

# Practical sessions overview

- **Practical 1: Databases, Alignment, Search** – 31/03
- Practical 2: Motifs, Python & assignment – 28/04
- Practical 3: Phylogenetic trees, protein structures and Gene Ontology – 5/05
- Practical 4: Next Generation Sequencing – 12/05
- Q&A session – 19/05
- EXAM – 26/5

# Practicum agreements

- Assignment:
  - **Mandatory**
  - Differs for computer scientists – biologists
  - Will be explained during practicum 2

- Biology students:  python<sup>TM</sup> ?

# Practical session 1

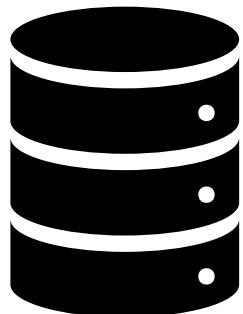
## Databases & Alignment

# Practical agreements

- There are 2 correction moments:
  - 15u: correction chapter 1
  - 17u: correction chapters 2-3
- All solutions will be provided afterwards
- Homework questions Q1-Q3: solutions will be provided afterwards.

# Overview of the practical session

NCBI &  
UniProt



gene/protein  
info

>Gene\_sequence  
TATCATCATATGGTAGTAGGTAAAT

>Protein\_sequence  
MKSGSGGGSPTSLW

Pairwise alignment

```
PDRRLRLPLCFLGVFVCYFYYGILQEKITRGRK  
| . . : | . | | : : . . : | | | . | | .  
PGLIQLAICVLGIVASFLSWGVLQEAITTVN
```

Multiple alignment

```
MSETAPVAQAASTATEKPAAAKKKPAK-AAAPF  
MSETAPVPQPASVAPEKPAATKKTRKPAK-AAVPF  
MSETVPPAPAASAAPEKPLAGKKAKKPAKAAAASK  
MSEVALPAPAASTSPEKPSAGKKAKKPAKAAAAAF  
*****. . . * . : . * * * * * * : * * * * . * . .
```

# Section 1

## Databases

# Exercise 1.1: NCBI databases (Gene DB)

(1) Look for human insulin in the **Gene database** and answer following questions:

- Which diseases are associated with problems of insulin metabolism?
- Find molecular components that interact with insulin.
- 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?

# Exercise 1.1: NCBI databases (Gene DB)

Gene (homo sapiens[Organism]) AND insulin[Gene/Protein Name] ✖ Search

Name/Gene ID	Description	Location	Aliases	MIM
<input type="checkbox"/> <a href="#">IGF1</a> ID: 3479	insulin like growth factor 1 [ <i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (102395860..102481839, complement)	IGF, IGF-I, IGFI, MGF	147440
<input type="checkbox"/> <a href="#">CTLA4</a> ID: 1493	cytotoxic T-lymphocyte associated protein 4 [ <i>Homo sapiens</i> (human)]	Chromosome 2, NC_000002.12 (203867771..203873965)	ALPS5, CD, CD152, CELIAC3, CTLA-4, GRD4, GSE, IDDM12	123890
<input type="checkbox"/> <a href="#">IGF1R</a> ID: 3480	insulin like growth factor 1 receptor [ <i>Homo sapiens</i> (human)]	Chromosome 15, NC_000015.10 (98648539..98964530)	CD221, IGFIR, IGFR, JTK13	147370
<input type="checkbox"/> <a href="#">INS</a> ID: 3630	insulin [ <i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (2159779..2161209, complement)	IDDM, IDDM1, IDDM2, ILPR, IRDN, MODY10	176730
<input type="checkbox"/> <a href="#">MAPK3</a> ID: 5595	mitogen-activated protein kinase 3 [ <i>Homo sapiens</i> (human)]	Chromosome 16, NC_000016.10 (30114105..30123309, complement)	ERK-1, ERK1, ERT2, HS44KDAP, HUMKER1A, P44ERK1, P44MAPK, PRKM3, p44-ERK1, p44-MAPK	601795
<input type="checkbox"/> <a href="#">IGFBP3</a> ID: 3486	insulin like growth factor binding protein 3 [ <i>Homo sapiens</i> (human)]	Chromosome 7, NC_000007.14 (45912245..45921272, complement)	BP-53, IBP3	146732
<input type="checkbox"/> <a href="#">INSR</a> ID: 3643	insulin receptor [ <i>Homo sapiens</i> (human)]	Chromosome 19, NC_000019.10 (7112257..7294414, complement)	CD220, HHF5	147670

# Exercise 1.1: NCBI databases (Gene DB)

The screenshot shows the NCBI Gene search results for the query "human insulin".

**Search Bar:** Gene dropdown, search term "human insulin", Search button, Create RSS, Save search, Advanced, Help.

**Display Options:** Tabular, 20 per page, Sort by Relevance.

**Content Area:**

- Gene Summary:** GENE, INS – insulin, [Homo sapiens \(human\)](#). Also known as: IDDM, IDDM1, IDDM2, ILPR, IRDN, MODY10, PNDM4. Gene ID: 3630.
- Buttons:** RefSeq products, Orthologs, Genome Data Viewer.
- Feedback:** Was this helpful? (thumbs up/down).
- Section:** [New - Visualize gene across multiple species](#).
- RefSeq Sequences:** A tabbed section showing RefSeq Sequences.
- Search Results:** Items: 1 to 20 of 55790.
- Search Details:** The search query is displayed as: (( "Homo sapiens" [Organism] OR human [All Fields]) AND insulin [All Fields]) AND alive [prop].

**Sidebar:** Send to, Filters: Manage Filters, Results by taxon, Top Organisms (Tree), Find related data, Database: Select, Find items, See more... .

# Exercise 1.1: NCBI databases (Gene DB)

Gene  Advanced  Help

Full Report

**INS insulin [ *Homo sapiens* (human) ]**

Gene ID: 3630, updated on 26-Mar-2025

**Summary**

**Official Symbol** INS provided by HGNC  
**Official Full Name** insulin provided by HGNC  
**Primary source** HGNC:HGNC:6081  
**See related** Ensembl:ENSG00000254647 MIM:176730; AllianceGenome:HGNC:6081  
**Gene type** protein coding  
**RefSeq status** REVIEWED  
**Organism** *Homo sapiens*  
**Lineage** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo  
**Also known as** IDDM; ILPR; IRDN; IDDM1; IDDM2; PNDM4; MODY10  
**Summary** This gene encodes insulin, a peptide hormone that plays a vital role in the regulation of carbohydrate and lipid metabolism. After removal of the precursor signal peptide, proinsulin is post-translationally cleaved into three peptides: the B chain and A chain peptides, which are covalently linked via two disulfide bonds to form insulin, and C-peptide. Binding of insulin to the insulin receptor (INSR) stimulates glucose uptake. A multitude of mutant alleles with phenotypic effects have been identified, including insulin-dependent diabetes mellitus, permanent neonatal diabetes mellitus, maturity-onset diabetes of the young type 10 and hyperproinsulinemia. There is a read-through gene, INS-IGF2, which overlaps with this gene at the 5' region and with the IGF2 gene at the 3' region. [provided by RefSeq, May 2020]  
**Expression** Restricted expression toward pancreas (RPKM 671.7) [See more](#)  
**Orthologs** mouse all

**NEW** Try the new Gene table  
Try the new Transcript table

**Table of contents**

Summary  
Genomic context  
Genomic regions, transcripts, and products  
Expression  
Bibliography  
Phenotypes  
Variation  
HIV-1 interactions  
Pathways from PubChem  
Interactions  
General gene information  
Markers, Readthrough INS-IGF2, Homology, Gene Ontology  
General protein information  
NCBI Reference Sequences (RefSeq)  
Related sequences  
Additional links  
Locus-specific Databases

Genome Browsers

# Exercise 1.1: NCBI databases (Gene DB)

- Which diseases are associated with problems of insulin metabolism?

**Phenotypes**

[Find tests for this gene in the NIH Genetic Testing Registry \(GTR\)](#)

[Review eQTL and phenotype association data in this region using PheGenI](#)

Associated conditions

Description
<a href="#">Diabetes mellitus, insulin-dependent, 2</a> MedGen: <a href="#">C1852092</a> , OMIM: <a href="#">125852</a> , GeneReviews: Not available
<a href="#">Hyperproinsulinemia</a> MedGen: <a href="#">C0342283</a> , OMIM: <a href="#">616214</a> , GeneReviews: Not available
<a href="#">Maturity-onset diabetes of the young, type 10</a> MedGen: <a href="#">C3150617</a> , OMIM: <a href="#">613370</a> , GeneReviews: Not available
<a href="#">Permanent neonatal diabetes mellitus</a> MedGen: <a href="#">C1833104</a> , OMIM: <a href="#">606176</a> , GeneReviews: <a href="#">Permanent Neonatal Diabetes Mellitus</a>

# Exercise 1.1: NCBI databases (Gene DB)

- Find molecular components that interact with insulin.

Interactions

Items 1 - 25 of 29

Products	Interactant	Other Gene	Complex	Source	Pubs	Description
P01308	<a href="#">P16870</a>	<a href="#">CPE</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
P01308	<a href="#">P07858</a>	<a href="#">CTSB</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
P01308	<a href="#">P07339</a>	<a href="#">CTSD</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
P01308	<a href="#">P14091</a>	<a href="#">CTSE</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
P01308	<a href="#">P35557</a>	<a href="#">GCK</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	
P01308	<a href="#">P01906</a>	<a href="#">HLA-DQA2</a>		<a href="#">HPRD</a>	<a href="#">PubMed</a>	

# Exercise 1.1: NCBI databases (Gene DB)

- 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?

## General gene information

Function	Evidence Code	Pubs
<a href="#">hormone activity</a>	IC	<a href="#">PubMed</a>
<a href="#">hormone activity</a>	IMP	<a href="#">PubMed</a>
<a href="#">hormone activity</a>	NAS	<a href="#">PubMed</a>
<a href="#">identical protein binding</a>	IPI	<a href="#">PubMed</a>

Process	Evidence Code	Pubs
<a href="#">ER to Golgi vesicle-mediated transport</a>	TAS	
<a href="#">G protein-coupled receptor signaling pathway</a>	IDA	<a href="#">PubMed</a>
<a href="#">activation of protein kinase B activity</a>	IDA	<a href="#">PubMed</a>
<a href="#">acute-phase response</a>	IDA	<a href="#">PubMed</a>

Component	Evidence Code	Pubs
<a href="#">Golgi lumen</a>	TAS	
<a href="#">Golgi membrane</a>	TAS	
<a href="#">endoplasmic reticulum lumen</a>	TAS	
<a href="#">endoplasmic reticulum-Golgi intermediate compartment membrane</a>	TAS	

# Exercise 1.1: NCBI databases (GenBank)

(2) Look for insulin in **GenBank** (Nucleotide) and answer the following questions:

- 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 data base?
- What is the version of the gene assembly?
- Who put this information on Genbank, and can you find out something more about the research that was done to produce this data?
- Find the FASTA sequence of the gene. What is a FASTA file and how is it structured?

# Exercise 1.1: NCBI databases (GenBank)

Nucleotide (homo sapiens[Organism]) AND insulin[Protein Name]

Create alert Advanced Help

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾ Filters: [Manage Filters](#)

**Items: 12**

- [Homo sapiens insulin \(INS\) mRNA, partial cds](#)
  - 1. 285 bp linear mRNA
    - Accession: JF909299.1 GI: 333826818
    - [Protein](#) [Taxonomy](#)
    - [GenBank](#) [FASTA](#) [Graphics](#)
- [Synthetic construct Homo sapiens clone IMAGE:100010743; FLH192918.01L; RZPDo839A1068D](#)
  - 2. [insulin \(INS\) gene, encodes complete protein](#)
    - 373 bp linear other-genetic
      - Accession: DQ896283.2 GI: 123999447
      - [Protein](#) [Taxonomy](#)
      - [GenBank](#) [FASTA](#) [Graphics](#)
- [Homo sapiens insulin \(INS\) gene, complete cds](#)
  - 3. 4,969 bp linear DNA
    - Accession: AH002844.2 GI: 1036032746
    - [Protein](#) [PubMed](#) [Taxonomy](#)
    - [GenBank](#) [FASTA](#) [Graphics](#)

**Results by taxon**

**Top Organisms [Tree]**

- Homo sapiens (9)
- synthetic construct (3)

**Analyze these sequences**

Run BLAST

Find in these sequences

**Find related data**

Database:

**Search details**

"Homo sapiens"[Organism] AND

# Exercise 1.1: NCBI databases (GenBank)

- 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 data base?

The screenshot shows the NCBI Nucleotide search results for the **Homo sapiens insulin (INS) gene, complete cds**. The accession number is AH002844.2. The page includes a search bar, navigation links, and various analysis tools like BLAST and primer picking.

**GenBank:** AH002844.2

**FASTA** **Graphics**

**Go to:**

LOCUS	AH002844	4969 bp	DNA	linear	PRI	10-JUN-2016
DEFINITION	Homo sapiens insulin (INS) gene, complete cds.					
ACCESSION	AH002844 J00265 J00268					
VERSION	AH002844.2					
KEYWORDS	GC rich region; insulin; polymorphic variation; tandem repeat.					
SOURCE	Homo sapiens (human)					
ORGANISM	<a href="#">Homo sapiens</a> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.					
REFERENCE	1 (bases 2414 to 2610)					
AUTHORS	Bell,G.I., Swain,W.F., Pictet,R., Cordell,B., Goodman,H.M. and Rutter,W.J.					
TITLE	Nucleotide sequence of a cDNA clone encoding human preproinsulin					
JOURNAL	Nature 282 (5738), 525-527 (1979)					
PUBMED	<a href="#">503234</a>					

**Analyze this sequence**

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

**Articles about the INS gene**

- Sepsis uncouples serum C-peptide and insulin levels in critically ill patients [Crit Care Resusc. 2019]
- Insulin and epithelial growth factor (EGF) promote programmed death [BMC Cancer. 2019]
- The association of plasma peroxiredoxin 3 with insulin in *Bianchini Bianchi et al. 2019*

# Exercise 1.1: NCBI databases (GenBank)

- What is the version of the gene assembly?

The screenshot shows the NCBI Nucleotide search results for the **Homo sapiens insulin (INS) gene, complete cds**. The page includes a header with 'Nucleotide' search dropdown, a search bar, and a 'Search' button. Below the search area, there are tabs for 'GenBank' and 'Change region shown'. On the left, detailed information about the sequence is listed:

LOCUS	AH002844	4969 bp	DNA	linear	PRI	10-JUN-2016
DEFINITION	Homo sapiens insulin (INS) gene, complete cds.					
ACCESSION	AH002844 100265 100269					
VERSION	AH002844.2					
KEYWORDS	GC rich region; insulin; polymorphic variation; tandem repeat.					
SOURCE	Homo sapiens (human)					
ORGANISM	<a href="#">Homo sapiens</a>					
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;						
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;						
Catarrhini; Hominidae; Homo.						
REFERENCE	1 (bases 2414 to 2610)					
AUTHORS	Bell,G.I., Swain,W.F., Pictet,R., Cordell,B., Goodman,H.M. and Rutter,W.J.					
TITLE	Nucleotide sequence of a cDNA clone encoding human preproinsulin					
JOURNAL	Nature 282 (5738), 525-527 (1979)					
PUBMED	<a href="#">503234</a>					

On the right side of the page, there are several interactive links: 'Customize view', 'Analyze this sequence' (with options for 'Run BLAST', 'Pick Primers', 'Highlight Sequence Features', and 'Find in this Sequence'), and 'Articles about the INS gene' (with links to a 2019 Crit Care Resusc paper and a 2019 BMC Cancer paper).

# Exercise 1.1: NCBI databases (GenBank)

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

The screenshot shows the NCBI Nucleotide search results for the **Homo sapiens insulin (INS) gene, complete cds**. The page includes a header with 'NCBI Resources How To' and a search bar. Below the header, there are tabs for 'GenBank' and 'Change region shown'. On the left, there's a 'Customize view' section with options like 'Run BLAST', 'Pick Primers', 'Highlight Sequence Features', and 'Find in this Sequence'. The main content area displays detailed genomic information for the gene, including its definition, accession numbers, keywords, source, organism, and references. A red box highlights the 'REFERENCE' section, which lists the reference ID (1), bases 2414 to 2610, and the authors (Bell, G.I., Swain, W.F., Pictet, R., Cordell, B., Goodman, H.M. and Rutter, W.J.). The 'TITLE' section indicates the nucleotide sequence of a cDNA clone encoding human preproinsulin from Nature 282 (5738), 525-527 (1979). The 'JOURNAL' section shows the reference ID 503234. The 'PUBMED' section is also present.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Advanced Search Help

GenBank Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

**Homo sapiens insulin (INS) gene, complete cds**

GenBank: AH002844.2

FASTA Graphics

Go to:

LOCUS AH002844 4969 bp DNA linear PRI 10-JUN-2016

DEFINITION Homo sapiens insulin (INS) gene, complete cds.

ACCESSION AH002844 J00265 J00268

VERSION AH002844.2

KEYWORDS GC rich region; insulin; polymorphic variation; tandem repeat.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 2414 to 2610)

AUTHORS Bell, G.I., Swain, W.F., Pictet, R., Cordell, B., Goodman, H.M. and Rutter, W.J.

TITLE Nucleotide sequence of a cDNA clone encoding human preproinsulin

JOURNAL Nature 282 (5738), 525-527 (1979)

PUBMED [503234](#)

Articles about the INS gene

Sepsis uncouples serum C-peptide and insulin levels in critically ill patients [Crit Care Resusc. 2019]

Insulin and epithelial growth factor (EGF) promote programmed death [BMC Cancer. 2019]

The association of plasma peroxiredoxin 3 with insulin in *Bianchini Bianchini Prog Commun. 2019*

# Exercise 1.1: NCBI databases (GenBank)

- Find the FASTA sequence of the gene.

NCBI Resources How To

Nucleotide Nucleotide Advanced

GenBank

### Homo sapiens insulin (INS) gene, complete cds

GenBank: AH002844.2

[FASTA](#) [Graphics](#)

Go to:

LOCUS	AH002844	4969 bp	DNA	linear	PRI 10-JUN-2016
DEFINITION	Homo sapiens insulin (INS) gene, complete cds.				
ACCESSION	AH002844 J00265 J00268				
VERSION	AH002844.2				
KEYWORDS	GC rich region; insulin; polymorphic variation; tandem repeat.				
SOURCE	Homo sapiens (human)				
ORGANISM	<a href="#">Homo sapiens</a> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 2414 to 2610)				
AUTHORS	Bell,G.I., Swain,W.F., Pictet,R., Cordell,B., Goodman,H.M. and Rutter,W.J.				
TITLE	Nucleotide sequence of a cDNA clone encoding human preproinsulin				
JOURNAL	Nature 282 (5738), 525-527 (1979)				
PUBMED	<a href="#">503234</a>				

### FASTA

#### Homo sapiens insulin (INS) gene, complete cds

GenBank: AH002844.2

[GenBank](#) [Graphics](#)

```
>AH002844.2 Homo sapiens insulin (INS) gene, complete cds
CTCGAGGGCCTAGACATTGCCCTCAGAGAGAGCACCCAAACCCCTCCAGGCTGACCGGCCAGGGTGT
CCCCCTCCTACCTTGAGAGAGCAGCCCAGGGCATCCTGCAGGGGTGCTGGGACACCAGCTGGCCTTC
AAGGTCTCTGCCCTCCCTCAGCCACCCACTACACGCTGCTGGATCTCAGCTCCCTGGCCGA
CAACACTGGCAAACCTCTACTCATCACGAAGGCCCTCTGGGATGGTGGCCTTCCAGCCTGGCAGT
CTGTCCTCACACACCTTGTAGTGCCTCAGGCCAGGGTGCAGCTGGGGTGTCTGAAGGGCTGTG
AGCCCCAGGAAGCCCTGGGAAGTGCCTGCCTTCCTCCCCCGGCCCTGCCAGGCCCTGGCTCTGCC
TCCTACCTGGGCTCCCCCATCCAGCCTCCCTCCACACACTCCTCTCAAGGAGGCACCCATGTCTCT
CCAGCTGCCGGCCTCAGAGCACTGTGGCGCTCTGGGCAGCCACCGCATGTCCTGCTGGCATGGCTC
AGGGTGGAAAGGGCGGAAGGGAGGGTCTGCAGATAGCTGGTCCCCTACAAACCCGCTGGGGCAG
GAGAGCCAAGGCTGGGTGTGCAGAGCGGCCCCGAGAGGTTCCGAGGCTGAGGCCAGGGTGGACATA
GGGATGCCAGGGGCCGGGACAGGATACTCCAACCTGCCCTCCCCCATGGTCTCATCCTCTGCTTCTG
GGACCTCCTGATCCTGCCCTGGTGCTAAGAGGCAGGTAAAGGGCTGCAGGCAGGCCAGGGCTCGGAGGCCA
TGCCCCCTCACCAGGGTCAGGCTGGACCTCCAGGTGCCTGTTCTGGGAGGCTGGGAGGCCGGAGGGGT
GTACCCCAGGGCTCAGCCAGATGACACTATGGGGTGTAGGGTGTCTGGGACCTGGCCAGGAGAGGGG
AGATGGGCTCCAGAAGAGGAGTGGGGCTGAGAGGGTGCCTGGGGGCCAGGACGGAGCTGGCCAGTG
CACAGCTTCCCACACCTGCCACCCCCAGAGTCCTGCCACCCCCAGATCACACGGAAAGATGAGGTCC
```

# Exercise 1.1: NCBI databases (GenBank)

- What is a FASTA file and how is it structured?

## The FASTA format

>Gene\_1 Very descriptive description

TATATATATGGTAGTAGGTAAAT

>Gene\_2 This gene has some unknown bases in the middle

TACGATCGTCGCTAGCTAGCTAC

TGATGTCTNNNNATCGATGCTT

CTAGCTAGCTAGCTGATCGATGC

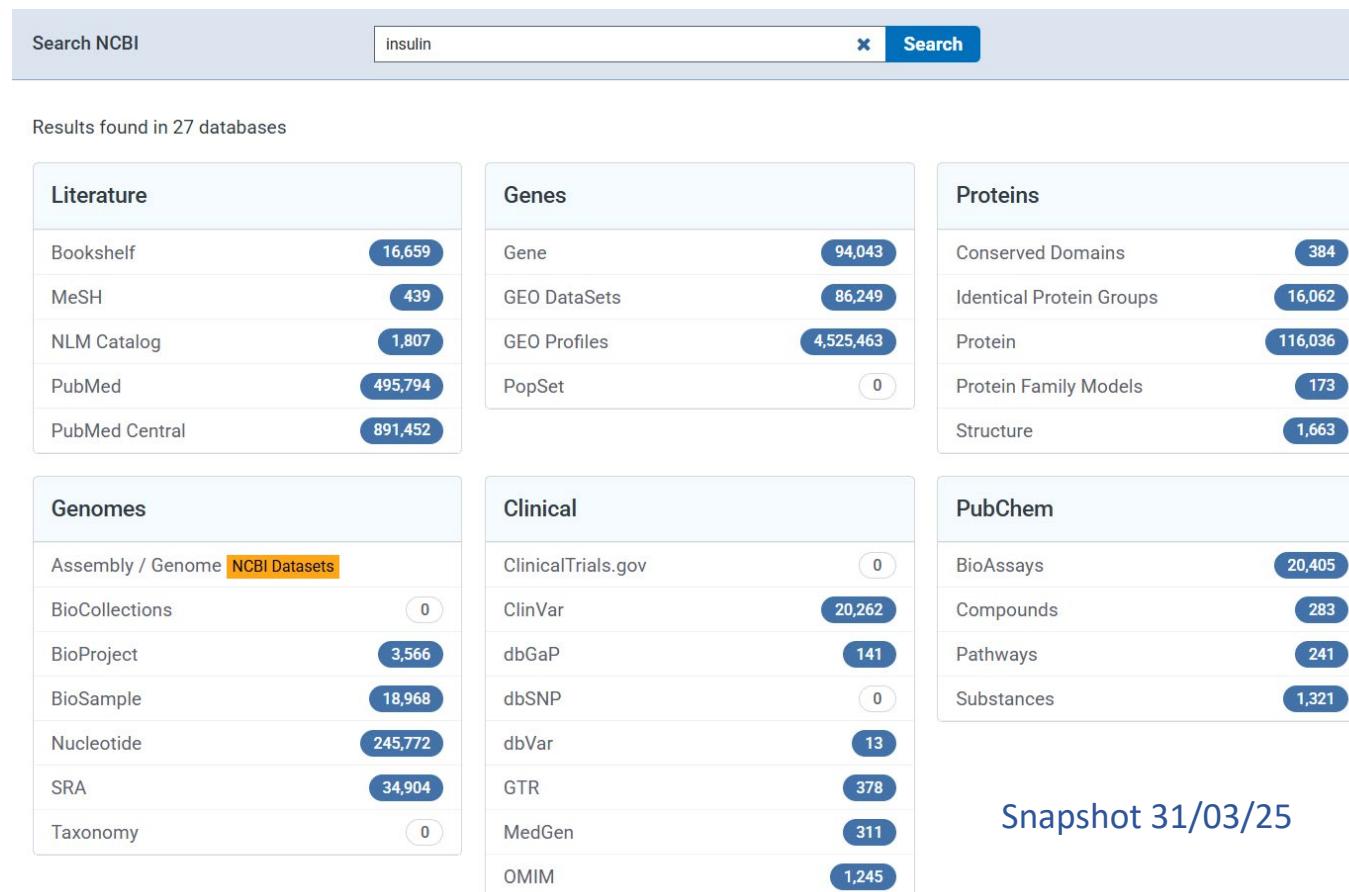
>Gene\_3

TATAGTCTGCTGCTGATCGACGA

TGCTAGTCGTGATCGATGCTAGC

# Exercise 1.1: NCBI databases (all databases)

(3) Finally, search for insulin in all NCBI databases. What other types of additional information can you find here?



# Exercise 1.1: NCBI databases (dbSNP)

**(4)** In this exercise, we will study the human BRCA1 gene. Mutations in this gene have been associated with breast cancer:

- 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

# Exercise 1.1: NCBI databases (dbSNP)

How many deletions for this gene can you find in dbSNP that are pathogenic?

dbSNP

SNP ▾ Homo sapiens BRCA1

Create alert Advanced Help

Clinical Significance likely pathogenic ✓ pathogenic pathogenic likely pathogenic

Search results Items 1 to 20 of 649

Display Settings: Summary, 20 per page, Sorted by SNP\_ID

Send to: Filters: Manage Filters

Find related data Database: Select

Find items

Search details

( "Homo sapiens" [Organism] AND BRCA1 [All Fields] ) AND ( pathogenic [Clinical\_Significance] AND del [SNP Class] )

Recent activity Turn Off Clear

Homo sapiens BRCA1 AND (pathogenic[Clinical\_Significance] AND d SNP)

Homo sapiens BRCA1 AND /pathogenic[Clinical\_Significance]

Annotation somatic

Variation Class clear

del

Snapshot 31/03/25

# Exercise 1.1: NCBI databases (dbSNP)

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

rs80357502

Organism	<i>Homo sapiens</i>	Clinical Significance	Reported in ClinVar
Position	chr17:43092551 (GRCh38.p14) 	Gene : Consequence	BRCA1 : Frameshift Variant
Alleles	delA	Publications	1 citation
Variation Type	Deletion	LitVar	 4
Frequency	None	Genomic View	See rs on genome

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

# Exercise 1.1: NCBI databases (dbSNP)

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

Variant Details	Allele: delA (allele ID: <a href="#">69404</a> )	?	
Clinical Significance	ClinVar Accession	Disease Names	Clinical Significance
	<a href="#">RCV000111974.3</a>	Breast-ovarian cancer, familial 1	Pathogenic
Aliases			
Submissions			
History			
Publications			

# Exercise 1.1: NCBI databases (dbSNP)

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

**Genomic regions, transcripts, and products** [Top](#) [?](#)

Choose placement GRCh38.p12 (NC\_000017.11) [▼](#)

[See rs80357502 in Variation Viewer](#)

The screenshot shows a genomic browser interface for the NC\_000017.11 genome. At the top, there's a navigation bar with a 'Find:' input field, search icons, and a 'Variation Viewer' link. Below the navigation is a sequence track showing the DNA sequence from 43,092,500 to 43,092,600. A specific SNP, rs80357502, is highlighted with a blue box and a lock icon. A red box highlights the 5' end of the sequence. A blue vertical bar is positioned over the sequence, indicating the location of the SNP. Below the sequence track, there are several tracks for 'Live RefSNPs, dbSNP b153 v2'. These tracks show various SNPs across the genome, each with its ID, allele, and frequency. The tracks are color-coded by allele (e.g., G/A, T/C). The bottom of the screen displays gene annotations and transcript details.

## Exercise 1.2: UniProt

- Navigate to <http://www.uniprot.org/> and search again for human insulin.  
What additional information can you find here compared to Genbank?
- Try to find out which protein is associated with Alexander disease.
- 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?

# Exercise 1.2: UniProt

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

What additional information can you find here compared to Genbank?

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB (protein\_name:insulin) AND (organism\_id:9606) Advanced List Search Help

Status

Reviewed (Swiss-Prot) (60)  
Unreviewed (TrEMBL) (147)

Popular organisms

Human (207)

Taxonomy

9606 X

Filter by taxonomy

Proteins with

3D structure (35)

Active site (8)

## UniProtKB 207 results

BLAST Align Map IDs Download Add View: Cards Table  Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P01308	INS_HUMAN	Insulin	INS	Homo sapiens (Human)	110 AA
P06213	INSR_HUMAN	Insulin receptor[...]	INSR	Homo sapiens (Human)	1,382 AA
Q9Y5Q6	INSL5_HUMAN	Insulin-like peptide INSL5[...]	INSL5, UNQ156/PRO182	Homo sapiens (Human)	135 AA
P14735	IDE_HUMAN	Insulin-degrading enzyme[...]	IDE	Homo sapiens (Human)	1,019 AA

# Exercise 1.2: UniProt

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

What additional information can you find here compared to Genbank?

 P01308 · INS\_HUMAN

Insulin · Homo sapiens (Human) · Gene: INS · 110 amino acids · Evidence at protein level · Annotation score: 5/5

---

Entry Feature viewer Publications External links History

---

BLAST Align [Download](#) [Add](#) Add a publication Entry feedback

### Function<sup>i</sup>

Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver.

### GO Annotations<sup>i</sup>

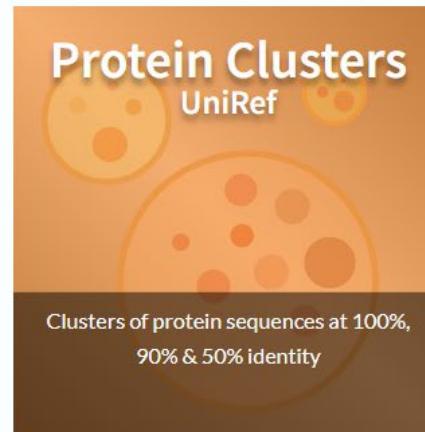
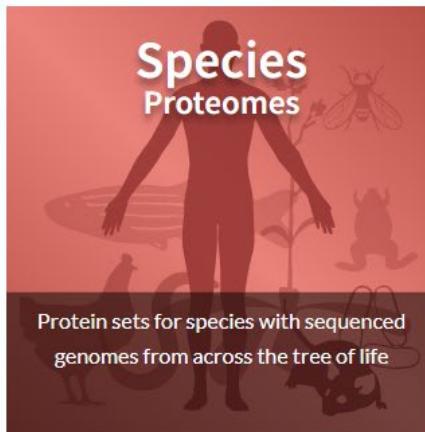
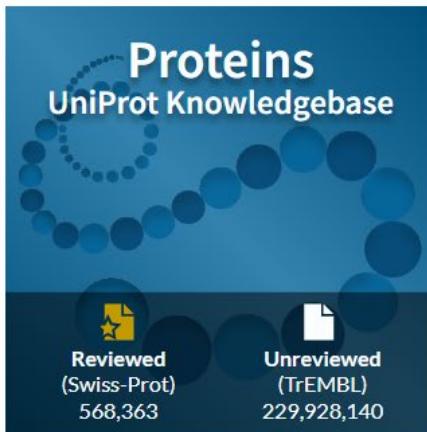
Slimming set:

generic ▾

[Feedback](#)  
[Help](#)

# Exercise 1.2: UniProt

- Try to find out which protein is associated with Alexander disease.



## AlphaFold structures

Search with all the power of the UniProt search engine for proteins with an AlphaFold prediction provided by DeepMind

## ProtNLM Predictions

Explore all the entries annotated with Google's ProtNLM predictions

## UniProt COVID-19 portal

UniProt portal for the latest SARS-CoV-2 coronavirus protein entries and receptors, updated independent of the general UniProt release cycle

## Supporting Data

Human diseases

Taxonomy

Keywords

Literature Citations

Cross-referenced databases

Subcellular locations

Automatic annotations: UniRule & ARBA

# Exercise 1.2: UniProt

- Try to find out which protein is associated with Alexander disease.

Alternative names	Alexander's disease
Keywords	Leukodystrophy
Cross references	MIM: 203450 ↗ (phenotype) MedGen: C0270726 ↗ MeSH: D038261 ↗

**Disclaimer**  
*Any medical or genetic information present in this entry is provided for research, educational and informational purposes only. It is not in any way intended to be used as a substitute for professional medical advice, diagnosis, treatment or care. Our staff consists of biologists and biochemists that are not trained to give medical advice.*

## Related UniProtKB entry

Browse 1 entry

### P14136 · GFAP\_HUMAN

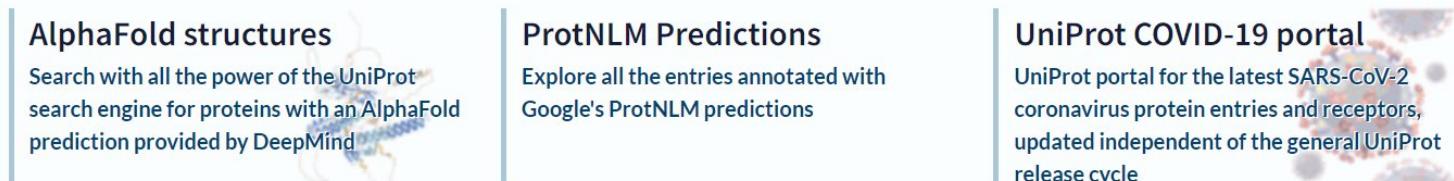
Glial fibrillary acidic protein · Homo sapiens (Human) · Gene: GFAP · 432 amino acids · Evidence at protein level · Annotation score: 5/5

#Disease variant #Leukodystrophy

1 domain · 15 PTMs · 70 reviewed variants · 3 isoforms · 208 interactions · 1 disease · 1 3D structure · 45 reviewed publications

# Exercise 1.2: UniProt

- 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?



# Exercise 1.2: UniProt

- 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?

The screenshot shows the UniProt search interface. At the top, there are three tabs: 'Search', 'ID mapping', and 'SPARQL'. Below these is a red button labeled 'Proteomes' with a dropdown arrow. To the right of the button is a search bar containing the text 'danio rerio'. The background of the search bar area is dark blue.

## Proteomes 1 result

[Download \(1\)](#) View: Cards  Table  Customize columns [Share](#)

Entry	Organism	Organism ID	Protein count	Snapshot
<input type="checkbox"/> <a href="#">UP000000437</a>	<a href="#">Danio rerio (Zebrafish) (Brachydanio rerio) (Tuebingen)</a>	7955	46,878	31/03/25

# Exercise 1.2: UniProt

- 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?

## Proteomes · *Danio rerio* (Zebrafish)

### Overview

Status  Reference proteome

Protein count<sup>i</sup> **46,691**

Gene count 20,358 [Download one protein sequence per gene \(FASTA\)](#)

Proteome ID<sup>i</sup> UP000000437

Snapshot 31/03/25



### Status

 Reviewed (Swiss-Prot)

**(3,351)**

 Unreviewed (TrEMBL)

(43,527)

### Popular organisms

At the end of this chapter, 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

# Section 2

## Pairwise alignment

# Exercise 2.1: Getting familiar with dot plots

- Use the emboss dottup tool:

[https://www.ebi.ac.uk/jdispatcher/seqstats/emboss dottup](https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_dottup)

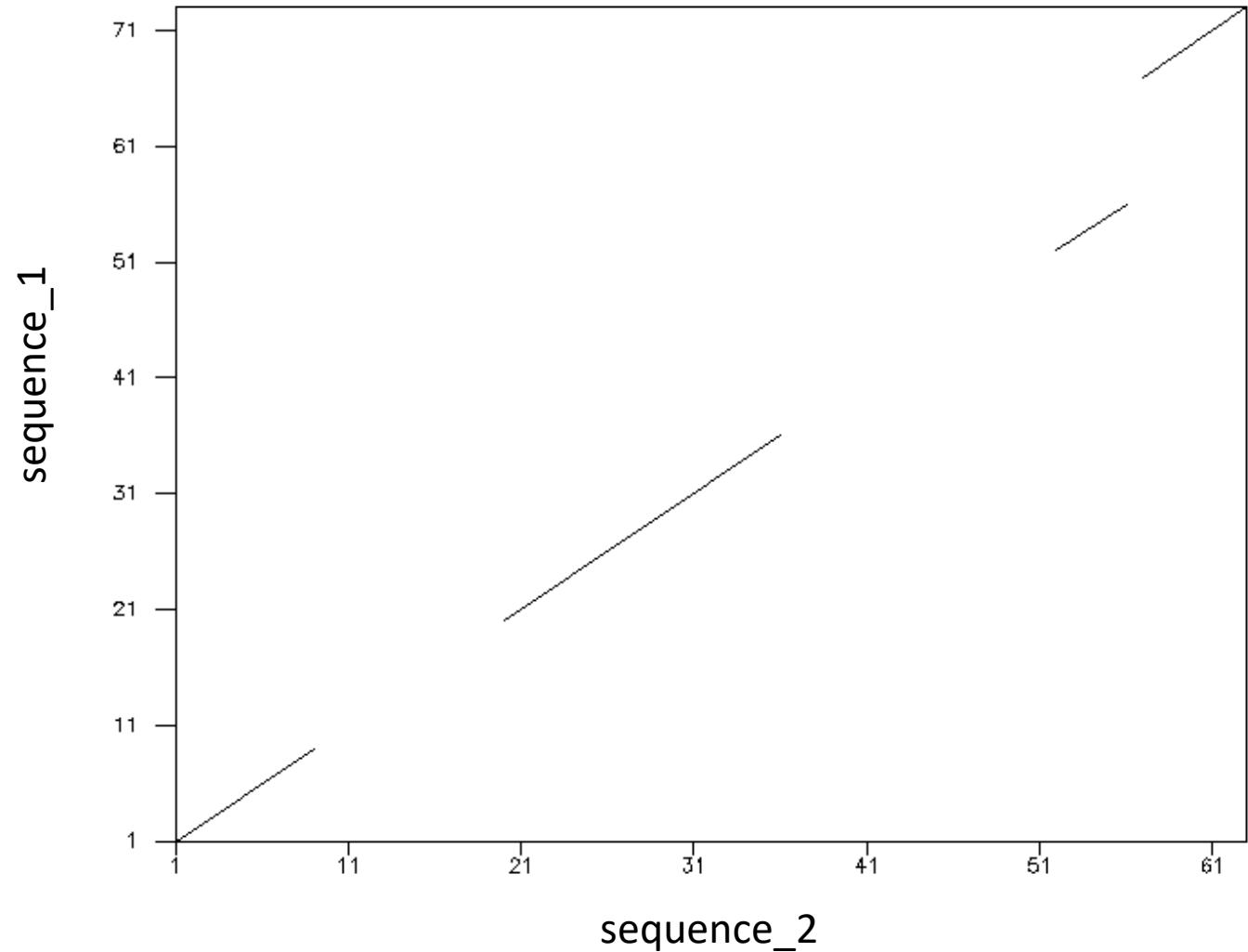
- Make the 4 different alignments described in the course notes (*use a wordsize of 5 for all plots*)

# Exercise 2.1: Getting familiar with dot plots

**1**

```
>sequence_1  
MSETAPAAPASAPAPAEEKTPVKK  
KARKSAGAAKRKASGPPVSELIT  
KAVAASKERSKKALAAAGYDGVS  
LAAL
```

```
>sequence_2  
MSETAPAAPKLIASSKPLPPVKKK  
ARKSAGAAKRKAAASWTRRPLA  
STVNWSKERSGVSLAAL
```

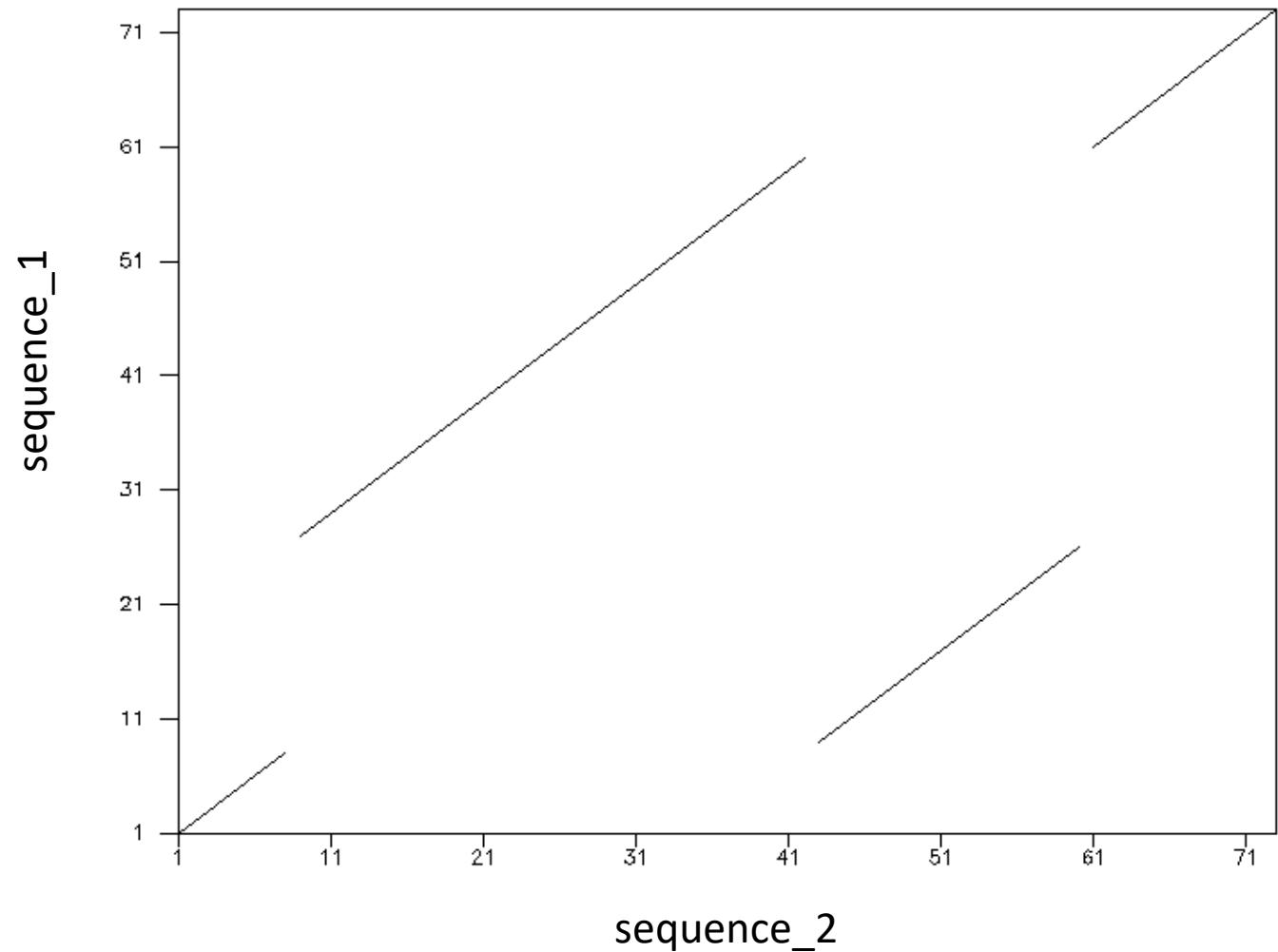


# Exercise 2.1: Getting familiar with dot plots

2

```
>sequence_1  
MSETAPAAPASAPAPAEEKTPVKKK  
ARKSAGAAKRKASGPPVSELITKAV  
AASKERSGVSLAALKKALAAAGYD
```

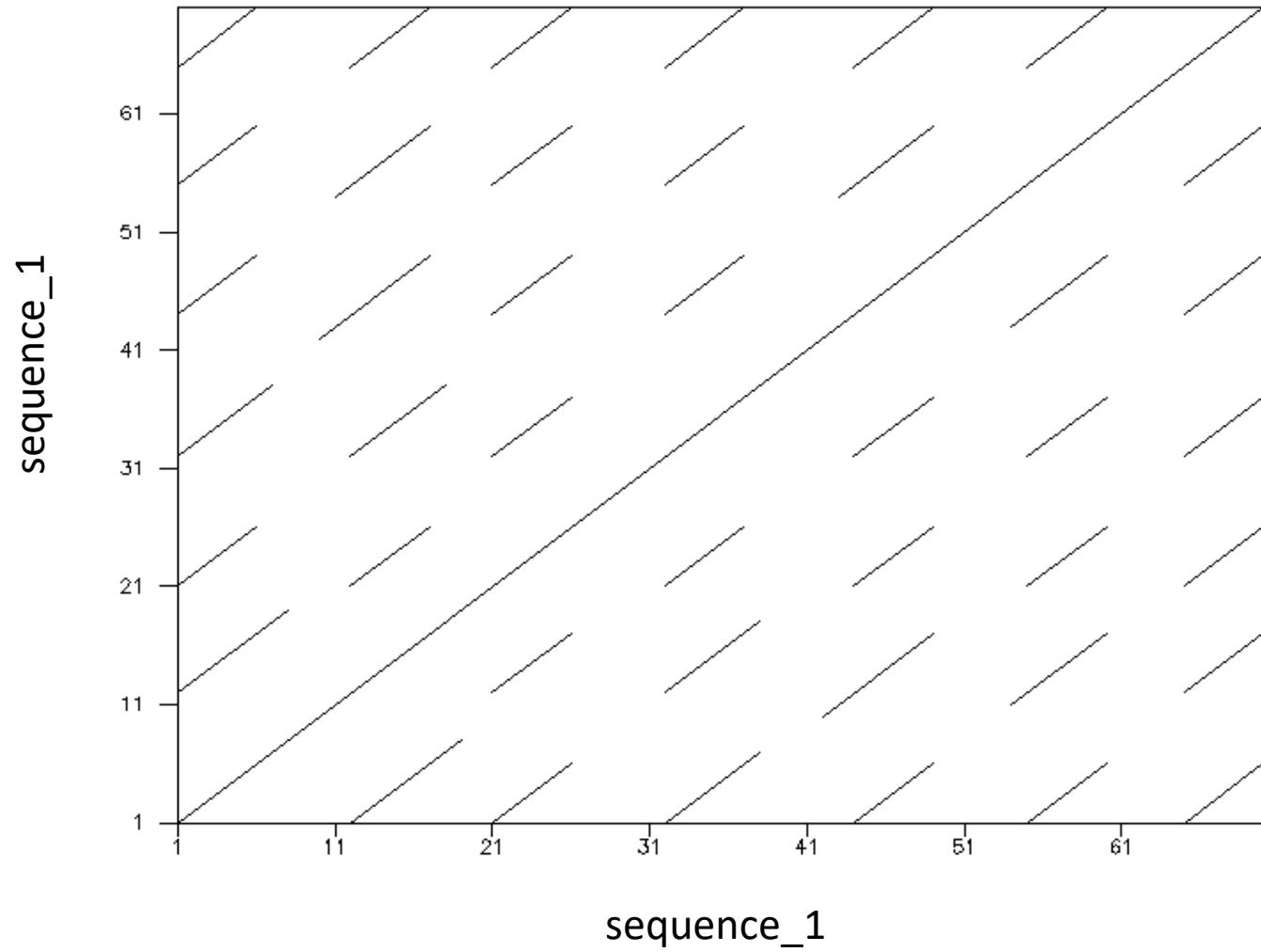
```
>sequence_2  
MSETAPAAKSAGAAKRKASGPPVS  
ELITKAVAASKERSGVSLPASAPAPA  
EKTPVKKKARAALKKALAAAGYD
```



# Exercise 2.1: Getting familiar with dot plots

3

```
>sequence_1  
MSETAPAAPASMSETAPAARMSETA  
PKSAGAMSETAPAKRKASMSETAPG  
PPVSMSETAPELITMSETAP
```

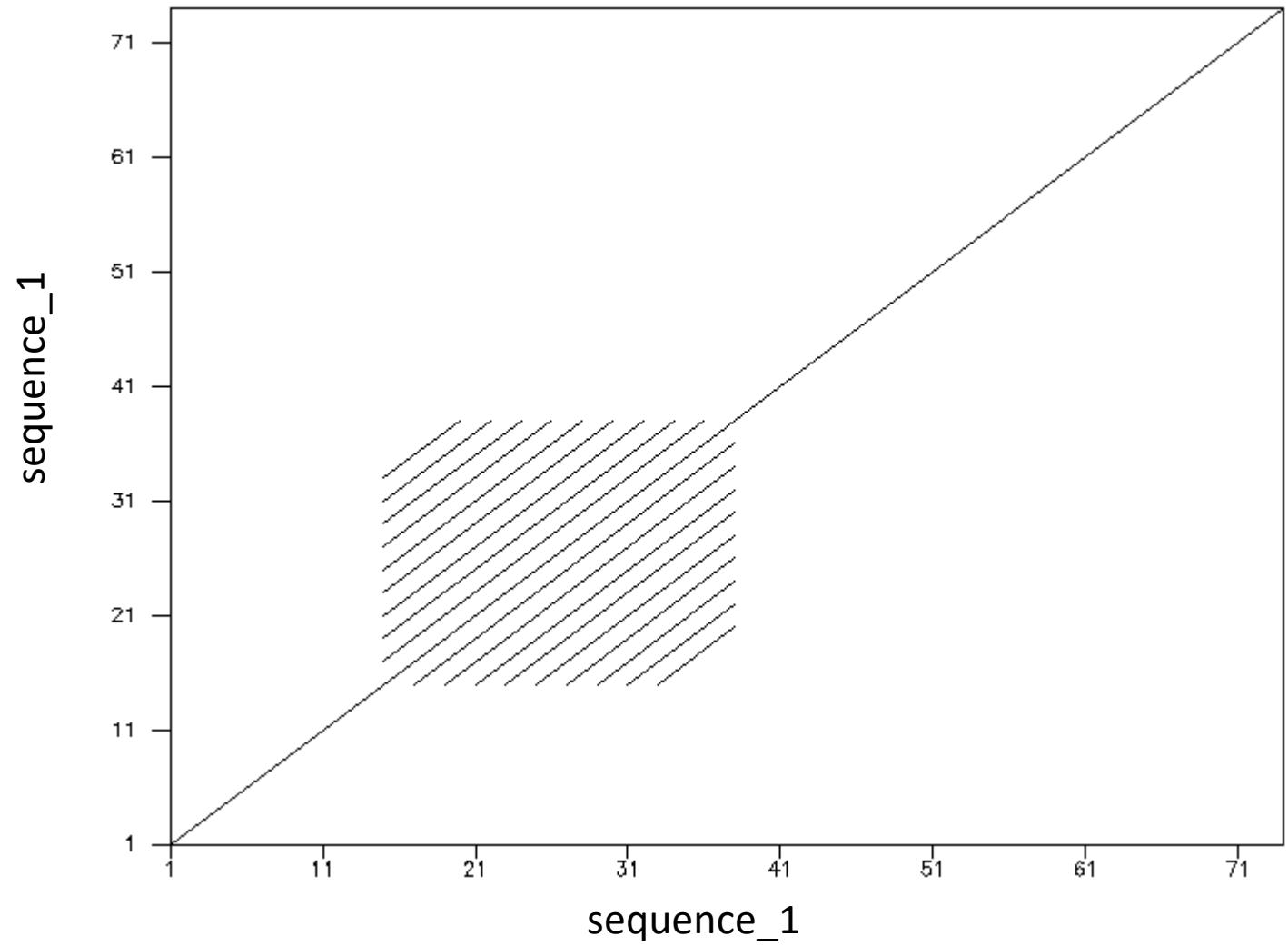


# Exercise 2.1: Getting familiar with dot plots

4

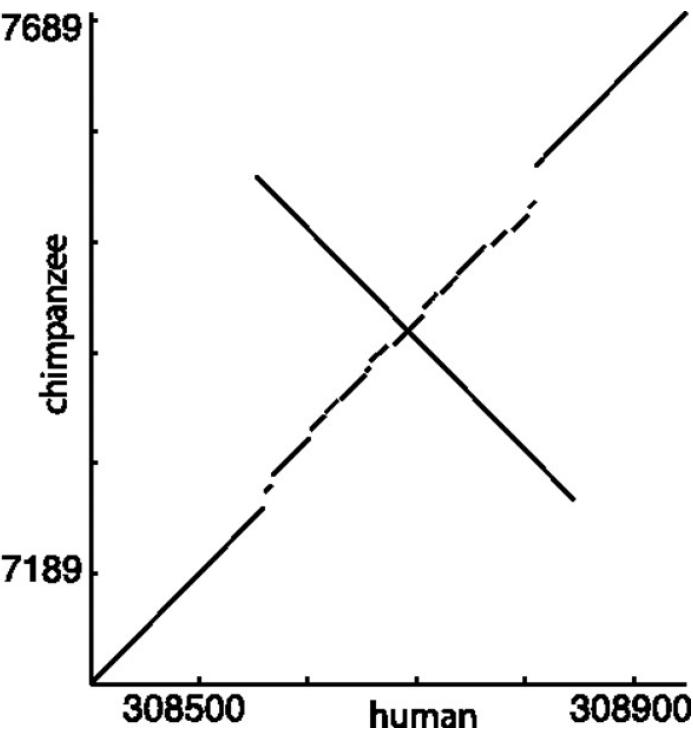
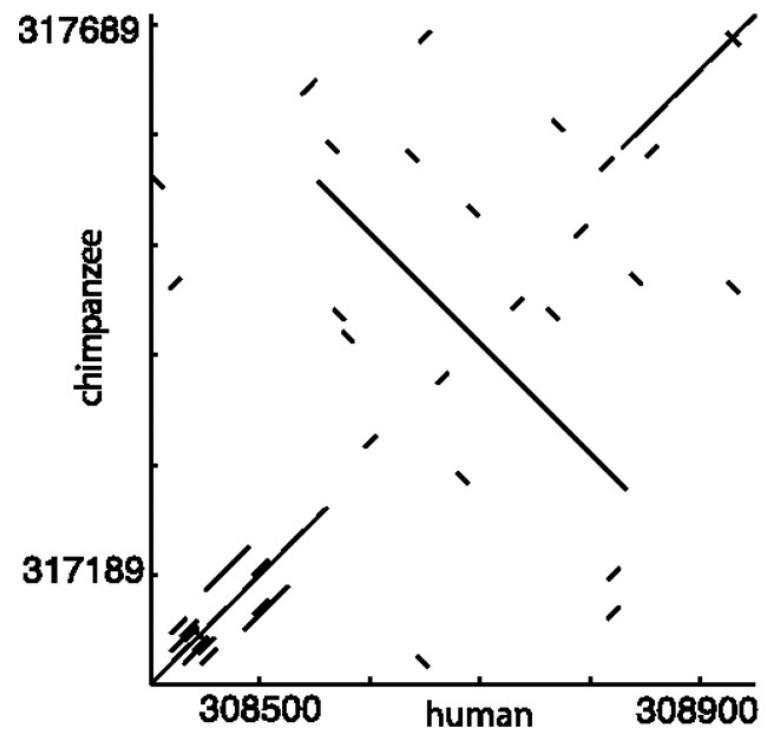
```
>sequence_1  
ASAPAPAEKTPVKKAKAKAKAKAKAKAKAK  
AKAKAKAKARKSAGAAKRKASGPPVSELIT  
KAVAASKERSGVSLA
```

```
>sequence_1  
ASAPAPAEKTPVKKAKAKAKAKAKAKAKAK  
AKAKAKAKARKSAGAAKRKASGPPVSELIT  
KAVAASKERSGVSLA
```



# Exercise 2.1: Getting familiar with dot plots

Example of a dot plot with inversion



# 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.
- Use the emboss dottup tool  
from [https://www.ebi.ac.uk/jdispatcher/seqstats/emboss\\_dottup](https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_dottup) to construct a dotplot. Do you see similarities between both sequences?
- Vary the word size from 5 to 10. What happens and why?

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

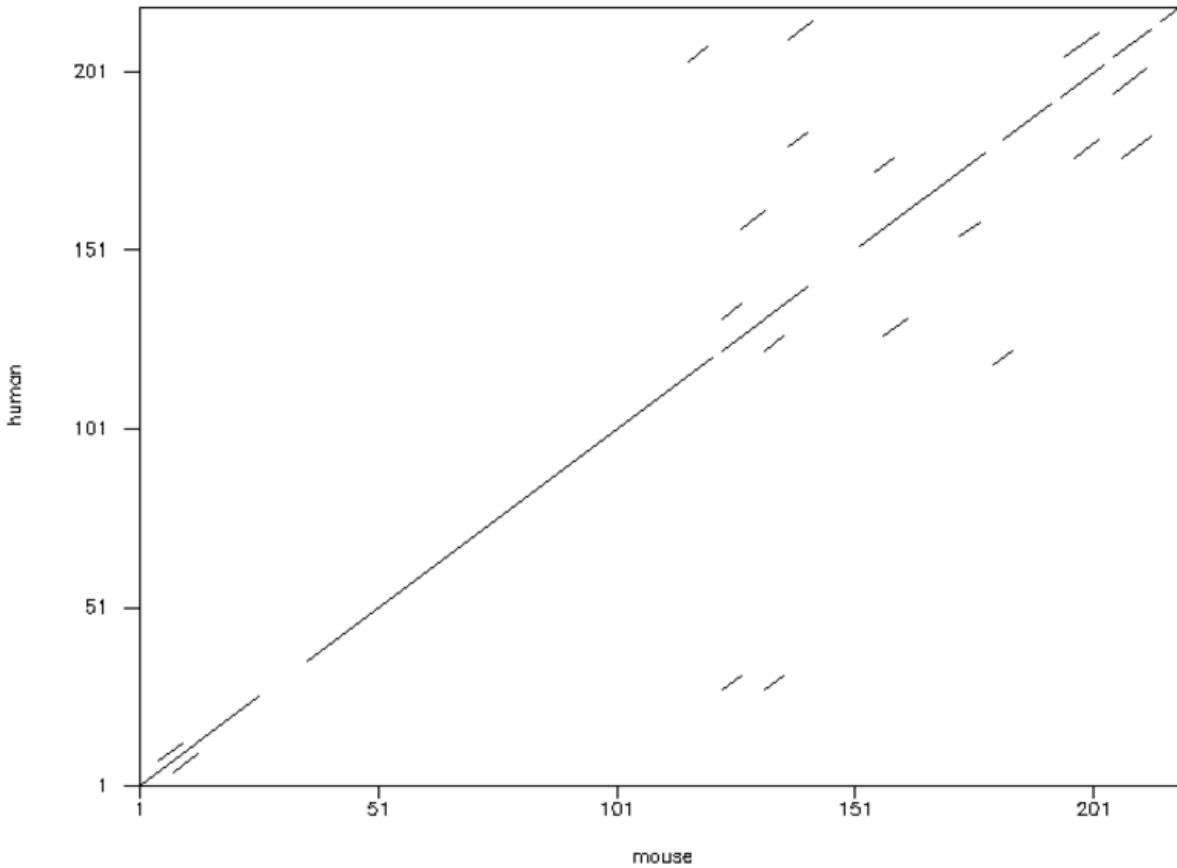
The screenshot shows the UniProtKB search interface. The top navigation bar includes links for Peptide search, ID mapping, SPARQL, and UniProtKB, with 'histone h1.4' entered in the search input field. To the right are Advanced, List, and Search buttons.

## UniProtKB 2,349 results

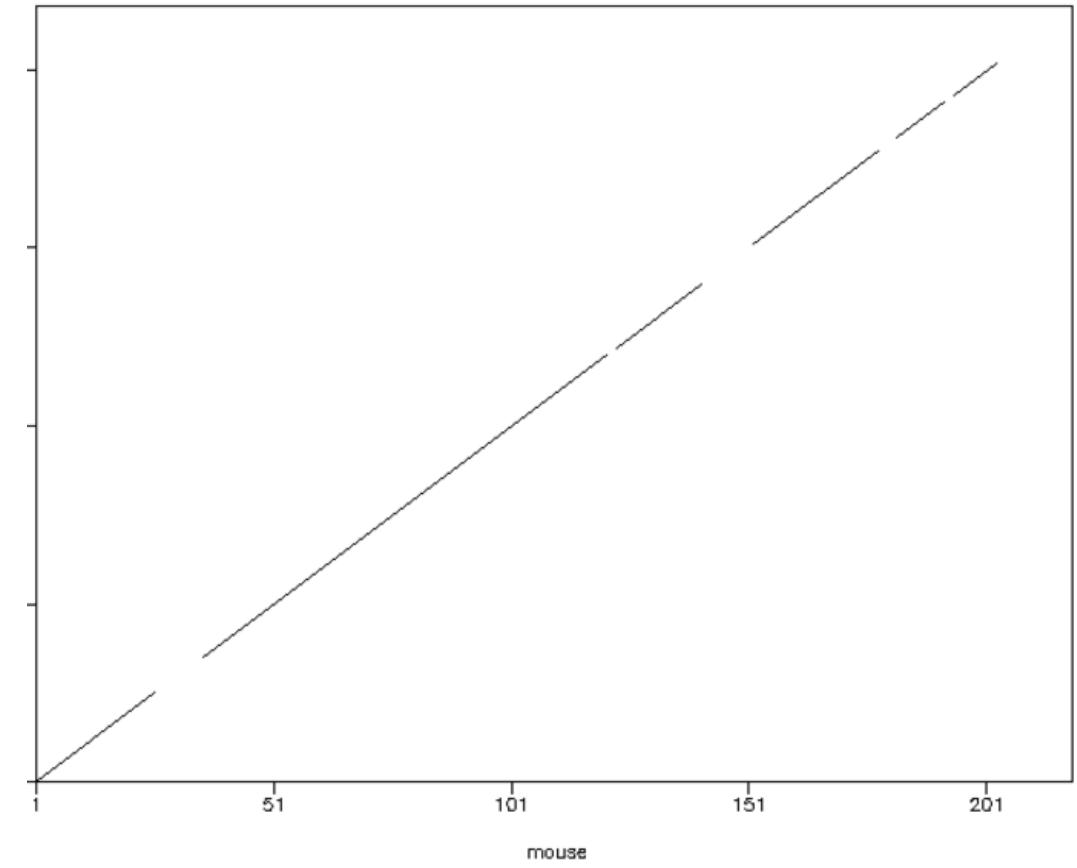
BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism
P10412	H14_HUMAN	Histone H1.4[...]	H1-4, H1F4, HIST1H1E	Homo sapiens (Human)
P15865	H14_RAT	Histone H1.4[...]	H1-4, H1f4	Rattus norvegicus (Rat)
P43274	H14_MOUSE	Histone H1.4[...]	H1-4, H1f4, Hist1h1e	Mus musculus (Mouse)

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



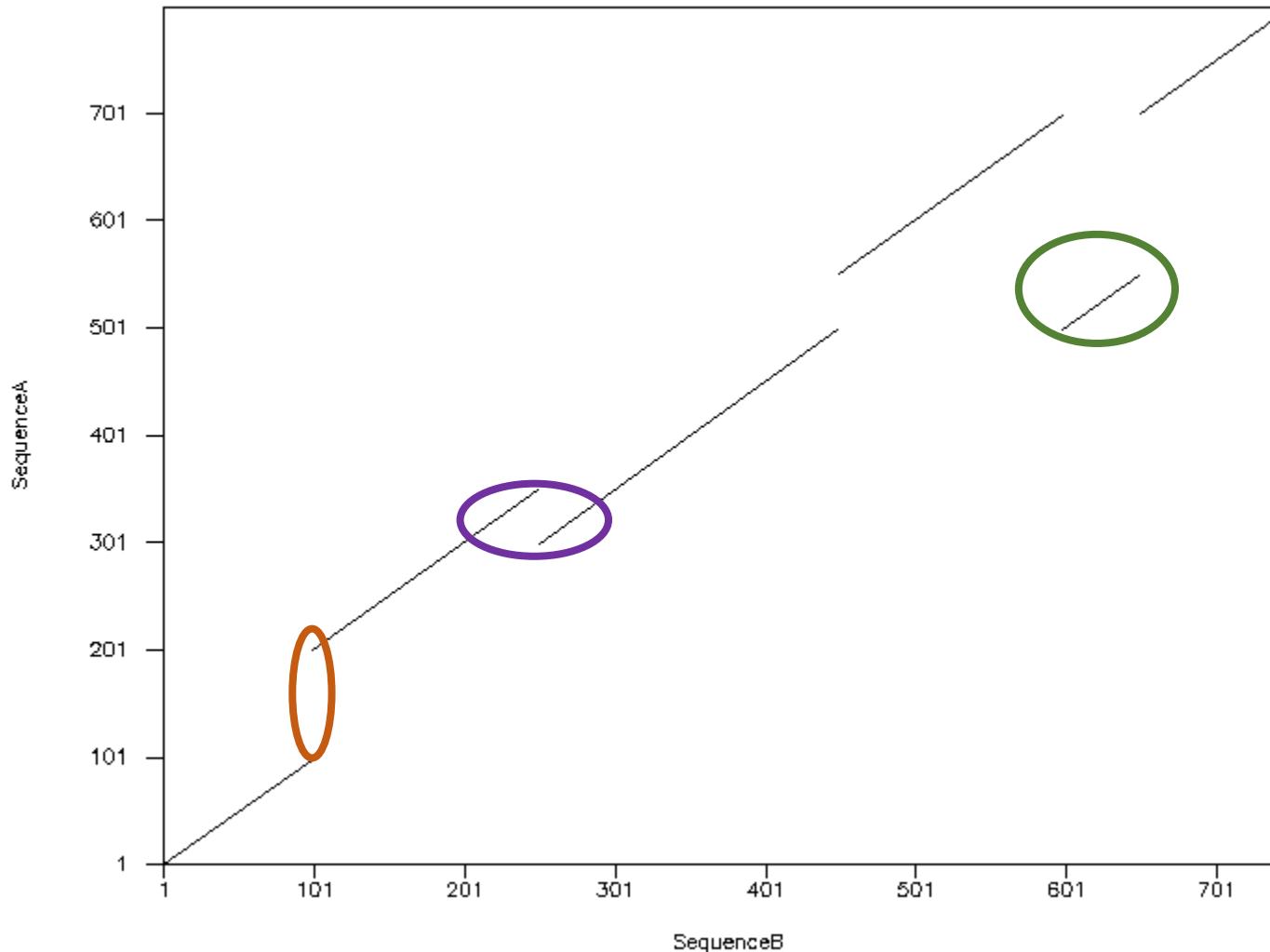
Wordsize 5



wordsize 10

# Exercise 2.3: Read a dotplot

Draw the dotplot of sequences A en B (wordsize 10) and describe what happened to sequence A to get to sequence B.



- Region 100-200 deleted in B
- Region 300-350 in A is duplicated in B
- Region 500-550 in A is translocated to 600-650 in B.

# Exercise 2.4: Exploring the NW and SW algorithms

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

>sequence 1

GLVGEIIGR

>sequence 2

GAVGLIIVR

Explanation of the algorithms can be found in the main course material. It is important that you understand these algorithms and are able to calculate the score of a global and local alignment.

# 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/psa>

(Here, you can use the default settings of the alignment tools. You will get more insight in the parameters in the following exercise).

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

# Exercise 2.5: Selecting the proper alignment tool (part 1)

Step 1: retrieve the protein sequences from Uniprot

<input type="checkbox"/> P78383	 S35B1_HUMAN	Solute carrier family 35 member B1, ATP/ADP exchanger ER, AXER, Endoplasmic reticulum ATP/ADP translocase, UDP-galactose transporter- related protein 1, UGTrel1	SLC35B1, UGTREL1	Homo sapiens (Human)
---------------------------------	---	---	---------------------	----------------------

■ Entry ▲	Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲	Length ▲
<input type="checkbox"/> Q4WJM7	 HUT1_ASFPFU	UDP-galactose transporter homolog 1	hut1, AFUA_1G05440	Neosartorya fumigata (strain ATCC MYA-4609 / Af293 / CBS 101355 / FGSC A1100) (Aspergillus fumigatus)	415 AA

# Exercise 2.5: Selecting the proper alignment tool (part 1)

## Step 2: choose the right alignment method

### Pairwise Sequence Alignment

EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences

STEP 1 - Enter your protein sequences

Enter a pair of

PROTEIN

sequences. Enter or paste your first **protein** sequence in any supported format:

```
>sp|P78383|S35B1_HUMAN Solute carrier family 35 member B1 OS=Homo sapiens OX=9606 GN=SLC35B1 PE=1 SV=1
MASSSSLPDPDRLLPLCFLGVFVCFYYGILQEKIRGKYGEGAKQETFFALTFLVFIQC
VINAFAKILQFFDTARVDRTRSWLYAACISYLGAMVSSNSALQFVNYPQTQVLGKSCK
PIPVMLLGVTLKKYPLAKYLCVLLIVAGVALFMYKPKVVVGIEEHTVGYGELLLLSL
TLDGLTGSQDHMRAYQTGSNHMMMLNINLWSTLLGMGILFTGELWEFLSFAERYPAII
YNILLFGLTSALGQSFIMFTVVFGLTCIITTRKFTTILASVILFANPISPMQWWGT
VLVFLGLGLDAKFGKGAKKTSH
```

Or, upload a file:

Use a example sequence | Clear sequence | See more example inputs

AND

Enter or paste your second **protein** sequence in any supported format:

```
>sp|Q4WJM7|HUT1_ASPFU UDP-galactose transporter homolog 1 OS=Neosartorya fumigata (strain ATCC MYA-4609 / Af293 / CBS 101355 / FGSC A1100)
OX=330879 GN=hut1 PE=3 SV=1
MHLVPEGSESMSTQQNGSAQKPVTLNGSASTKGQAPEAPLETGPLIQLAICVLGIYASFL
SWGVLQEAITTVNFPVRPPTAEENPPTTERFTSIVLNTIQSTFAITGFLYFSTPAG
KKVPSIFPTRKILFPLLVSISSSLASPFYASLAHIDYLTFILAKSCKLLPVMFLHLTI
FRKTYPLYKYGVVLLVTLGVATFTLHHPGTSKKVAASAAKNQSGSSLYGIFLSSINLLLD
GLTNTTQDHVFSSPQIYTRFTGPQMVAQNILSTILTTYLLVMMPHSSTGALHALLPIP
IPRSTETELASAVGELSPRUEVNLVWCEACACACOLEEVTLSPESLUVVWTATPK
```

Or, upload a file:

# Exercise 2.5: Selecting the proper alignment tool (part 1)

## Step 3: interpret the results

S35B1_HUMAN	9 PDRRLRLPLCFLGVFVCYFYGYILQEKITRGKY-----GEGAKQETFT	50
HUT1_ASPPU	43 PGLIQLAICVLGIYASFLSWGVLQEAITTVNFPVRPPTAEPPNPPTERFT	92
S35B1_HUMAN	51 FALTLVFIQCVINAVFAKILIQFFDTA----RVDRTRSWLYAACSISYL	95
HUT1_ASPPU	93 FSIVLNTIQSTFAAITGFLYLYFSTPAGKKVPSIFPTRKILFPLLLVSIS	142
S35B1_HUMAN	96 GAMVS--SNSALQFVNYPQTQVLGKSCKPIPVMILLGVTLKKYPLAKYLC	143
HUT1_ASPPU	143 SSLASPFPGYASLAHIDYLTIFLAKSKLLPVMFLHLTIFRKTYPLYKYGV	192
S35B1_HUMAN	144 VLLIVAGVALF-MYKP---KKV-VGIEEHTVG---YGELLLLLSLTLDGL	185
HUT1_ASPPU	193 VLLVTLGVATFTLHHPGTSKKVAASAACKNQSGSSLYGIFLLSINLLDGL	242
S35B1_HUMAN	186 TGVSQDHMRAHYQ----TGSNHMMLNINLWSTLLLGMGILF-----TG	224
HUT1_ASPPU	243 TNTTQDHVFSSPQIYTRFTGP-QMMVAQNILSTILTTYLLVMPHLSSTG	291
S35B1_HUMAN	225 -----ELWEFLSFAERYPAIIYNILLFGLTSALGQSFIG	258
HUT1_ASPPU	292 ALHALLPIPIPPSTETELASAVSFLSRHPEVMKNVLGFAACGAIGQLFIF	341
S35B1_HUMAN	259 MTVVYFGPLTCIITTRKFFTILASVILFANPISMQWVGTVLVFLGLG	308
HUT1_ASPPU	342 YTLSRFSSLLLVTVTVTRKMLTMLLSVFWFGHTLSAGQWLGIQLVFGGIG	391
S35B1_HUMAN	309 LDAKFGKGAKKT 320	
HUT1_ASPPU	392 AEAVVQKREKQS 403	

# Exercise 2.5: Selecting the proper alignment tool (part 2)

Step 1: retrieve the protein sequences from Uniprot

The screenshot shows the UniProtKB search interface. The search bar contains the query "b-cell translocation gene 1". Below the search bar are buttons for "Advanced", "List", and "Search". To the right of the search bar is a download icon.

## UniProtKB 14,584 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism
P62324	BTG1_HUMAN	Protein BTG1[...]	BTG1	Homo sapiens (Human)
P41182	BCL6_HUMAN	B-cell lymphoma 6 protein[...]	BCL6, BCL5, LAZ3, ZBTB27, ZNF51	Homo sapiens (Human)
Q4VC05	BCL7A_HUMAN	B-cell CLL/lymphoma 7 protein family member A	BCL7A	Homo sapiens (Human)
Q8WV28	BLNK_HUMAN	B-cell linker protein[...]	BLNK, BASH, SLP65	Homo sapiens (Human)
P62325	BTG1_MOUSE	Protein BTG1[...]	Btg1	Mus musculus (Mouse)

# Exercise 2.5: Selecting the proper alignment tool (part 2)

## Step 2: choose the right alignment method

### Pairwise Sequence Alignment

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

STEP 1 - Enter your protein sequences

Enter a pair of

PROTEIN

sequences. Enter or paste your first **protein** sequence in any supported format:

```
>sp|P62324|BTG1_HUMAN Protein BTG1 OS=Homo sapiens OX=9606 GN=BTG1 PE=1 SV=1
MHPFYTRAATMIGEIAAAVSFISKFLRTKGLTSERQLQTFSQSLQELLAEHYKHHWFPEK
PCKGSGYRCIRINHKMDPLIGQAQRIGLSSQELFRLLPSELTWVDPYEVSYRIGEDGS
ICVLYEASPAGGSTQNSTNVQMVDSRISCKEELLGRTSPSKNYNMMTVSG
```

Or, upload a file:  Geen bestand gekozen

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

**AND**

Enter or paste your second **protein** sequence in any supported format:

```
>sp|P62325|BTG1_MOUSE Protein BTG1 OS=Mus musculus OX=10090 GN=Btg1 PE=1 SV=1
MHPFYTRAATMIGEIAAAVSFISKFLRTKGLTSERQLQTFSQSLQELLAEHYKHHWFPEK
PCKGSGYRCIRINHKMDPLIGQAQRIGLSSQELFRLLPSELTWVDPYEVSYRIGEDGS
ICVLYEASPAGGSTQNSTNVQMVDSRISCKEELLGRTSPSKNYNMMTVSG
```

Or, upload a file:  Geen bestand gekozen

# Exercise 2.5: Selecting the proper alignment tool (part 2)

## Step 3: interpret the results

```
#=====
#  
# 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 ELTLWVDPYEVSYRIGEDGSICVLYEASPAGGSTQNSTNVQMVDUSRISCK	150
BTG1_MOUSE	101 ELTLWVDPYEVSYRIGEDGSICVLYEASPAGGSTQNSTNVQMVDUSRISCK	150
BTG1_HUMAN	151 EELLLGRTSPSKNYNMMTVSG 171	
BTG1_MOUSE	151 EELLLGRTSPSKNYNMMTVSG 171	

# Exercise 2.5: Selecting the proper alignment tool

## Summary

Global alignment:

- align two sequences in their entirety (start to end)
- search for differences in similar sequences with similar length

Local alignment:

- searches for a partial overlap somewhere in the sequences
- 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 using a local alignment tool.
- Different sets of parameters
- Influence of gap opening penalty, gap extension penalty, substitution matrix?

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

- Sequences:

```
>sp|Q02539|H11_HUMAN Histone H1.1 OS=Homo sapiens OX=9606 GN=HIST1H1A PE=1 SV=3  
MSETVPPAPAASAAPEKPLAGKKAKKPAKAAAASKKPAGPSVSELIVQAASSSKERGGVSLAALKALAAAGY  
DVEKNNSRIKLGKSLVSKGTLVQTKGASGSFKLNKKASSVETK PGASKVATKTKATGASKKLKKATGASKKSV  
KTPKKAKKPAATRKSSKNPKPKTVPKKKV AKSPAKAKAVPKAAKARVTKPKTAKPKKAAPKKK
```

```
>sp|P0ACF8|HNS_ECOLI DNA-binding protein H-NS OS=Escherichia coli (strain K12) OX=83333 GN=hns PE=1 SV=2  
MSEALKILNNIRTLRAQARECTLETLEEMLEKLEVVVNERREESAAAAEVEERTRKLQQYREMLIADGIDPNEL  
LNSLAAVKSGTKAKRAQRPAKYSYVDENGETKTWTGQGRTPAVIK KAMDEQGKSLDDFLIKQ
```

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

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

```
=====
#
# 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 ATRKSSKNPKPKTVKPKKVAKSPAKAKAVKPKAAKA	197
	.... . .....: .....:::  .. .....: :	
HNS_ECOLI	93 RPAKYSYVDENGETKTWTGQGRTPAVIKKAMDEQGKS	129

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

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

```
#=====
H11_HUMAN      45 ELIVQ-----AASSSKERGGVSLAALKKALAAAGYDVEKNNSRIKLG      86
                  |::|.|    ||:...||.. .|....:|.|| ..|   :|
HNS_ECOLI      34 EVVVNERREEEESAAAAEVEERTR-KLQQYREMLIADG--IDPN---EL-      75
                  " # Length: 122
H11_HUMAN      87 IKSLVSKGTLVQTKGAGGSFKLNKKASSVETKPGA-SKV----ATKTK      131
                  :..||..|  |:: ||.|       |:|.  :|.. |.|   .|||
HNS_ECOLI      76 LNSLAA---VKS-GTKA-----KRAQ---RPAKYSYVDENGETKT-      108
                  " # Identity:      33/122 (27.0%)
H11_HUMAN      132 ATGASKK---LKKATGASKKSV      150
                  .|||...|  :|||.....||:
HNS_ECOLI      109 WTGQGRTPAVIKKAMDEQGKSL      130
                  " # Similarity:   51/122 (41.8%)
                  " # Gaps:        41/122 (33.6%)
                  " # Score:       75.0
```

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

blosum62, gap opening penalty of 5, **gap extension penalty of 5**

H11_HUMAN	70 AAAGYDVEKNNSRIKLGKSLVKGTQVTKGTGASGSFKLNKKASSVET	119	# Length: 88
	: ... :..... ...  :..... ...  ....		# Identity: 21/88 (23.9%)
HNS_ECOLI	45 SAAAAEVEERTRKLQQYREMLIADG-IDPNELLNSLAALKSGTKAKRAQ-	92	# Similarity: 40/88 (45.5%)
H11_HUMAN	120 KPGA-SKV-AT-KTKA-TGASK--K-LKKATGASKKSV	150	# Gaps: 9/88 (10.2%)
	: .. . ...:  .  ... .::   ....  :		# Score: 44.0
HNS_ECOLI	93 RPAKYSYVDENGETKTWTGQGRTPAVIKKAMDEQGKSL	130	

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

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

H11_HUMAN	79	NNSRIKLGKSLVSKGTLVQTKGTGASGSFKLNKK-----ASSVETKPGAA   . . . ::::: .   .....: .: .  : .  : .  : .  :	123	
HNS_ECOLI	9	NNIRT---LRAQARECTLETLEEMLEKLEVNVNERREEESAAAEEVERT	55	# Length: 124
H11_HUMAN	124	SKVATKTKATGASKKLKKATGASKKSVKTPKKAKKPAATRKSSKNPKPK .  : : .: : .  . .: .: .  .:     .    : .: .  .  .: .: .	173	# Identity: 23/124 (18.5%) # Similarity: 58/124 (46.8%) # Gaps: 8/124 ( 6.5%)
HNS_ECOLI	56	RKLQQYREMLIADGIDPNELLNSLAALKSGTKAKRAQRPAKYSYVDENGE	105	# Score: 83.0
H11_HUMAN	174	TVKPKKVAKSPAKAKAVKPKAAKA   .. .: .:     ..   .: .: .  :	197	
HNS_ECOLI	106	TKTWTGQGRTPAVIKKAMDEQGKS	129	

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

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

H11_HUMAN	118	ETKPGASKVATKTAKTGA  ..... ... ...	ATG-----ASKKS :  :: . .  .. ..	151	# Length: 59	15/59 (25.4%)
HNS_ECOLI	43	EESAAAEEVEERT-----RKLQQYREMLIADGIDPNELLNSLAAVKSGT		86	# Identity: 24/59 (40.7%)	24/59 (40.7%)
H11_HUMAN	152	TPKKAKKPA 160			# Similarity: 22/59 (37.3%)	22/59 (37.3%)
HNS_ECOLI	87	KAKRAQRPA 95			# Gaps: # Score: 39.5	

At the end of this chapter, 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)

# Section 3

## Multiple sequence alignment

# Exercise 3.1: Progressive versus iterative methods

Let's try to make a multiple sequencing alignment of 5 sequences.

Use Clustal Omega, T-Coffee, MUSCLE and MAFFT to perform a multiple sequencing alignment. 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?**

**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 3.1: Progressive versus iterative methods

## Progressive methods

### T-Coffee

CLUSTAL W (1.83) multiple sequence alignment

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

### Clustal Omega

CLUSTAL O(1.2.4) multiple sequence alignment

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

## Iterative methods

### MUSCLE

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

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

### MAFFT

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

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

## 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?

# Exercise 3.2: Comparison of orthologous gene products

Step 1: search the protein sequences on Uniprot

The screenshot shows the UniProtKB search interface. At the top, there are tabs for "Peptide search", "ID mapping", "SPARQL", and "UniProtKB". A search query "(protein\_name:'histone h1.1') AND (reviewed:true)" is entered in the search bar. Below the search bar, the results are displayed under the heading "UniProtKB 9 results". The results table has columns: "Entry", "Entry Name", "Protein Names", "Gene Names", and "Organism". The entries are:

Entry	Entry Name	Protein Names	Gene Names	Organism
Q02539	H11_HUMAN	Histone H1.1[...]	H1-1, H1F1, HIST1H1A	Homo sapiens (Human)
P10771	H24_CAEEL	Histone 24[...]	his-24, H1.1, HH1, M163.3	Caenorhabditis elegans
D4A3K5	H11_RAT	Histone H1.1[...]	H1-1, Hist1h1a	Rattus norvegicus (Rat)
P43275	H11_MOUSE	Histone H1.1[...]	H1-1, H1a, H1f1, Hist1h1a	Mus musculus (Mouse)
G3N131	H11_BOVIN	Histone H1.1[...]	H1-1, HIST1H1A	Bos taurus (Bovine)

Step 2: Select the Human, Rat, Mouse and Bovin results and Download as Fasta (canonical) uncompressed

# Exercise 3.2: Comparison of orthologous gene products

## Step 2: Perform multiple alignment

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

sp P43275 H11_MOUSE	MSETAPVAQAASTATEKPAAKTKKPAK-AAAPRKKPAGPSVSELIVQAVSSSKERSGV
sp D4A3K5 H11_RAT	MSETAPVPQPASVAPEKPAATKKTRKPAK-AAVPRKKPAGPSVSELIVQAVSSSKERSGV
sp Q02539 H11_HUMAN	MSETVPPAPAASAAPEKPLAGKKAKKPAKAAAASKKPAGPSVSELIVQAASSSKERGGV
sp G3N131 H11_BOVIN	MSEVALPAPAASTSPEKPSAGKKAKKPAKAAAAAKKKPAGPSVSELIVQAVSSSKERSGV

sp P43275 H11_MOUSE	SLAALKS LAAGYDVEKNNSRIKGLKL SVNKG TLVQT KGTGAAGSF KLN KKA --- ESK
sp D4A3K5 H11_RAT	SLAALKS LAAGYDVEKNNSRIKGLKL SVNKG TLVQT KGTGAAGSF KLN KKA --- ESK
sp Q02539 H11_HUMAN	SLAALKK ALAAGYDVEKNNSRIKLG IKS L VSK G TLVQT KGTGASGSF KLN KKASS VETK
sp G3N131 H11_BOVIN	SLAALKK ALAAGYDVEKNNSRIKGLKSLVGK GTL VQT KGTGASGSF KLN KKVAS VDAK

sp P43275 H11_MOUSE	AITTKV SVKAKASGA AKKPKKT AG-AAAKKT V KTPKKPKPAV SKK-TSKSPKKPV VKA
sp D4A3K5 H11_RAT	ASTTKV TVKAKASGA AKKPKKT AG-AAAKKT V KTPKKPKPAV SKKTSSKSPKKPV VKA
sp Q02539 H11_HUMAN	PGASKV ATKT KATG ASK KKL KKAT G-- ASKKS V KTPKKAKKPA ATRK-SSKNP KKPKTV KP
sp G3N131 H11_BOVIN	PTATKV ATKT KV TSAS KKP KK ASG AAAKKS V KTPKKARKS VL T KK-SSKSPKKPKA V KP

sp P43275 H11_MOUSE	KKVAK SPAKAKA KV PK KASKA KV T KPK TPAKPKKA APKKK
sp D4A3K5 H11_RAT	KKVAK SPAKAKA KV PK KAA KV VT KPK TPAKPKKA APKKK
sp Q02539 H11_HUMAN	KKVAK SPAKAKA KV PK KAA KAR V T KPK T-AKP KKA APKKK
sp G3N131 H11_BOVIN	KKVAK SPAKAKA KV PK GAK KV VT KPK TAA KPK KA APKKK

At the end of this chapter, you should:

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

# 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 Overview\_question. In this exercise, you will examine these sequences to see whether his son has inherited the mutation.

# Overview question

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?

# Overview question

(1) Identify all the SNPs in the sequences

Step 1: Get the reference sequence of the FGF8 gene

FASTA ▾

Send to: ▾

## Homo sapiens fibroblast growth factor 8 (FGF8), RefSeqGene on

NCBI Reference Sequence: NG\_007151.1

[GenBank](#) [Graphics](#)

>NG\_007151.1:4957-10940 Homo sapiens fibroblast growth factor 8 (FGF8), RefSeqGene  
on chromosome 10

```
AGCCAGCGGCCACCGCTCCGGCACAGCGATTGGTGCAGCGGCGAGCACGACGTTCCAC
GGGACCCGGAGCCGGTGTGATGCCGCCCTCCGCACCCGACCCCTCTCCGCTCCGCCCTGC
TCAGCGCGTCCTCCGGCCGGGGACGGCGTGACCCGGGGCTCTGGTCCCCGGGGCGCG
CGCCATGGCAGCCCCCGCTCCGCCTGAGCTGCGTGTGACTACCGCGCCGCCCTGCCCCGCCACCCGCC
CCCCGGCCCTCGCCTCACCCGCCTCTCTCTCCGCCCTTTGTCTCCCACAGGCTGTTGCACTT
GCTGGTCTCTGCCCTCCAAGCCCAGGTGAGGAGGGTGCGCCGAGGCGGGGGCCGGCGCCGGTGTGA
GACCCGGGTGGGCAGGCCGGTGGGGGACCGGGACTGACTCTGGCCGGGGAGGGCTGAGGGC
ACCTTAGAAACCCAGCCCCGAGCCACCCGGAGGAGGGAGCTGAGGCACAGAGAGGTAGCACCCCTCTGA
GGTCACACAGCGACTGACTGCCAGGATAAACGAAGGTCTGGAGCCAGCACTGTCCCCATGCATC
```

Complete Record  
 Coding Sequences  
 Gene Features

### Choose Destination

File  Clipboard  
 Collections  Analysis Tool

Download 1 item.

Format

Show GI

# Overview question

(1) Identify all the SNPs in the sequences

Step 2: Find the genetic variations through a local alignment

Alignment of allele 1 with the reference gene

NG_007151.1	5251	CCTCACCCGCGCCTCTCTCTCCCCGGCCGCTTTGTCTCCCACAGGGCTGTT	5300
allele1	1	CCTCACCCGCGCCTCTCTCTCCCCGGCCGCTTTGTCTCCCACAGGGCTGTT	50
NG_007151.1	5301	GCACATTGCTGGTCCCTCTGCC	5320
allele1	51	GCACATTGCTGGTCCCTCTGCC	70

Alignment of allele 2 with the reference gene

NG_007151.1	5251	CCTCACCCGCGCCTCTCTCTCCCCGGCCGCTTTGTCTCCCACAGGGCTGTT	5300
allele2	1	CCTCACCCGCGCCTCTCTCTCCCCGGCCGCTTTGTCTCCCACAGGGCTGTT	50
NG_007151.1	5301	GCACATTGCTGGTCCCTCTGCC	5320
allele2	51	GAACATTGCTGGTCCCTCTGCC	70

# Overview question

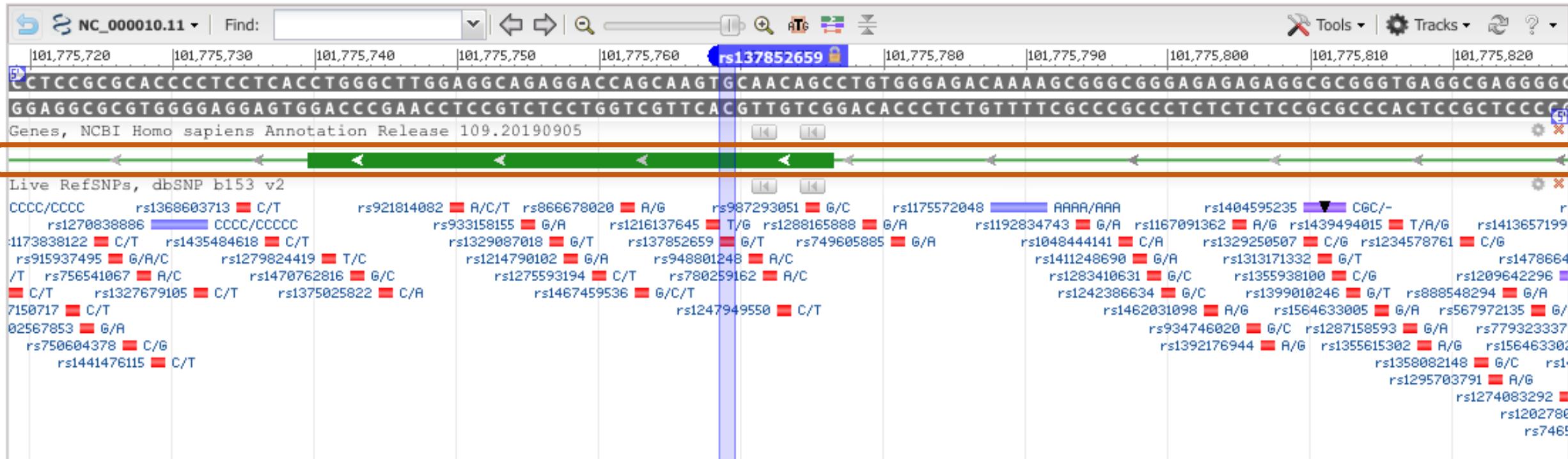
(2) Check whether the mutation in allele 2 was inherited from the father



# Overview question

## (3) Short questions:

*On which strand is the gene located?*



# Overview question

## (3) Short questions:

*On which strand is the gene located?*

Gene: <a href="#">FGF8</a> , 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

# Overview question

## (3) Short questions:

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

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

<b>Clinical Significance</b>	Reported in <a href="#">ClinVar</a>
<b>Gene : Consequence</b>	FGF8 Missense Variant
<b>Publications</b>	1 citation
<b>Genomic View</b>	<a href="#">See rs on genome</a>

# Overview question

## (3) Short questions:

*Can you find other genomic variations in this gene that are associated with a disease?*

The screenshot shows the NCBI dbSNP search interface. The search query is '(FGF8[Gene Name]) AND Homo Sapiens[Organism]'. The results are filtered by 'pathogenic' clinical significance. There are 7 items found.

**Search results**  
Items: 7

Filters activated: pathogenic. Clear all to show 3302 items.

1. rs137852659 [Homo sapiens]

Variant type:	SNV
Alleles:	G>T [Show Flanks]
Chromosome:	10:101775769
Gene:	FGF8 (Varview)
Functional Consequence:	coding_sequence_variant, 5_prime_UTR_variant, missense_variant
Clinical significance:	pathogenic
Validated:	by cluster
HGVS:	NC_000010.11:g.101775769G>T, NC_000010.10:g.103535526G>T, NG_007151.1:g.5302C>A, NM_006119.4:c.40C>A, NM_006119.5:c.40C>A, NM_033163.3:c.40C>A, NM_033163.4:c.40C>A, NM_033164.3:c.40C>A, NM_033164.4:c.40C>A, NM_033165.3:c.40C>A, NM_033165.4:c.40C>A, NM_001206389.1:c.-153C>A, NP_006110.1:p.His14Asn, NP_149353.1:p.His14Asn, NP_149354.1:p.His14Asn, NP_149355.1:p.His14Asn

Send to: Filters: Manage Filters

Find related data  
Database: Select

Find Items

Search details  
FGF8[Gene Name] AND "Homo sapiens"  
[Organism] AND  
pathogenic[Clinical\_Significance]

Search See more...

Recent activity Turn Off Clear

Search (FGF8[Gene Name]) AND Homo sapiens[Organism] AND /pathogenic[Clinical\_Significance]