1. What is an Open Reading Frame? Why is it easier to find or predict genes in prokaryotic genomes than in eukaryotic genomes?

An open reading frame is a stretch of codons, which starts with a start codon (ATG/AUG) ends with a stop codon (e.i. TGA, TAA). This stretch has the possibility to be translated, as long as the amount of nucleotides in the open reading frame can be divided by three.

It is easier to predict genes in prokaryotes because there are, in comparison to ORFs in eukaryotes, no introns in prokaryotic genes.

2. Give example of a few key features used to find genes in eukaryotic genomes.

- CpG-islands
- it is important to distinguish between introns and exons, which can be found by splice sites
- binding sites of poly-A-tails

3. Attached you find a part of a sequence from E.coli (E.coli.txt). Use ORF finder to find open reading frames longer than 300 nucleotides in that sequence. How many do you find (if does not matter if they are overlapping)?

17.

4. Attached you find an unknown genomic sequence (unknown-seq.txt). Use ORF finder again to find ORFs larger than 300 nucleotides in this sequence. How many do you find?

11.

5. Based on your findings, do you think that the sequence is prokaryotic or eukaryotic? Motivate your answer.

I am convinced, that the unknown sequence is eukarzotic, since there are a lot of untranslated regions between the open reading frames. In prokaryotic sequences there are nearly no untranslated regions.

6. How many known structures of Bacteriorhodopsin (and its variants) did you find with entrez ?

215.

7. How many Bacteriorhodopsin proteins and variants are found in each of the following: Bacteria, Eukaryotes and Archaea? Describe briefly how you performed the search.

Bacteria:	4842
Eukaryotes:	1832
Archea:	1745

8. How many articles are there about "molecular dynamics"?

60498

9. How many articles have an author named Lindahl?

3507

10. How many articles about "molecular dynamics" have an author named Lindahl?

45

11. How many review articles about "protein structure prediction" can you find?

122

12. How many sequences are there in SwissProt and in TrEMBL?

Swissprot:	556388
TrEMBL:	102248261

13. Why does TrEMBL have considerably more sequences than SwissProt?

Sequences from SwissProt are manually annotated and reviewed, whereas sequences from TrEMBL are only automatically annotated and not reviewed.

14. Search for CCR5 (which is the C-C chemokine receptor type 5) in Uniprot. How long is the CCR5 protein in human?

352

15. Check out CCR5 in human. How many scientific articles are referring to this protein (excluding computationally mapped references)?

32

16. Which disease is it associated with?

Diabetis mellitus, insulin dependent

17. What kind of databases are they?

International Nucleotide Sequence Database Collaboration (INSDC), basically databases which show nucleotide sequences.

18. How are they different from SwissProt and TrEMBL?

DDBJ, EMBL-ENA and GenBank collect data about nucleotide sequences and not about proteins

19. What does kegg stand for? What is KEGG?

KEGG: Kyoto Encyclopedia of Genes and Genomes

It is a collection of databases which consists of genomes, biological pathways, diseases, drugs and chemical substances. It aims to educate and research in bioinformatics.

20. Sucrose is involved in a number of pathways, search for it in KEGG and please find out which.

Galactose metabolism

Starch and sucrose metabolism

Metabolic pathways

ABC transporters

Phosphotransferase system (PTS)

Taste transduction

Carbohydrate digestion and absorption

21. What does OMIM stand for? What is OMIM?

OMIM: Online Mendelian Inheritance in Man

It is a collection of human genes and genetic phenotypes connected to genetic disorders.

22. Look up oculotaneous albinism (type II) in OMIM. What eye-colour can this lead to? Please provide the link you used.

https://omim.org/clinicalSynopsis/203200 eyes: blue-gray to light brown

23. What gene is thought to be connected with this disorder and where is it located? Please provide the link you used.

https://omim.org/entry/203200   OCA2 gene

24. What is Gene Ontology (GO)?

GO is a initiative working in bioinformatics, which aims to unify the representation of gene and gene product preferences accross all species.

25. What three categories are at the first classification level in GO?

- cellular component
- molecular function
- biological process

26. Search for alcohol dehydrogenase in GO. What definition for the alcohol dehydrogenase activity is there for the first hit?

Catalysis of the reaction: sinapaldehyde + NADPH + H+ = sinapyl-alcohol + NADP+.