Dear teacher/student

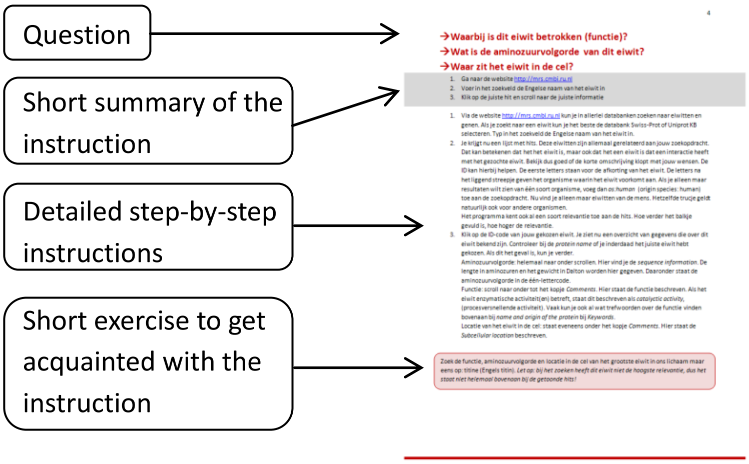
This booklet is the NAVIGENE. It is a navigation tool developed for biology teachers at secondary schools, developed by the Netherlands Bioinformatics Centre and the Freudenthal Institute for Science and Mathematics Education. Bioinformatics is a field of research that focusses on DNA and proteins. You can use bioinformatics tools and databases to demonstrate authentic research in your classroom. Using this software, you can improve your students' insight in DNA and proteins by giving them a visual representation of genetics and protein structure.

Under construction

Please note that the instrument is still evolving: bioinformatics is a dynamic field of research, so hyperlinks to websites and website lay-outs may change. We'd like to further develop NAVIGENE with your help. Should you find a corrupted hyperlink, a changed website lay-out or an error in the instructions, please let us know. Other feedback, comments or wishes are always welcome. Together we can improve NAVIGENE.

Using NAVIGENE

NAVIGENE consists of a cover page and an instructional guide. On the cover page you can find several questions. The main structure is the "phenomenon-protein-gene" triangle: following your discovery of a protein or cellular process in the biology handbook that you want to clarify, you may choose where to start. From this point on, you can simply follow the arrows and the corresponding instructions. At the bottom of the scheme, you will find questions which can be investigated using bioinformatics. For each arrow, there is a question and a number, for example P.4. This means that you can answer this question with the instructions on page 4. The tutorial is structured as shown in the figure.


Question
Short summary of the instruction
Detailed step-by-step instructions
Short exercise to get acquainted with the instruction
4

→Waarbij is dit eiwit betrokken (functie)?
→Wat is de aminozuurvolgorde van dit eiwit?
→Waar zit het eiwit in de cel?

1. Ga naar de website http://mrs.cmbi.ru.nl
2. Voer in het zoekveld de Engelse naam van het eiwit in
3. Klik op de juiste hit en scroll naar de juiste informatie

1. Via de website http://mrs.cmbi.ru.nl kun je in allerlei databanken zoeken naar eiwitten en genen. Als je zoekt naar een eiwit kun je het beste de databank Swiss-Prot of Uniprot KB selecteren. Typ in het zoekveld de Engelse naam van het eiwit in.
2. Je krijgt nu een lijst met hits. Deze eiwitten zijn allemaal gerelateerd aan jouw zoekopdracht. Dat kan betekenen dat het het eiwit is, maar ook dat het een eiwit is dat een interactie heeft met het gezochte eiwit. Bekijk dus goed of de korte omschrijving klopt met jouw wensen. De ID kan hierbij helpen. De eerste letters staan voor de afkorting van het eiwit. De letters na het liggend streepje geven het organisme waarin het eiwit voorkomt aan. Als je alleen maar resultaten wilt zien van één soort organisme, voeg dan *os:human* (origin species: human) toe aan de zoekopdracht. Nu vind je alleen maar eiwitten van de mens. Hetzelfde trucje geldt natuurlijk ook voor andere organismen.
   Het programma kent ook al een soort relevantie toe aan de hits. Hoe verder het balkje gevuld is, hoe hoger de relevantie.
3. Klik op de ID-code van jouw gekozen eiwit. Je ziet nu een overzicht van gegevens die over dit eiwit bekend zijn. Controleer bij de *protein name of* je inderdaad het juiste eiwit hebt gekozen. Als dit het geval is, kun je verder.
   Aminozuurvolgorde: helemaal naar onder scrollen. Hier vind je de *sequence information*. De lengte in aminozuren en het gewicht in Dalton worden hier gegeven. Daaronder staat de aminozuurvolgorde in de één-lettercode.
   Functie: scroll naar onder tot het kopje *Comments*. Hier staat de functie beschreven. Als het eiwit enzymatische activiteit(en) betreft, staat dit beschreven als *catalyctic activity*, (procesversnellende activiteit). Vaak kun je ook al wat trefwoorden over de functie vinden bovenaan bij *name and origin of the protein* bij *Keywords*.
   Locatie van het eiwit in de cel: staat eveneens onder het kopje *Comments*. Hier staat de *Subcellular location* beschreven.

Zoek de functie, aminozuurvolgorde en locatie in de cel van het grootste eiwit in ons lichaam maar eens op: titine (Engels titin). *Let op: bij het zoeken heeft dit eiwit niet de hoogste relevantie, dus het staat niet helemaal bovenaan bij de getoonde hits!*

To get acquainted with NAVIGENE, you can scan through this instruction guide. The red or green text blocks at the bottom of each tutorial contain exercises which will help you understand the purpose of the tutorial. You can also contact us if you wish to have more information about our next NAVIGENE workshop. We can also provide a custom workshop for your school.

We wish you many useful discoveries and valuable surprises when using NAVIGENE. Suggestions, questions, information about workshops and other comments can be addressed to onderwijs@nbic.nl.

The NAVIGENE is developed by Hienke Sminia in collaboration with Dirk Jan Boerwinkel.

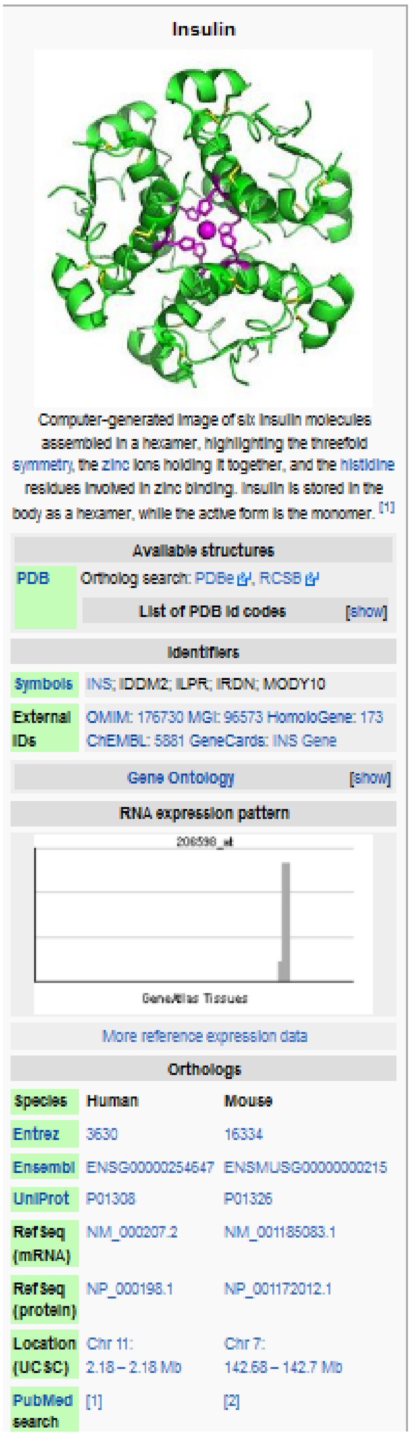
nbic netherlands bioinformatics centre

[Faculteit Bètawetenschappen
FIsME]

→ Which protein or gene is involved in this biological phenomenon?

1. Use Google (www.google.com) to search for information on a biological phenomenon.
2. Scan this information for genes.
3. You may want to use Wikipedia to get additional information.

1. Proteins and genes play a major role in determining characteristics such as the color of your hair and eyes, the development of syndromes and illnesses (Huntingtons disease, sickle cell anemia, color blindness) and biological processes (photosynthesis, digestion, insulin production). As of now there exists no single database where one can easily find genes and proteins that are involved in a disease, a process or a characteristic body feature. This information is currently scattered among many different databases, websites of research institutes and online encyclopedias. Search engines like Google enable us to search all these sources simultaneously. Visit http://www.google.com.
2. Use the desired phenomenon as a query in Google. You may want to specify the search term with terms such as 'gene' or 'protein'. Multiple word queries can be submitted using quotation marks (example: "Huntington's disease"). Click 'Google Search' and scan the resulting websites for information regarding the genes and proteins that are involved. This information can be checked by comparing it to the information that is stored in a relevant database: see also page 5 *What is the function of this protein?* or page 6 *In which cellular processes is this gene involved and what is its function?*
3. Wikipedia can also be a valuable information source. One can search through this online encyclopedia by adding the term 'wiki' to a Google query or by using Wikipedia's own search engine. This engine can be accessed directly on http://en.wikipedia.org. You may find several pages that seem relevant. Often, the first one is the best hit.
4. A Wikipedia entry on a single protein often contains a list of pathways, processes and reactions in which the protein is involved. On the right side of the page a table is displayed. The figure on this page is an example of such a table. This table often contains the following information:
   - An image of the 3D-structure of the protein with a description. The description can contain, for example, the name of the organism from which the protein or 3D structure originates.
   - *Available structures*: (click *show*) PDB is the abbrevation for Protein DataBank (see also page 13: *What does the 3D-structure of the protein look like?*). Here you can find all ID-codes for pdb-files that contain the structure of the protein. Different files can contain



Insulin

Computer-generated image of six insulin molecules assembled in a hexamer, highlighting the threefold symmetry, the zinc ions holding it together, and the histidine residues involved in zinc binding. Insulin is stored in the body as a hexamer, while the active form is the monomer. [1]

Available structures

PDB	Ortholog search: PDBe, RCSB List of PDB id codes [show]

Identifiers

Symbols	INS; IDDM2; ILPR; IRDN; MODY10
External IDs	OMIM: 176730 MGI: 96573 HomoloGene: 173 ChEMBL: 5881 GeneCards: INS Gene

Gene Ontology [show]

RNA expression pattern

More reference expression data

Orthologs

Species	Human	Mouse
Entrez	3630	16334
Ensembl	ENSG00000254647	ENSMUSG00000000215
UniProt	P01308	P01326
RefSeq (mRNA)	NM_000207.2	NM_001185083.1
RefSeq (protein)	NP_000198.1	NP_001172012.1
Location (UCSC)	Chr 11: 2.18 – 2.18 Mb	Chr 7: 142.68 – 142.7 Mb
PubMed search	[1]	[2]

different configurations, mutated forms or different protein complexes. What structure should be used or viewed depends entirely on your purpose. Thus, there is no rule-of-thumb which structure should be chosen.
- *Identifiers*: Below the header *symbols* you can find several ID-codes. Although referring to the same protein, the ID-codes vary among different databases. Often, the first ID-code is the one that is used most often. You should use this one when working with MRS (see page 5).
- *Gene Ontology*: (click *show*) This is a list of molecular functions, biological processes and cellular components that are somehow associated with the protein.
- *RNA expression pattern*: This graph shows the abundance of the protein in different tissues. Higher bars in the graph indicate higher expression of the protein in this tissue. Multiple graphs point to different expression patterns in individuals or organisms. The graph can be enlarged by clicking on it.
- *Orthologs:* Summarizes information concerning the gene and contains a comparison between the human form and that of another organism, mostly the mouse.

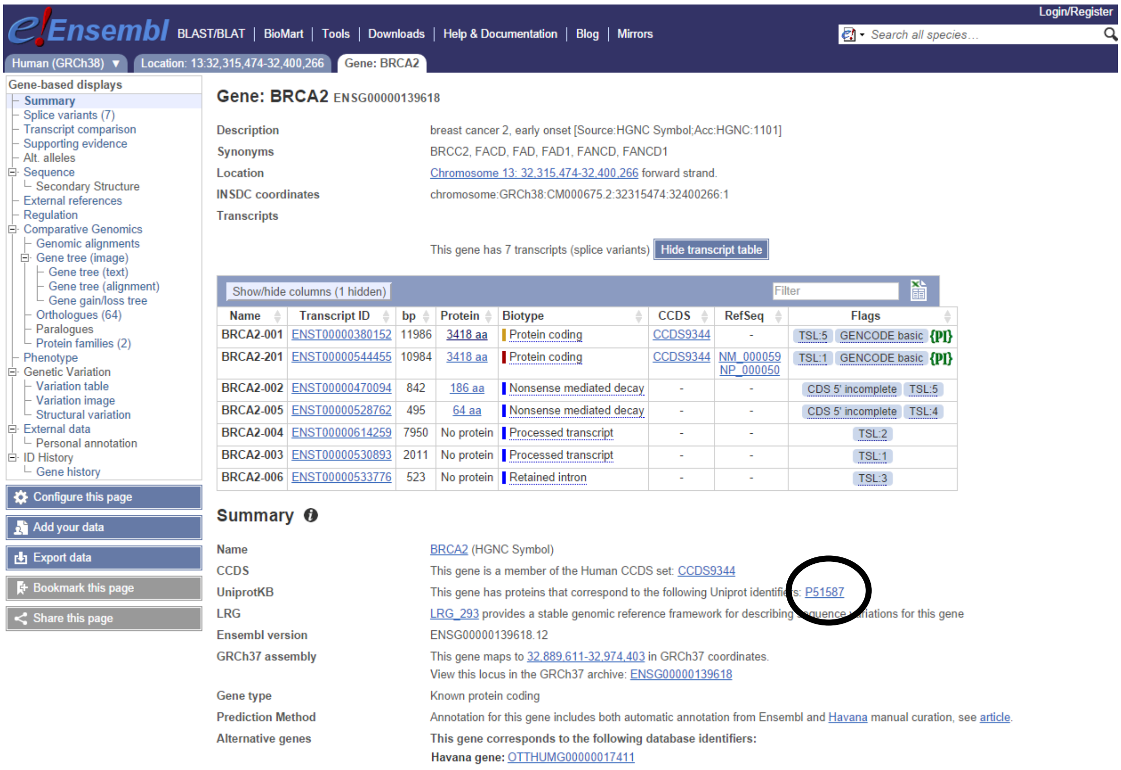
Entrez – a search engine for medical databases
Ensembl – database which contains genomes from multiple organisms (see page 14)
Uniprot – A database that combines the data of different protein databases.
RefSeq (mRNA) en *RefSeq (protein)* – Referential sequences for the mRNA and the protein
Location (UCSC) – The location of the gene on the chromosome (chromosome number and coordinates)
PubMed search – Search a database with articles of several different scientific journals. To view the article, you often need a subscription to the journal. Most universities have these subscriptions.

Try to find the enzyme that is involved in the secretion of gastric acid in your stomach during digestion.
Or you can try to find which gene is probably involved in your eye color. Can you also find in which other colorful phenomenon this gene is involved?

→ What protein does this gene encode?

1. Go to http://www.ensembl.org
2. Find the gene of interest
3. Click *Uniprot Identifiers*

1. Ensembl is a genome browser in which you can find all sorts of information about genes, for example: which protein does a gene encode? On page 18, you will find other applications of this genome browser. Note that you can only use Ensembl to search in vertebrates and other eukaryotes. Plant genes, for example, are not available in this database.
   Go to http://www.ensembl.org
2. Use the search function on the Ensembl home page to find your gene of interest. You can use different queries such as the name of the gene, the gene symbol or the coordinates of the gene's location. Click 'Go' to start the query.
3. You are now presented an overview of the hits. Search for the name of your gene of interest, followed by "*Human Gene*". If you are searching for a gene in a different organism, you should locate the name followed by "*[Species name] Gene*". Click this hit for more information: you will be directed to the gene's information page.
4. The information page shows a lot of information, for example the location or the amount of transcript of the gene etc. To find out what protein this gene encodes, find *UniprotKB* under the *Summary* header. Uniprot is a big protein database (analogue to Swissprot). If your gene of interest encodes a well-known protein, it will have an 'Identifier' in Uniprot. Clicking the link will directly send you to Uniprot, however, this doesn't always work. Alternatively, you can go to www.uniprot.org and copy-paste the Identifier code in the search field. This is then your query, which will lead you directly to the right protein information.



Gene: BRCA2 ENSG00000139618

Description: breast cancer 2, early onset [Source:HGNC Symbol;Acc:HGNC:1101]

Synonyms: BRCC2, FACD, FAD, FAD1, FANCD, FANCD1

Location: Chromosome 13: 32,315,474-32,400,266 forward strand.

INSDC coordinates: chromosome:GRCh38:CM000675.2:32315474:32400266:1

Transcripts

This gene has 7 transcripts (splice variants)

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq	Flags
BRCA2-001	ENST00000380152	11986	3418 aa	Protein coding	CCDS9344	-	TSL:5 GENCODE basic {PI}
BRCA2-201	ENST00000544455	10984	3418 aa	Protein coding	CCDS9344	NM_000059 NP_000050	TSL:1 GENCODE basic {PI}
BRCA2-002	ENST00000470094	842	186 aa	Nonsense mediated decay	-	-	CDS 5' incomplete TSL:5
BRCA2-005	ENST00000528762	495	64 aa	Nonsense mediated decay	-	-	CDS 5' incomplete TSL:4
BRCA2-004	ENST00000614259	7950	No protein	Processed transcript	-	-	TSL:2
BRCA2-003	ENST00000530893	2011	No protein	Processed transcript	-	-	TSL:1
BRCA2-006	ENST00000533776	523	No protein	Retained intron	-	-	TSL:3

Summary

Name: BRCA2 (HGNC Symbol)

CCDS: This gene is a member of the Human CCDS set: CCDS9344

UniprotKB: This gene has proteins that correspond to the following Uniprot identifiers: P51587

LRG: LRG_293 provides a stable genomic reference framework for describing sequence variations for this gene

Ensembl version: ENSG00000139618.12

GRCh37 assembly: This gene maps to 32,889,611-32,974,403 in GRCh37 coordinates.
View this locus in the GRCh37 archive: ENSG00000139618

Gene type: Known protein coding

Prediction Method: Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.

Alternative genes: **This gene corresponds to the following database identifiers:**
Havana gene: OTTHUMG00000017411

Find the protein that is encoded by the BRCA2 gene.

→ What gene encodes this protein?
→ What is the function of the protein?
→ What is the protein's primary structure?
→ In which place in the cell can the protein be found?

1. Visit http://mrs.cmbi.ru.nl
2. Enter the name of the protein in the search bar.
3. Select the best hit and scroll down to get to the information.

1. The website http://mrs.cmbi.ru.nl serves as a portal to search for genes and proteins in many different databases. When looking for proteins, the best databases are Swiss-Prot and Uniprot KB. Enter the name of the protein in the search bar.
2. You will probably end up with several hits. All proteins in this list are somehow related to the protein in your query. Use the description to determine if a protein is the one that you are looking for, or if it only interacts with the protein that you are interested in. The ID can also give you some clues. The first letters are an abbreviation of the name of the protein and the ones after the bar are related to the organism where that specific protein is found. By extending your query with *os:human* (origin species: human) you can look specifically for human proteins. The same goes for other organisms. The software gives a score to each hit, the larger the bar, the more relevant the hit.
3. Click the ID-code of your protein of interest. All information found in the database is listed. Check *protein name* to make sure that you have selected the right protein. If this is the case, move on. Depending on what you're looking for, look in these sections of the information page:
   Name of the encoding gene: in the second section (*'Name and origin of the protein'*), you will find the *Gene names*. The *Name* gives the most common name for the gene, *Synonyms* shows other names for the gene. Sometimes, only a gene symbol is given (an abbreviation of letters and numbers).
   Primary structure: scroll to the bottom of the page. Here you can find the tab *sequence information*. The proteins weight and length (in amino acids) are listed together with the amino acids composition.
   Function: scroll down until you find the tab *Comments*. Here you can find the function and enzymatic properties (indicated often as *function* or *catalytic activity*). The *Keywords* at the top of the page may also contain useful information.
   Location in the cell: the tab *Comments* also features a header *Subcellular location*.

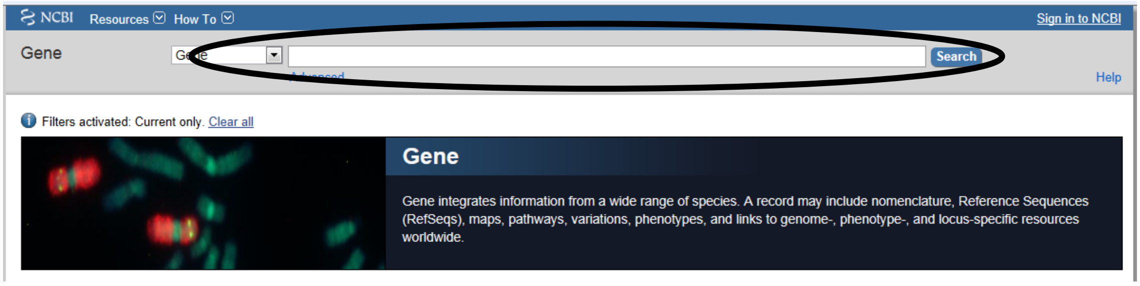
Find the function, amino acid sequence and location of the protein in a cell of the largest protein in our body: titin.

Please note: this protein doesn't have the highest relevance when searching with MRS, so it may not appear on top in the hitlist.

→ In which cellular processes is this gene involved and what is its function?

1. Go to http://www.ncbi.nlm.nih.gov/gene
2. Search for the desired gene.
3. Look up the genes 'General gene information'

1. The NCBI (National Centre for Biotechnology Information) takes care of several databases and offers many tools for searching them. ‘Gene’ is a database with a lot of information about genes such as their function, their location on the chromosomes, the pathways in which they are involved, their phenotypes and the variations in the gene.
2. Type your query in the search bar, which is indicated below. You can narrow down your results by adding the name of your species of interest.



Filters activated: Current only. Clear all

Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

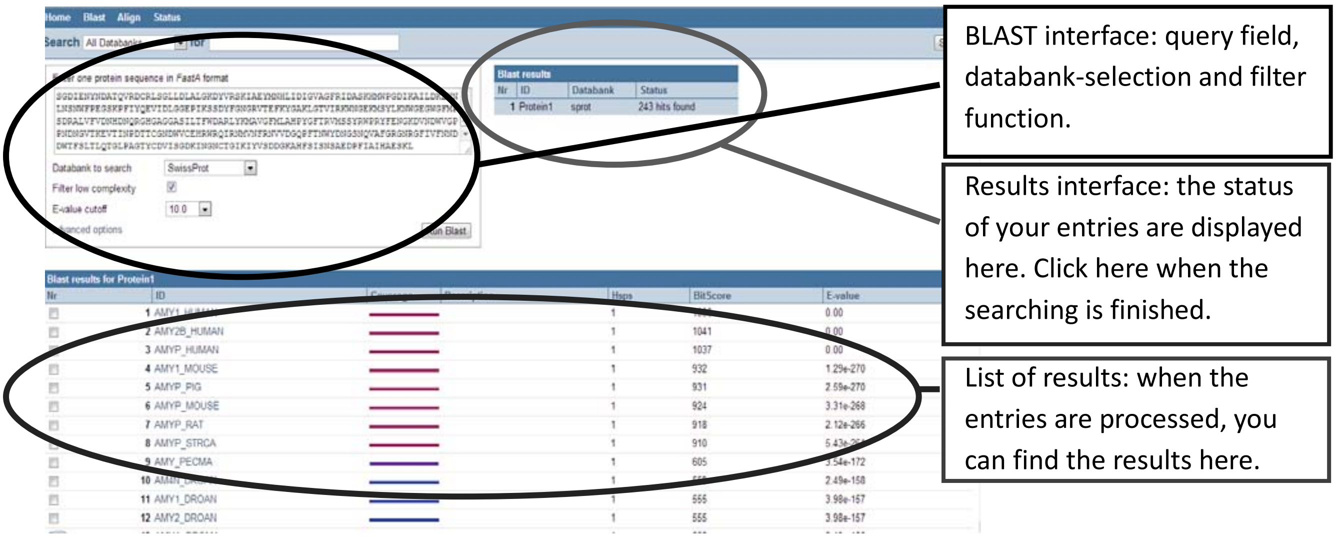
3. You will find a table with the gene symbol (a unique, universal code for every gene which consists of letters and numbers), a short description of the gene, its location in the genome and alternative names for the gene. Click the gene symbol of your gene of interest to view the full information page.
4. The information page of your gene of interest consists of different sections. To find out the function of the gene, look at ‘Summary’. For detailed information on the processes in which the gene is involved, look at the section ‘General gene information’.

Find the function of the *CONSTANS* gene in the *Arabidopsis thaliana* (thale cress).

→ What is the name of the protein?

1. Use the BLAST software at http://mrs.cmbi.ru.nl
2. Copy the amino acid sequence preceded by the query name (starting with a '>' sign) in the appropriate box.
3. Click the first hit and then this proteins ID code.

1. A great variety of bioinformatics tools can easily be found on the internet. For identifying an amino acid sequence one can use BLAST. This is essentially a search engine that can search through a number of databases and compare the submitted sequence to the ones that are stored there. It assigns a score to all alignments and the ones with the highest scores end up at the top of the search report. Beware, the sequence of the first hit is not always completely equal to the one that you submitted! Through the search report one can easily access a form with information on the protein and even links to other databases and literature.
2. Multiple BLAST tools are available. When looking for amino acid sequences, you should use the one that is developed by the Radboud University. It can be found at http://mrs.cmbi.ru.nl. Ensembl's tool is the most suitable when looking for proteins using DNA sequences. It can be accessed through http://www.ensembl.org/Multi/blastview (select *peptide queries* and then *peptide database*). However, the following instructions assume you are using Radboud University's BLAST tool.
3. Copy your amino acid sequence to the search field. Start the sequence query with a line *>nameofyoursequence.* You are now using the so called FastA-format, without which the search engine will not work. It is important to be precise, as the BLAST software is prone to inconsistent input.
4. Proceed by selecting the database that you would like to search. SwissProt (Swiss protein) is the most well-known, but you can also use Uniprot (Universal Protein).
5. Be aware of the 'Filter sequence' option. When this option is checked, BLAST filters low complexity sequences which are essentially large repeats of short sequences. This will result in less hits, because proteins with similarities in this domain only will not be taken into account. When searching for a well known protein you can safely uncheck the option.
6. Click 'BLAST' at the upper right corner of the screen. Your query can take a few minutes, especially if you submitted a very short sequence. BLAST automatically shows a 'finished' sign when it is finished. Click on the proper query, multiple ones can be displayed, to see the results.
7. Here, all hits, proteins that contain or are roughly equal to the the sequence that you have submitted, are listed. Hits are accompanied by a number of scores. The lower the E-value, the better the match. For the BitScore it is the other way around. Additional information can be



er one protein sequence in FastA format

```
SGDIENYNDATQVRDCRLSGLLDLALGKDYVRSKIAEYMNHLIDIGVAGFRIDASKHMWPGDIKAILDK
LNSNWFPEGSKPFIYQEVIDLGGEPIKSSDYFGNGRVTEFKYGAKLGTVIRKWNGEKMSYLKNWGEGWGF
SDRALVFVDNHDNQRGHGAGGASILTFWDARLYKMAVGFMLAHPYGFTRVMSSYRWPRYFENGKDVNDWVGP
PNDNGVTKEVTINPDTTCGNDWVCEHRWRQIRNMVIFRNVVDGQPFTNWYDNGSNQVAFGRGNRGFIVFNND
DWTFSLTLQTGLPAGTYCDVISGDKINGNCTGIKIYVSDDGKAHFSISNSAEDPFIAIHAESKL
```

Databank to search: SwissProt
Filter low complexity: ☑
E-value cutoff: 10.0
anced options

Blast results

Nr	ID	Databank	Status
1	Protein1	sprot	243 hits found

Blast results for Protein1

Nr	ID	Coverage	Description	Hsps	BitScore	E-value
□	1 AMY1_HUMAN			1	[illegible]	0.00
□	2 AMY2B_HUMAN			1	1041	0.00
□	3 AMYP_HUMAN			1	1037	0.00
□	4 AMY1_MOUSE			1	932	1.29e-270
□	5 AMYP_PIG			1	931	2.59e-270
□	6 AMYP_MOUSE			1	924	3.31e-268
□	7 AMYP_RAT			1	918	2.12e-266
□	8 AMYP_STRCA			1	910	[illegible]
□	9 AMY_PECMA			1	605	3.54e-172
□	10 [illegible]			1	[illegible]	2.49e-158
□	11 AMY1_DROAN			1	555	3.98e-157
□	12 AMY2_DROAN			1	555	3.98e-157

BLAST interface: query field, databank-selection and filter function.

Results interface: the status of your entries are displayed here. Click here when the searching is finished.

List of results: when the entries are processed, you can find the results here.

protein from the database. Clicking again shows you the alignment with 'q' standing for query and 's' for sequence. When an amino acid occurs in both sequences BLAST shows it between the 'q' and 's' line. A gap indicates the amino acid is missing in one of the sequences, a '+' indicates the amino acids differ, but that their characteristics are similar. All amino acids that are left out of the alignment are crossed out.

8. Click on the 'ID' of the protein that is most probably the one that you are looking for. Most of the time it is simply the first one. All information concerning this protein is listed. You can find its function at the 'Comments' tab (*function* or *catalytic activity*).

Try to find out which protein is written here:

```
>Protein1
QYSSNTQQGR TSIVHLFEWR WVDIALECER YLAPKGFGGV QVSPPNENVA IHNPFRPWWE
RYQPVSYKLC TRSGNEDEFR NMVTRCNNVG VRIYVDAVIN HMCGNAVSAG TSSTCGSYFN
PGSRDFPAVP YSGWDFNDGK CKTGSGDIEN YNDATQVRDC RLSGLLDLAL GKDYVRSKIA
EYMNHLIDIG VAGFRIDASK HMWPGDIKAI LDKLHNLNSN WFPEGSKPFI YQEVIDLGGE
PIKSSDYFGN GRVTEFKYGA KLGTVIRKWN GEKMSYLKNW GEGWGFMPSD RALVFVDNHD
NQRGHGAGGA SILTFWDARL YKMAVGFMLA HPYGFTRVMS SYRWPRYFEN GKDVNDWVGP
PNDNGVTKEV TINPDTTCGN DWVCEHRWRQ IRNMVNFRNV VDGQPFTNWY DNGSNQVAFG
RGNRGFIVFN NDDWTFSLTL QTGLPAGTYC DVISGDKING NCTGIKIYVS DDGKAHFSIS
NSAEDPFIAI HAESKL
```

→ Are there any organisms with similar proteins?

1. Visit http://mrs.cmbi.ru.nl
2. Search for the protein that you would like to use

3a. Click *'Find similar'*

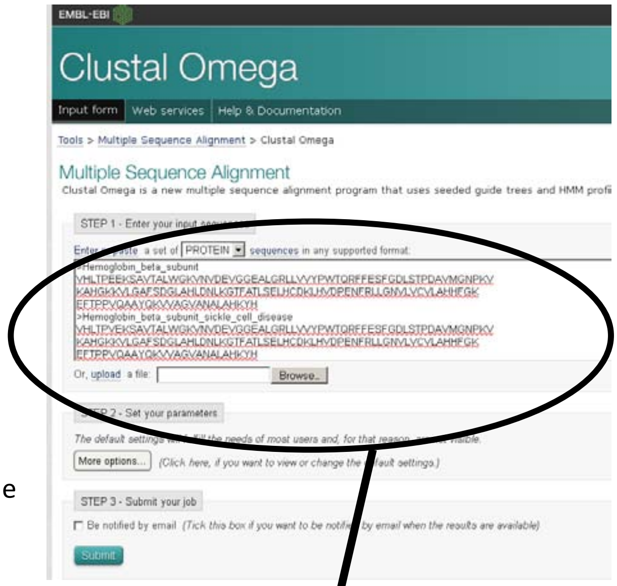
3b. Click *'Blast'*

1. The website http://mrs.cmbi.ru.nl enables you to search for genes in proteins in numerous databases. When looking for proteins the databases Swiss-Prot and Uniprot KB offer the best, most thoroughly checked sequences. Search using the name of the protein.
2. You will get a list of hits: the hits can be exactly the protein that you are looking for, possibly originating from different organisms, or proteins that somehow interact with your protein. Thus, make sure to check the description. The ID can also aid you: the first letters are an abbreviation of the name of the protein, the last one an abbreviation of the name of the organism. When looking for protein from a single organism, add *os:human* (origin species: human, other organisms are possible) to your query. The software assigns a score to each hit. The larger part of the bar is colored, the better the score.
3. Click the ID-code of the desired protein. This gets you a list of information on this protein. Check whether you've found the right protein by looking at the *protein name*. When looking for similar proteins two methods can be used. The results can be highly similar, but occasionally quite different.
   a. Click '*Find similar*'. The proteins in this list are found by comparing their descriptions and key words. Again, you can click the ID for the proteins information sheet.
   b. Click *'Blast*' and then '*Run Blast*'. The result shows you proteins that are selected based on the similarity of their amino acid sequence.
4. The most relevant information on the sheet can be found on the lines *'Protein name*', en '*From*', '*Keywords'* and '*Function*'.

→ In what way do the proteins differ?

→ In what way do the genes differ?

1. Visit http://www.ebi.ac.uk/Tools/clustalw2/index.html
2. Copy both amino acid sequences into the text box.
3. Click ‘Submit'



Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profi

STEP 1 - Enter your input

Enter a set of PROTEIN sequences in any supported format:

```
>Hemoglobin_beta_subunit
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
>Hemoglobin_beta_subunit_sickle_cell_disease
VHLTPVEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
```

Or, upload a file: Browse..

2 - Set your parameters

The default settings the needs of most users and, for that reason, visible.

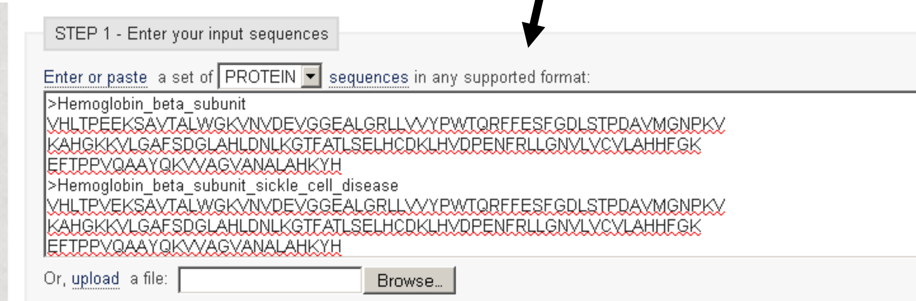
More options... (Click here, if you want to view or change the fault settings.)

STEP 3 - Submit your job

□ Be notified by email (Tick this box if you want to be notifi by email when the results are available)

Submit

1. An alignment-tool compares the amino acid or nucleotide sequences of your proteins or genes of interest. Visit Clustal Omega’s website: http://www.ebi.ac.uk/Tools/msa/clustalo/
2. The software needs at least two amino acid sequences to make a comparison. First you should enter the name of the sequence in a *>nameofyoursequence* format. Be aware that the name can consist of a single word only. Copy the sequence to the lines below the name. Repeat the process for all other sequences.
3. Although you can adjust a number of parameters, the normal configuration will suffice for a simple alignment. When you are finished copying your sequences, click 'Submit'. The calculations may take a while, depending on the number and length of the sequences submitted. However, be sure to set the box above the search field to ‘PROTEIN’ or ‘DNA’, depending on what sequences you entered.
4. The result consists of a number of different pages. The first one, titled 'Alignments' shows the actual alignments. An asterisk (*) indicates that the amino acids of the proteins (or the nucleotides of the genes) are identical. A gap indicates a difference between the sequences. Finally you can encounter a colon (:) or a dot (.), which both mean that although the amino acids differ, their properties are similar. This can happen when, for example, both amino acids are positively charged. By clicking *Show Colors* you can make the alignment a bit more clear.
5. The second page is called 'Result Summary'. Here, you can see the score of the alignment. The more identical the sequences are, the higher the score will be. You can also sort the alignments by their score. If you click *Jalview*, you will get a different view which enables you to judge where the conserved regions in the proteins are.
6. The third page is called ‘Guide Tree’. This page will be further explored and explained on page 12, *What does the phylogenetic tree of the protein look like?*



STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN sequences in any supported format:

```
>Hemoglobin_beta_subunit
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
>Hemoglobin_beta_subunit_sickle_cell_disease
VHLTPVEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
```

Or, upload a file: Browse..

Try to find the differences between the beta-subunit of healthy hemoglobin and the beta-subunit of a patient with sickle cell anemia.

```
>Hemoglobin_beta_subunit
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
>Hemoglobin_beta_subunit_sickle_cell_disease
VHLTPVEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
```

→ What does the phylogenetic tree of the protein look like?

→ What does the phylogenetic tree of the gene look like?

See also 'In what way do the proteins differ?, p.10'. After you have made the alignment (point 4), continue here.

Click on the 'Phylogenetic Tree' tab.
Under the header *'Phylogram'*, you can choose between a *"Cladogram"* or *"Real"*. The Cladogram is the default setting. The way these trees are made up differs and this can affect the actual results.
A cladogram is made by calculating the smallest number of changes to get from one sequence to the other. By calculating this for all proteins in the alignment, the tree is made. The *"Real"* function will give you a phylogram, which is made by calculating the 'evolutionary distance' between a pair of proteins. Thus, when two proteins are almost equal, the branches of the tree will be shorter. You can get the distances between proteins by clicking '*Show distances*'.
Notice that for both trees, adding or deleting a single protein from the alignment can have a profound effect on the resulting tree.

Try to draw a phylogenetic tree of the protein myoglobin

```
>human
MGLSDGEWQL VLNVWGKVEA DIPGHGQEVL IRLFKGHPET LEKFDKFKHL KSEDEMKASE
DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI PVKYLEFISE CIIQVLQSKH
PGDFGADAQG AMNKALELFR KDMASNYKEL GFQG
>rabbit
MGLSDAEWQL VLNVWGKVEA DLAGHGQEVL IRLFHTHPET LEKFDKFKHL KSEDEMKASE
DLKKHGNTVL TALGAILKKK GHHEAEIKPL AQSHATKHKI PVKYLEFISE AIIHVLHSKH
PGDFGADAQA AMSKALELFR NDIAAQYKEL GFQG
>shark
MABWDKVNSV WSAVEQNITA IGQNILLRLF EQYPESEDYF PKLKNKSLGE LKDTADIKAQ
ADTVLRALGN IVKKKGDHSQ PVKALAATHI TTHKIPPHYF TKITTIAVGV LSEMYPSEMN
AQAQAAFSGA FKNICSDIEK EYKAANFQG
>tuna
MADFDAVLKC WGPVEADYTT MGGLVLTRLF KEHPETQKLF PKFAGIAQAD IAGNAAISAH
GATVLKKLGE LLKAKGSHAA ILKPLANSHA TKHKIPINNF KLISEVLVKV MHEKAGLDAG
GQTALRNVMG IIIADLEANY KELGFSG
>gibbon
MGLSDGEWQL VLNVWGKVEA DIPSHGQEVL IRLFKGHPET LEKFDKFKHL KSEDEMKASE
DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI PVKYLEFISE CIIQVLQSKH
PGDFGADAQG AMNKALELFR KDMASNYKEL GFQG
>baboon
MGLSDGEWQL VLNVWGKVEA DIPSHGQEVL IRLFKGHPET LEKFDKFKHL KSEDEMKASE
DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI PVKYLELISE SIIQVLQSKH
PGDFGADAQG AMNKALELFR NDMAAKYKEL GFQG
>common carp
MHDAELVLKC WGGVEADFEG TGGEVLTRLF KQHPETQKLF PKFVGIASNE LAGNAAVKAH
GATVLKKLGE LLKARGDHAA ILKPLATTHA NTHKIALNNF RLITEVLVKV MAEKAGLDAG
GQSALRRVMD VVIGDIDTYY KEIGFAG
>zebra
MGLSDGEWQQ VLNVWGKVEA DIAGHGQEVL IRLFTGHPET LEKFDKFKHL KTEAEMKASE
DLKKHGTVVL TALGGILKKK GHHEAELKPL AQSHATKHKI PIKYLEFISD AIIHVLHSKH
PGDFGADAQG AMTKALELFR NDIAAKYKEL GFQG
```

→ What does the 3D structure of the protein look like?

1. Download the pdb-file of the protein at http://www.pdb.org
2. Use Yasara to open the pdb-file

1. A pdb-file is the most common format of 3D protein structures. The Protein Data Bank is a large database where all kinds of protein structures are stored. Besides directly downloading the protein file from the actual PDB database, there are some other possibilities to obtain pdb-files.
   Google (www.google.com): Search for the desired protein and add 'pdb' to your query.
   The next part will show you how to search through the PDB, http://www.pdb.org.
2. Enter the name of the desired protein in the search bar. You can refine your query by adding more words (e.g. '*lipase human'*) or by selecting for certain organisms or publications at '*Query refinements'*. Since the majority of proteins hasn't had its structure determined it is perfectly possible that you cannot find a protein in the PDB.
   Each file has its own ID-code. For transferine, a protein that binds iron ions in the blood, the code is 1H76. When searching for this code you are immediately directed to the corresponding file.
3. Click the desired file in the list of results. A small image of the structure on the right side of the page can help you to determine if you found the right protein. The tab *'Molecular description*' contains the information on the structure (*Molecule*). Click 'Download files' and subsequently 'PDB file (text)' to download the file.
4. Start Yasara and load the pdb-file. For information how to obtain and configure Yasara, see page 14.

Tip: Have a look at section 101 of the PDB database (http://www.pdb.org/pdb/101/structural_view_of_biology.do). Here, you can find a variety of interesting proteins with detailed but comprehensible explanations. For example, DNA ligase: http://www.pdb.org/pdb/101/motm.do?momID=55

→ Now that I've identified my protein, I want to take a look at its structure.

Yasara manual

Yasara is used to view and manipulate protein structures in 3D. When the sofware isn't available on your computer, you can download and use it for free.

1 Visit www.yasara.org and click 'Products' in the menu.
2 Then click the 'freely download now' button next to 'Yasara View'.
3 Fill in the form. Enter the name of your school in the 'department' field. The submitted e-mail address will only be used to send you the download link.
4 The download link will be delivered to your mailbox. Now, you can install Yasara in any desired directory on your computer.
5 Follow the instructions to install Yasara.
6 Additional information on Yasara can be found at: http://www.cmbi.ru.nl/~hvensela/yasara/

These are the most frequently used options:

Rotation and zooming
Turn the molecule by holding the left mouse button and moving the mouse.
Hold the right mouse button to zoom in (moving the mouse forward) or out (moving the mouse backwards).
The arrow keys on your keyboard can be used to move the molecule across the screen.

Load files
Yasara is able to load a number of different files. These files have different extensions, such as .pdb, .sce, .yob etc. The most commonly used ones are *.pdb* (PDB file) and *.sce* (Yasara scene) files.
To load a file in Yasara, click 'File' > 'Load'. You can now choose which type of file you wish to open. For example, click 'PDB file' to load a .pdb file or 'Yasara scene' to load a .sce file.
If you are done looking at your protein and you wish to look at a different molecule, load a second file. The new molecule will be displayed in the same screen, so you may want to select 'File' and 'New' first to start with an empty screen again.

Different views
Yasara has a number of different views which all have their own advantages and drawbacks. Use the keys F1 to F8 to switch between these views.

F1	Ball
F2	Ball-and-stick
F3	Stick
F4	Trace
F5	Tube
F6	Ribbon
F7	Cartoon

The F8 key can be used in all these views to show or hide amino acid side chains (residues). Some files have parts of their structures highlighted or colored by default. This will be lost when you switch between different views. It can be retrieved by reloading the file.

Additional options

Color negatively charged residues	- Display hydrogen atoms (*Edit > Add > Hydrogens to All*). - Select *view > color > residue* - In the third column, select *(belongs to or has) Charge < 0* and click *Ok*. - Choose your color and hit *Apply Unique color*.
Color positively charged residues	- Display hydrogen atoms (*Edit > Add > Hydrogens to All*). - Select *view > color > residue* - In the third column, select *(belongs to or has) Charge > 0* and click *Ok*. - Choose your color and hit *Apply Unique color*.
Color hydrophylic residues	- Select *view > color > residue* - In the second column, select Arg, Asp, Asn, Glu, Gln, His, Lys, Ser and Thr (while holding Ctrl) and click *Ok*. - Choose your color and hit *Apply Unique color*.
Color hydrophobic residues	- Select *view > color > residue* - In the second column, select Ile, Leu, Met, Phe en Val (while holding Ctrl) and click *Ok*. - Choose your color and hit *Apply Unique color*.
Show hydrogen atoms	Select *Edit > Add > Hydrogens to All*
Show hydrogen bonds	- Select *Edit > Add > Hydrogens to All* - Select *View > Show hydrogen bonds > All*
Show secondary structures	F6
Show sidechains	F8
Delete water molecules	Select *Edit > Delete > Waters*

The next part elaborates on the questions:

- ➔ *(enzyme) Where does it bind its substrate? – p.16*
- ➔ *(receptor/channel) How is the protein bound in the membrane? – p.17*

→ (enzyme) Where does it bind its substrate?

1. Look up the 3D structure of the substrate
2. Load the protein structure in Yasara
3. Look for cavities or other locations on the protein where the substrate can possibly bind
4. Scan these sites for possible interactions between the enzyme and the substrate

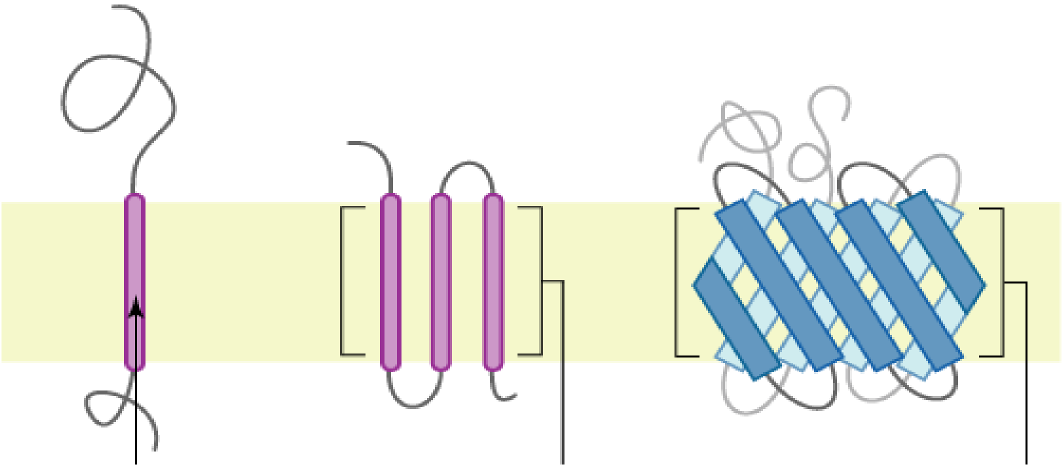
1. Every enzyme binds to a certain substrate/ligand. Use Google, Wikipedia or your biology handbook to determine with which substrate your enzyme of interest interacts. Wikipedia often shows you the structure of its substrate, but Google images may also give some results. Make sure to obtain at least an estimate of the size and structure of the substrate.
2. Load the 3D structure of the enzyme in Yasara.
3. Very often, a substrate binds to a cavity in the enzyme because it creates a bigger surface area to bind with the enzyme. Therefore, a cavity is what you want to look for first. If you can’t find a cavity, look for uncommon parts of a structure. Finding the active site often isn't as straightforward as simply using the structure of the enzyme, so you will probably have to use alternative methods.
4. Check if the cavity is really used to bind a substrate by checking it for possibilities for interactions. The presence of ions or charged amino acid residues can be an indication, especially if your substrate has a net charge too. Also, multiple possibilities for hydrogen bonds, hydrophobic interactions or disulfide bonds (SS-bond) can be a feature of an active site.
5. The PDB database contains a few files of well-known and well-studied proteins that show these proteins along with its substrate. Again, you can find these using PDB or Google.

Try to find the active site of the lactase protein. This protein cleaves lactose. You can find the protein in the PDB database with the ID-code: 3E1F

→ (receptor/channel) How is the protein bound in the membrane?

1. Open the pdb-file in Yasara
2. Press F6
3. Look for secondary structures that can cross the membrane

1. Load the structure of the receptor or ion channel in Yasara
2. Press F6 to switch to the cartoon view. This view enables you to trace α-helices and β-sheets easily. The helices are colored blue, the sheets red.
3. Three structures are known to be able to cross the membrane. These are single helices, bundles of helices and β-barrels (see figure below).
   Helices with hydrophobic residues will automatically stick together to form a bundle. Since the membrane consists largely of lipids these bundles prefer getting incorporated in the membrane. The number of helices in a bundle can range from three to dozens.
   A β-barrel consists of multiple β-strands, which are interwoven to form a pore in the membrane.



Schematic display of a single helix, a helix bundle and a β-barrel

Source: http://en.wikipedia.org/wiki/Transmembrane_protein

Find the bundle of helices in the acetylcholine receptor. PDB ID-code: 2BG9

→ What is the name of the gene?

1. Visit http://www.ensembl.org/Multi/blastview
2. Enter the sequence and hit RUN
3. Analyse the contigview of the first 'hit' to find the corresponding gene

1. Several bioinformatics-tools are freely available on the internet. To identify a DNA sequence one can use Blast. Blast is a search engine that compares your sequence to a vast amount of sequences in a database. It identifies 'hits': genes that have at least a part of their sequence in common with the one you submitted. The first hit listed is the best hit, which is most similar to your submitted gene. It can be the exact gene, but it isn't always that straightforward, since a multitude of exons, introns and reading frames can lead to different transcripts and thus different results. This manual gives an overview of the basic use of Blast.
2. There are several different Blast-tools available. The one developed by Ensembl is the most suitable when searching for DNA sequences, but the Blast-tool found at http://mrs.cmbi.ru.nl is preferable when you are searching for amino acid sequences of proteins.
   Visit the Ensembl webpage: http://www.ensembl.org/Multi/blastview
3. Copy your DNA nucleotide sequence to the search field. Start with a line *>nameofyoursequence.* You are now using the so called FastA-format without which the search engine will not work. It is important to be precise, as the BLAST software is prone to inconsistent input.
   This Blast tool also enables you to search using reference codes (ID codes) from other databases. You can enter such an ID at *Enter a sequence ID or accession (EMBL, UniProt, RefSeq)*. Continue by selecting *Retrieve*.
4. You now have the following options:
   - *Select the databases to search against* – This tool is also able to make alignments. You then need to select the databases which genes you would like to align. This option is not necessary for gene identification.
   -*Select the Search Tool* – Different Blast-tools work slightly different and hence come up with slightly different results. The default selection is BLAT (Blast Like Alignment Tool). This is the fastest one and it is perfectly suitable for this purpose. Other tools include BLASTN (Blast Nucleotides) and TBLASTX (Translate Blast X). Both are relatively slow because of the optimalisation calculations they perform.
   -*Search sensitivity* – for gene identification the *Near-exact matches* option is fine. If this results in a large number of hits you might want to select *Exact matches*. In case of hardly any result the *No optimisation* option is recommended, although you might end up with a number of totally irrelevant hits.
5. When you've configured all the settings to your preference, hit RUN. The query might take a while because of the enormous amount of sequences that have to be compared. If your query is finished the tool switches to a new screen.
   This screen starts with *Alignment Locations vs. Karyotype*. The red arrows indicate locations on the chromosomes that contain a gene that is similar to the one that you submitted. The one outlined with a red box is the best hit.
   The *Alignment Locations vs. Query* shows you the locations of HSPs (High-scoring Sequence Pairs). These are the parts of the sequence that Blast used in the search process. For our

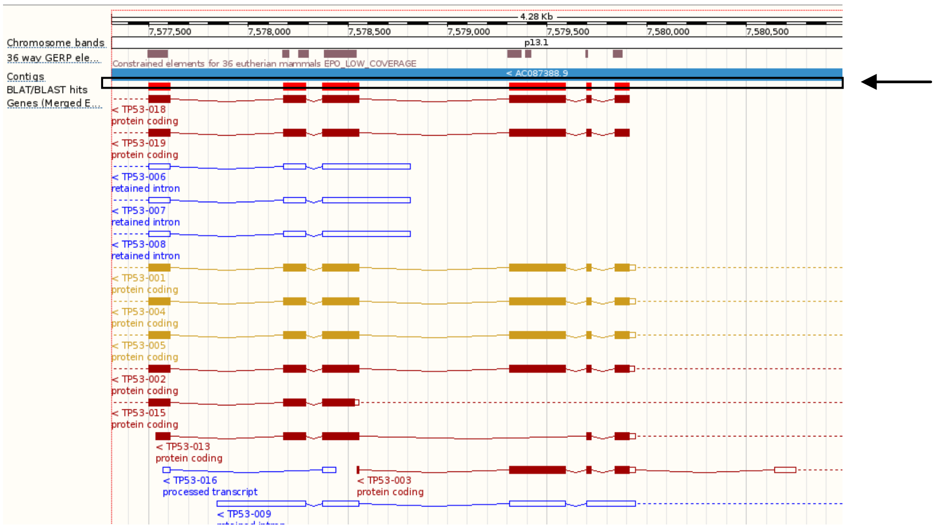
purpose, this information is irrelevant and can be discarded if desired.

The last data page is *Alignment Summary*. At the top of the diagram, you will find the best hit. You can verify its similarity at *Stats* at the far right of your screen. The higher the *Score*, the higher the similarity. *%ID* indicates that amount of identity as a percentage, *Length* gives you the length of the identical area.

At the far left of the screen you will find the letters A, S, G en C.

Click the C (Contigview) of the first hit.

6. You will now get a screen that shows you the surrounding area on the chromosome.



If you scroll down you will find a list of comparable genes. The red bar is the gene that you just selected (indicated with the arrow in the diagram shown here), the other ones are (parts of) other genes. Dark red ones are genes that encode proteins. At the left you can find the gene name. Click the desired gene and then click the code next to *Gene*.

7. The next page shows you the gene information, including its name, its location on the chromosome and the transcripts that are known to originate from this gene.

<u>Try this for:</u>

```
>Nucleotidesequence1
ATGGAGGAGCCGCAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAGGAAACATTTTCAGACCTATGGA
AACTACTTCCTGAAAACAACGTTCTGTCCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCC
GGACGATATTGAACAATGGTTCACTGAAGACCCAGGTCCAGATGAAGCTCCCAGAATGCCAGAGGCTGCT
CCCCCCGTGGCCCCTGCACCAGCAGCTCCTACACCGGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGT
CATCTTCTGTCCCTTCCCAGAAAACCTACCAGGGCAGCTACGGTTTCCGTCTGGGCTTCTTGCATTCTGG
GACAGCCAAGTCTGTGACTTGCACGTACTCCCCTGCCCTCAACAAGATGTTTTGCCAACTGGCCAAGACC
TGCCCTGTGCAGCTGTGGGTTGATTCCACACCCCCGCCCGGCACCCGCGTCCGCGCCATGGCCATCTACA
AGCAGTCACAGCACATGACGGAGGTTGTGAGGCGCTGCCCCCACCATGAGCGCTGCTCAGATAGCGATGG
TCTGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTATTTGGATGACAGAAAC
ACTTTTCGACATAGTGTGGTGGTGCCCTATGAGCCGCCTGAGGTTGGCTCTGACTGTACCACCATCCACT
ACAACTACATGTGTAACAGTTCCTGCATGGGCGGCATGAACCGGAGGCCCATCCTCACCATCATCACACT
GGAAGACTCCAGTGGTAATCTACTGGGACGGAACAGCTTTGAGGTGCGTGTTTGTGCCTGTCCTGGGAGA
GACCGGCGCACAGAGGAAGAGAATCTCCGCAAGAAAGGGGAGCCTCACCACGAGCTGCCCCCAGGGAGCA
CTAAGCGAGCACTGCCCAACAACACCAGCTCCTCTCCCCAGCCAAAGAAGAAACCACTGGATGGAGAATA
TTTCACCCTTCAGATCCGTGGGCGTGAGCGCTTCGAGATGTTCCGAGAGCTGAATGAGGCCTTGGAACTC
AAGGATGCCCAGGCTGGGAAGGAGCCAGGGGGGAGCAGGGCTCACTCCAGCCACCTGAAGTCCAAAAAGG
GTCAGTCTACCTCCCGCCATAAAAAACTCATGTTCAAGACAGAAGGGCCTGACTCAGACTGA
```

Note: this is only the coding part of the gene (only exons)

→ What is the nucleotide sequence of this gene?
→ How many different transcripts originate from this gene?
→ What is the intron/exon composition of this gene?
→ Where on the chromosome is this gene located?

1. Visit the Ensembl genome database: http://www.ensembl.org/index.html
2. Enter the name of the gene in the searchfield
3. Select 'gene' of the best hit
4. For exons/introns: click the desired transcript

1. Ensembl is a genome browser, in which you can find all sorts of information about genes. Note that you can only use Ensembl to search in vertebrates and other eukaryotes. Plant genes, for example, are not available in its database.
   Go to http://www.ensembl.org
2. Use the search function on the Ensembl home page to find your gene of interest. You can use different queries such as the name of the gene, the gene symbol or the coordinates of the gene's location. Click 'Go' to start the query.
3. You are now presented an overview of the hits. Search for the name of your gene of interest, followed by "*Human Gene"*. If you are searching for a gene in a different organism, you should locate the name followed by "*[Species name] Gene"*. Click this hit for more information: you will be directed to the gene's information page.
4. The information page shows a lot of information. The top section gives a description of the gene, alternative names, and its location.
   Below this, a table is displayed with all the known transcripts of the gene. Here, you can find how many base pairs (bp) the gene consists of and whether or not it encodes a protein (*gene encoding*). If the gene encodes a protein, the amount of amino acids of this protein is given. If it doesn't, it reads *No protein*, followed by a description of what happens to the gene transcript. Point the mouse on the text to get more information.
   Under the table, you will find the *Summary*, starting with the gene symbol. There are also multiple links referring to other databases or indicating which methods have been used to identify the coding of the gene.
   At the bottom of the page, you will find a schematic representation of the locations of the transcripts. Your gene of interest and the genes surrounding it are displayed here. The gene is displayed as a line with blocks, in which the lines are introns and the blocks are exons.
5. To find the nucleotide sequence of a transcript, first click the transcript ID of the desired transcript. Then, click *cDNA*, located under the header *Sequence* in the menu at the left side of the screen. At the bottom of the page, you can now find the nucleotide sequence and its corresponding amino acid sequence.

Try to find the human gene for *tumor protein p53* (*TP53*). Find out its nucleotide sequence, amount of transcripts, amount of exons and its location on the chromosome.

→ In which tissue is the gene expressed?

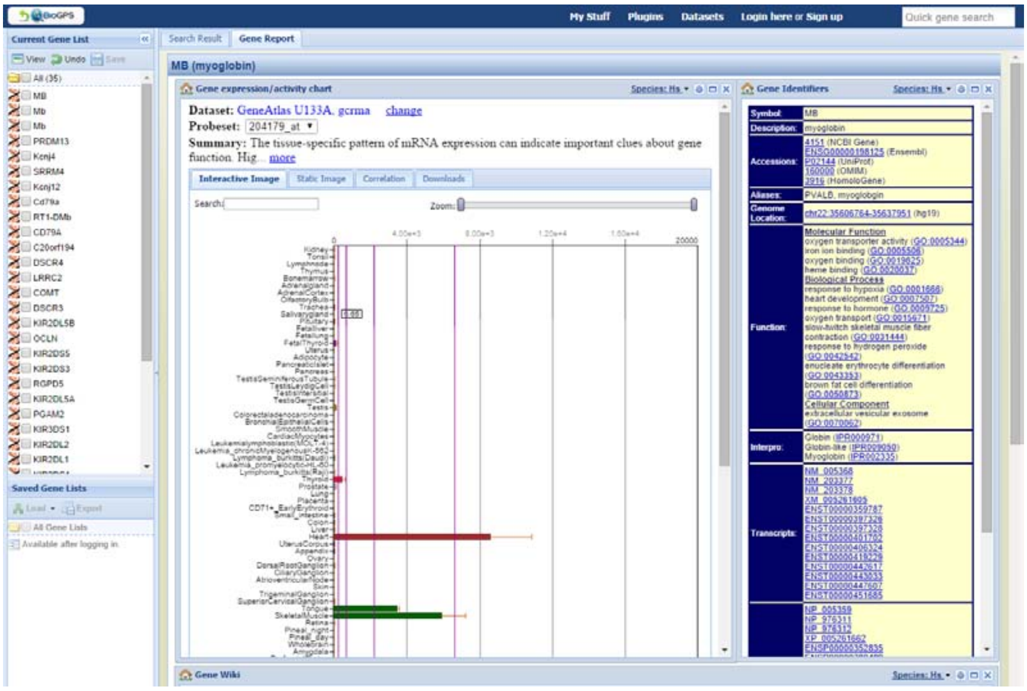
1. Visit http://biogps.org
2. Search for your gene.
3. You can view the expression pattern of the gene by clicking the graph.

1. If we want to find the activity of a gene, we look at the presence of its RNA in certain areas of the body. If there's a lot of RNA of this gene in a certain tissue, this means that this gene has been translated here. We call this gene expression.

   You can make a gene expression profile to map expression levels in the body. These profile can be found on many locations. On a Wikipedia page of a gene, for example, this profile is often found under RNA expression pattern, located in the column on the right side of the screen.

   You can also find these profiles in BioGPS. Visit http://biogps.org.
2. Use the gene symbol of your gene of interest as your query. If you don't know this symbol, use www.google.com to find it, by searching for "[name of your gene] gene symbol". Usually, you can find the gene symbol in one of the first Google hits. It is an abbreviation of letters and numbers.

   Once you inserted your gene symbol in the BioGPS search field, press Search. A table will appear: in the last column, you can find the species name of the gene. Locate your gene of interest and click it.
3. A graph with bars appears: on the Y-axis, you will find a list of tissues in which the expression of this gene was measured. On the X-axis, you will find the amount of gene expression. From this graph, you can derive the gene expression and thus its activity.



MB (myoglobin)

Dataset: GeneAtlas U133A, gcrma change
Probeset: 204179_at
Summary: The tissue-specific pattern of mRNA expression can indicate important clues about gene function. Hig... more

Symbol:	MB
Description:	myoglobin
Accessions:	4151 (NCBI Gene) ENSG00000198125 (Ensembl) P02144 (UniProt) 160000 (OMIM) 3916 (HomoloGene)
Aliases:	PVALB, myoglobgin
Genome Location:	chr22:35606764-35637951 (hg19)
Function:	Molecular Function oxygen transporter activity (GO:0005344) iron ion binding (GO:0005506) oxygen binding (GO:0019825) heme binding (GO:0020037) Biological Process response to hypoxia (GO:0001666) heart development (GO:0007507) response to hormone (GO:0009725) oxygen transport (GO:0015671) slow-twitch skeletal muscle fiber contraction (GO:0031444) response to hydrogen peroxide (GO:0042542) enucleate erythrocyte differentiation (GO:0043353) brown fat cell differentiation (GO:0050873) Cellular Component extracellular vesicular exosome (GO:0070062)
Interpro:	Globin (IPR000971) Globin-like (IPR009050) Myoglobin (IPR002335)
Transcripts:	NM_005368 NM_203377 NM_203378 XM_005261605 ENST00000359787 ENST00000397326 ENST00000397328 ENST00000401702 ENST00000406324 ENST00000419229 ENST00000442617 ENST00000443033 ENST00000447607 ENST00000451685
	NP_005359 NP_976311 NP_976312 XP_005261662 ENSP00000352835

In which tissue is the myoglobin gene (symbol: MB) mainly active?

→ Are there any organisms with similar genes?

1. Ga naar http://www.ensembl.org
2. Type de naam van het gen in het tekstveld
3. Bekijk de beste hit onder 'gene'
4. Klik op ‘Orthologues’ in het menu aan de linkerkant
5. Selecteer van welke soortgroepen je gedetailleerde informatie wilt zien

1. Ensembl is a genome browser, in which you can find all sorts of information about genes. Note that you can only use Ensembl to search in vertebrates and other eukaryotes. Plant genes, for example, are not available in its database.
   Go to http://www.ensembl.org
2. Use the search function on the Ensembl home page to find your gene of interest. You can use different queries such as the name of the gene, the gene symbol or the coordinates of the gene’s location. Click ‘Go’ to start the query.
3. You are now presented an overview of the hits. Search for the name of your gene of interest, followed by “*Human Gene”*. If you are searching for a gene in a different organism, you should locate the name followed by “*[Species name] Gene”*. Click this hit for more information: you will be directed to the gene’s information page.
4. The information page shows a lot of information. To find which organisms have similar genes, click *Orthologues*, located under *Comparative Genomics* in the menu on the left side.
5. At the bottom of the page, you will now find a table with groups. For each group, you can opt to show more information (*Show details*). If you check a *Show details* box, another table appears with a list of species that have a similar gene to your chosen gene.

Try to find which organisms have the gene for rhodopsin, which is involved in detecting light (vision).