347750..24348619 (+ strand) class=mRNA length=870
GCTAAGAAAACAAAATCGGCTTATATCACTATAGAAGAATAATGGGTTTGAAAATGTC AAG
CAATGCACTTCTCTATCTTTGTTTCTCTACTCTTTGTCTCTTTCTGAAAATTGGAGGGA
GTGAGCACTCACTGGAAAATAGGTGAGTGCCTCTCTATCTCACAAATAGCTCAAGCTAT
CAATGGATTTTCTCCCTAAACCAAATTTAGCTTATTTAAGAATATGGTTGTTTGGGA
AACAAATTAATACTGTAGGATTAGTAATTCGGGTTATTTTGTGCTTTGTATTGTATAA
AATAGATTTTATTCATCATTTTGAACATATATCTCTATCTGATACACATTCAATCTTCAG
TAGAAGACAGTAAAGGGCAAAATGCAACCCCTCCATCTCTAACATGCGGAGGCCAAGGA
CTTGAGAGCCACAAACACGTTTAAGTCCATGCCCCGCTCCACGGCCAAAGGCCACGGCCAG
TACAGGCTCTTAAGAAGATTACTACCGACCTTGAAGAAGAAGAAATAAGAATATAAAT
ATTCTATCTTTGTGCTGTAACCTTGATTACATTTTGTAGAGACAGTACAAACAAAGTTT
```

Primary Data

Name	AT1G65484:CDS:1
Type	CDS
Position	Chr1:24347792..24347897 (+ strand)
Length	106 bp

Attributes

Id AT1G65484:CDS:1
Phase 0
Seq_id Chr1
Source Araport11

Region sequence

FASTA

```
>Chr1 Chr1:24347792..24347897 (+ strand) class=CDS length=106
ATGGGTTTGAAAATGTC AAGCAATGCACTTCTCTATCTTTGTTTCTTCTACTTCTTTGTCT
TTTTCTGAAATTGGAGGGAGTGAGACAACCTCACTGGAAAATAG
```

Primary Data

Name	AT1G65484:CDS:2
Type	CDS
Position	Chr1:24348122..24348258 (+ strand)
Length	137 bp

Attributes

Id AT1G65484:CDS:2
Phase 2
Seq_id Chr1
Source Araport11

Region sequence

FASTA

```
>Chr1 Chr1:24348122..24348258 (+ strand) class=CDS length=137
TAGAAGAACCAGTAAGAGGGCAAAATGCCACCCCTCCATCTCTAACATGCGGAGGCCAAGAC
TTGAGGACCAACACACGTTTAAGTCCATGCCCGCTCCACGGCCAAAGGCCACGGCCACGTA
CAGGCTCTTAA
```

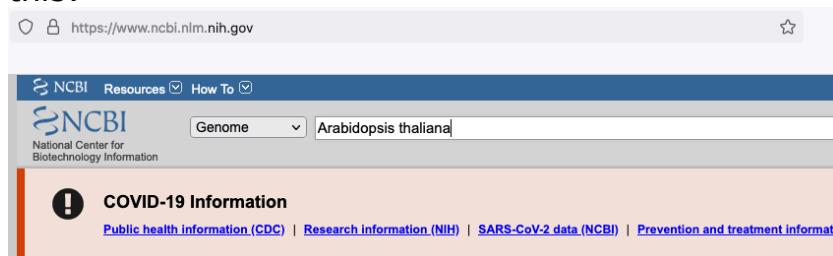
cds.fasta

```
cds
ATGGGTTTGAAAATGTC AAGCAATGCACTTCTCTATCTTTGTTTCTTCTACTTCTTTGTCT
CTTTTCTGAAATTGGAGGGAGTGAGACAACCTCACTGGAAAATAG
TAGAAGAACCAGTAAGAGGGCAAAATGCAACCCCTCCATCTCTAACATGCGGAGGCCAAGGA
CTTGAGGACCACACACGTTTAAGTCCATGCCCGCTCCACGGCCAAAGGCCACGGCCACG
TACAGGCTCTTAA
```

- Yes, $106+137=243$, $243/3 = 81$ codons and the last codon is a stop codon.
- Yes, MGL...
- Yes, about 24 nucleotides before the transcription start. Transcription starts at the 5' UTR of AT1G65484.1 (same as start of the gene locus). The sequence before that is

CTTC**TATATAA**ACCGGTCCAGTATTATT. The bold/underlined bases form the TATA box in line with the definition (Fig 1.5 in the book, see also knowledge clip on DNA, slide 16).

11. GC content
 - a. Possible tools to use are
<http://www.endmemo.com/bio/gc.php> or
<https://www.sciencebuddies.org/science-fair-projects/references/genomics-g-c-content-calculator>
 GC content for transcript: 34.137931 (length 870)
 GC content for CDS: 46.91358 (length 243)
 The GC content in the CDS is much higher.
 - b. The GC content of chromosome 1 is 35.9
 It is known that coding regions have higher GC than the background genome, consistent with our observation.
 The GC content can be found by searching NCBI like this:



You should get to this page:

<https://www.ncbi.nlm.nih.gov/genome/?term=Arabidopsis+thaliana>

12. Viruses do contain DNA or RNA genomes, but they can only replicate inside a living cell of another organism, and thus are not considered cellular life forms.
<https://www.nature.com/scitable/content/viruses-and-the-tree-of-life-14465158/>

13.

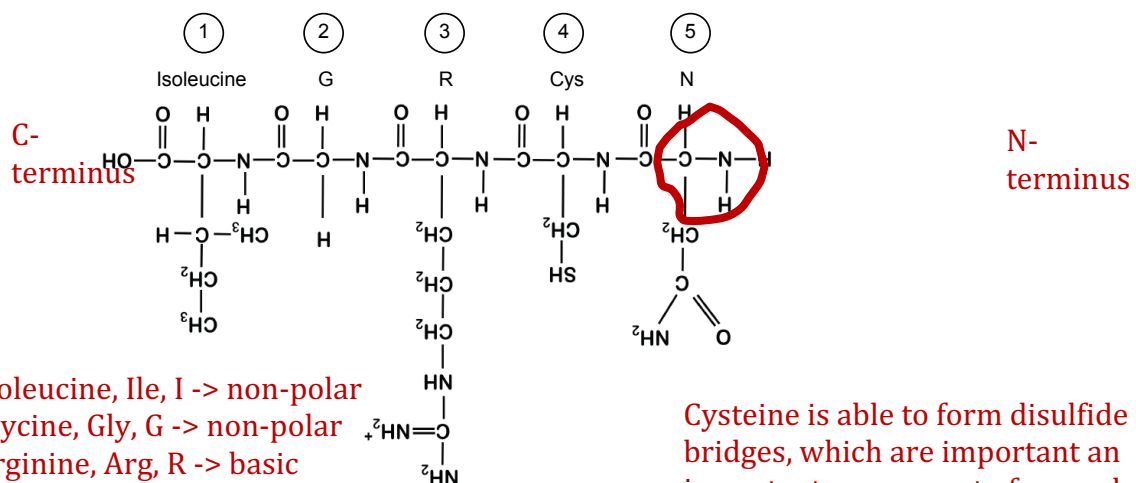
	Domain	Family
Moraxella catarrhalis	Bacteria	Moraxellaceae
Haloarcula quadrata	Archaea	Haloarculaceae
Loxodonta cyclotis	Eukaryota	Elephantidae

Solutions II. Proteins

1. Glycine is the smallest amino acid; it only has one hydrogen (H) atom as its side chain.
2. The nonpolar amino acids are generally hydrophobic. So, you could have listed any of glycine, alanine, valine, leucine, Isoleucine, proline, phenylalanine, methionine, tryptophan, cysteine.

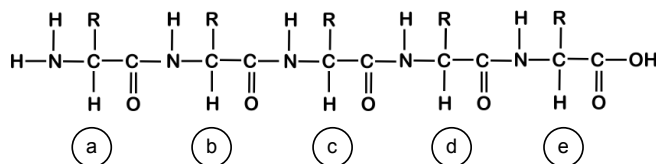
3. Aspartic acid and glutamic acid
4. Answer a is incorrect, A stands for Alanine
5. The nonpolar amino acids are hydrophobic (not liking water), and therefore tend to be buried inside the protein surrounded by other hydrophobic amino acids.

6. Side chain activities



- 1: Isoleucine, Ile, I -> non-polar
- 2: Glycine, Gly, G -> non-polar
- 3: Arginine, Arg, R -> basic
- 4: Cysteine, Cys, C -> non-polar
- 5: Asparagine, Asn, N -> uncharged polar

Cysteine is able to form disulfide bridges, which are important an important component of secondary and tertiary structures.



7. Side chains and their activities
 - a & b: other non-polar aa such as Val/Ala/Ieu → hydrophobic
 - c: Asp -> electrostatic
 - d: Cys -> disulfide
 - e: Tyr (or any other polar amino acid)-> hydrogen bond
8. Protein structures
 - a. there are 2 times 4 and 3 large beta strands (organized in 2 anti-parallel beta-sheets). Furthermore the

structure contains 1 alpha-helix and 3 3/10 helices, the fourth most common type of protein secondary structures.

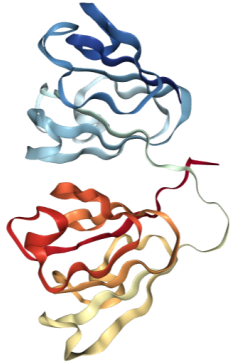
- b. Likely two identical domains. If you look under "Annotations" in the menu, you see various different annotation sources that mention a gamma-crystallin domain. Under the "Sequence" tab you can see where these domains are on the protein, for example the PFAM or SCOP annotations clearly show two domains.

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

Structure Summary **3D View** Annotations Sequence Sequence Similarity Structure Similarity Experiment Literature

1AMM
1.2 ANGSTROM STRUCTURE OF GAMMA-B CRYSTALLIN AT 150K

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds.



Display Options

- Assembly: Bioassembly 1
- Model: Model 1
- Symmetry: None
- Interaction: None
- Style: Cartoon
- Color: Rainbow
- Ligand: Ball & Stick
- Quality: Automatic
- ☐ Water ☐ Ions
- ☒ Hydrogens ☐ Clashes

Viewer Options

- ☐ Spin
- ☐ Fullscreen
- Focus:

NGI is a WebGL based 3D viewer powered by MMTF. Select a Viewer NGI (WebGL)

9. Amino acid quiz

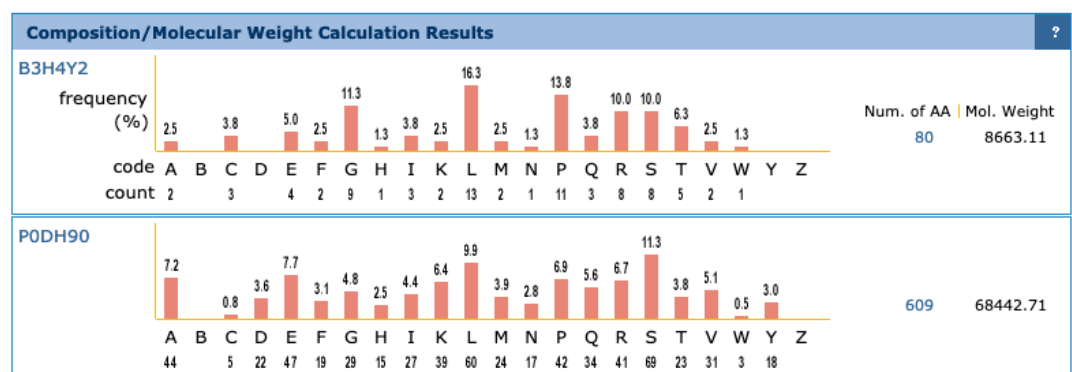
	1-letter	3-letter	Full name	Class
1	E	Glu	Glutamic acid	Nonpolar/Polar/ Acidic /Basic
2	F	Phe	Phenylalanine	Nonpolar /Polar/Acidic/Basic
3	T	Thr	Threonine	Nonpolar/ Polar /Acidic/Basic
4	P	Pro	Proline	Nonpolar /Polar/Acidic/Basic
5	S	Ser	Serine	Nonpolar/ Polar /Acidic/Basic
6	K	Lys	Lysine	Nonpolar/Polar/Acidic/ Basic
7	I	Ile	Isoleucine	Nonpolar /Polar/Acidic/Basic
8	N	Asn	Asparagine	Nonpolar/ Polar /Acidic/Basic
9	M	Met	Methionine	Nonpolar /Polar/Acidic/Basic
10	A	Ala	Alanine	Nonpolar /Polar/Acidic/Basic
11	P	Pro	Proline	Nonpolar /Polar/Acidic/Basic
12	H	His	Histidine	Nonpolar/Polar/Acidic/ Basic

Solutions III. Databases

1. As a result of this question you should have explored a few databases. On the exam you could be asked to mention a few databases with biological data, so it's good to get a feeling for how much and what kind of data is out there and how the data is organized and searchable.
2. Redundancy
 - a. Redundancy in a database means that the database contains multiple entries with identical data. In a sequence database it could be that a certain protein sequence of a species has been submitted by several labs.
 - b. Each of the UniProt databases is non-redundant (<https://www.uniprot.org/help/redundancy>). Definitions of redundancy differ.
 - c. GenBank is a sequence database, containing sequences submitted by individual labs or large-sequencing projects. GenBank is redundant and can be very redundant for certain loci. RefSeq is the non-redundant version of GenBank where (near-)identical entries are merged. If multiple GenBank submissions represent the same molecule for an organism, the "best" sequence is chosen to represent as the RefSeq record.
3. Ontology
 - a. An ontology is a formal specification of used terms and their connections. A set of concepts and categories in a subject area or domain that shows their properties and the relations between them.
 - b. Biological process, molecular function, cellular component
 - c. B3H4Y2: You find 12 annotations. There are several GO-terms assigned by UniProt, the others are assigned by TAIR (the arabidopsis information resource). GO:0016021 suggests this is an integral component of membrane; However, the molecular function annotation are assigned through sequence similarity only (i.e. there is no experimental evidence that verifies this). The linked processes are "response to salicylic acid, ethylene, and water deprivation" (based on expression pattern evidence).
 - d. Biological processes: flower development, cell differentiation. Cellular components: nuclear speck. All terms are assigned by uniprot through 'electronic annotation'.

4.

- Both proteins are in swissprot. This can be seen through the gold/silver image next to the ID at the top of the info page. From this perspective both protein annotations are equally trustworthy.
- B3H4Y2 is linked to 3 publications: 2 that describe the Arabidopsis genome (unspecific) and 1 (recent paper) about secreted transmembrane peptides (seems relevant). PODH90 is linked to 8 specific publications
- For B3H4Y2 it is e.g. possible to look up RefSeq, we find the genbank file of the protein. For a novel DB it is interesting to look up EnsemblPlants, which covers information on the transcript and on known variation. For PODH90 it is also possible to look up Pfam. For a novel DB, it is interesting to look up Expression Atlas, where the expression data from several papers is presented.
- We find a very high fraction of Proline in B3H4Y2, this is unusual since the residue is also bound to the amino group, which has important effects on the protein structure. The proline-rich C-terminus is also noted in Uniprot (under Compositional bias).



- Findable:** data is well annotated, persistent identifiers, indexed

Accessible: standardized communication protocols

Interoperable: allow linking/exchange/import of data from different sources through accepted data representations and ontologies.

Reusable: others should be able to use it, clear, standardized descriptions, proper license, etc.

Solutions IV. Genome annotation

- Many annotated prokaryotic genomes are present in databases. For a novel genome, you can search against these databases (either nucleotide sequences of genes or protein sequences) to see if these genes/proteins are also present in

this novel genome. Typically more than 50% of the genes can be identified this way. For eukaryotes this is much more complex because of introns. There it is more useful to align known proteins to a novel genome as evidence.

2. RNA-seq is a direct read-out of transcription and as such is very useful in finding splice-sites. Not all genes are always expressed, so typically RNA-seq is not sufficient to identify all protein-coding genes.
3.
 - a. 6 ORFs are detected, the reading frames are different (1, 1, 2, 2, 2, 3). Note: if you copied the gene sequence from Araport, you will likely have found 6 ORFs in frames 1, 1, 2, 2, 2, 2. No, because the gene contains an intron, so there's not 1 continuous ORF in the gene sequence. After splicing, the ORF should be in the coding sequence (CDS).
 - b. ORF1 corresponds to the start of the gene, but it is too long. The gene has an intron at some point, which is not translated.
 >|c|ORF1
MGLKMSSNALLLSLFLLLLCLFSEIGGSETTHWKIGQCLPISH
 NSSSYQWIFFSPKPNLAYLRIWLFLETNYNRRISNSALFLLLLY
 CIK
 The first CDS has a length of 106 (producing 35 AAa), the second part 137, together 243. The second part of the protein does not start with a start codon, so won't be found by ORFFINDER.
 - c. In prokaryotes, when there are no introns in the genes, it is very useful. In eukaryotes containing many genes with introns, it typically does not work
4.
 - a. 6572 protein-coding genes
 - b. 10 tRNAs on Chr 3, 24 on mt.
 - c. 85,779 nt (85.8 Kb), circular genome
 - d. <https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG3>. 6 codons have a different meaning.
https://en.wikipedia.org/wiki/Yeast_mitochondrial_code also mentions that 2 codons are absent.
 - e. ---
 - f. When you click on the image from the web server, or look in the GFF file, you can count 24 tRNAs.
 - g. Yes: the coordinates of the first tRNA match the reference annotation (but for example in the second one the stop position is different, so the coordinates do not

always match). The tRNA corresponds to Proline, 3 GO terms: GO:0005739, GO:0006414, GO:0030533

- h. The protein seems to be some sort of polymerase according to for example the 'homologous superfamily' section of the output. Associated GO term: mRNA processing (GO:0006397)

5.

- a. 143 predicted, 9 with an intron (search for the word intron, look at them and count)
- b. Using InterProScan, the information you find is limited, for example there are no associated GO terms. There is a domain of an ATP-dependent chromatin-remodelling protein, which could provide hints for further research into this proteins function. If you used BLASTP on UniProt, you could have found a lot more information: it seems this protein is <https://www.uniprot.org/uniprotkb/P25632/entry>, and there is for example some literature on this class of proteins. NOTE: using BLASTP was not part of this week's assignment, so you are not expected to have found this information at this point in the course.

- 6. Structural annotation is the identification of genome features in the genome, e.g. start and stop positions of gene regions, exons, introns, UTRs, CDS. Functional annotation is assigning biological information to the genome features, e.g. protein function, domains, enzyme codes, type of transposon, etc.