# NAVIGENE

*(This is a prototype-version!)*

netherlands bioinformatics centre

bioinformatics @ school

## Phenotype

→(cell) Process
→Feature
→Disease / Syndrome

*Which phenomenon is involved with this protein (function)? P.4*

*Which protein is involved in this phenomenon? P.2*

*Which phenomenon is involved with this gene (function)? P.5*

*Which gene is involved in this phenomenon? P.2*

## Protein

*For which protein encodes this gene? P.xx*

*Which gene encodes for this protein? P.xx*

## DNA

I know the **structure** of the protein *(P.13)*

*What does the 3D structure of the protein look like? P.12*

*What type of protein is this? P.xx*

*(Enzym) Where does the substrate bind? P.15*

*(receptor/channel) With what does the protein attach itself to the membrane? P.16*

I know the **name** of the protein

*What is the amino acid sequence of this protein?P.4*

*What is the name of the protein? P.6*

*Where in the body is the protein located? P.xx*

*Where in the cell is the protein lacated? P.4*

I know the **amino acid sequence** of the protein

*Do other organisms have similar proteins? P.8*

*In what way do these proteins differ when compared to each other? P.9*

*What does a phylogenetic tree look like based on this protein? P.11*

I know the **name** of the gene

*What is the nucleotide sequence of this gene? P.17*

*What is the name of this gene? P.18*

*How many transcripts of this gene exist? P.21*

*How many coding regions does this gene have (introns/exons)? P.21*

*What is the chromosome location of this gene? P.21*

*In which tissue is the gene operating? P.22*

I know the **nucleotide sequence** of the gene

*Do other organisms have similar genes? P.xx*

*In what way do these genes differ when compared to each other?*

*What does a phylogenetic tree look like based on this gene? P.xx*

bioinformatics @ school

## Dear teacher/student

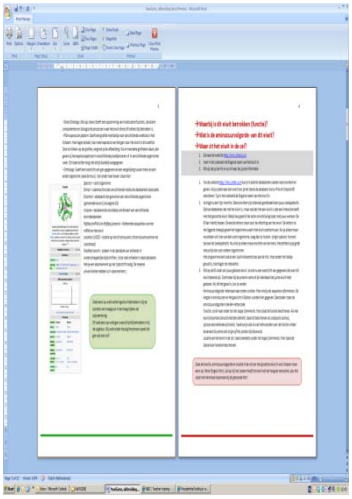
This booklet is the NAVIGENE. It is a navigation tool developed for biology teachers at secondary schools, developed by the Netherlands Bioinformatics Centre and the Freudenthal Institute for Science and Mathematics Education.

**Under construction**

The tool consists of a cover page (diagram) and an accompanying instructional guide. Please note that the instrument is still evolving: some instructions are missing, questions are incomplete and not all of the instructions are tested. And since bioinformatics is a dynamic research field, links to websites can change. We like to further develop NAVIGENE with your help. So please let us know if you have any wishes or comments. Together we can improve NAVIGENE.

**Using NAVIGENE**

There are several ways to get acquainted with NAVIGENE:
*Participation in a bioinformatics@school teacher day* – During this day we will work with bioinformatics in the classroom. We will discuss NAVIGENE in detail and together with other teachers and students you will create your own specific assignments that you can use in class.
*Participation in a workshop bioinformatics@school* – At various teacher meetings bioinformatics@school provides workshops. During these workshops you will get a brief introduction in the use of NAVIGENE.
*Contact bioinformatics@school* – When you independently want to discover how NAVIGENE works, you can get help from Hienke Sminia. The proper contact information is shown at the bottom of this page. You can obtain the data and locations of the teacher days and workshops through our newsletter which appears about every two months. Subscribe to the newsletter via hienke.sminia@nbic.nl.



An on-site training is also an option, provided that there are enough participants. You can discuss the possibilities with Hienke Sminia.

**The structure of NAVIGENE**

NAVIGENE consists of a cover page and an instructional guide. On the cover page you can find several questions. There is a number for each question, for example P. 4. This means that you can answer this question with the instructions on page 4. The instruction is structured as shown in the figure.

We wish you many useful discoveries and valuable surprises when using NAVIGENE.

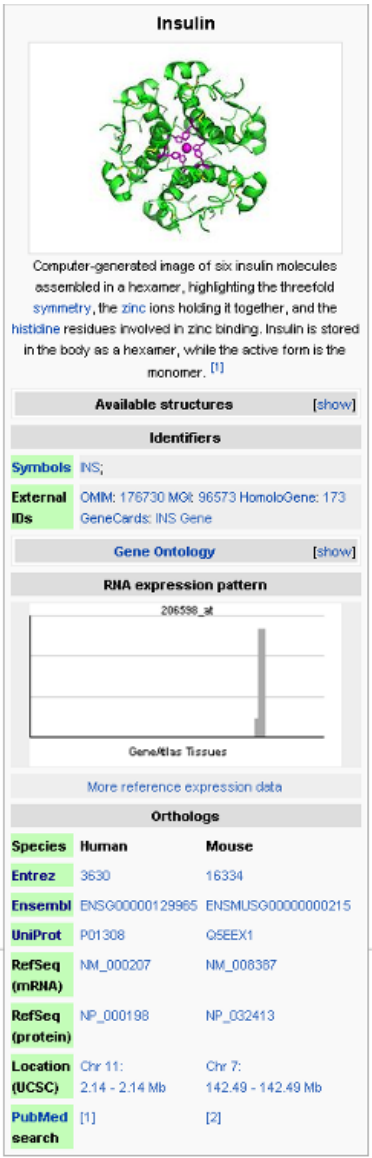
Hienke Sminia - Bioinformatics@school from the Netherlands Bioinformatics Centre
hienke.sminia@nbic.nl / phone +31 243619501
Dirk Jan Boerwinkel - Freudenthal Institute for Didactics of Mathematics and Sciences

# → 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. Protein 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) and biological processes (photosynthesis, digestion).
   As of now there exists no single databank where one can find genes and protein that are involved in a characteristic, disease or process. This information is scattered among many different databanks, websites of research institutes and encyclopedias. Search engines like Google enable to search all these sources simultaneously. Visit http://www.google.com.
2. Use the desired phenomenon as a query. You may want to extend the search term with terms as ‘gene’ or ‘protein’. Multiple word queries can be submitted using quotation marks. 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. For further information on this check see page 5. (*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 wherein the protein is involved. On the right side of the page a table is displayed. This table contains the following information:
   - An image of the 3D-structure of the protein with a description. The description can for example name the organism from which the protein or 3D structure originates.
   - *Available structures*: (click *show*) PDB is the abbrevation for Protein DataBank (see also page 11 – *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 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 frequently. You should use this one when working with MRS (see page 4).

- *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. 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* – databank which contains genomes from multiple organisms (see page 18)
*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 databank 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 is the function of the protein?
## →What is the proteins 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.
   Click on the ID-code of the protein that you prefer. All information found in the database is listed. Check *protein name* to make sure that you have selected the right protein.
   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 (*catalyctic 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 has not 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. Surf to http://www.ncbi.nlm.nih.gov/gene
2. Search for the desired gene.
3. Look up the genes 'General gene information'

Find the function of the gene *AGAMOUS* in the *Arabidopsis thaliana*.

## →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 with a line
   *>nameofyoursequence*
   You are now using the so called FastA-format without which the search engine will not work. 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).
4. 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. When searching for a well known protein you can safely check the option.
5. 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.
6. 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 obtained by clicking the colored bar. The identity and similarity both describe the similarity of your query and the 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.

7. 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'*

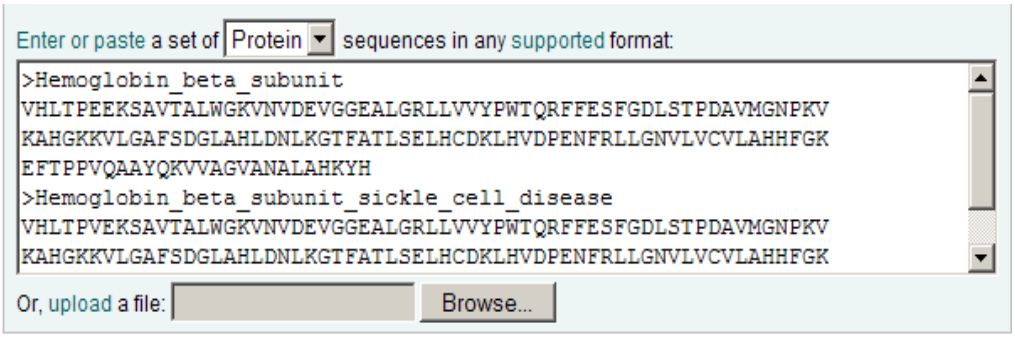
3b. Click *'Blast'*

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. 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.
   Click on the ID-code of the desired protein. This gets you a list of information on this protein. 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.
3. 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?

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

1. An alignment-tool compares the primary structure of a multitude of proteins. Visit ClustalW2's website: http://www.ebi.ac.uk/Tools/clustalw2/index.html
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.
4. The result consists of a number of different pages. The first one, titled 'Alignments' shows the actual alignments. A '*' sign indicates that the amino acids of the proteins are equal. A gap points to a difference between the sequences. Finally you can encounter a ':' 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. 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. For explanations on the 'Guide Tree' page, see page 11.

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?

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

Click on the 'Guide Tree' tab.
Here, you can choose between a cladogram (*Cladogram Tree*) or a phylogenetic tree (*Phylogram Tree)*. The last one 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. A phylogram 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

```
>mens
MGLSDGEWQL VLNVWGKVEA DIPGHGQEVL IRLFKGHPET LEKFDKFKHL KSEDEMKASE
DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI PVKYLEFISE CIIQVLQSKH
PGDFGADAQG AMNKALELFR KDMASNYKEL GFQG
>konijn
MGLSDAEWQL VLNVWGKVEA DLAGHGQEVL IRLFHTHPET LEKFDKFKHL KSEDEMKASE
DLKKHGNTVL TALGAILKKK GHHEAEIKPL AQSHATKHKI PVKYLEFISE AIIHVLHSKH
PGDFGADAQA AMSKALELFR NDIAAQYKEL GFQG
>haai
MABWDKVNSV WSAVEQNITA IGQNILLRLF EQYPESEDYF PKLKNKSLGE LKDTADIKAQ
ADTVLRALGN IVKKKGDHSQ PVKALAATHI TTHKIPPHYF TKITTIAVGV LSEMYPSEMN
AQAQAAFSGA FKNICSDIEK EYKAANFQG
>tonijn
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
>baviaan
MGLSDGEWQL VLNVWGKVEA DIPSHGQEVL IRLFKGHPET LEKFDKFKHL KSEDEMKASE
DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI PVKYLELISE SIIQVLQSKH
PGDFGADAQG AMNKALELFR NDMAAKYKEL GFQG
>karper
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
```

# →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, there are some other possibilities to obtain pdb-files.
   Google (www.google.com): Search for the desired protein and add 'pdb' to your query.
   MRS (http://mrs.cmbi.ru.nl) (see page 4): Click the protein code at *Cross-references* and select *Download*. When saving the file, change its extension to *.pdb*.
   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 (‘*lipase human*’) or by selecting for certain organism 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 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 and where to obtain Yasara, see page 13.

# →Now that I 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 your schools name in the ‘departement’ field. The submitted email adress 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 a desired directory.
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 it in the desired direction. Use the right mouse button to zoom in (moving the mouse forward) or out (by 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 differ by their extension. Be careful, because when you choose to load a PDB file all other files won't be displayed in folders you search through and henceforth cannot be loaded. The most common ones are *.pdb* (PDB file) and *.sce* (Yasara scene) files. If you 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&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 sidechains. Some files have parts of their structures highlighted or colored. This will be lost when you switch to another view. It can be retrieved by reloading the file.

*Additional options*

| | |
|---|---|
| Color negatively charged residues | - Select *Edit > Add > Hydrogens to All*<br>- Select *view > color > residue*<br>- Select in the third column *(belongs to or has) Charge < 0* and click *Ok*.<br>- Choose your color and hit *Apply Unique color*. |
| Color positively charged residues | - Select *Edit > Add > Hydrogens to All*<br>- Select *view > color > residue*<br>- Select in the third column *(belongs to or has) Charge > 0* and click *Ok*.<br>- Choose your color and hit *Apply Unique color*. |
| Color hydrophylic residues | - Select *view > color > residue*<br>- In the second column, select Arg, Asp, Asn, Glu, Gln, His, Lys, Ser and Thr (while pressing ctrl) and click *Ok*.<br>- Choose your color and hit *Apply Unique color*. |
| Color hydrophobic residues | - Select *view > color > residue*<br>- In the second column, select Ile, Leu, Met, Phe en Val (while pressing ctrl) and click *Ok*.<br>- 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*<br>- 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.14*
- ➔ *(receptor/channel) How is the protein bound in the membrane? – p.15*

# →(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 locations on the protein where the substrate can possibly bind
4. Scan these sites for possible interactions between the enzyme and the substrate

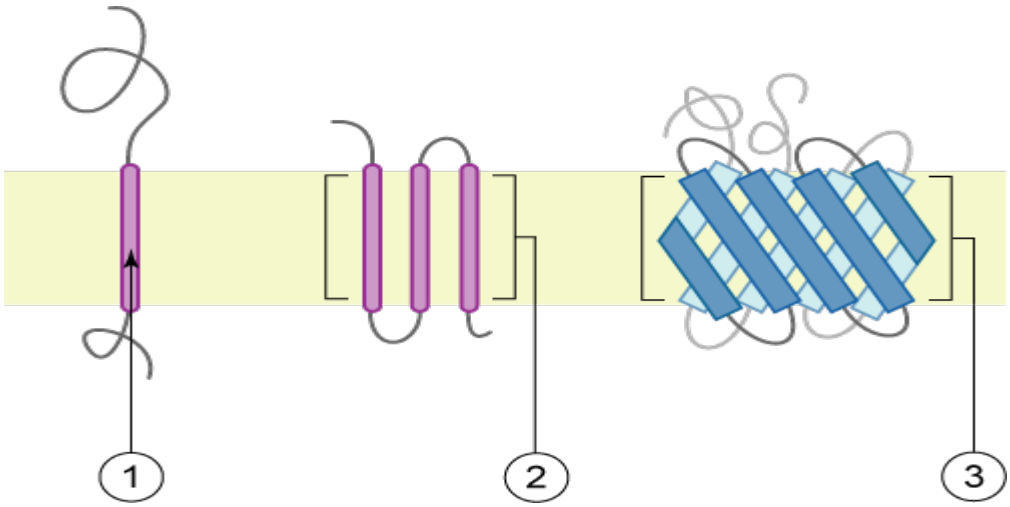
1. Every enzyme binds to a certain substrate. Use google, wikipedia or your biology handbook to determine with which substrate the enzyme interacts. Wikipedia often shows you the structure of its substrate, but Google images may also give some results. Make sure to obtain at least an estimate of the 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, so you probably want to start with looking for one. When the enzyme has no cavity, look for uncommon parts of a structure. Finding the active site by using just a structure isn't straightforward most of the time to say the least, so you will probably have to check other sources.
4. Check if the cavity is really used to bind a substrate by checking it for possibilities for interactions. The appearance of ions or charged amino acids can be an indication just as multiple possibilities for hydrogen bonds or hydrophobic interactions.
5. The PDB contains a few files of well-known and well-studied proteins that show these proteins along with its substrate.

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 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 helices (1), bundles of helices (2) and β-barrels (3).
   Helices with hydrophobic residues will automatically stick together. 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 that are interwoven and thus form a pore in the membrane.




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

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

## →What is the sequence of the gene?

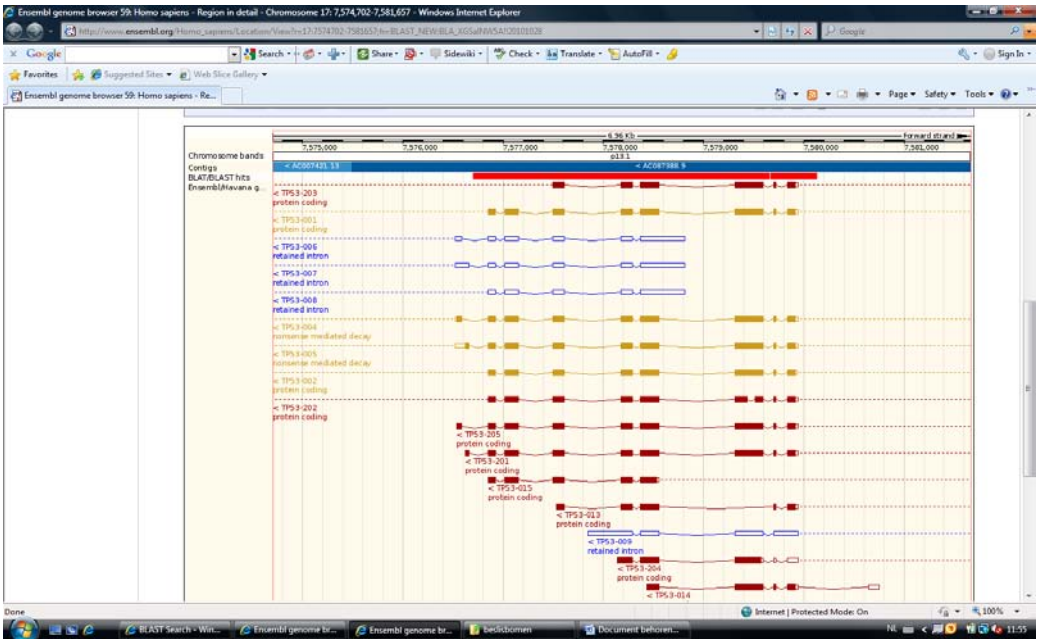
1. Visit http://mrs.cmbi.ru.nl
2. Look up the gene in the EMBL database
3. Select the best ‘hit’ and scroll to the bottom of the page

## →What is the name of the gene?

4. Visit http://www.ensembl.org/Multi/blastview
5. Enter the sequence and hit 'RUN'
6. 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 youre 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 that you submitted. The first hit can be the exact gene, but it isn't always that straightforward because of a multitude of exons, introns and reading frames can lead to different results.
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. Visit the Ensembl webpage: http://www.ensembl.org/Multi/blastview
3. Copy your amino acid sequence to the search field. Start with a line *>nameofyoursequenc.*You are now using the so called FastA-format without which the search engine will not work.
   This tool also enables you to search using reference 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 redundant 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 selected all the desired options, 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 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 with a red outlining is the best hit.
   *Alignment Locations vs. Query* show you the locations of HSPs or High-scoring Sequence Pairs. These are the parts of the sequence that Blast used in the search process. For this purpose the information is irrelevant.

The last one is *Alignment Summary*. The best hit, which is outlined in red in the karyogram is at the top. You can check this at *Stats* at the far right of your screen. The higher the *Score*, the higher the similarity. *%ID* indicates that amount of similarity in percents, *Length* gives you the length of the corresponding 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, the other ones are (parts of) other genes. Dark red ones are genes that code for proteins. At the left you can find the gene name. Click on the desired gene and then click on the code next to *Gene*.
7. This page shows you the gene, the location on the chromosome and the transcripts that are known to originate from this gene.

Try this for:

>nucleotidesequence

```
ATGGAGGAGCCGCAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAGGAAACATTTTCAGACCTATGGAAACTACTTCC
TGAAAACAACGTTCTGTCCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCCGGACGATATTGAACAATGGT
TCACTGAAGACCCAGGTCCAGATGAAGCTCCCAGAATGCCAGAGGCTGCTCCCCCCGTGGCCCCTGCACCAGCAGCTCCT
ACACCGGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTTCTGTCCCTTCCCAGAAAACCTACCAGGGCAGCTA
CGGTTTCCGTCTGGGCTTCTTGCATTCTGGGACAGCCAAGTCTGTGACTTGCACGTACTCCCCTGCCCTCAACAAGATGT
TTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTGTGGGTTGATTCCACACCCCCGCCCGGCACCCGCGTCCGCGCCATG
GCCATCTACAAGCAGTCACAGCACATGACGGAGGTTGTGAGGCGCTGCCCCCACCATGAGCGCTGCTCAGATAGCGATGG
TCTGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTATTTGGATGACAGAAACACTTTTCGAC
ATAGTGTGGTGGTGCCCTATGAGCCGCCTGAGGTTGGCTCTGACTGTACCACCATCCACTACAACTACATGTGTAACAGT
TCCTGCATGGGCGGCATGAACCGGAGGCCCATCCTCACCATCATCACACTGGAAGACTCCAGTGGTAATCTACTGGGACG
GAACAGCTTTGAGGTGCGTGTTTGTGCCTGTCCTGGGAGAGACCGGCGCACAGAGGAAGAGAATCTCCGCAAGAAAGGGG
AGCCTCACCACGAGCTGCCCCCAGGGAGCACTAAGCGAGCACTGCCCAACAACACCAGCTCCTCTCCCCAGCCAAAGAAG
AAACCACTGGATGGAGAATATTTCACCCTTCAGATCCGTGGGCGTGAGCGCTTCGAGATGTTCCGAGAGCTGAATGAGGC
CTTGGAACTCAAGGATGCCCAGGCTGGGAAGGAGCCAGGGGGGAGCAGGGCTCACTCCAGCCACCTGAAGTCCAAAAAGG
GTCAGTCTACCTCCCGCCATAAAAAACTCATGTTCAAGACAGAAGGGCCTGACTCAGACTGA
```

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

## →How many different transcripts originate from this gene?
## →What is the introns/exon composition of the 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

Try to find the human gene for *tumor protein p53*.

## →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.

In which tissue is the myoglobin gene active?