Practicum Bioinformatics - answer key

Prof. Kris Laukens - Adrem Data Lab
Department of Computer Science, University of Antwerp

Academic year 2024-2025

Contents

Note about the answers

These are just example solutions. For many of the exercises there will be other correct solutions as well. The main goal of these practicals is to become acquainted with the different databases, techniques and principles of bioinformatics. Try to understand why we are using a specific technique, the different aspects of the outputs of the tools and how they work, rather than learning everything by heart. You will need to be able to interpret related questions to the ones you've solved here while using this text as a quick reference, but you will need to know the major sections and themes of the practicals if you want to be able to retrieve them in a timely fashion.

1 Databases & Alignment

Section 1 – Databases

Learning goals

At the end of this section, you should:

- know the databases discussed in the syllabus and what information they contain
- be able to access the databases and retrieve the desired information
- understand how the different databases relate to each other

Before starting the exercises on databases, note that a database is something dynamic. Records of your search results could vary over time. Therefore, the screen shots in this tutorial could already be outdated while writing this document. When doing research it is always important to write down the database version and the date and time you have accessed it.

Exercise 1.1: NCBI Databases:

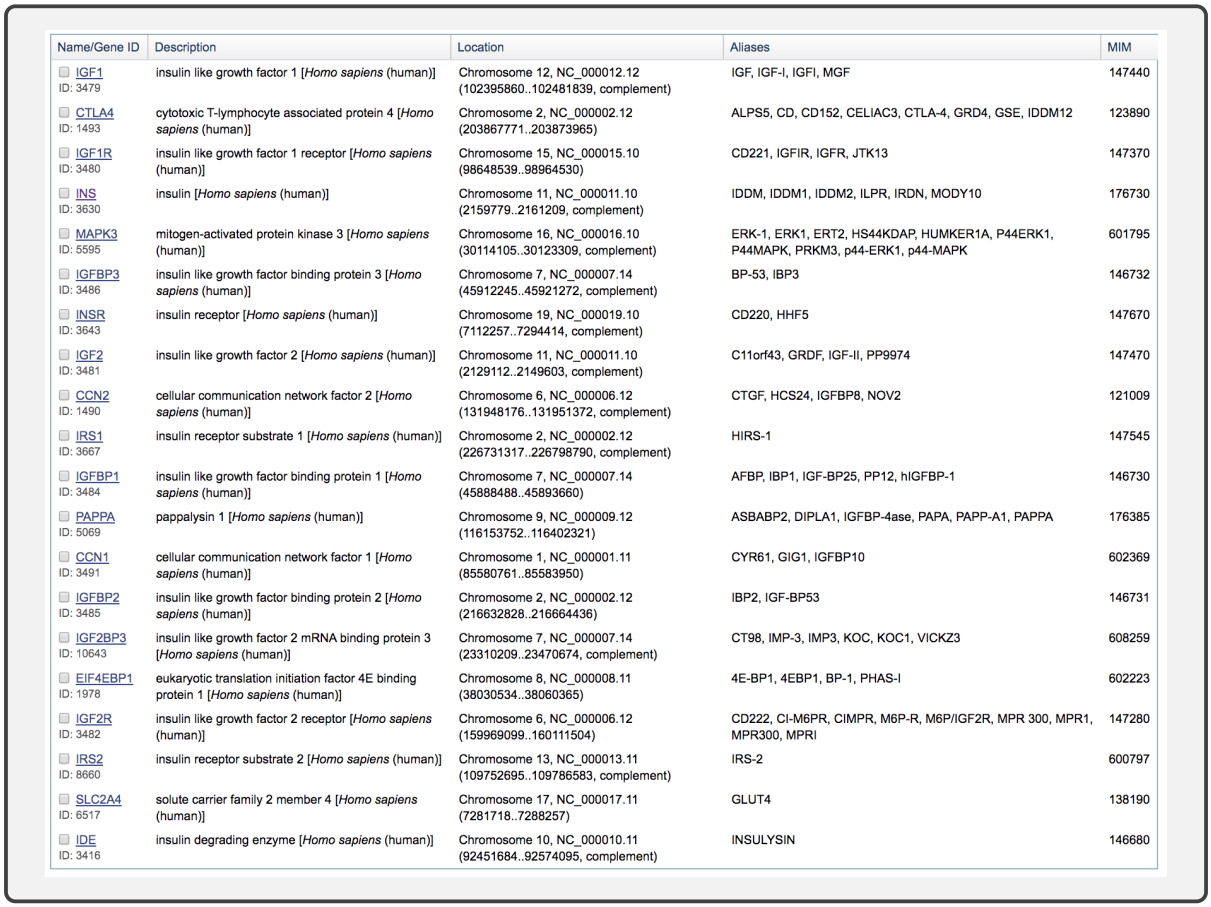
1.1.1 Gene database

Suppose we want to study the human insulin gene. The first step is to find out what is already known about this gene.

Go to: `http://www.ncbi.nlm.nih.gov/`. NCBI is a US-based organisation that maintains several reference databases for biological and molecular data (an overview of all the databases they maintain can be found here: `https://www.ncbi.nlm.nih.gov/guide/all/`). We are interested in the *Gene* database right now, which you can reach by navigating to *Genes & Expression* and then *Gene*.

Underneath the search box you can click `Advanced` to create a custom query. Try to look for the **human** version of the **insulin** gene.

Below you can see a screenshot from the results. The first column gives you the ID of the gene in the database, the second the description of the gene and the third the chromosome and location within the chromosome. Apparently, our search was not very specific and only the 4th record is the one we are looking for. The rest of the terms also contain `"insulin"` and `"Homo sapiens"` like the insulin receptor, insulin like growth factor 1, etc.

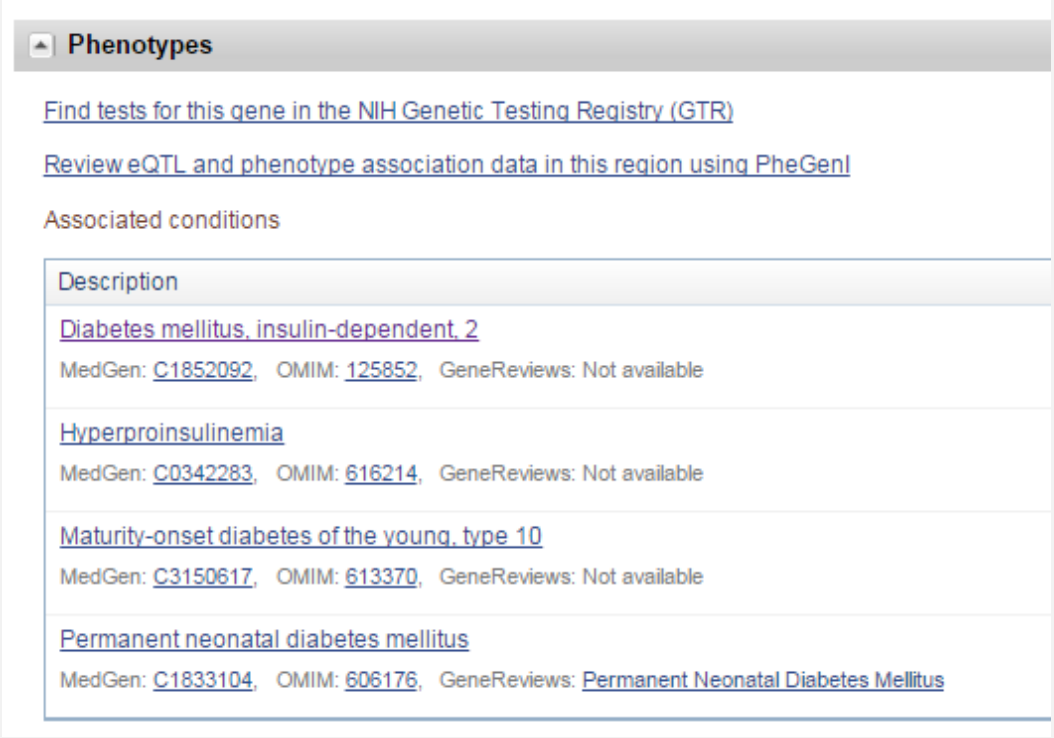


Name/Gene ID	Description	Location	Aliases	MIM
IGF1 ID: 3479	insulin like growth factor 1 [*Homo sapiens* (human)]	Chromosome 12, NC_000012.12 (102395860..102481839, complement)	IGF, IGF-I, IGFI, MGF	147440
CTLA4 ID: 1493	cytotoxic T-lymphocyte associated protein 4 [*Homo sapiens* (human)]	Chromosome 2, NC_000002.12 (203867771..203873965)	ALPS5, CD, CD152, CELIAC3, CTLA-4, GRD4, GSE, IDDM12	123890
IGF1R ID: 3480	insulin like growth factor 1 receptor [*Homo sapiens* (human)]	Chromosome 15, NC_000015.10 (98648539..98964530)	CD221, IGFIR, IGFR, JTK13	147370
INS ID: 3630	insulin [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (2159779..2161209, complement)	IDDM, IDDM1, IDDM2, ILPR, IRDN, MODY10	176730
MAPK3 ID: 5595	mitogen-activated protein kinase 3 [*Homo sapiens* (human)]	Chromosome 16, NC_000016.10 (30114105..30123309, complement)	ERK-1, ERK1, ERT2, HS44KDAP, HUMKER1A, P44ERK1, P44MAPK, PRKM3, p44-ERK1, p44-MAPK	601795
IGFBP3 ID: 3486	insulin like growth factor binding protein 3 [*Homo sapiens* (human)]	Chromosome 7, NC_000007.14 (45912245..45921272, complement)	BP-53, IBP3	146732
INSR ID: 3643	insulin receptor [*Homo sapiens* (human)]	Chromosome 19, NC_000019.10 (7112257..7294414, complement)	CD220, HHF5	147670
IGF2 ID: 3481	insulin like growth factor 2 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (2129112..2149603, complement)	C11orf43, GRDF, IGF-II, PP9974	147470
CCN2 ID: 1490	cellular communication network factor 2 [*Homo sapiens* (human)]	Chromosome 6, NC_000006.12 (131948176..131951372, complement)	CTGF, HCS24, IGFBP8, NOV2	121009
IRS1 ID: 3667	insulin receptor substrate 1 [*Homo sapiens* (human)]	Chromosome 2, NC_000002.12 (226731317..226798790, complement)	HIRS-1	147545
IGFBP1 ID: 3484	insulin like growth factor binding protein 1 [*Homo sapiens* (human)]	Chromosome 7, NC_000007.14 (45888488..45893660)	AFBP, IBP1, IGF-BP25, PP12, hIGFBP-1	146730
PAPPA ID: 5069	pappalysin 1 [*Homo sapiens* (human)]	Chromosome 9, NC_000009.12 (116153752..116402321)	ASBABP2, DIPLA1, IGFBP-4ase, PAPA, PAPP-A1, PAPPA	176385
CCN1 ID: 3491	cellular communication network factor 1 [*Homo sapiens* (human)]	Chromosome 1, NC_000001.11 (85580761..85583950)	CYR61, GIG1, IGFBP10	602369
IGFBP2 ID: 3485	insulin like growth factor binding protein 2 [*Homo sapiens* (human)]	Chromosome 2, NC_000002.12 (216632828..216664436)	IBP2, IGF-BP53	146731
IGF2BP3 ID: 10643	insulin like growth factor 2 mRNA binding protein 3 [*Homo sapiens* (human)]	Chromosome 7, NC_000007.14 (23310209..23470674, complement)	CT98, IMP-3, IMP3, KOC, KOC1, VICKZ3	608259
EIF4EBP1 ID: 1978	eukaryotic translation initiation factor 4E binding protein 1 [*Homo sapiens* (human)]	Chromosome 8, NC_000008.11 (38030534..38060365)	4E-BP1, 4EBP1, BP-1, PHAS-I	602223
IGF2R ID: 3482	insulin like growth factor 2 receptor [*Homo sapiens* (human)]	Chromosome 6, NC_000006.12 (159969099..160111504)	CD222, CI-M6PR, CIMPR, M6P-R, M6P/IGF2R, MPR 300, MPR1, MPR300, MPRI	147280
IRS2 ID: 8660	insulin receptor substrate 2 [*Homo sapiens* (human)]	Chromosome 13, NC_000013.11 (109752695..109786583, complement)	IRS-2	600797
SLC2A4 ID: 6517	solute carrier family 2 member 4 [*Homo sapiens* (human)]	Chromosome 17, NC_000017.11 (7281718..7288257)	GLUT4	138190
IDE ID: 3416	insulin degrading enzyme [*Homo sapiens* (human)]	Chromosome 10, NC_000010.11 (92451684..92574095, complement)	INSULYSIN	146680

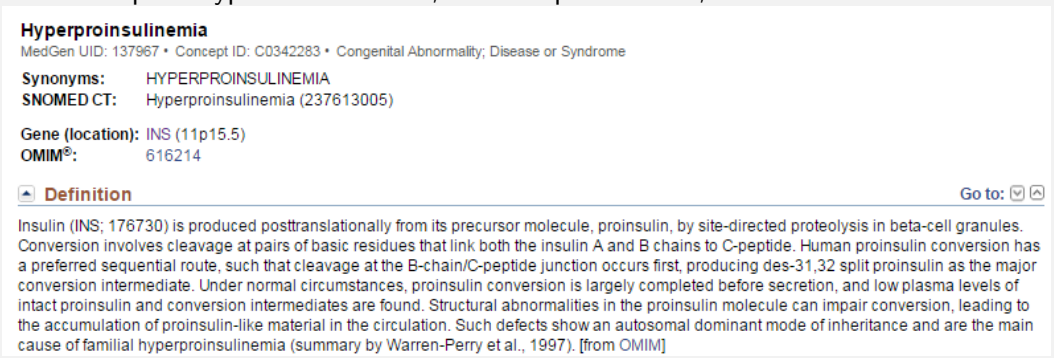
Select insulin out of the search result list and explore the information on the page. Find the answers to following questions:

- Which diseases are associated with problems of insulin metabolism?

This information can be found under `"Phenotypes"` and then `"Associated conditions"`. Not surprisingly, several types of diabetes turn up:

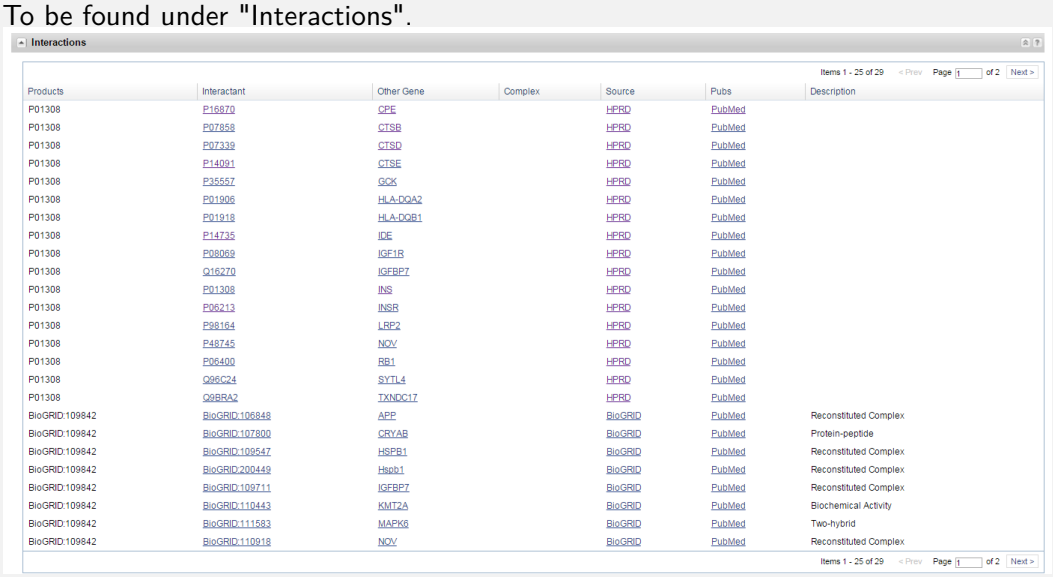


If you want to know more about the associated conditions, you can follow the indicated hyperlinks.
For example, MedGen, another NCBI database will provide you more information about cause and phenotype of the disease, relevant publications, etc.



- Find molecular components that interact with insulin.

To be found under "Interactions".



Interactions

Items 1 - 25 of 29 < Prev Page 1 of 2 Next >

Products	Interactant	Other Gene	Complex	Source	Pubs	Description
P01308	P16870	CPE		HPRD	PubMed	
P01308	P07858	CTSB		HPRD	PubMed	
P01308	P07339	CTSD		HPRD	PubMed	
P01308	P14091	CTSE		HPRD	PubMed	
P01308	P35557	GCK		HPRD	PubMed	
P01308	P01906	HLA-DQA2		HPRD	PubMed	
P01308	P01918	HLA-DQB1		HPRD	PubMed	
P01308	P14735	IDE		HPRD	PubMed	
P01308	P08069	IGF1R		HPRD	PubMed	
P01308	Q16270	IGFBP7		HPRD	PubMed	
P01308	P01308	INS		HPRD	PubMed	
P01308	P06213	INSR		HPRD	PubMed	
P01308	P98164	LRP2		HPRD	PubMed	
P01308	P48745	NOV		HPRD	PubMed	
P01308	P06400	RB1		HPRD	PubMed	
P01308	Q96C24	SYTL4		HPRD	PubMed	
P01308	Q9BRA2	TXNDC17		HPRD	PubMed	
BioGRID:109842	BioGRID:106848	APP		BioGRID	PubMed	Reconstituted Complex
BioGRID:109842	BioGRID:107800	CRYAB		BioGRID	PubMed	Protein-peptide
BioGRID:109842	BioGRID:109547	HSPB1		BioGRID	PubMed	Reconstituted Complex
BioGRID:109842	BioGRID:200449	Hspb1		BioGRID	PubMed	Reconstituted Complex
BioGRID:109842	BioGRID:109711	IGFBP7		BioGRID	PubMed	Reconstituted Complex
BioGRID:109842	BioGRID:110443	KMT2A		BioGRID	PubMed	Biochemical Activity
BioGRID:109842	BioGRID:111583	MAPK6		BioGRID	PubMed	Two-hybrid
BioGRID:109842	BioGRID:110918	NOV		BioGRID	PubMed	Reconstituted Complex

Items 1 - 25 of 29 < Prev Page 1 of 2 Next >

The column `"interactant"` gives you the link to the protein (NCBI Protein database) and `"other gene"` to the corresponding gene (NCBI Gene database). The source database is mentioned in the `"source column"` (NCBI summarizes the results from several databases) and `"Pubs"` gives you the link to the relevant publication to support this interaction with evidence.

- Gene Ontology: Take a look at the 3 gene ontology classes and the associated terms for insulin. What can this tell you about the function of insulin?

There are three types of gene ontology: cellular component, molecular function and biological process. All three can be found under the section "General gene information". For more information about gene ontology we refer to the third practicum ("Phylogenetic trees, protein structure & gene ontology").

1.1.2 GenBank

Look for the same gene in GenBank, an annotated collection of all publicly available DNA sequences. You can do this by searching for "insulin" using the same search box as before, but this time selecting `"Nucleotide"` instead of `"Gene"` from the dropdown in front of it. The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq[1] and PDB.

GenBank provides you with the most crucial information in a very simple text format. Therefore, it is easy to copy and paste from this page to use the acquired information in other tools.

Some general information about GenBank entries can be found here: `https://www.ncbi.nlm.nih.gov/genbank/`

Answer the following questions by browsing the database results:

[1] RefSeq is a curated version of GenBank. Researchers can upload their nucleotide sequencing data to GenBank. Starting from this information, the NCBI then creates a single database record for each molecule of each major organism. They revise this sequence as new information emerges.

- What is the accession number of insulin. (And what is an accession number?!) What is the point of this identifier? What would happen if a new (or corrected) version of the same sequence is added to the database?

  `AH002844 J00265 J00268`
  An accession number is a unique identifier given to a database record. It is crucial to mention this whenever you use the corresponding database information. For example, different versions of genes can exist, or they can become updated over time. Therefore, when you write a paper and you mention something about say position 50 in a specific gene, it is crucial to report also the accession number, otherwise position 50 might refer to a different location in the sequence than you intended.
  More information about the different type of identifiers used by NCBI can be found here: `https://www.ncbi.nlm.nih.gov/genbank/sequenceids/`
  Do not worry if you wrote down another accession number. It could be that you ended up with another variant/version/database entry of insulin. The version we used, you can find here: `https://www.ncbi.nlm.nih.gov/nuccore/AH002844.2`

- What is the version of the gene assembly?

  `AH002844.2` In case of GenBank, genes can also be updated WITHOUT changing the accession number. However, the version number does receive an increment of 1 then (*accession dot version*). Therefore, when reporting about a gene taken from GenBank, it is important to mention both the accession and version number.

- Who put this information on Genbank, and can you find out something more about the research that was done to produce this data?

  See `REFERENCE` blocks: these contain information about published research pertaining to the sequence.

- Find the FASTA sequence of the gene. What is a FASTA file and how is it structured?

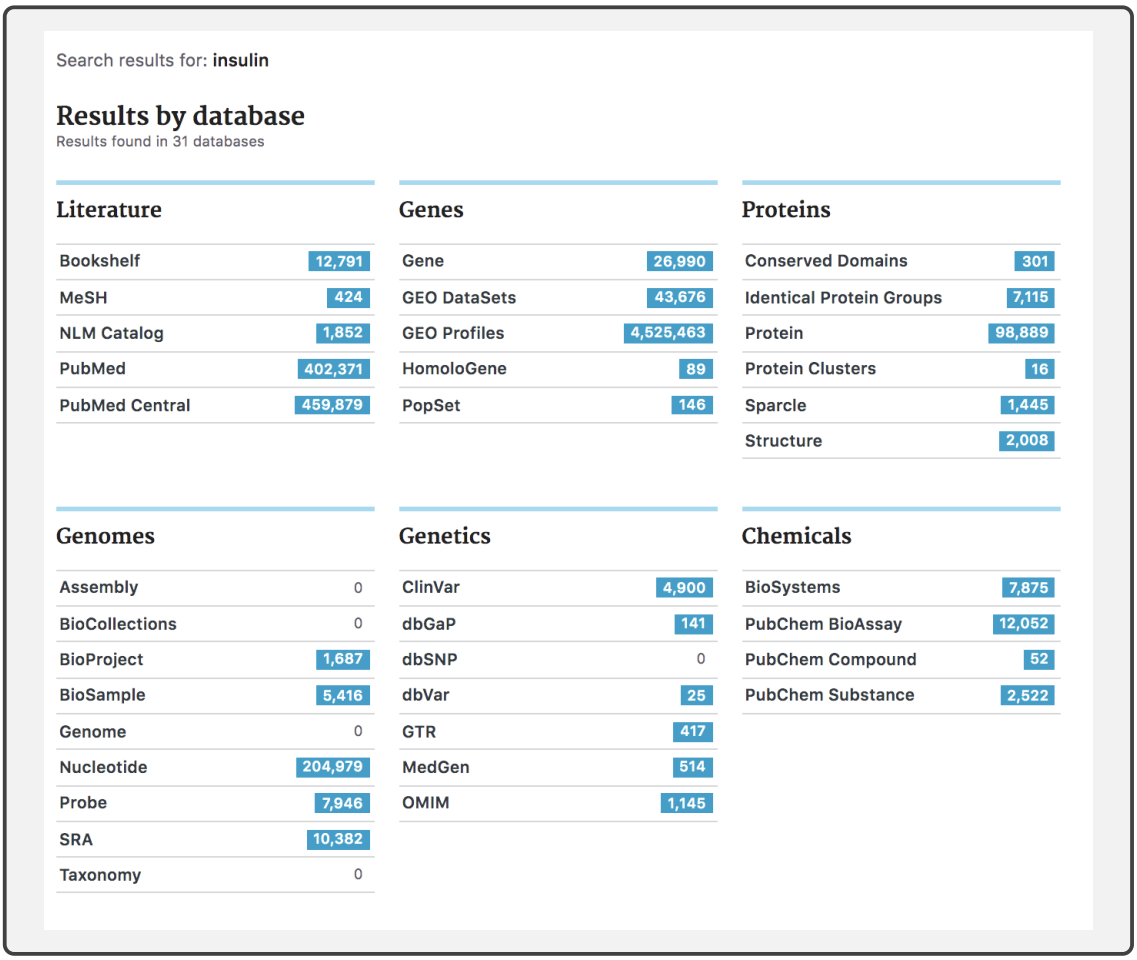
A FASTA file is a simple text file with following properties: The first line for each gene contains "> NAME_OF_THE_ENTITY (SPACE+OPTIONAL DESCRIPTION)" and is followed by (on a new line!) the nucleotides for a gene or transcript, and the amino acids for a protein (both in 1 symbol abbreviations). The nucleotide sequence can be spread over several (or hundreds of) lines.
Example:

```
>Gene_1 Very descriptive description
TATATATATGGTAGTAGGTAAAT
>Gene_2 This gene has some unknown bases in the middle
TACGATCGTCGCTAGCTAGCTAC
TGATGTCTNNNNGATCGATGCTT
CTAGCTAGCTAGCTGATCGATGC
>Gene_3
TATAGTCTGCTGCTGATCGACGA
TGCTAGTCGTGATCGATGCTAGC
```

1.1.3 All databases

Finally, search for insulin in all NCBI databases:

- What other types of additional information can you find here?



1.1.4 dbSNP

dbSNP is world's largest database for nucleotide variations, and is part of the National Center for Biotechnology Information (NCBI). dbSNP is comprised of a large cluster of species-specific databases that contain over 12 million non-redundant sequence variations (single nucleotide polymorphisms (SNPs[2], insertion/deletions, and short tandem repeats) and over 1 billion individual genotypes from HapMap and other large-scale genotyping activities.

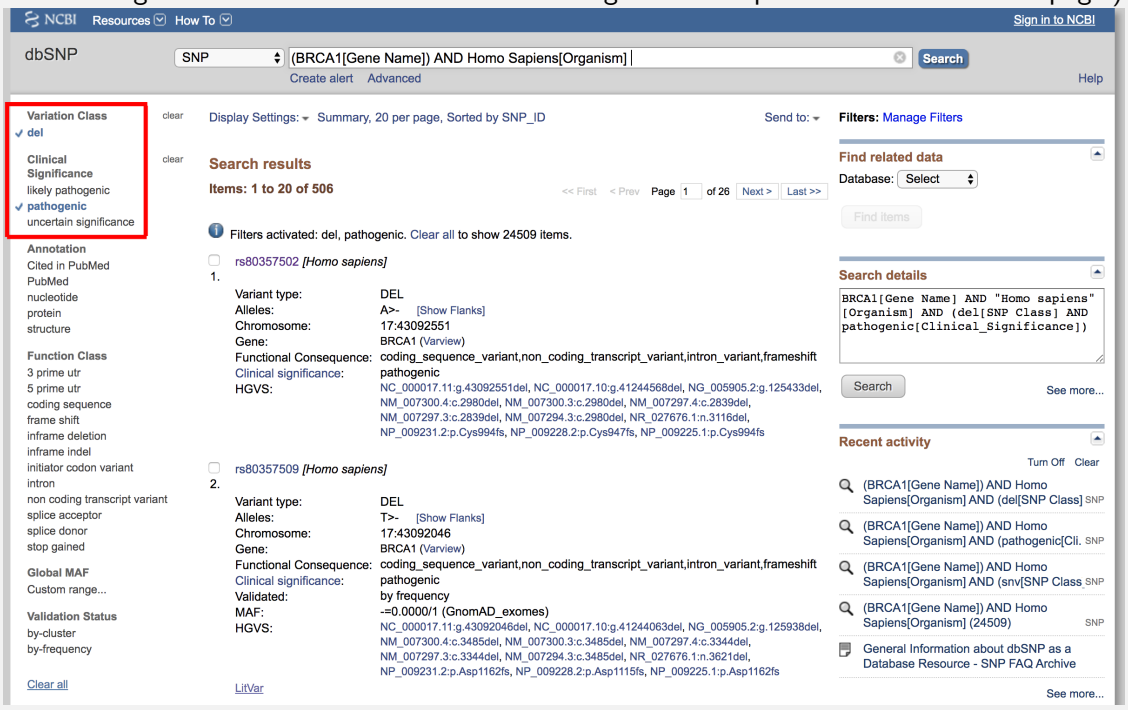
In this exercise, we will study the human BRCA1 gene[3]:

- How many deletions for this gene can you find in dbSNP that are pathogenic?
- Take a closer look at one of the selected deletions to see what kind of information is stored in dbSNP: what kind of effect does the deletion have, and what disease is this SNP associated with?

[2]A single nucleotide polymorphism or SNP (pronounced [snip]) is a variation at a single position in a DNA sequence among individuals.

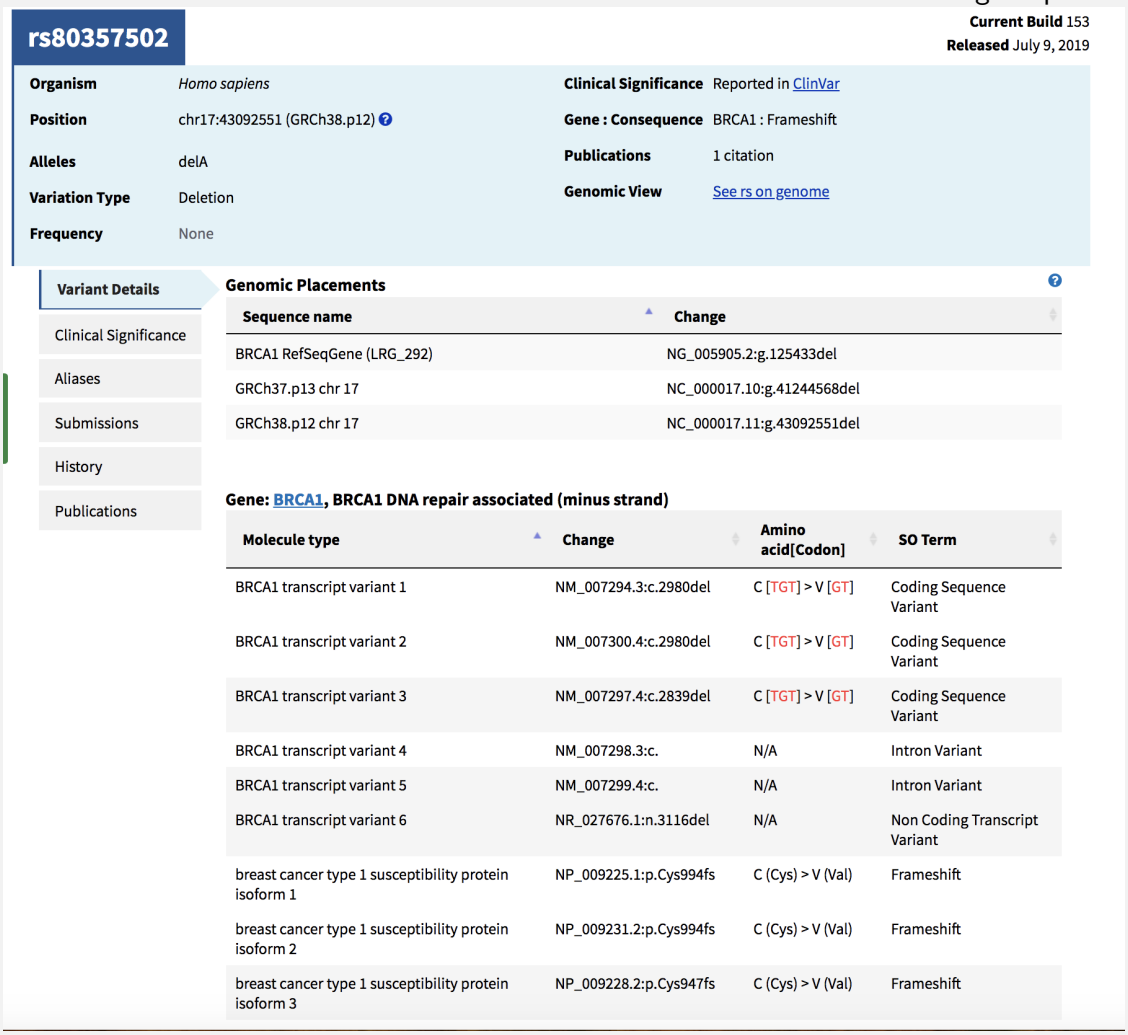
[3]BRCA1 is one of the 2 BRCA genes, which are tumor suppressor genes involved in DNA repair. BRCA is pronounced as [bra-ka].

How many deletions for this gene can you find in dbSNP that are pathogenic?
With the advanced search option, you can easily search for all genetic variations in the human BRCA1 gene. Use the filter options to retrieve only the pathogenic deletions.(This can be done through the advanced search or later on using the filter options listed on the web page.)



Take a closer look at one of the selected deletions to see what kind of information is stored in dbSNP.

The beginning of the page gives an overview on the position, alleles, variation type, frequency and consequence of the genetic variations. Underneath, you get a detailed overview of the codon and amino acid changes. According to this table the deletion results in a frameshift when the deletion occurs in the coding sequence.



rs80357502

Current Build 153
Released July 9, 2019

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr17:43092551 (GRCh38.p12)	**Gene : Consequence**	BRCA1 : Frameshift
Alleles	delA	**Publications**	1 citation
Variation Type	Deletion	**Genomic View**	See rs on genome
Frequency	None		

Variant Details
Clinical Significance
Aliases
Submissions
History
Publications

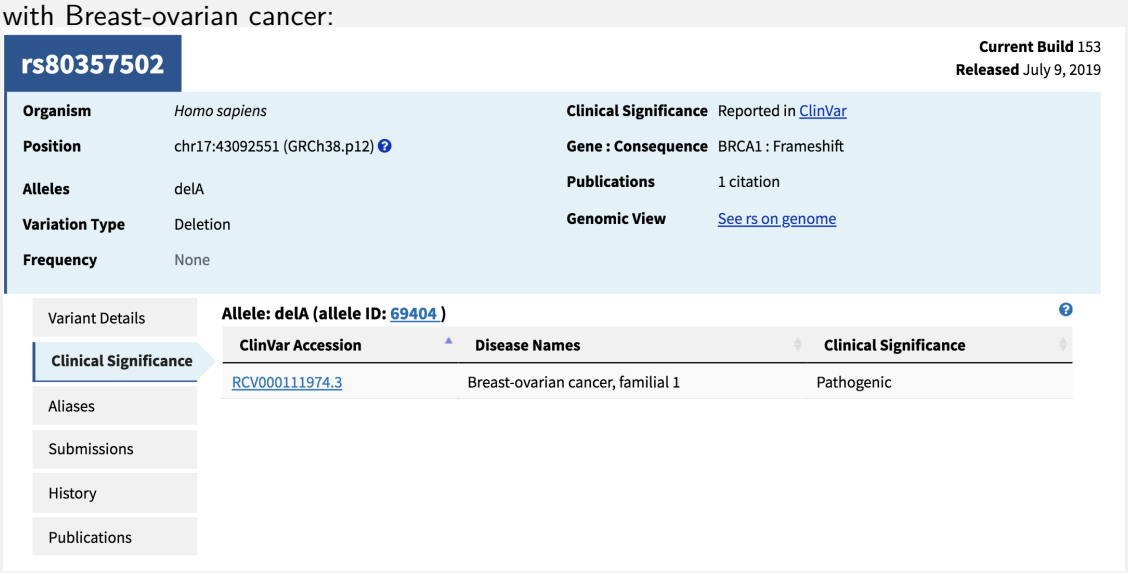
Genomic Placements

Sequence name	Change
BRCA1 RefSeqGene (LRG_292)	NG_005905.2:g.125433del
GRCh37.p13 chr 17	NC_000017.10:g.41244568del
GRCh38.p12 chr 17	NC_000017.11:g.43092551del

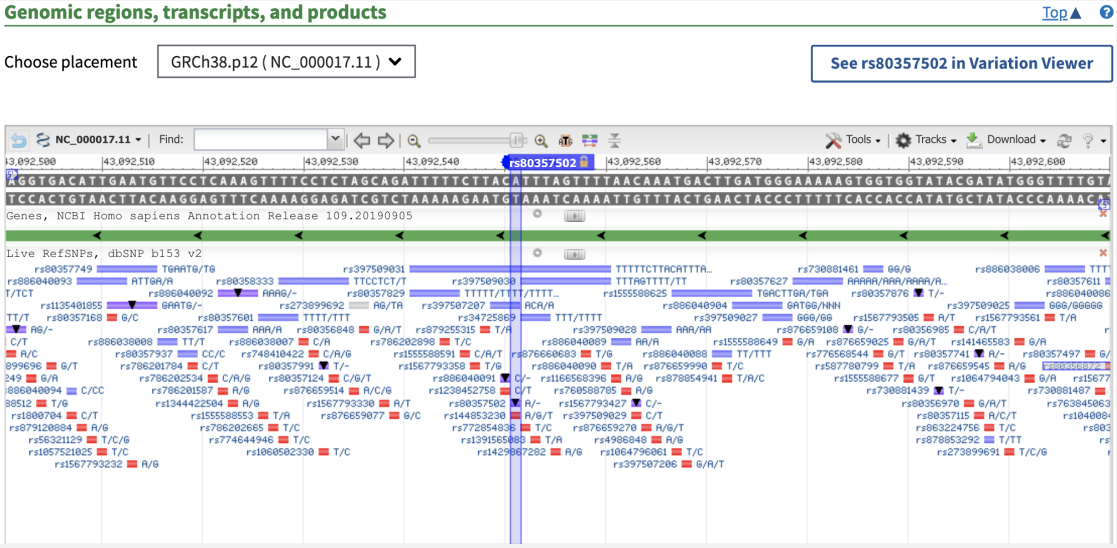
Gene: BRCA1, BRCA1 DNA repair associated (minus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
BRCA1 transcript variant 1	NM_007294.3:c.2980del	C [TGT] > V [GT]	Coding Sequence Variant
BRCA1 transcript variant 2	NM_007300.4:c.2980del	C [TGT] > V [GT]	Coding Sequence Variant
BRCA1 transcript variant 3	NM_007297.4:c.2839del	C [TGT] > V [GT]	Coding Sequence Variant
BRCA1 transcript variant 4	NM_007298.3:c.	N/A	Intron Variant
BRCA1 transcript variant 5	NM_007299.4:c.	N/A	Intron Variant
BRCA1 transcript variant 6	NR_027676.1:n.3116del	N/A	Non Coding Transcript Variant
breast cancer type 1 susceptibility protein isoform 1	NP_009225.1:p.Cys994fs	C (Cys) > V (Val)	Frameshift
breast cancer type 1 susceptibility protein isoform 2	NP_009231.2:p.Cys994fs	C (Cys) > V (Val)	Frameshift
breast cancer type 1 susceptibility protein isoform 3	NP_009228.2:p.Cys947fs	C (Cys) > V (Val)	Frameshift

Click on 'Clinical significance' to find associations with diseases. This deletion is associated with Breast-ovarian cancer:



At the bottom of the web page you get an overview of the genomic location of the genetic variant. Here, you can see that the deletion is located on the antisense strand.



Exercise 1.2: UniProt

UniProt is one of the main protein databases and is a combination of the previous Swiss-Prot and TrEMBL (Translated EMBL Nucleotide Sequence Data Library) databases. While Swiss-Prot only contained manually curated data of very high quality (i.e. experimentally proven or computer-predicted data that has been reviewed, verified, catalogued or updated by humans), it was impossible to keep up with the high throughput methods like whole genome sequencing and proteomics. Consequently, automated (lower quality) protein annotation is provided by TrEMBL. When returning search results from UniProt, the results page still enables to filter between the curated (Swiss-Prot) and non-curated

results (TrEMBL).

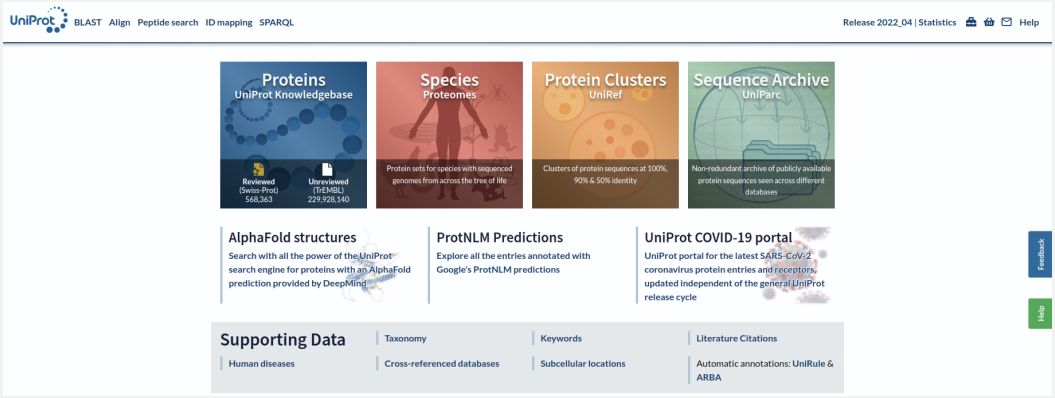
- Navigate to `http://www.uniprot.org/` and search again for human insulin.

  What additional information can you find here compared to Genbank?

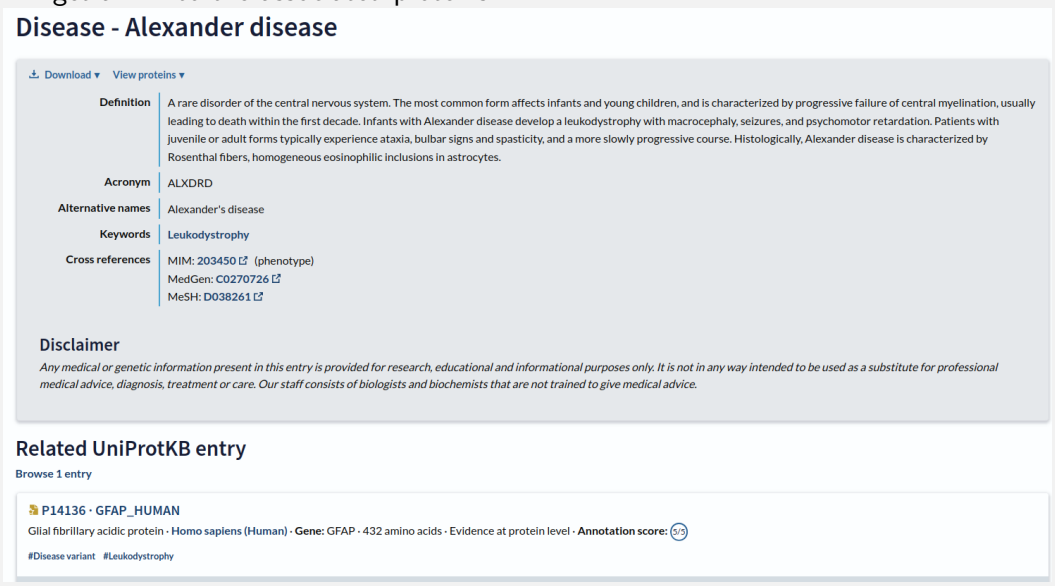
  > UniProt entry: `http://www.uniprot.org/uniprot/P01308`
  > The information is (obviously) more protein oriented. There is a lot of information about post-translational modifications and protein structure available.

- Try to find out which protein is associated with Alexander disease.
  HINT: UniProt has a separate section for human diseases.

On the UniProt home page, select `"Human diseases"` under the `"Supporting Data"` title. Type in `"Alexander disease"` and select the result.



You will get some information about the disease, but on the bottom of the page you will get a link to the associated proteins.



In this case, there was 1 associated protein: the glial fibrillary acidic protein

- UniProt isn't only able to return single protein entries, you can also look up whole proteomes if you want. Search for the proteome of the zebrafish (Danio rerio) or your favorite model organism. How many proteins does it contain? How many of those are reviewed?

`http://www.uniprot.org/proteomes/UP000000437` Click on the Protein count under Overview to get a list of all proteins. On the left you see that the reference proteome of the zebrafish on UniProt contains 46 691 proteins, of which 3 248 are reviewed.

Section 2 – Pairwise alignment

Learning goals

At the end of this section, you should:

- be able to construct and interpret dot plots
- understand the algorithms of pairwise alignment and the influence of the different parameters
- understand the fundamental differences between the pairwise alignment methods
- perform and interpret pairwise alignments (with online tools and manually)

Exercise 2.1: Getting familiar with dot plots

Dot plots are a graphical representation of sequence identity (i.e. similarity). They show all the possible comparisons that can be made between two sequences by sliding one over the other and checking if two characters (or multiple characters for word sizes > 1) match. Whenever there is a match (or hit), a dot is drawn on the corresponding position of the sequences where the x- and y-axes represent the index of the characters (or words) in each sequence.

To get familiar with the dot matrix method, we will make a couple dot plots that allow you to interpret the (dis)similarity between two sequences. Visualize the alignment of the sequences using the emboss dottup tool from `https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_dottup`. Try to explain what is happening in the alignments based on what you have been taught in past genetics courses. **Use a wordsize of 5 for all plots.**

- Align following sequences to each other:

  ```
  >sequence_1
  MSETAPAAPASAPAPAEKTPVKKKARKSAGAAKRKASGPPVSELITKAVAASKERSKKALAAAGYDGVSLAAL
  >sequence_2
  MSETAPAAPKLIASSKPLPPVKKKARKSAGAAKRKAAASWTRRPLASTVNWSKERSGVSLAAL
  ```

- Align following sequences to each other:

  ```
  >sequence_1
  MSETAPAAPASAPAPAEKTPVKKKARKSAGAAKRKASGPPVSELITKAVAASKERSGVSLAALKKALAAAGYD
  >sequence_2
  MSETAPAAKSAGAAKRKASGPPVSELITKAVAASKERSGVSLPASAPAPAEKTPVKKKARAALKKALAAAGYD
  ```

- Align following sequence to itself (put it in both input fields):

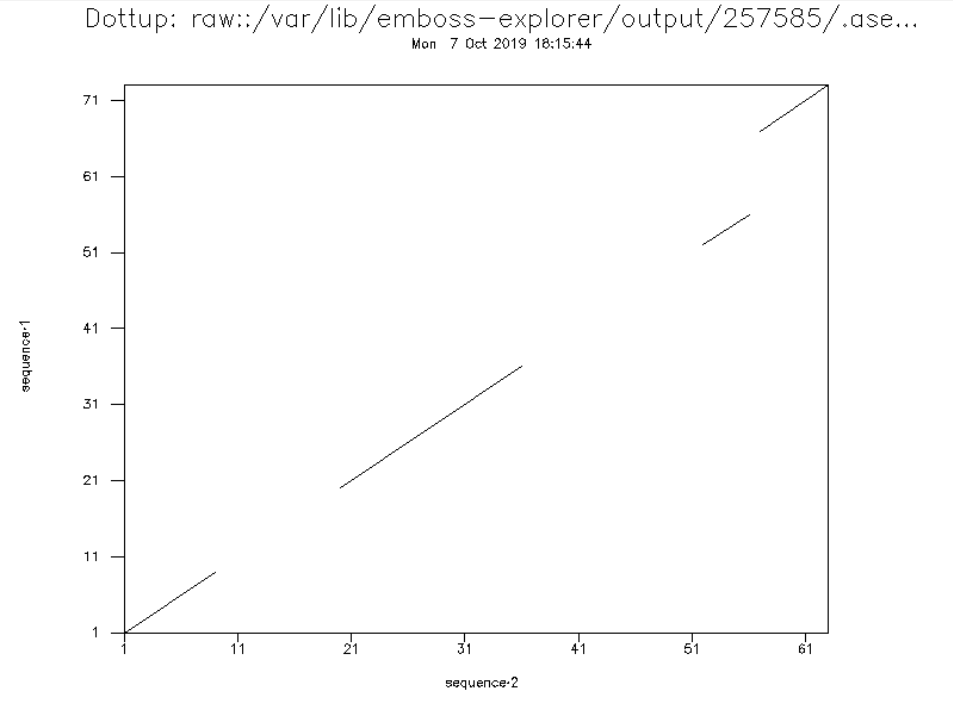
  ```
  >sequence_1
  MSETAPAAPASMSETAPAARMSETAPKSAGAMSETAPAKRKASMSETAPGPPVSMSETAPELITMSETAP
  ```

- Align following sequence to itself:

  ```
  >sequence_1
  ASAPAPAEKTPVKKKAKAKAKAKAKAKAKAKAKAKAKAKARKSAGAAKRKASGPPVSELITKAVAASKERSGVSLA
  ```

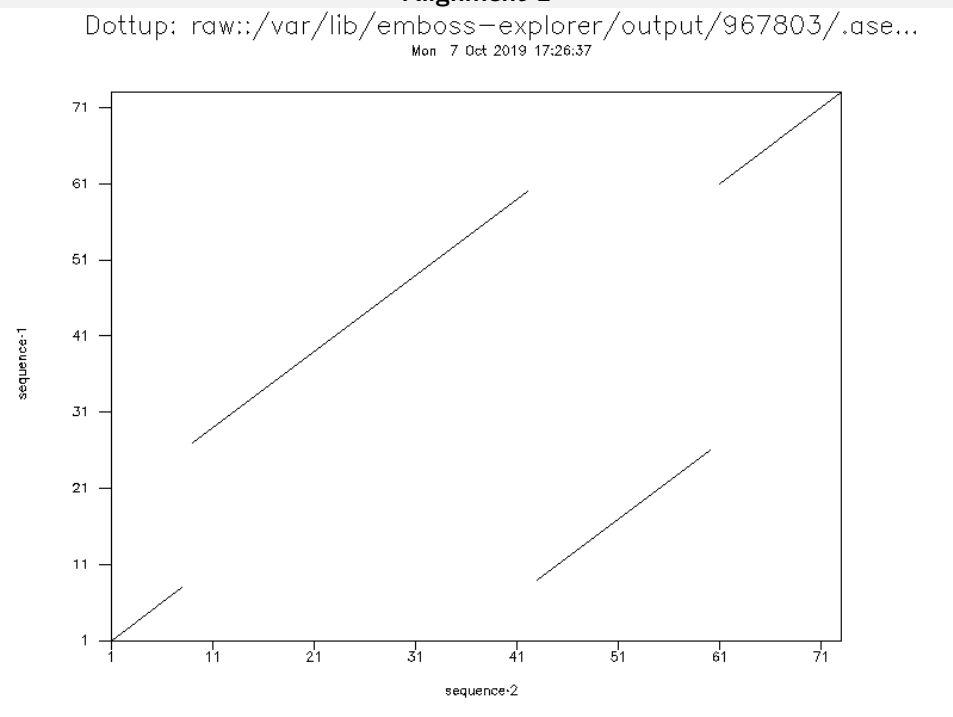
- One of the limitations of the Emboss dot plot tool is the inability to show is an *inversion*. What do you think an inversion looks like on a dot plot? Try to draw one (pen and paper) using sequences and parameters of your choice.

Alignment 1



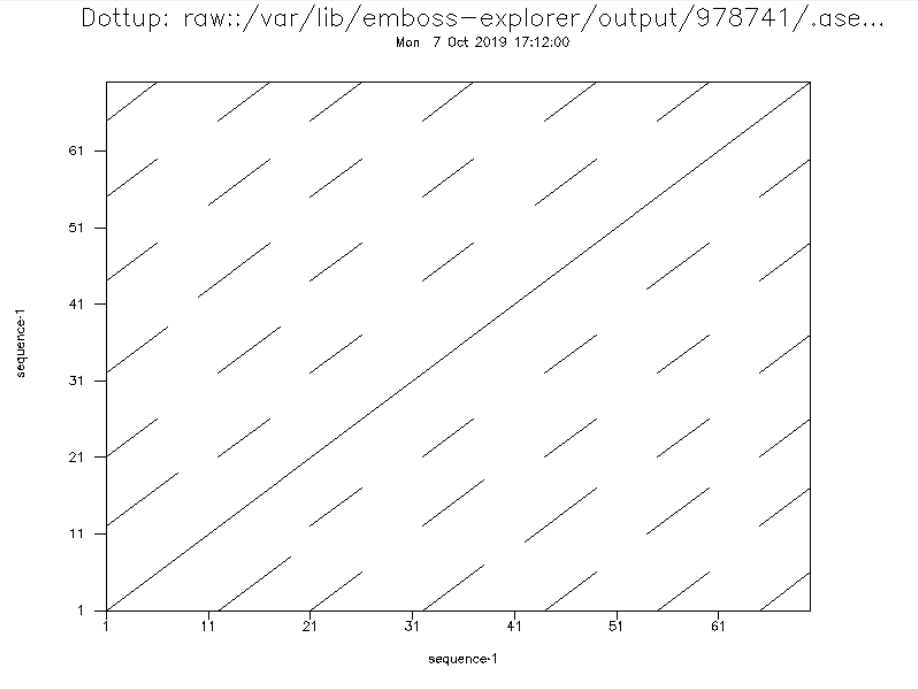
This is an example of a very simple alignment. The diagonal parts indicate that the sequences match in those regions. The gaps indicates that the sequences do not match in those regions. At the end of the alignment, the vertical movement indicates either a *deletion* in sequence 2 or an *insertion* in sequence 1. In any case, sequence 1 contains an additional piece, resulting in a vertical 'hole' in the dot plot.

Alignment 2



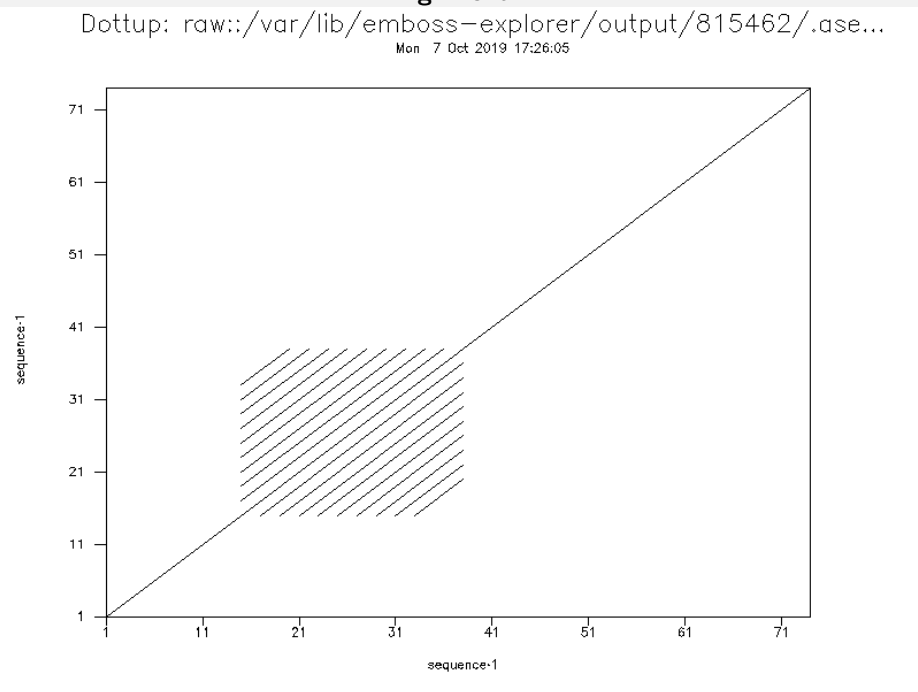
In this dot plot something weird is going on. Part of sequence 1 matches with sequence 2 on a completely different location. Other than this, the sequences are identical. This indicates a *translocation*. You can interpret this either as a part from sequence 1 that translocated downstream in the sequence, or a part of sequence 2 that translocated upstream.

Alignment 3

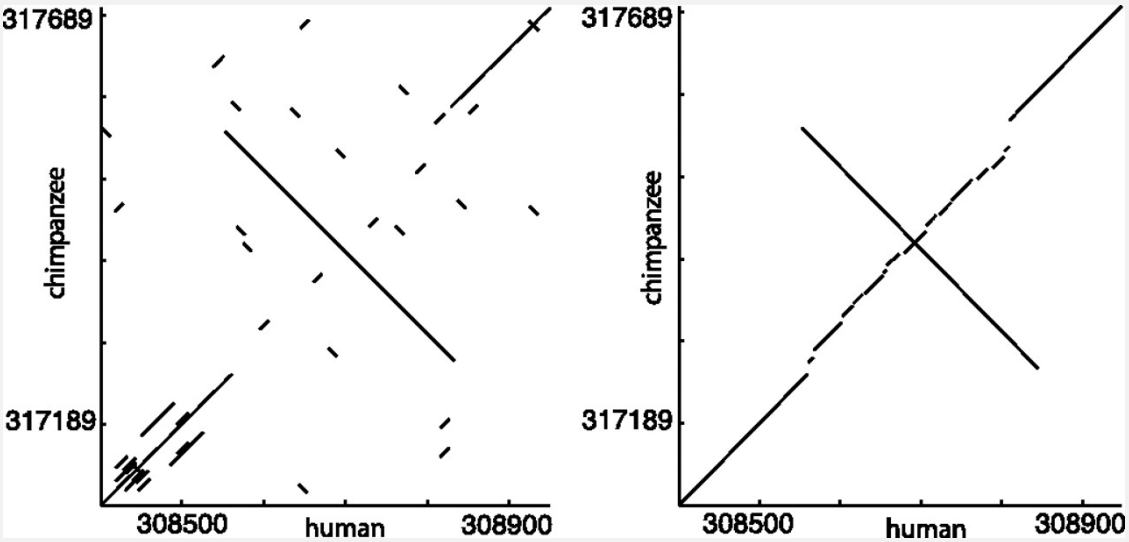


When we align the sequence to itself we see a lot of parallel diagonal lines. Firstly, the middle diagonal, starting at the bottom left corner all the way up to the top right corner, indicates that we are aligning identical sequences. The other diagonals, spread out in the dot plot, indicate that our sequence contains one or more repeats. These repeats match multiple times, resulting in a bunch of smaller lines.

Alignment 4



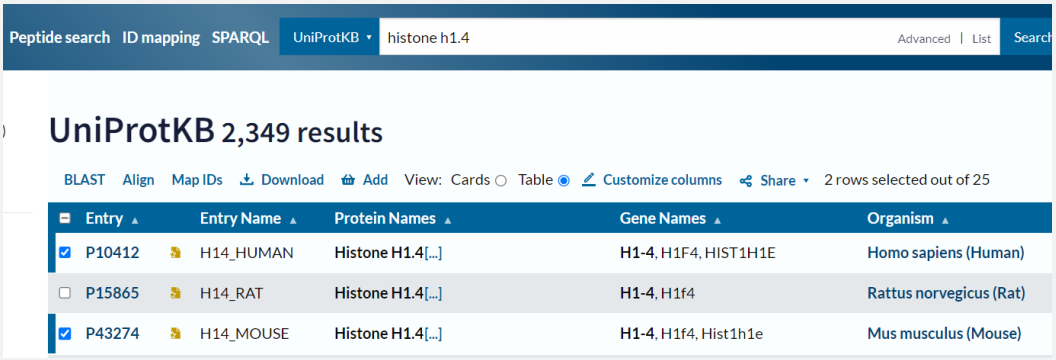
Like in the previous alignment, you would think that this sequence would contain a repeat, resulting in the pattern seen. This is partly true. While it could represent a repeat, we are actually looking at a *low complexity region*. This is the result of redundancy in the sequence, i.e. a region where we see a lot of the same amino acid (short) patterns occurring next to each other. As mentioned, you can look at it as a repeat occurring multiple times in a row.



Example of a dot plot containing inversion from a publication:
Microinversions in mammalian evolution by Chaisson et al.
`https://doi.org/10.1073/pnas.0603984103`
Inversions result in perpendicular diagonals.

Exercise 2.2: Studying the influence of the word size on dot plots

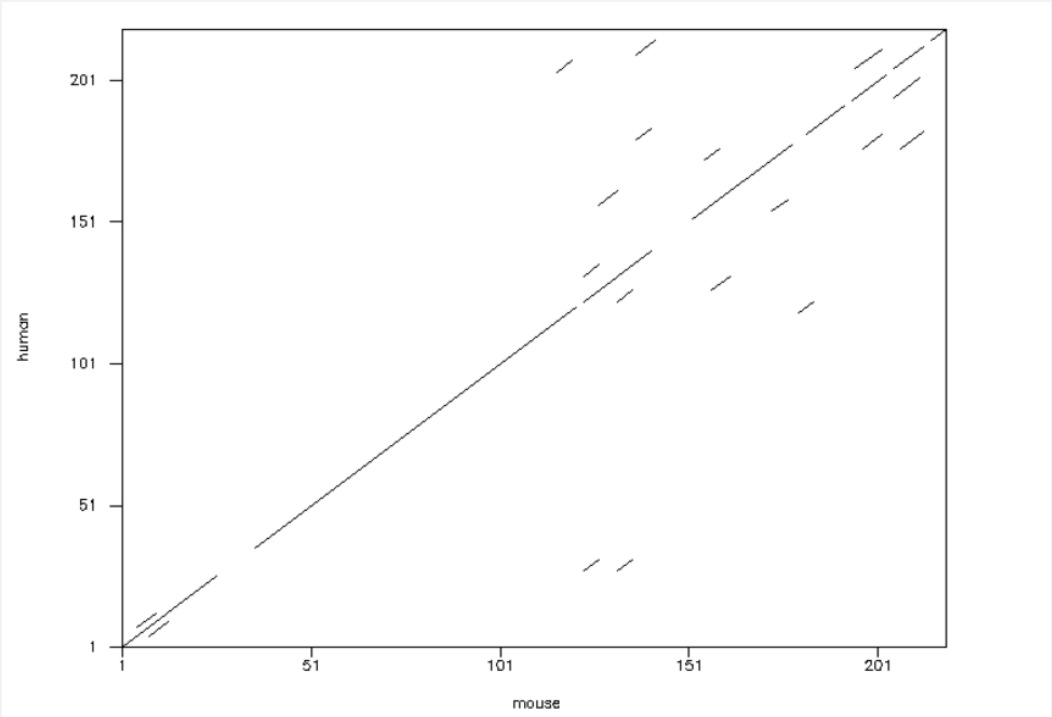
- From the UniProt website, retrieve the protein sequence from histone H1.4 from human and mouse in `FASTA` format.

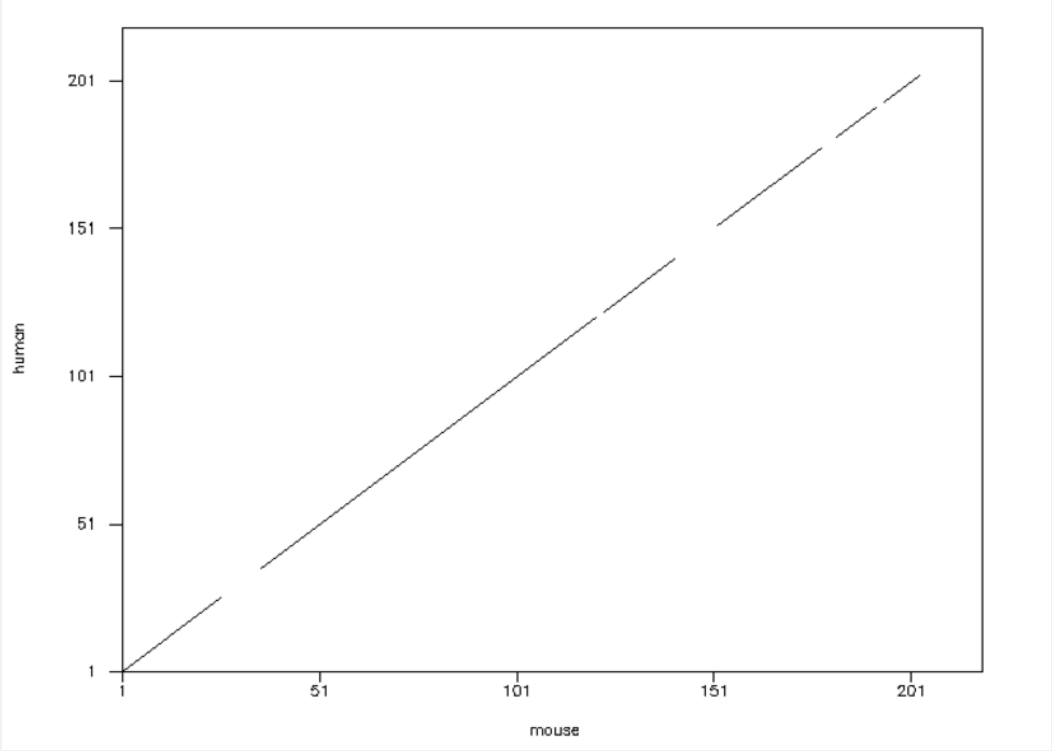


You should get something like the screenshot above. You can just select the database hits you need and click the download button. You should get following FASTA sequences, which you can just copy paste for the next exercise.

```
>sp|P10412|H14_HUMAN Histone H1.4 OS=Homo sapiens GN=HIST1H1E PE=1 SV=2
MSETAPAAPAAPAPAEKTPVKKKARKSAGAAKRKASGPPVSELITKAVAASKERSGVSLA
ALKKALAAAGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKLNKKAASGEAKPKA
KKAGAAKAKKPAGAAKKPKKATGAATPKKSAKKTPKKAKKPAAAAGAKKAKSPKKAKAAK
PKKAPKSPAKAKAVKPKAAKPKTAKPKAAKPKKAAAKKK

>sp|P43274|H14_MOUSE Histone H1.4 OS=Mus musculus GN=Hist1h1e PE=1 SV=2
MSETAPAAPAAPAPAEKTPVKKKARKAAGGAKRKTSGPPVSELITKAVAASKERSGVSLA
ALKKALAAAGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKLNKKAASGEAKPKA
KRAGAAKAKKPAGAAKKPKKAAGTATAKKSTKKTPKKAKKPAAAAGAKKAKSPKKAKATK
AKKAPKSPAKAKTVKPKAAKPKTSKPKAAKPKKTAAKKK
```

- Use the emboss dottup tool from `https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_dottup` to construct a dot plot. Do you see similarities between both sequences?
- Vary the word size from 5 to 10. What happens and why?



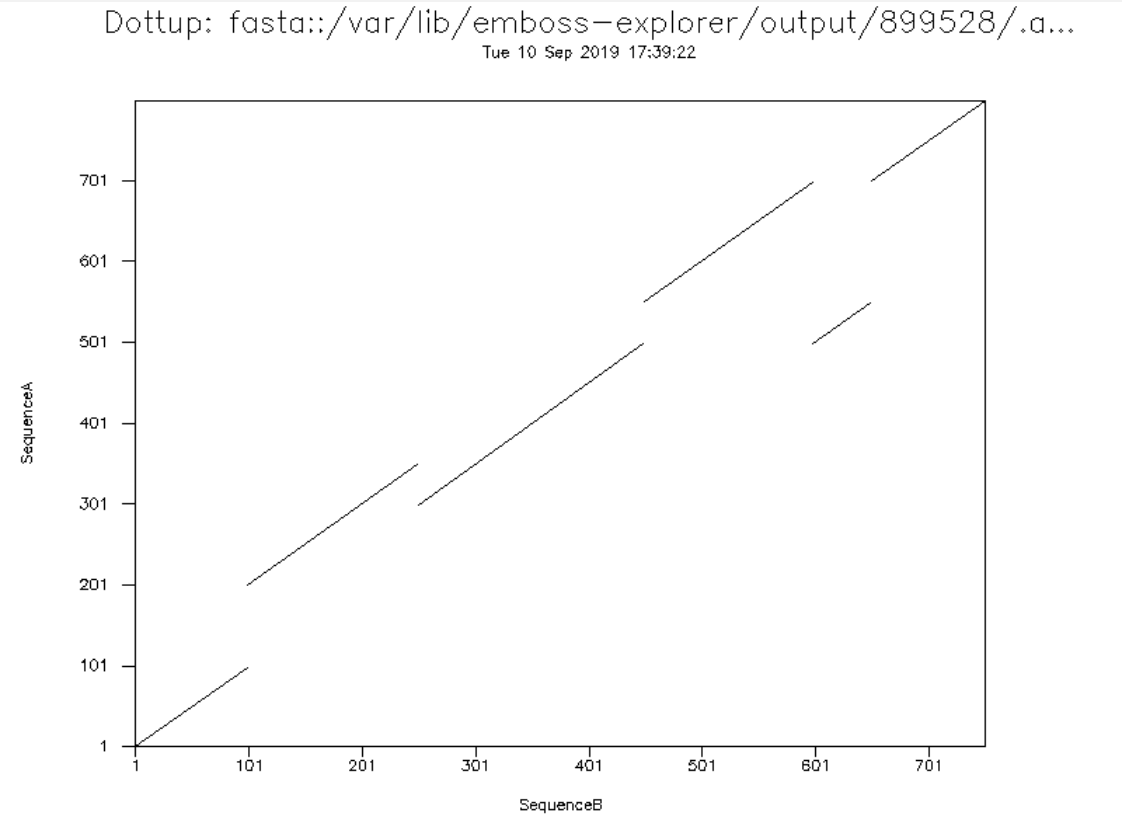


The word size determines how long an identical stretch of amino acids or nucleotides has to be in order to generate a dot on the diagram. Naturally, longer perfect matches are rarer than short ones, but short hits might just be noise. Word size 10 clearly shows the sequences are very similar, since there is a lot of identity. Most regions are very conserved with some very local exceptions (where the line is interrupted). These are called mismatches. Reducing the word size makes the data more noisy because some amino acid combinations seem to occur quite often in the genes.

Exercise 2.3: Read a dot plot

Of the following sequences, sequence B evolved from sequence A. Draw the dot plot of sequences A and B (wordsize 10) and based on the dot plot describe what happened to sequence A to get to sequence B.

```
>sequence A
MSSLGASFVQIKFDDLQFFENCGGGSFGSVYRAKWISQDKEVAVKKLLKIEKEAEILSVL
SHRNIIQFYGVILEPPNYGIVTEYASLGSLYDYINSNRSEEMDMDHIMTWATDVAKGMHY
LHMEAPVKVIHRDLKSRNVVIAADGVLKICDFGASRFHNHTTHMSLVGTFPWMAPEVIQS
LPVSETCDTYSYGVVLWEMLTREVPFKGLEGLQVAWLVVEKNERLTIPSSCPRSFAELLH
QCWEADAKKRPSFKQIISILESMSNDTSLPDKCNSFLHNKAEWRCEIEATLERLKKLERD
LSFKEQELKERERRLKMWEQKLTEQSNTPLLPSFEIGAWTEDDVYCWVQQLVRKGDSSAE
MSVYASLFKENNITGKRLLLLEEEDLKDMGIVSKGHIIHFKSAIEKLTHDYINLFHFPPL
IKDSGGEPEENEEKIVNLELVFGFHLKPGTGPQDCKWKMYMEMDGDEIAITYIKDVTFNT
NLPDAEILKMTKPPFVMEKWIVGIAKSQTVECTVTYESDVRTPKSTKHVHSIQWSRTKPQ
DEVKAVQLAIQTLFTNSDGNPGSRSDSSADCQWLDTLRMRQIASNTSLQRSQSNPILGSP
FFSHFDGQDSYAAAVRRPQVPIKYQQITPVNQSRSSSPTQYGLTKNFSSLHLNSRDSGFS
SGNTDTSSERGRYSDRSRNKYGRGSISLNSSPRGRYSGKSQHSTPSRGRYPGKFYRVSQS
ALNPHQSPDFKRSPRDLHQPNTIPGMPLHPETDSRASEEDSKVSEGGWTKVEYRKKPHRP
SPAKTNKERARGDHRGWRNF

>sequence B
MSSLGASFVQIKFDDLQFFENCGGGSFGSVYRAKWISQDKEVAVKKLLKIEKEAEILSVL
SHRNIIQFYGVILEPPNYGIVTEYASLGSLYDYINSNRSTREVPFKGLEGLQVAWLVVEK
NERLTIPSSCPRSFAELLHQCWEADAKKRPSFKQIISILESMSNDTSLPDKCNSFLHNKA
EWRCEIEATLERLKKLERDLSFKEQELKERERRLKMWEQKLTEQSNTPLLPSFEIGAWTE
DDVYCWVQQDLSFKEQELKERERRLKMWEQKLTEQSNTPLLPSFEIGAWTEDDVYCWVQQ
LVRKGDSSAEMSVYASLFKENNITGKRLLLLEEEDLKDMGIVSKGHIIHFKSAIEKLTHD
YINLFHFPPLIKDSGGEPEENEEKIVNLELVFGFHLKPGTGPQDCKWKMYMEMDGDEIAI
TYIKDVTFNTNLPDAEILKMTKPPFVMEKQTLFTNSDGNPGSRSDSSADCQWLDTLRMRQ
IASNTSLQRSQSNPILGSPFFSHFDGQDSYAAAVRRPQVPIKYQQITPVNQSRSSSPTQY
GLTKNFSSLHLNSRDSGFSSGNTDTSSERGRYSDRSRNKYGRGSISLNSSPRGRYSGKWI
VGIAKSQTVECTVTYESDVRTPKSTKHVHSIQWSRTKPQDEVKAVQLAISQHSTPSRGRY
PGKFYRVSQSALNPHQSPDFKRSPRDLHQPNTIPGMPLHPETDSRASEEDSKVSEGGWTK
VEYRKKPHRPSPAKTNKERARGDHRGWRNF
```



The dotplot figure shows the following:

- Sequence B does not have the amino acids that appear in region 100-200 in sequence A. Hence, this region was deleted in sequence B.
- The amino acids in region 300-350 in sequence A appears two consecutive times in sequence B (at regions 200-250 and 250-300). Hence, this region was duplicated in sequence B.
- The amino acids that appear in region 500-550 in sequence A are present in region 600-650 in sequence B. This is a clear example of a translocation.

Exercise 2.4: Exploring the NW and SW algorithms

Pairwise sequence alignment methods are used to find the best-matching (local or global) alignments of two query sequences. Two well-known algorithms include Needleman-Wunsch and Smith-Waterman.

In this exercise, you will get familiar with the algorithms behind global and local alignment. Following sites give a nice visualization of these algorithms:
`http://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Needleman-Wunsch`
`http://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Smith-Waterman`

- Using the default parameters, perform a global and local alignment of following two sequences.

  ```
  >sequence 1
  GLVGEIIGR
  >sequence 2
  GAVGLIIVR
  ```

  Make sure you understand the calculation of the score at each step in the process.

> Explanation of the algorithms can be found in the main course material were the same sequences were used to demonstrate the algorithms. It is important that you understand these algorithms and are able to calculate the score of a global and local alignment. If you still have problems with this, use these sites to get some additional practice.

Exercise 2.5: Selecting the proper alignment tool

For this exercise you will need the Needle and Water tools at `https://www.ebi.ac.uk/jdispatcher/` (Here, you can use the default settings of the alignment tools. You will get more insight in the parameters in following exercise).

- Search for conserved residues in the human UDP-galactose transporter-related protein 1 (Solute carrier family 35 member B1) and UDP-galactose transporter homolog 1 from *Neosartorya fumigata*.
- Compare the protein sequences of the human and mouse B-cell translocation gene 1 protein using a pairwise alignment. How do these sequences differ?

(1) Alignment of human UDP-galactose transporter-related protein 1 and UDP-galactose transporter homolog 1 from *Neosartorya fumigata*

In this first exercise, we want to align two proteins that do not seem to be similar: a protein from a fungi having 415 amino acids and a human protein having 322 amino acids. We therefore choose to align them locally.

```
########################################
#
# Aligned_sequences: 2
# 1: S35B1_HUMAN
# 2: HUT1_ASPFU
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 362
# Identity:     121/362 (33.4%)
# Similarity:   179/362 (49.4%)
# Gaps:          51/362 (14.1%)
# Score: 433.0
#
#
########################################

S35B1_HUMAN         9 PDRLRLPLCFLGVFVCYFYYGILQEKITRGKY--------GEGAKQETFT     50
                      |..::|.:|.||::..:..:|:|||.||...:        ......|.||
HUT1_ASPFU         43 PGLIQLAICVLGIYASFLSWGVLQEAITTVNFPVRPPTAEEPNPPTERFT     92

S35B1_HUMAN        51 FALTLVFIQCVINAVFAKILIQFFDTA-----RVDRTRSWLYAACSISYL     95
                      |::.|..||....|:...:.:.|...|     .:..||..|:....:|..
HUT1_ASPFU         93 FSIVLNTIQSTFAAITGFLYLYFSTPAGKKVPSIFPTRKILFPLLLVSIS    142

S35B1_HUMAN        96 GAMVS--SNSALQFVNYPTQVLGKSCKPIPVMLLGVTLLKKKYPLAKYLC    143
                      .::.|  ..::|..::|.|.:|.||||.:|||.|.:|:.:|.|||.||..
HUT1_ASPFU        143 SSLASPFGYASLAHIDYLTFILAKSCKLLPVMFLHLTIFRKTYPLYKYGV    192

S35B1_HUMAN       144 VLLIVAGVALF-MYKP---KKV-VGIEEHTVG---YGELLLLLSLTLDGL    185
                      |||:..|||.| ::.|   ||| ....::..|   ||..||.::|.||||
HUT1_ASPFU        193 VLLVTLGVATFTLHHPGTSKKVAASAAKNQSGSSLYGIFLLSINLLLDGL    242

S35B1_HUMAN       186 TGVSQDHMRAHYQ-----TGSNHMMLNINLWSTLLLGMGILF------TG    224
                      |..:|||:.:..|     ||. .||:..|:.||:|....:|.      ||
HUT1_ASPFU        243 TNTTQDHVFSSPQIYTRFTGP-QMMVAQNILSTILTTTYLLVMPHLSSTG    291

S35B1_HUMAN       225 ----------------ELWEFLSFAERYPAIIYNILLFGLTSALGQSFIF    258
                                      ||...:||..|:|.::.|:|.|....|:||.|||
HUT1_ASPFU        292 ALHALLPIPIPPSTETELASAVSFLSRHPEVMKNVLGFAACGAIGQLFIF    341

S35B1_HUMAN       259 MTVVYFGPLTCSIITTTRKFFTILASVILFANPISPMQWVGTVLVFLGLG    308
                      .|:..|..|....:|.|||..|:|.||..|.:.:|..||:|..|||.|:|
HUT1_ASPFU        342 YTLSRFSSLLLVTVTVTRKMLTMLLSVFWFGHTLSAGQWLGIGLVFGGIG    391

S35B1_HUMAN       309 LDAKFGKGAKKT    320
                      .:|...|..|::
HUT1_ASPFU        392 AEAVVQKREKQS    403
```

In general the markup line uses a space for a mismatch or a gap, '.' for any small positive score, ':' for a similarity which scores more than 1.0, and '|' for an identity where both sequences have the same residue regardless of its score. These can be conserved or be present by chance. As you can see, the alignment doesn't give a clear conserved regio between the two sequences.

(2) Alignment of human and mouse B-cell translocation gene 1 protein
In this second exercise, we want to align two proteins that seem to be more similar at first sight: they both contain 171 amino acids and are both derived from mammals.

```
#=======================================
#
# Aligned_sequences: 2
# 1: BTG1_HUMAN
# 2: BTG1_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 171
# Identity:     171/171 (100.0%)
# Similarity:   171/171 (100.0%)
# Gaps:           0/171 ( 0.0%)
# Score: 892.0
#
#
#=======================================

BTG1_HUMAN          1 MHPFYTRAATMIGEIAAAVSFISKFLRTKGLTSERQLQTFSQSLQELLAE     50
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
BTG1_MOUSE          1 MHPFYTRAATMIGEIAAAVSFISKFLRTKGLTSERQLQTFSQSLQELLAE     50

BTG1_HUMAN         51 HYKHHWFPEKPCKGSGYRCIRINHKMDPLIGQAAQRIGLSSQELFRLLPS    100
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
BTG1_MOUSE         51 HYKHHWFPEKPCKGSGYRCIRINHKMDPLIGQAAQRIGLSSQELFRLLPS    100

BTG1_HUMAN        101 ELTLWVDPYEVSYRIGEDGSICVLYEASPAGGSTQNSTNVQMVDSRISCK    150
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||
BTG1_MOUSE        101 ELTLWVDPYEVSYRIGEDGSICVLYEASPAGGSTQNSTNVQMVDSRISCK    150

BTG1_HUMAN        151 EELLLGRTSPSKNYNMMTVSG    171
                      |||||||||||||||||||||
BTG1_MOUSE        151 EELLLGRTSPSKNYNMMTVSG    171
```

From the alignment we can conclude that the two proteins sequences are identical.

Summary:
Global alignment methods attempt to align two sequences in their entirety (start to end), whereas a local alignment searches for a partial overlap somewhere in the sequences (i.e. similar sequence motifs or domains). Hence global alignments can be used to search for differences in similar sequences with similar length, while local alignments are useful to find conserved residues in less similar sequences.

Exercise 2.6: Studying the effect of the parameters on pairwise alignments

- Align the human Histone H1.1 protein and the Histone-like nucleoid-structuring protein (DNA-binding protein H-NS) from *E.coli* with a local alignment tool. Retrieve both protein sequences from Uniprot in the fasta format. Make different pairwise sequence alignments using following sets of parameters:

  1) blosum62, gap opening penalty of 10, gap extension penalty of 0.5

  2) blosum62, gap opening penalty of 5, gap extension penalty of 0.5

  3) blosum62, gap opening penalty of 5, gap extension penalty of 5

4) blosum30, gap opening penalty of 10, gap extension penalty of 0.5

5) blosum90, gap opening penalty of 10, gap extension penalty of 0.5

- Based on the different alignments you have generated: what is the influence of the gap opening penalty, the gap extension penalty and the score matrix on the pairwise alignment?

```
#=======================================
#
# Aligned_sequences: 2
# 1: H11_HUMAN
# 2: HNS_ECOLI
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 87
# Identity:      19/87 (21.8%)
# Similarity:    35/87 (40.2%)
# Gaps:           7/87 ( 8.0%)
# Score: 41.0
#
#
#=======================================

H11_HUMAN        118 ETKPGASKVATKTKATGASKKLKKATGA-------SKKSVKTPKKAKKPA    160
                     |....|::|..:|:.....:::..|.|.       |..:||:..|||:..
HNS_ECOLI         43 EESAAAAEVEERTRKLQQYREMLIADGIDPNELLNSLAAVKSGTKAKRAQ     92

H11_HUMAN        161 ATRKSSKNPKKPKTVKPKKVAKSPAKAKAVKPKAAKA    197
                     ...|.|...:..:|.......::||..|....:..|:
HNS_ECOLI         93 RPAKYSYVDENGETKTWTGQGRTPAVIKKAMDEQGKS    129
```

This is what the normal local pairwise sequence alignment looks like using the default parameters. Overall, the alignment looks not that good, meaning that the two sequences are very different.

```
#=======================================
#
# Aligned_sequences: 2
# 1: H11_HUMAN
# 2: HNS_ECOLI
# Matrix: EBLOSUM62
# Gap_penalty: 5.0
# Extend_penalty: 0.5
#
# Length: 122
# Identity:      33/122 (27.0%)
# Similarity:    51/122 (41.8%)
# Gaps:          41/122 (33.6%)
# Score: 75.0
#
#
#=======================================

H11_HUMAN         45 ELIVQ--------AASSSKERGGVSLAALKKALAAAGYDVEKNNSRIKLG     86
                     |::|.        ||:..:||.. .|...::.|.|.|  ::.|    :|
HNS_ECOLI         34 EVVVNERREEESAAAAEVEERTR-KLQQYREMLIADG--IDPN----EL-     75

H11_HUMAN         87 IKSLVSKGTLVQTKGTGASGSFKLNKKASSVETKPGA-SKV----ATKTK    131
                     :.||.:    |:: ||.|       |:|.    :|.. |.|    .|||
HNS_ECOLI         76 LNSLAA----VKS-GTKA-------KRAQ----RPAKYSYVDENGETKT-    108

H11_HUMAN        132 ATGASKK---LKKATGASKKSV    150
                     .||..:.   :|||.....||:
HNS_ECOLI        109 WTGQGRTPAVIKKAMDEQGKSL    130
```

Here we have decreased the gap *opening* penalty from 10 to 5. This means that the alignment score will suffer less from opening a gap. As a consequence, you can see more gaps in the alignment. Getting the right gap penalty is important. If you set it too high, you will not have any gaps, even in locations where there should be (and consequently you'll end up with a poor alignment). A gap penalty of 0 means there is no cost of placing a gap.

```
#=======================================
#
# Aligned_sequences: 2
# 1: H11_HUMAN
# 2: HNS_ECOLI
# Matrix: EBLOSUM62
# Gap_penalty: 5.0
# Extend_penalty: 5.0
#
# Length: 88
# Identity:      21/88 (23.9%)
# Similarity:    40/88 (45.5%)
# Gaps:           9/88 (10.2%)
# Score: 44.0
#
#
#=======================================

H11_HUMAN         70 AAAGYDVEKNNSRIKLGIKSLVSKGTLVQTKGTGASGSFKLNKKASSVET    119
                     :||..:||:...:::...:.|::.| :...:...:..:.|...||...:
HNS_ECOLI         45 SAAAAEVEERTRKLQQYREMLIADG-IDPNELLNSLAAVKSGTKAKRAQ-     92

H11_HUMAN        120 KPGA-SKV-AT-KTKA-TGASK--K-LKKATGASKKSV    150
                     :|.. |.| .. :||. ||..:  . :|||.....||:
HNS_ECOLI         93 RPAKYSYVDENGETKTWTGQGRTPAVIKKAMDEQGKSL    130


#---------------------------------------
#---------------------------------------
```

In this alignment we have drastically increased the *gap extension* penalty from 0.5 to 5. Remember that the gap opening penalty is a penalty given only when opening a gap (i.e. matching a character to a gap), while introducing gaps immediately after other gaps is influenced by the gap extension penalty. Increasing the gap extension penalty clearly results in a lot less *longer* gaps. Occasionally we still see shorter gaps, as the gap opening penalty has remained unchanged. Using these parameters, the alignment score is better off by introducing mismatches than long gaps.

```
#=======================================
#
# Aligned_sequences: 2
# 1: H11_HUMAN
# 2: HNS_ECOLI
# Matrix: EBLOSUM30
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 124
# Identity:      23/124 (18.5%)
# Similarity:    58/124 (46.8%)
# Gaps:           8/124 ( 6.5%)
# Score: 83.0
#
#
#=======================================

H11_HUMAN          79 NNSRIKLGIKSLVSKGTLVQTKGTGASGSFKLNKK-----ASSVETKPGA    123
                      ||.|.   :::.:.:.||...:.......:.:|::     |:::|:::.:
HNS_ECOLI           9 NNIRT---LRAQARECTLETLEEMLEKLEVVVNERREEESAAAAEVEERT     55

H11_HUMAN         124 SKVATKTKATGASKKLKKATGASKKSVKTPKKAKKPAATRKSSKNPKKPK    173
                      .|::...:..|....:.....|..:||:..|||:.:...|.|...:..:
HNS_ECOLI          56 RKLQQYREMLIADGIDPNELLNSLAAVKSGTKAKRAQRPAKYSYVDENGE    105

H11_HUMAN         174 TVKPKKVAKSPAKAKAVKPKAAKA    197
                      |.......::||..|.::.::.|:
HNS_ECOLI         106 TKTWTGQGRTPAVIKKAMDEQGKS    129
```

For this alignment we have used a different substitution matrix. The blosum30 matrix is built using sequences with approximately 30% identity. This means that this substitution matrix is more suited to compare divergent sequences than, let's say, a blosum62 or blosum90. As you can see from the alignment and the alignment score, the blosum30 is very 'forgiving' and will allow for a lot of mismatches as these are not really penalized.

```
#=======================================
#
# Aligned_sequences: 2
# 1: H11_HUMAN
# 2: HNS_ECOLI
# Matrix: EBLOSUM90
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 59
# Identity:      15/59 (25.4%)
# Similarity:    24/59 (40.7%)
# Gaps:          22/59 (37.3%)
# Score: 39.5
#
#
#=======================================

H11_HUMAN         118 ETKPGASKVATKTKATGASKKLKK------ATG----------ASKKSVK    151
                      |....|:.|..:|      :||::      |.|          |:.||..
HNS_ECOLI          43 EESAAAAEVEERT------RKLQQYREMLIADGIDPNELLNSLAAVKSGT     86

H11_HUMAN         152 TPKKAKKPA    160
                      ..|:|::||
HNS_ECOLI          87 KAKRAQRPA     95
```

When using the blosum90 matrix (built using sequences with 90% identity) we get a completely different alignment. This matrix is more suited to align sequences that are more identical, as it penalizes mismatches harshly. As a consequence, our local alignment looks at smaller regions and we can see more gaps in the alignment. Based on the results generated here and keeping into account that we are comparing divergent sequences, it might be a better idea to use a less stringent substitution matrix. These parameters always need some fine tuning. However, remember that starting with the default parameters (blosum62) is always a good idea.

Section 3: Multiple sequence alignment

Now that we have tackled pairwise sequence alignment, where we align two sequences, we will go deeper into *multiple sequence alignment*. This allows us to align more than 2 sequences so that we can compare several sequences in one go.

Learning goals

At the end of this section, you should:

- be able to perform and interpret multiple alignments
- understand the fundamental differences between the discussed alignment methods

Exercise 3.1: Progressive versus iterative methods

Let's try to make a multiple sequencing alignment of the following sequences:

```
>Sequence1
GARFIELDTHELASTFATCAT
>Sequence2
GARFIELDTHEFASTCAT
>Sequence3
GARFIELDTHEVERYFASTCAT
>Sequence4
THEFATCAT
>Sequence5
GARFIELDTHEVASTCAT
```

Navigate to `https://www.ebi.ac.uk/jdispatcher/msa`. You can find many multiple sequence alignment tools here.

Use Clustal Omega, T-Coffee, MUSCLE and MAFFT to perform a multiple sequencing alignment. Make sure to enter the above sequences in the FASTA format! Take your course notes and review the differences between these MSA methods. You might notice differences between the alignments depending on the tool you use. Can you explain this? Which one do you think is the best alignment?

Progressive methods

T-Coffee

```
  CLUSTAL W (1.83) multiple sequence alignment

  Sequence1        GARFIELDTHELASTF-ATCAT
  Sequence2        GARFIELDTHEFAS----TCAT
  Sequence3        GARFIELDTHEVERYFASTCAT
  Sequence4        --------THEFA-----TCAT
  Sequence5        GARFIELDTHEVAS----TCAT
                           ***.      ****
```

Clustal Omega

```
CLUSTAL O(1.2.4) multiple sequence alignment


Sequence4       --------THEFAT-----CAT    9
Sequence3       GARFIELDTHEVERYFASTCAT    22
Sequence1       GARFIELDTHELASTFAT-CAT    21
Sequence2       GARFIELDTHEFASTCAT----    18
Sequence5       GARFIELDTHEVASTCAT----    18
                         ***.
```

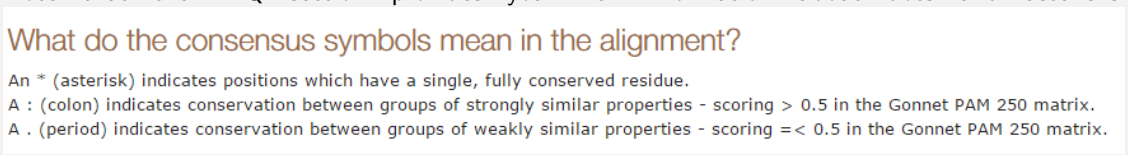
Iterative methods

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Sequence4        --------THE----FA-TCAT
Sequence2        GARFIELDTHE----FASTCAT
Sequence3        GARFIELDTHEVERYFASTCAT
Sequence1        GARFIELDTHELASTFA-TCAT
Sequence5        GARFIELDTHE----VASTCAT
                         ***     .* ****
```

```
CLUSTAL format alignment by MAFFT FFT-NS-i (v7.429)


Sequence1        GARFIELDTHELASTFA-TCAT
Sequence2        GARFIELDTHE----FASTCAT
Sequence5        GARFIELDTHE----VASTCAT
Sequence3        GARFIELDTHEVERYFASTCAT
Sequence4        --------THE----FA-TCAT
                         ***     .* ****
```

Muscle and MAFFT, the iterative methods, seem to do a better job than the progressive alignment tools T-Coffee and Clustal omega. One problem that progressive methods suffer from is that they can depend on the order in which the sequences are aligned. Once two sequences are combined into a subsequence, they are frozen in that conformation, so any errors that were introduced can no longer be corrected in the future. If you compare the output of T-Coffee with Clustal Omega, you can see that T-Coffee works better for this example. One big difference between these two methods is that T-Coffee makes use of additional information while performing the progressive alignment, namely a library with all the pairwise alignments of your sequences. This information is used to guide the progressive alignment of T-Coffee which reduces the number of errors that would be made by a standard progressive alignment method.
Iterative methods do also have a pitfall: they are much slower than the progressive alignment tools. In this small alignment it does not matter, but for larger alignments it can become a big issue.

Note that the FAQ section provides you with information about dots and asterisks:

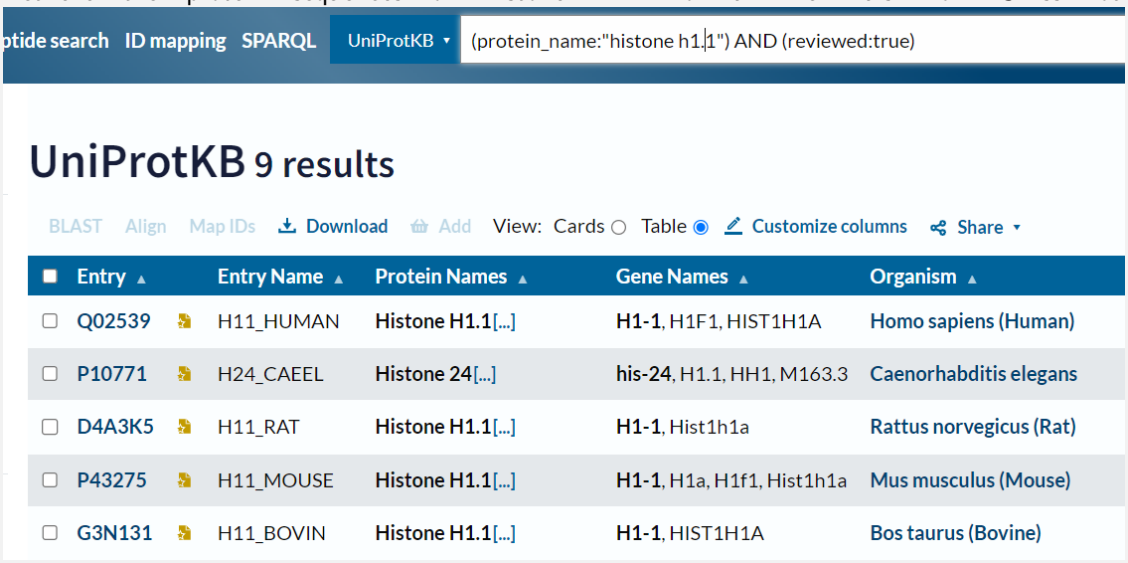

What do the consensus symbols mean in the alignment?

An * (asterisk) indicates positions which have a single, fully conserved residue.
A : (colon) indicates conservation between groups of strongly similar properties - scoring > 0.5 in the Gonnet PAM 250 matrix.
A . (period) indicates conservation between groups of weakly similar properties - scoring =< 0.5 in the Gonnet PAM 250 matrix.

Exercise 3.2: Comparison of orthologous gene products

Perform a multiple alignment to compare the Histone H1.1 proteins from four mammals (i.e. mouse, rat, human and cow). Are the sequences very similar or dissimilar?

Retrieve the protein sequences of Histone H1.1 for all mammals from Swiss-Prot.



To align the four proteins you must choose one of the multiple alignment methods discussed above. Since we will align four short sequences, we don't expect this to take a long time. Therefore, we choose an iterative alignment as these produces better results. Following figures shows the multiple alignment using Muscle:

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


sp|P43275|H11_MOUSE      MSETAPVAQAASTATEKPAAAKKTKKPAK-AAAPRKKPAGPSVSELIVQAVSSSKERSGV
sp|D4A3K5|H11_RAT        MSETAPVPQPASVAPEKPAATKKTRKPAK-AAVPRKKPAGPSVSELIVQAVSSSKERSGV
sp|Q02539|H11_HUMAN      MSETVPPAPAASAAPEKPLAGKKAKKPAKAAAASKKKPAGPSVSELIVQAASSSKERGGV
sp|G3N131|H11_BOVIN      MSEVALPAPAASTSPEKPSAGKKAKKPAKAAAAAKKKPAGPSVSELIVQAVSSSKERSGV
                         ***..  . .**.:.*** * **:.**** **...***************.******.**

sp|P43275|H11_MOUSE      SLAALKKSLAAAGYDVEKNNSRIKLGLKSLVNKGTLVQTKGTGAAGSFKLNKKA---ESK
sp|D4A3K5|H11_RAT        SLAALKKSLAAAGYDVEKNNSRIKLGLKSLVNKGTLVQTKGTGAAGSFKLNKKA---ESK
sp|Q02539|H11_HUMAN      SLAALKKALAAAGYDVEKNNSRIKLGIKSLVSKGTLVQTKGTGASGSFKLNKKASSVETK
sp|G3N131|H11_BOVIN      SLAALKKALAAAGYDVEKNNSRIKLGLKSLVGKGTLVQTKGTGASGSFKLNKKVASVDAK
                         *******:******************:****.************:********.   ::*

sp|P43275|H11_MOUSE      AITTKVSVKAKASGAAKKPKKTAG-AAAKKTVKTPKKPKKPAVSKK-TSKSPKKPKVVKA
sp|D4A3K5|H11_RAT        ASTTKVTVKAKASGAAKKPKKTAG-AAAKKTVKTPKKPKKPAVSKKTSSKSPKKPKVVKA
sp|Q02539|H11_HUMAN      PGASKVATKTKATGASKKLKKATG--ASKKSVKTPKKAKKPAATRK-SSKNPKKPKTVKP
sp|G3N131|H11_BOVIN      PTATKVATKTKVTSASKKPKKASGAAAAKKSVKTPKKARKSVLTKK-SSKSPKKPKAVKP
                         . ::***:.*:*.:.*:** **::*  *:**:*******..*.. :.* :**.*****.**.

sp|P43275|H11_MOUSE      KKVAKSPAKAKAVKPKASKAKVTKPKTPAKPKKAAPKKK
sp|D4A3K5|H11_RAT        KKVAKSPAKAKAVKPKAAKVKVTKPKTPAKPKKAAPKKK
sp|Q02539|H11_HUMAN      KKVAKSPAKAKAVKPKAAKARVTKPKT-AKPKKAAPKKK
sp|G3N131|H11_BOVIN      KKVAKSPAKAKAVKPKGAKVKVTKPKTAAKPKKAAPKKK
                         ****************.:*..****** ***********
```

The studied proteins seem very similar: most residues are conserved across the four sequences.

Overview question

Kallmann syndrome 6 is a condition that causes hypogonadotropic hypogonadism (HH) and an impaired sense of smell. Mutations in multiple genes have been associated with this disease, including FGF8. Imagine a family of three where the father has already been diagnosed with this syndrome. A genetic analysis concluded that he has a SNP in one of his FGF8 alleles (rs137852659). To see whether his son also carries this mutation, the region around this SNP was sequenced for both alleles. These sequences are stored in FGF8_sequences.fasta in files_P1.zip. In this exercise, you will examine these sequences to see whether his son has inherited the mutation:

1. First, check whether the sequences contain any SNPs.
   tip: to find a SNP, you will have to compare the sequences to the reference sequence
2. If you find a SNP, check whether it is identical to the SNP that was detected in one of the alleles of the father.
3. In addition, try to answer following questions using dbSNP:

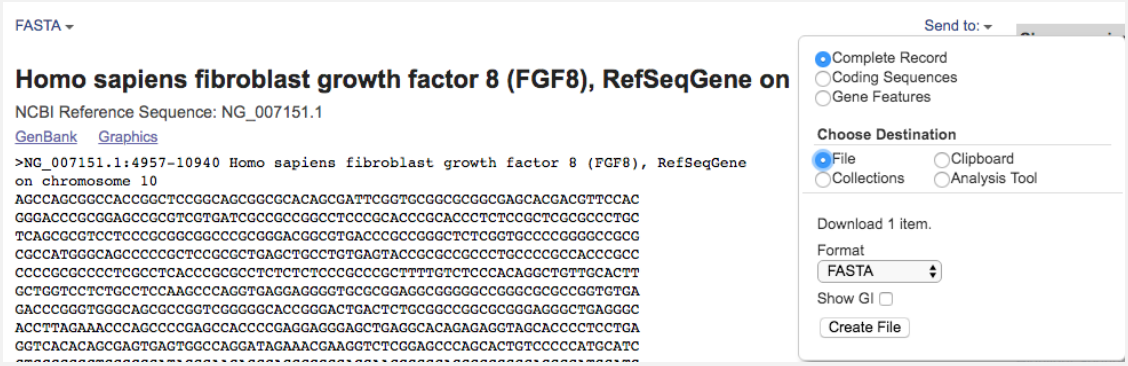
- On which strand is the gene located?
- What is the influence of the SNP on the protein level?
- Can you find other genomic variations in this gene that are associated with a disease?

(1) Identify all the SNPs in the sequences

To find genomic variations in the sequences, we align them with the reference sequence of the gene using a local alignment algorithm. (Make sure you understand why we choose a local alignment here.)

Step 1: Get the reference sequence of the FGF8 gene

Downloading a reference sequence is done preferably from the RefSeq database, because this database contains curated nucleotide sequences. If the sequence you're searching for is not available there, you can search for it in GenBank. FGF8 is available on RefSeq, so you can search RefSeq for the RefSeqGene of FGF8. This will send you to a record containing the DNA sequence of the entire gene. To download this gene in FASTA format, you can click on *Send to*.



Step 2: Find the genetic variations through a local alignment

(1) Alignment of allele 1 with the reference gene

```
NG_007151.1     5251 CCTCACCCGCGCCTCTCTCTCCCGCCCGCTTTTGTCTCCCACAGGCTGTT   5300
                     |||||||||||||||||||||||||||||||||||||||||||||||||||
allele1            1 CCTCACCCGCGCCTCTCTCTCCCGCCCGCTTTTGTCTCCCACAGGCTGTT     50

NG_007151.1     5301 GCACTTGCTGGTCCTCTGCC   5320
                     ||||||||||||||||||||
allele1           51 GCACTTGCTGGTCCTCTGCC     70
```

Allele 1 does not contain any SNPs.

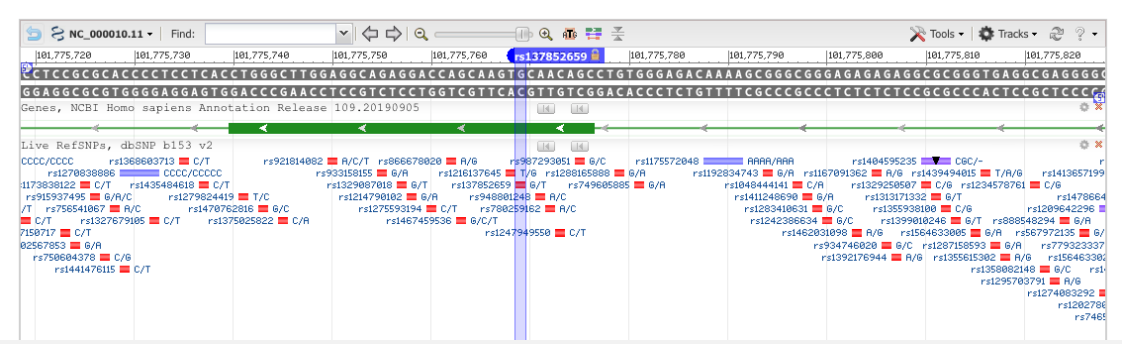
(2) Alignment of allele 2 with the reference gene

```
NG_007151.1     5251 CCTCACCCGCGCCTCTCTCTCCCGCCCGCTTTTGTCTCCCACAGGCTGTT   5300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
allele2            1 CCTCACCCGCGCCTCTCTCTCCCGCCCGCTTTTGTCTCCCACAGGCTGTT     50

NG_007151.1     5301 GCACTTGCTGGTCCTCTGCC   5320
                     |.||||||||||||||||||||
allele2           51 GAACTTGCTGGTCCTCTGCC     70
```

Allele 2 contains 1 SNP: a C in the reference genome was replaced by an A in allele 2.

(2) Check whether the mutation in allele 2 was inherited from the father

It is possible that the mutation in allele 2 is different from the father (and inherited from the mother). To check whether this mutation is the same as rs137852659, we can examine the region around the SNP in dbSNP. dbSNP is a large database containing different human genetic variants including SNPs and small INDELs (i.e. insertions and deletions). You can search this database using different filters including genomic position, gene name, ... For this exercise, we use the accession number: rs137852659.



When comparing the reported sequences from the two alleles with the region around the SNP in dbSNP, you can see that these are identical. Therefore, we can say that the boy carries the same mutation as his father.

(3) Short questions

On which strand is the gene located?

You can already see from the visualization of the genomic region that the gene is located on the antisense strand. You could also conclude this from the table that lists the SNPs in the different gene variants (Variant Details).

Gene: FGF8, fibroblast growth factor 8 (minus strand)

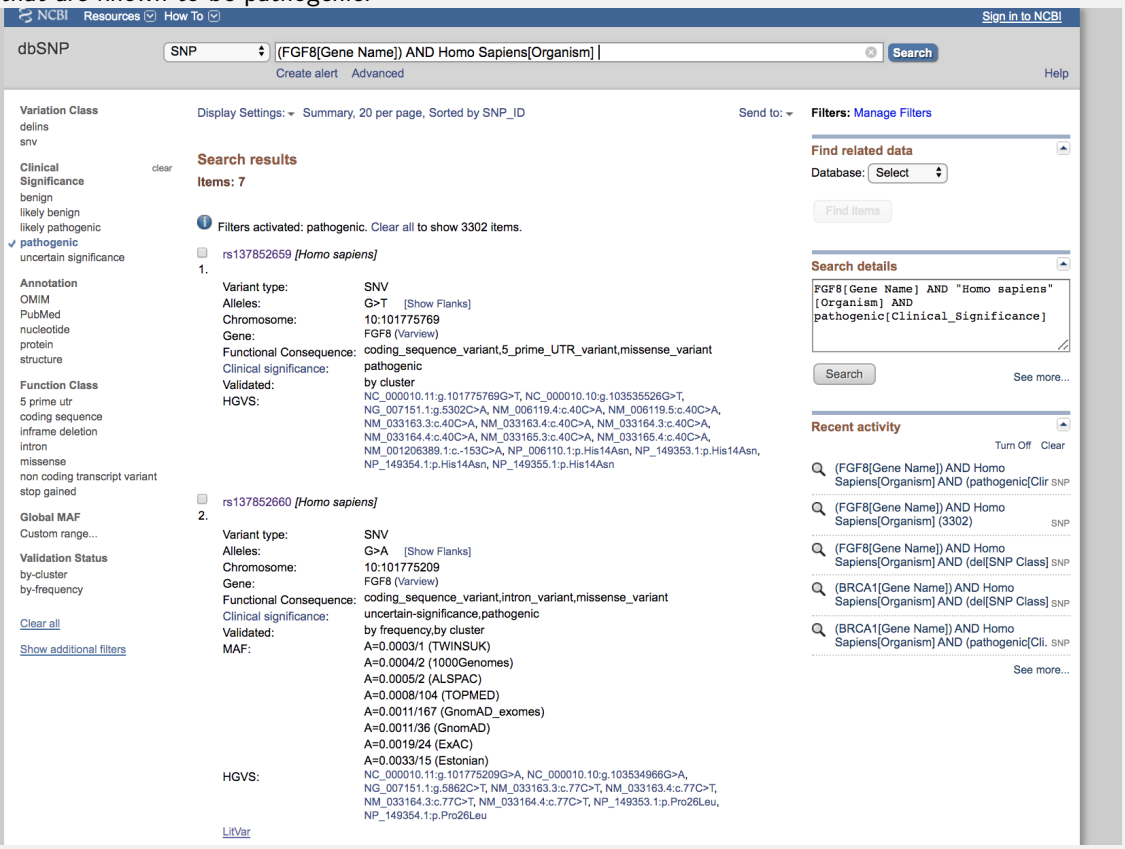
Molecule type	Change	Amino acid[Codon]	SO Term
FGF8 transcript variant A	NM_033165.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant B	NM_006119.5:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant E	NM_033164.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant F	NM_033163.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant G	NM_001206389.1:c.	N/A	5 Prime UTR Variant
fibroblast growth factor 8 isoform A precursor	NP_149355.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform B precursor	NP_006110.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform E precursor	NP_149354.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform F precursor	NP_149353.1:p.His14Asn	H (His) > N (Asn)	Missense Variant

What is the influence of the SNP on the protein level?

According to dbSNP, this SNP is a missense variant: the SNP results in the change of one amino acid. Take a look at previous figure if you want to know which amino acid is changed by the SNP.

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr10:101775769 (GRCh38.p12)	**Gene : Consequence**	FGF8 : Missense Variant
Alleles	G>T	**Publications**	1 citation
Variation Type	SNV Single Nucleotide Variation	**Genomic View**	See rs on genome
Frequency	None		

Can you find other genomic variations in this gene that are associated with a disease?
Search for all variations in the human FGF8 gene that has been reported in dbSNP and select those that are known to be pathogenic.



In total, dbSNP reports 7 variations in the FGF8 gene that are associated with a disease.

Homework

Q1: Extracting phenotype information from DNA

The eye color of humans is determined by SNPs in a number of genes (including HERC2, IRF4, OCA2 and TYR). In this exercise, you will predict the eye color of an unborn using DNA sequences from these genes:

- The site `https://hirisplex.erasmusmc.nl/` (final prediction table) enables the prediction of eye color using SNPs in 6 genes.
- For four of these genes, the DNA around the SNPs was identified (see homework_Q1 in files_P1.zip). Using these short sequences, make a prediction of the eye color of the unborn.

Use the default parameters to perform pairwise alignments between the reference sequences and the gene sequences from the Homework_Q1 folder. (For this, you will need to download the reference sequences in the same manner as previous question.) If you do not see any differences between the short reference sequence and the gene sequences, you need to use dbSNP to identify the reference nucleotide at the SNP position. Be aware that genes can lie on the sense or antisense strand

Alignments for HERC2

```
NG_016355.1   206668 GAACTTGACATTTAATGCTCA 206688
                     |||||||||||||||||||||
allele1            1 GAACTTGACATTTAATGCTCA     21

NG_016355.1   206668 GAACTTGACATTTAATGCTCA 206688
                     |||||||||||||||||||||
allele2            1 GAACTTGACATTTAATGCTCA     21
```

Here we are looking for SNP `https://www.ncbi.nlm.nih.gov/snp/rs12913832`: in the reference genome T is present at chr15:28120472 (on antisense strand). The fetus has this reference nucleotide at both of its HERC2 alleles.

Alignments for IRF4

```
NG_027728.1     9573 TAAAAGAAGGCAAATTCCCCT   9593
                     |||||||||||||||||||||
allele1            1 TAAAAGAAGGCAAATTCCCCT     21

NG_027728.1     9573 TAAAAGAAGGCAAATTCCCCT   9593
                     |||||||||||||||||||||
allele2            1 TAAAAGAAGGCAAATTCCCCT     21
```

This corresponds to SNP `https://www.ncbi.nlm.nih.gov/snp/rs12203592`: in the reference genome C is present at chr6:396321 (on sense strand). The fetus has this reference nucleotide in both of its IRF4 alleles.

Alignments for OCA2

```
NG_009846.1    119131 CGGCTCTCCCGGGGACGGGTG 119151
                      ||||||||||.||||||||||
allele1             1 CGGCTCTCCCAGGGACGGGTG     21
```

```
NG_009846.1    119131 CGGCTCTCCCGGGGACGGGTG 119151
                      ||||||||||.||||||||||
allele2             1 CGGCTCTCCCAGGGACGGGTG     21
```

This corresponds to SNP `https://www.ncbi.nlm.nih.gov/snp/rs1800407`: in the reference genome G is present at chr15:27985172 (on antisense strand). The fetus has A instead of G in both of its OCA2 alleles.

Alignments for TYR

```
NG_008748.1    104997 TCTCTGCAACGAAATCTGTGT 105017
                      |||||||||||||||||||||
allele1             1 TCTCTGCAACGAAATCTGTGT     21
```

```
NG_008748.1    104997 TCTCTGCAACGAAATCTGTGT 105017
                      ||||||||||.||||||||||
allele2             1 TCTCTGCAACAAAATCTGTGT     21
```

This corresponds to SNP `https://www.ncbi.nlm.nih.gov/snp/rs1393350`: in the reference genome G is present at chr11:89277878 (on sense strand). The fetus has one A and one G located at this position.

Based on the information of the studied genes, the tool predicts a blue eye color.

The IrisPlex System

	Gene	SNP	Allele	No. of Alleles
1	*HERC2*	rs12913832	T	0 1 **2** NA
2	*OCA2*	rs1800407	A	0 1 **2** NA
3	*LOC105370627*	rs12896399	T	0 1 2 **NA**
4	*SLC45A2*	rs16891982	C	0 1 2 **NA**
5	*TYR*	rs1393350	T	**0** 1 2 NA
6	*IRF4*	rs12203592	T	**0** 1 2 NA

Display Predicted Phenotype | Download Predicted Phenotype

Predicted phenotype

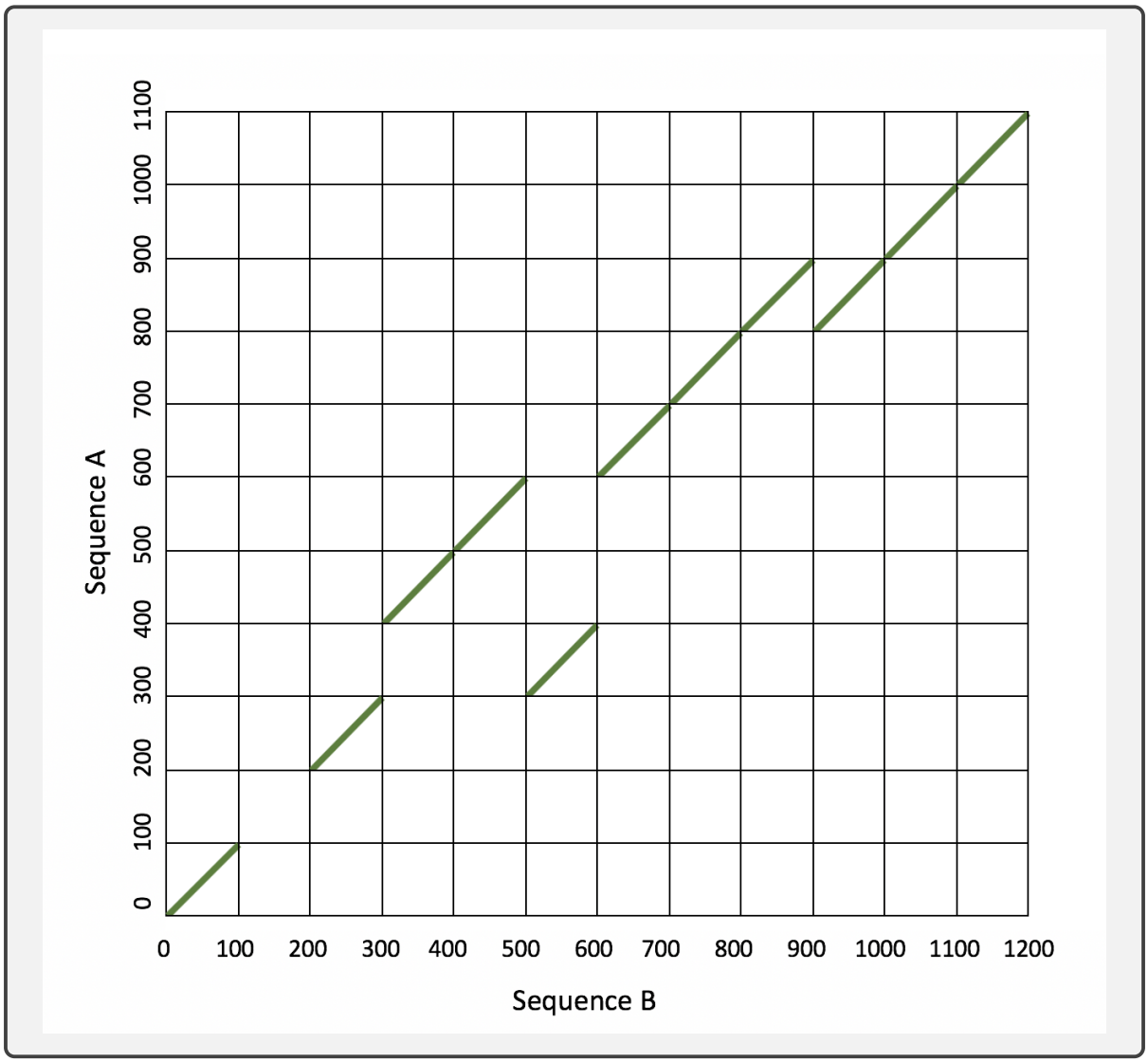
	p-value	AUC Loss
blue eye	0.007	0.015
intermediate eye	0.096	0.041
brown eye	0.897	0.01

Q2: Drawing a dotplot by hand

Given gene A (1100 nucleotides) and gene B. Imagine that gene B was derived from gene A as follows:

- Region 100-200 was changed.
- Region 300-400 was translocated to region 500-600.
- Region 800-900 was duplicated.

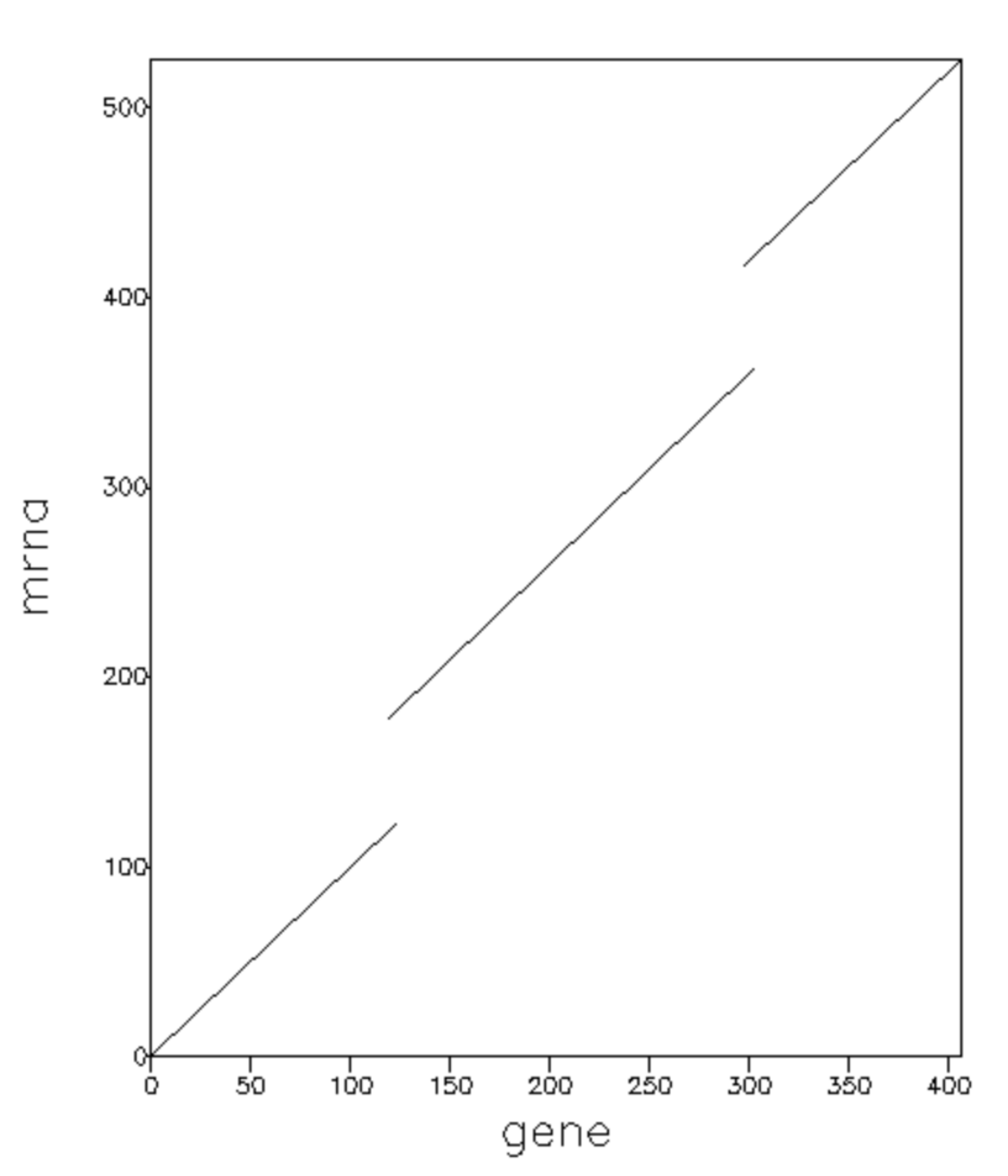
Draw a dotplot of gene A and B on paper.



Q3: Describe briefly how you would solve following problems

a) Identify/visualize the introns and exons from a gene starting from the gene sequence and corresponding mRNA sequence (without using any databases).

You can do this very easily with a dotplot of the mRNA sequence and gene sequence. The introns are represented as deletions in the mRNA while the exons are represented by straight lines.



b) Find conservative residues in the protein sequences from orthologous genes.

Align all protein sequences using a multiple alignment approach. Depending on the number of protein sequences and their lengths, you might want to use an iterative method (these are slower than the progressive methods, but attain better alignments) or T-coffee (a progressive alignment achieving better results due to the inclusion of information during the alignment.)

c) Find the chromosomal location of the human apolipoprotein E gene.

You can find this in the NCBI Gene database:

- Search term: apolipoprotein E [Protein Full Name] AND homo sapiens[Organism]
- Go to genomic context

d) Allocate differences in the amino acid sequences of the progranulin protein of *Mus musculus* and *Rattus norvegicus*.

Since these two proteins are very similar, you can use a global alignment method to find the differences (i.e. the Needleman-Wunsch approach)