Week 1 Practical Guide

This guide contains questions and exercises to help you process the study materials of Week 1. You have 2 mornings to work your way through the exercises. In a single session you should aim to get about halfway through this guide (i.e., day 1: assignment 1-3, day 2: assignment 4 and project preparation exercise). Answers to assignment 1-3 will be provided at the end of the morning on day 1, the remaining answer will become available at the end of the morning on day 2.

These practical exercises offer you the best preparation for the project. Especially the **project preparation exercise** at the end is a good reflection of the level that is required to write a good project report. Make sure that you develop your practical skills now, in order to apply them during the project.

Assignment I: **Genes/DNA**

Time indication: ~45min.

1. How do you distinguish a ribose sugar from a deoxyribose?

2. Which bases are purines?

3. What is the complementary base of A?   C?   G?   T?

4. What is the reverse complement of sequence ACGGTGATC?

5. What is the GC content of sequence ATCGATCGGC?

6. Which is correct? A nucleotide sequence is written from:
   a. 5’ to 3’
   b. 3’ to 5’

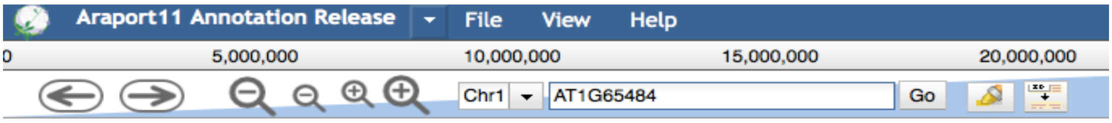
7. In a DNA sequence the G stands for
   a. Glycine
   b. Guanine
   c. Glucose
   d. Glutamic acid

8. Given a coding DNA strand. Write down the non-coding strand, the transcribed sequence, and the resulting chain of amino acids. You may use Fig. 1.3.

```
Coding strand:      5’ ATGGTTTTACTTGAA 3’
Non-coding strand:  ......................
mRNA:               ......................
Amino acids:        ......................
```

9. On the computer, browse to www.uniprot.org and search for UniProt ID B3H4Y2.
   a. In which organism is this protein found? What is the length of this protein? What is the corresponding gene ID?
   b. Write down the first 5 and last 5 amino acids of the protein

10. Browse to www.araport.org and click on "TAIR JBrowse" (Firefox or Chrome recommended). This will take you to a genome browser of the Arabidopsis genome. Search for the gene ID from question 9 (see screenshot below).



Under "Help" -> "General" you can find some information to help you understand what you are looking at.

a. You can see that this gene produces two different mRNA transcripts (indicated by .1 and .2) and thus 2 different proteins. How many exons do these transcripts contain? How many introns?
b. Turn on the track "Light grown seedling" under "RNA-seq based evidence"/"Aligned reads". Do you recognize the splice sites? Are the first two and last two bases of the intron as expected (based on the description in the book, Fig. 1.5)?
c. Save the data for transcript 1. Save a fasta file for the whole transcript and one for each coding sequence (CDS). Create a fasta file on your computer that contains the complete coding sequence of the protein.
d. Is the length of the coding sequence in line with your expectation (based on your findings in question 9a)?
e. Translate the first and last few codons to compare them against the protein sequence (Q.9). Do they match?
f. Look upstream of the gene. Can you find the TATA box? How many nucleotides before the start of transcription?

11. GC content

a. Find a tool on the internet to calculate the GC content of a gene. Which tool did you find? Use it to calculate the GC content for the whole transcript and for the coding sequence that you created in the previous task. What do you observe?
b. Look up the GC content of the chromosome where this gene is located (Hint: Search NCBI Genome for the species).
Read about GC content in coding sequences https://en.wikipedia.org/wiki/GC-content Which of the information presented here agrees with your analysis?

12. Why are viruses not represented in the tree of life? Take a look at this site (https://www.nature.com/scitable/content/viruses-and-the-tree-of-life-14465158)

13. Browse to the NCBI taxonomy. Look up the domain and family of the following species:

	Domain	Family
Moraxella catarrhalis		
Haloarcula quadrata		
Loxodonta cyclotis		

Assignment II: **Proteins**

Time indication: ~45min.

1. What is special about the amino acid glycine?

2. List three hydrophobic amino acids

3. Which amino acids are acidic? Color or underline them.

alanine	glutamine	leucine	serine
arginine	glutamic acid	lysine	threonine
asparagine	glycine	methionine	tryptophan
aspartic acid	histidine	phenylalanine	tyrosine
cysteine	isoleucine	proline	valine

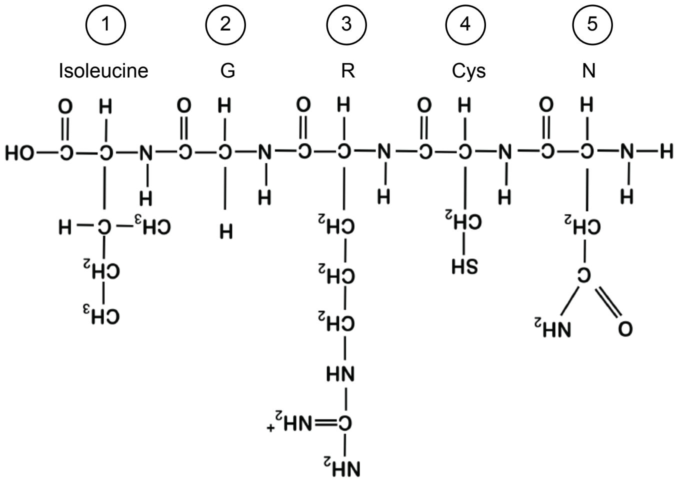
4. Which is incorrect?
   a. A = Arginine
   b. V = Valine
   c. Q = Glutamine
   d. T = Threonine

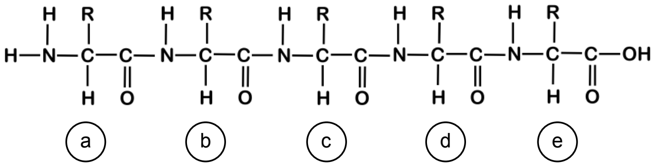
5. In a folded protein, the nonpolar amino acids tend to be:
   a. On the inside of the protein
   b. At the surface of the protein
   c. Randomly distributed

6. The side chains of amino acids play important roles in the folding and the function of proteins. On the next page you can see a short peptide that has been formed by five amino acids (labeled from 1 to 5).

   a) Indicate in blue and red the N-terminus and the C-terminus of the peptide, respectively, and highlight all peptide bonds in green.

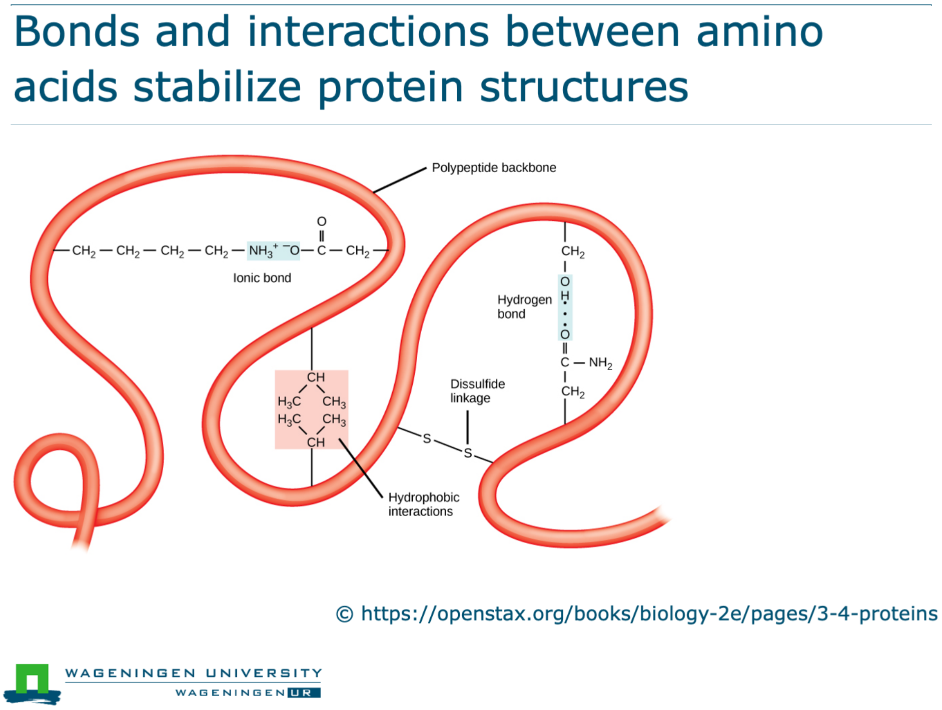
b) For each of the five amino acids (1-5), give either the name, the three-letter or the one-letter code, depending on the information lacking (for example, for amino acid 1, give the three- and the one-letter code, while for amino acid 2 give the name and the three-letter code).
c) Indicate for each of the amino acids (1 to 5) its physiochemical properties (nonpolar, polar, acidic, basic).
d) Describe in one sentence what specific property the side chain of amino acid 4 has, and why this property is important to form protein structures.





7. Amino acids and their side chains can interact with other amino acids and form bonds and interactions. Interactions between amino acids and their side chains play important roles in stable folded proteins structures. Revisit figure 14 from the reader. With this information in mind, take another look at the peptide sequence with five amino acids (see Q6). Below this peptide, you will find the backbone of another peptide in which the side chains have been only indicated with R (labeled *a*-*e*).
   Look at the 20 amino acids in the amino acid table in the reader. Discuss with your neighbour which of the 20 possible amino acids could be placed as a side chains R (*a*-*e*) such that

these can likely interact, i.e. forms bonds or other interactions, with the corresponding amino acids (1-5) in the upper peptide (i.e. *a* interacts with 1, *b* with 2, and so on). Indicate for each which type of interaction (e.g. hydrogen bonds) are occurring between your proposed pair of amino acids.



Bonds and interactions between amino acids stabilize protein structures

© https://openstax.org/books/biology-2e/pages/3-4-proteins

8. Proteins fold into compact structures, and this structure is important for proteins to have biological functional activity. In folded proteins (tertiary structure), the secondary structure is often still visible, i.e. helices and beta sheets are still visible. Sometimes proteins are not only formed by a single structural unit, so-called domain, but by multiple domains that can either be the same type or of different types. Sometimes, it can be useful to look at the tertiary structure of proteins with known fold (either experimentally or *in silico* determined), e.g. to see where mutations in the structure occur. We will have a look at the protein structure of Gamma B-Crystallin. Go to the website of PDB (https://www.rcsb.org), which is a resource for protein structures, and search on the main page for Gamma B-Crystallin, with the ID "1AMM". Click on '3D View' to see a three-dimensional model of the structure. On the bottom right, change the viewer to NGL.

   a) Color the structure by Secondary structure (see 'Color' under 'Structure View'). Under 'Structure View Documentation' you can find the meaning of each color.

Can you identify the number of secondary structure elements (helix, sheet) you can observe in the structure?

b) How many domains does this protein have?

9. Amino acid quiz: you have now worked extensively with amino acids and you should know the relation between the 1- and 3-letter code, the name of the amino acid and its biochemical properties. To test this knowledge once more, perform this small test by filling in the missing information in the table (do not look at the reader before finalizing the quiz).

	1-letter	**3-letter**	**Full name**	**Class**
1			Glutamic acid	Nonpolar/Polar/Acidic/Basic
2		Phe		Nonpolar/Polar/Acidic/Basic
3	T			Nonpolar/Polar/Acidic/Basic
4		Pro		Nonpolar/Polar/Acidic/Basic
5			Serine	Nonpolar/Polar/Acidic/Basic
6	K			Nonpolar/Polar/Acidic/Basic
7			Isoleucine	Nonpolar/Polar/Acidic/Basic
8		Asn		Nonpolar/Polar/Acidic/Basic
9			Methionine	Nonpolar/Polar/Acidic/Basic
10	A			Nonpolar/Polar/Acidic/Basic
11	P			Nonpolar/Polar/Acidic/Basic
12		His		Nonpolar/Polar/Acidic/Basic

Assignment III: **Databases**

Time indication: ~45min.

1. In a web browser, navigate to the Molecular Biology Database Collection of the journal *Nucleic Acids Research* (NAR): http://www.oxfordjournals.org/nar/database/c/
   Pick three databases from the list that draw your attention, preferably from different categories, and explore them (approx. 5 min each):
   a. What type of data is in there?
   b. What would it be used for? Highly specialized or broad applications?
   c. How can you search the database?
   d. Does it look up-to-date and regularly maintained?

2. Redundancy
   a. What does redundancy in a database mean? Give an example of redundancy in a sequence database.
   b. Are the UniProt databases redundant or non-redundant?
   c. What is the difference between RefSeq and GenBank in terms of redundancy?

3. Ontology
   a. Describe what an ontology is (use the reading material and/or google to find information).
   b. The Gene Ontology is one of the most important ontologies in bioinformatics. Which biological domains are covered in the Gene Ontology?
   c. Look up the Arabidopsis protein from the Chapter 1 exercises and look it up in UniProt (Accession B3H4Y2). Which information do you find about the GO terms associated with this protein?
   d. Now look up the famous Arabidopsis gene FRIGIDA (Accession P0DH90). Which GO terms are associated with this gene? In which cellular component is this protein found and which biological process is it involved in?

4. UniProt
   a. Look up the two proteins from Q3 in UniProt again. In which of the sections of UniProt is each of them deposited. Which of the two has a higher annotation quality?
   b. How many publications are linked to each of these proteins? Which of these publications contains specific information on the protein (based on the title)?
   c. For each protein, look up at least one cross-reference to a database that you know and to a database that you do not

yet know. Spend a few minutes to browse the information that you can gain in this way.

d. Calculate the frequency of individual amino acids in both protein sequences using the PIR website (http://pir.georgetown.edu/pirwww/search/comp_mw.shtml). Do you notice something remarkable (Hint: look at relative abundance of various amino acids)? Can you relate this to information that is present in Uniprot (Hint: look at family/domains)?

5. A hot topic in biological data management is “FAIR” data. What do the letters in FAIR stand for and what do those terms mean?

Assignment IV: **Genome annotation**

Time indication: ~120min.

1. Explain how homology searches can be useful in genome annotation and why it is more complex for eukaryotes than for prokaryotes.

2. How does RNA-sequencing data help gene prediction? Is RNA-sequencing data on its own sufficient to annotate a genome?

3. In the uniprot databse (www.uniprot.org) look up the Arabidopsis protein with identifier B3H4Y2. The corresponding gene ID is AT1G65484. In a second tab, look up this gene in www.araport.org JBrowse. The sequence of this gene can be found on BrightSpace. In a third tab, go to https://www.ncbi.nlm.nih.gov/orffinder/ and paste in the gene sequence in FASTA format and hit submit.

   a. How many ORFs are found? Are they all in the same reading frame?
   b. Does any of the ORFs correspond to the ORF in the annotated gene? Why or why not?
   c. When is a simple ORF detection tool useful? When is it insufficient?

4. Yeast (*Saccharomyces cerevisiae*) is a well-studied model organism. A lot of information about yeast is stored in the Saccharomyces Genome Database (SGD) (https://www.yeastgenome.org). You will work on the reference strain S288C. The relatively small genome of yeast gives us the opportunity to explore some online genome annotation tools (within reasonable time) and compare our findings to the high-quality annotation that is available. In the

menu on SGD click on Sequence -> Reference Genome -> Genome Snapshot. Under “Features by Type” explore both the graph and the table view.

a. How many genes does the yeast genome contain?
b. How many tRNAs have been annotated on chromosome 3 and how many on the mitochondrial genome?

First you will annotate the mitochondrial genome of yeast, using the MITOS Web Server (http://mitos2.bioinf.uni-leipzig.de).

c. Use the mitochondrial genome, provided on Brightspace or download the sequence from SGD. What is the length of the mitochondrial (MT) genome? Be creative or use google to find the answer. Is the MT genome linear or circular?
d. The yeast MT genome does not use the Standard Genetic Code. Which one does it use? Use https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi to find the answer. How many codons have a different meaning?
e. Go to the MITOS web server and provide your credentials (Name, Email). Select a relevant Reference and Genetic Code for yeast and upload the fasta file. Hit submit. This annotation can take about 10 minutes, so continue with the next exercises (6 and 7) until the results are done.
f. Browse the results by clicking around. How many tRNAs are predicted?
g. Look up the first predicted tRNA in the genome browser on SGD. Does the prediction match the known tRNA? To which amino acid does this tRNA correspond? How many GO terms are associated with this tRNA gene?
h. What can you find out about the “giy” gene? In the FAA fasta text file (download from the menu on the left) you can find the protein sequence produced by the gene. Go to https://www.ebi.ac.uk/interpro/search/sequence and use the protein sequence to search for known protein domains/functions using InterProScan. What functional information do you find for this protein? (Running InteProScan can take a few minutes, in the meantime continue with the next question).

5. Next, we will predict protein-coding genes on chromosome 3 using the widely used Augustus ab-initio gene predictor (http://bioinf.uni-greifswald.de/augustus/submission.php). Augustus outputs its predictions in the GFF3 text-based file format. To be able to make sense of the Augustus output, familiarize yourself with this format here:

https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md. The Augustus gene predictor has already been trained on many organisms.

a. Upload the sequence of chromosome 3, choose the right Organism and "run AUGUSTUS" (predict genes on both strands). How many genes are predicted? How many of those contain an intron?
b. What can you say about the function, domains, etc. of gene g100? Which databases did you use to find this information?

6. You have now seen several examples of tools used in genome annotation. Describe structural and functional genome annotation in your own words and give an example of each.

Project Preparation Exercise

Time indication: ~45min.

We want to obtain insights into members of the ARF gene family in *Arabidopsis thaliana*. ARF5 (UniProt ID P93024) and IAA5 (UniProt ID P33078) are two well-studied *A. thaliana* proteins that play a role in auxin-mediated regulation of gene expression. They are therefore chosen here as the starting points for exploring the plant ARF gene family. Perform a small background study on ARF5 and IAA5. Explore the protein sequences, properties (e.g., length, composition, etc.), interaction partners, and functional regions of ARF5 and IAA5. Finally, explore the genes encoding ARF5 and IAA5 in *A. thaliana* (genomic location, exon structure, expression, etc.).

Describe the following items in a few bullet points each. You might include up to two figures or tables.

1. **Materials & Methods** What did you do? Which data, databases and tools did you use, and why did you choose these? What important settings did you select?
2. **Results** What did you find, what are the main results? Report the relevant data, numbers, tables/figures, and clearly describe your observations.
3. **Discussion & Conclusion** Do the results make sense? Are they according to your expectation or do you see something surprising? What do the results mean, how can you interpret them? Do different tools agree or not? What can you conclude? Make sure to describe the expectations and assumptions underlying your interpretation.