

# **Bringing Bioinformatics to the Biology Classroom**

**Science Teachers' Association of Ontario 2014 Conference**  
Thursday November 8, 2018

**Presented by:**



**Global Organization for Bioinformatics  
Learning, Education and Training**

[mygoblet.org](http://mygoblet.org)

# Organizations working together on Education and Training

## GOBLET



Global Organization for Bioinformatics  
Learning, Education and Training

<http://mygoblet.org>



## ISCB

International Society for  
Computational Biology

<http://www.iscb.org>

Home / Training portal

### Training portal

This portal allows you to browse our course pages and training materials, and to download content of interest. For GOBLET members, the portal also allows *uploading* of course pages and training materials. If you would like to know how to upload *your* training materials and courses, please join us!

List of training materials and course pages

Displaying 1 - 30 of 88

Updated date **	Type	Topic	Audience
Plant and Pathogen Bioinformatics Updated 4 weeks 1 day ago	Training material	AlIBio, bioinformatics, Biological databases, Genome sequence analysis, plants, phytopathogens, pathogenesis	beginner bioinformaticians, early stage phytopathogen researchers
RNA-seq data analysis workshop Updated 1 month 1 day ago	Course page	RNA-seq	Biologists, bioinformaticians
ChIP- and DNase-seq data analysis workshop Updated 1 month 1 day ago	Course page	ChIP-seq, DNase-seq	Life Science Researchers, bioinformaticians, Biologists
Metagenome data analysis Workshop, May 21-23, 2014 Updated 1 month 1 week ago	Training material	AlIBio, bioinformatics, metagenomics, Data analysis, assembly, binning	
Presentation About GOBLET Portal Updated 1 month 1 week ago	Training material	Training portal, GOBLET	General Interest, Bioinformatics, Trainers, Users

<http://mygoblet.org>

**Training portal**

- Trainers and organisers
- Training materials
- Course pages
- FAQ

**Filter by audience**

Anyone wants to start using the Unix/Linux OS Bachelor students  
**beginner bioinformaticians**  
**Beginners** Bench biologists  
 biocurators  
**bioinformaticians**  
 Bioinformatics **Biologists**  
**Biologists, Genomicists, Computer Scientists** biology and bioinformatics sophomore undergraduates Biomedical researchers Clinical Bioinformaticians Clinical Scientists computational biologists computational scientists Computer scientists early stage phytopathogen researchers  
 Educators experimental biologist researchers experimentalists and bioinformaticians working on EST, NextGeneration Sequencing and microarray design projects, specially (but not exclusively) of non-model species. No programming knowledge is required. Field biologist researchers General Interest **Graduate Students** healthcare professionals





# Bioinformatics Resources for High School Teacher

hosted by ISCB Education Committee

- Bioinformatics is an integral part of biology
- ISCB is a resource for high school teacher
  - Toolbox for curriculum development - Lesson Plans and Hands-on Activities
- Career paths in Bioinformatics

<http://www.iscb.org/bioinformatics-resources-for-high-schools>

## ISCB Education

ISCB Education Committee

What is Bioinformatics?

Bioinformatics Resources for High Schools

Lesson Plans & Hands-on Activities

What is Bioinformatics?

Careers in Bioinformatics

Discussions on Bioinformatics in High Schools

Curriculum Guidelines for Colleges & Universities

Online Courses in Bioinformatics

Degree & Certificate Programs in Bioinformatics

Articles on Bioinformatics Education

Contact Education Committee

## Lesson Plans and Hands-on Activities for Bioinformatics Curriculum

The following may prove useful to secondary school educators and students looking for information on Bioinformatics teaching tools

- Bioinformatics @ School, Netherlands
- Bioinformatics at Schools, Portugal
- [BSCS: A Science Education Curriculum Study](#)
- [Center for Computational Research: University of Buffalo](#)
- CusMiBio, University of Milan, Italy
- DNA Learning Center at Cold Spring Harbor Laboratory
- [EMBL Learning Laboratory for the Life Sciences](#)
- Harvard University Life Sciences/HHMI (see Microbiology--Lesson Plans--Recreating the Tree of Life Using Bioinformatics)
- High School Bioinformatics Labs @ Whitehead Institute
- ISCB/GOBLET Workshop for High School Teachers - ISMB 2014
- Northwest Association for Biomedical Research
- The Educational Facilities of the Michael Smith Labs, Vancouver BC

[Back to ISCB Education Homepage](#)

<http://www.iscb.org/bioinformatics-resources-for-high-schools>

# File Location

All slide decks, guides, exercises and associated files  
can be found at: <https://tinyurl.com/ybbcqctm>

# Lesson Plan #1 – Bioinformatics.ca

Michelle Brazas



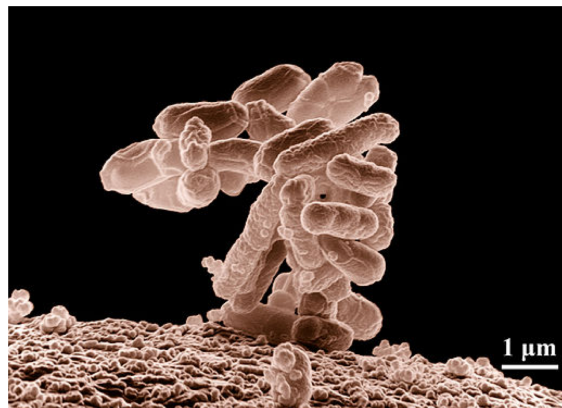
# Microbiome Bioinformatics and Health



## What is a microbiome?

A **microbiome** is the full collection of genes of all the **microbes** in a community. The size of the human microbiome (all of the genes from all of our microbes) outnumbers the size of the human genome (all of our genes in a human) by about 100 to 1.

A **microbe** is very small living organism, such as bacteria, protozoa, fungi, algae, amoebas, and slime molds.



## Microbiomes in Health

- The microbes in our microbiome are important both inside and outside our bodies
  - Microbe genes code for enzymes that break down food we cannot digest on our own
  - Microbe genes code for proteins that build vital nutrients needed by our bodies
  - Microbe genes code for molecules that kill other harmful bacteria
  - Microbe genes code for enzymes that convert skin oils to natural moisturizers that keep skin soft, flexible and crack free
- Our microbiome is involved in our nutrition, immunity, protection from infection, maintenance of protective barriers, organ development and more.

**Your microbiome is important to your health!**

3

Collaborate. Translate. Change lives.

## Microbiomes in Disease

- Microbes such as harmful viruses and bacteria cause diseases
  - Harmful bacteria cause strep throat and food poisoning
  - Harmful viruses cause measles and the flu
- Disruption to our microbiome can cause disease because a disrupted microbiome allows harmful bacteria to cause harm.
  - Acne
  - Dental cavities
  - Cancer
  - Gastric ulcers
  - Inflammatory bowel disease (IBD)

**Maintaining a healthy microbiome can help us avoid some diseases!**

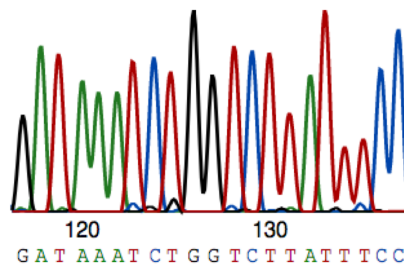
4

Collaborate. Translate. Change lives.



## Sequencing the microbiome

- DNA Sequencing is used to identify the microbes in a sample
- Sequence a marker or short, unique DNA sequence to help identify the microbial population
  - One common DNA marker is the gene that codes for the 16S subunit of ribosomal RNA (16s rRNA), an important part of the cell's protein-building machinery
  - All bacteria have the 16S rRNA gene, but the exact DNA sequence is unique to each species



7

Collaborate. Translate. Change lives.

## Role for Bioinformatics in microbiome research

- The sequence from each 16s rRNA is searched against a database (DB) of all known sequences
  - If they find a match, they can identify the species
  - If they do NOT find a match, they have discovered a new species
- This search involves aligning the test sequence against all other sequences in the DB, and calculating a sequence similarity score for each pair of sequences
  - The sequence pair with the highest score means that this sequence pair shares the highest similarity or relatedness
- The bioinformatics tool used for this pairwise alignment is called **BLAST** - Basic Local Alignment Search Tool



8

Collaborate. Translate. Change lives.

## Whose microbiome is unhealthy?

### Exercise:

Gut samples from Romeo and Juliet were amplified and their 16S rRNA marker genes sequenced. We found lots of different sequences but chose one very common one from each sample. One of them is healthy, but the other has lots of abdominal pain. Which person might need to see a doctor?

**Link to Microbiome Bioinformatics Exercise:** <https://tinyurl.com/yad6zb2v>

9

Collaborate. Translate. Change lives.

## References and Additional Resources

1. Material adapted from Learn.Genetics - <https://learn.genetics.utah.edu/content/microbiome/>
2. NCBI BLAST - <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
3. Helicobacter pylori infection - <https://www.mayoclinic.org/diseases-conditions/h-pylori/symptoms-causes/syc-20356171>

10

Collaborate. Translate. Change lives.





Funding for the Ontario Institute for Cancer Research  
is provided by the Government of Ontario





## Microbiome Bioinformatics for High Schools

---

**Developed by:** Ontario Institute for Cancer Research and bioinformatics.ca with acknowledgements for original work from Bioinformatics and Research Computing, Whitehead Institute

**Text adaptation:** Michelle Brazas

**License:** All the included material is protected by the Creative Commons ShareAlike 2.5 Canada license CC BY-SA 2.5 CA (<https://creativecommons.org/licenses/by-sa/2.5/ca/>)

For any questions or comments, please contact the bioinformatics.ca at [course\\_info@bioinformatics.ca](mailto:course_info@bioinformatics.ca)

---

### INTRODUCTION

#### **What can 16S rRNA from the gut tell us about one's health?**

The human microbiota consists of the 10-100 trillion symbiotic microbial cells harbored by each person, primarily bacteria in the gut; the human microbiome consists of the genes these cells harbor. Microbiome projects worldwide have been launched with the goal of understanding the roles that these symbionts play and their impacts on human health (PMCID: PMC3426293).

The National Center for Biotechnology Information (NCBI) has a lot of bioinformatics resources, including the Basic Local Alignment Search Tool (BLAST). This program lets you very quickly compare a DNA or protein sequence to many (millions) known DNA or protein sequences. The top hit has the highest similarity to the input sequence.

After profiling the guts of Romeo and Juliet by amplifying and sequencing 16S rRNA marker genes, we found lots of different sequences but chose one very common one in each sample. One of them is healthy but the other has lots of abdominal pain. Which person might need to see a doctor?

## CLASSROOM HANDS-ON EXERCISE

Requirements: A computer with internet access

1. To start, go to the NCBI page: <https://www.ncbi.nlm.nih.gov/>
2. Under "Popular Resources" (top right), click on "BLAST".
3. Since we will be comparing DNA sequences to DNA sequences, click on the "Nucleotide BLAST" box.
4. In the box under "Enter accession number(s), gi(s), or FASTA sequence(s)", copy and paste all lines (including the lines beginning with >) in the file "16S\_rRNA\_sequences\_for\_Microbiome\_Bioinformatics.txt". Note that we have two sequences, one for 16S rRNA from ROMEO's gut and one for 16S\_rRNA from JULIET's gut.  
- Sequence file can be found here: <https://tinyurl.com/y7vm3dqj>
5. Next to "Database ", note that we're going to compare Romeo and Juliet's RNA to a "Nucleotide collection" of many genes in many species.
6. Click on the BLAST button at the bottom left and wait 15-60 seconds for the results to return.
7. When the results page appears, next to "Results for", select Juliet (if not already selected).
8. Scroll below the "Graphic Summary" section and look at the first entry in the "Descriptions" table.
9. What species is the first sequence from?
10. What do you know about that species? Search the Web to find out more.
11. Does Juliet appear to be healthy?
12. Go back to "Results for" at the top of the page and select Romeo.
13. What species is the first sequence from for Romeo?
14. What do you know about that species? Search the Web to find out more.
15. Does Romeo appear to be healthy?



## Answer Key for Microbiome Bioinformatics for High Schools

---

**Developed by:** Ontario Institute for Cancer Research and bioinformatics.ca with acknowledgements for original work from Bioinformatics and Research Computing, Whitehead Institute

**Text adaptation:** Michelle Brazas

**License:** All the included material is protected by the Creative Commons ShareAlike 2.5 Canada license CC BY-SA 2.5 CA (<https://creativecommons.org/licenses/by-sa/2.5/ca/>)

For any questions or comments, please contact the bioinformatics.ca at [course\\_info@bioinformatics.ca](mailto:course_info@bioinformatics.ca)

---

### INTRODUCTION

#### **What can 16S rRNA from the gut tell us about one's health?**

The human microbiota consists of the 10-100 trillion symbiotic microbial cells harbored by each person, primarily bacteria in the gut; the human microbiome consists of the genes these cells harbor. Microbiome projects worldwide have been launched with the goal of understanding the roles that these symbionts play and their impacts on human health (PMCID: PMC3426293).

The National Center for Biotechnology Information (NCBI) has a lot of bioinformatics resources, including the Basic Local Alignment Search Tool (BLAST). This program lets you very quickly compare a DNA or protein sequence to many (millions) known DNA or protein sequences. The top hit has the highest similarity to the input sequence.

After profiling the guts of Romeo and Juliet by amplifying and sequencing 16S rRNA marker genes, we found lots of different sequences but chose one very common one in each sample. One of them is healthy but the other has lots of abdominal pain. Which person might need to see a doctor?

## CLASSROOM HANDS-ON EXERCISE

Requirements: A computer with internet access

1. To start, go to the NCBI page: <https://www.ncbi.nlm.nih.gov/>
2. Under "Popular Resources" (top right), click on "BLAST".
3. Since we will be comparing DNA sequences to DNA sequences, click on the "Nucleotide BLAST" box.
4. In the box under "Enter accession number(s), gi(s), or FASTA sequence(s)", copy and paste all lines (including the lines beginning with >) in the file "16S\_rRNA\_sequences\_for\_Microbiome\_Bioinformatics.txt". Note that we have two sequences, one for 16S rRNA from ROMEO's gut and one for 16S\_rRNA from JULIET's gut.  
- Sequence file can be found here: <https://tinyurl.com/y7vm3dql>
5. Next to "Database ", note that we're going to compare Romeo and Juliet's RNA to a "Nucleotide collection" of many genes in many species.
6. Click on the BLAST button at the bottom left and wait 15-60 seconds for the results to return.
7. When the results page appears, next to "Results for", select Juliet (if not already selected).
8. Scroll below the "Graphic Summary" section and look at the first entry in the "Descriptions" table.
9. **Bacteroides fragilis**
10. What do you know about that species? Search the Web to find out more. **Bacteroides fragilis is an obligately anaerobic, Gram-negative, rod-shaped bacterium. It is part of the normal microbiota of the human colon**
11. Does Juliet appear to be healthy? **Yes**
12. Go back to "Results for" at the top of the page and select Romeo.
13. What species is the first sequence from for Romeo? **Helicobacter pylori**
14. What do you know about that species? Search the Web to find out more. **Helicobacter pylori is a type of bacteria. These germs can enter your body and live in your digestive tract. After many years, they can cause sores, called ulcers, in the lining of your stomach or the upper part of your small intestine. For some people, an infection can lead to stomach cancer.**
15. Does Romeo appear to be healthy? **No**

# Lesson Plan #2 – Bioinformatics.ca

Ann Meyer

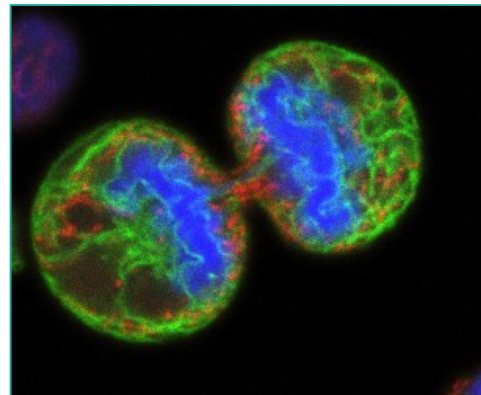


# Bioinformatics in Cancer Research



## What is cancer?

- Cancers are diseases of uncontrolled cell growth
- Cancer cells grow and divide continually because of changes in the DNA sequence of key genes, known as cancer genes



Human melanoma cell undergoing cell division  
Image credit: Paul Smith & Rachel Errington, Wellcome Images

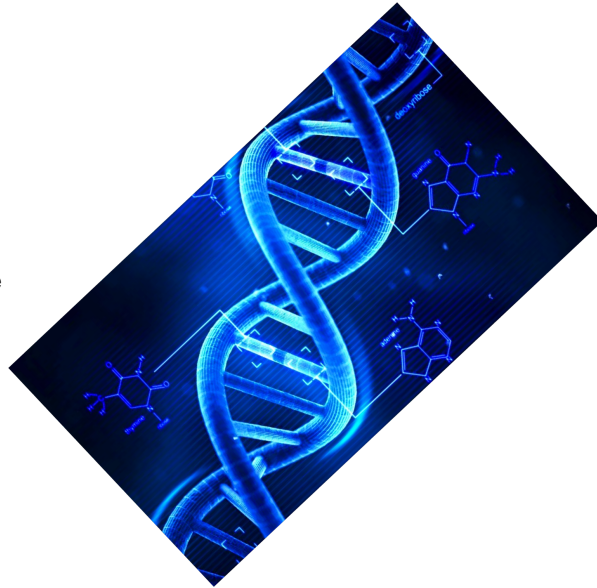
## What is DNA?

The blueprint that makes you!

Double helix structure

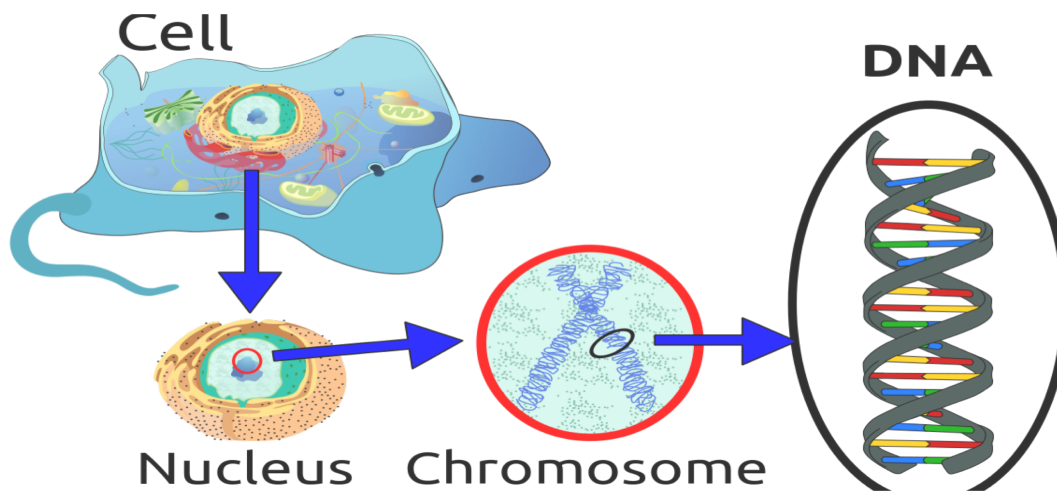
Information stored as 4 nucleotide bases:

Adenine, Thymine, Cytosine, and Guanine



Collaborate. Translate. Change lives.

## Where is DNA?

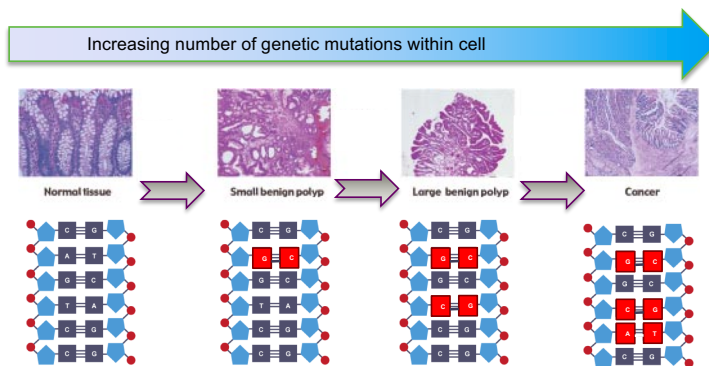


By Eukaryote\_DNA.svg

Collaborate. Translate. Change lives.



## How does a normal cell change into a cancer cell?



<http://biologywriter.com/wp-content/uploads/2015/03/figure7.jpg>

Collaborate. Translate. Change lives.

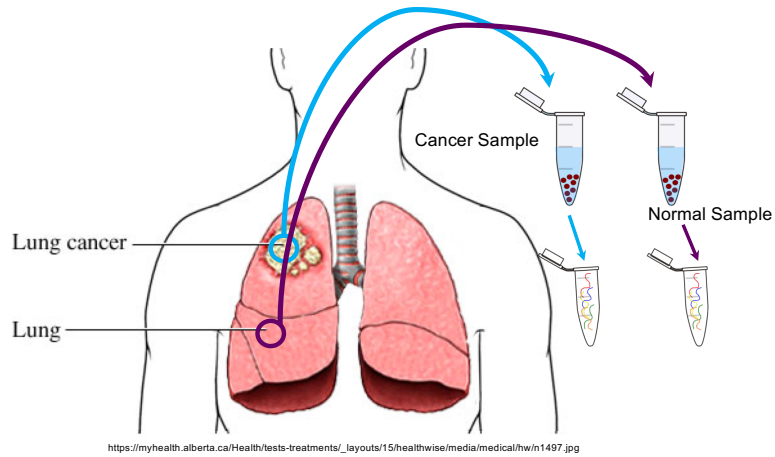
## Where do mutations come from?

1. Family genetics → you are born with these
2. Random mutation events in DNA replication
3. Lifestyle and habits
4. Environmental

Mutations in DNA accumulate over time to cause cancer

Collaborate. Translate. Change lives.

## How do we find cancer mutations?



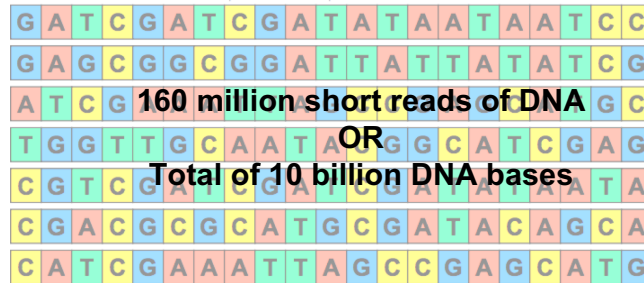
Collaborate. Translate. Change lives.

## We can sequence the DNA of the cancer cells!



Collaborate. Translate. Change lives.

## DNA sequencers produce millions of short DNA reads!



To understand the information in the DNA sequence reads, we need to **assemble** the fragments

This will let us compare the normal DNA sequence to the cancer DNA sequence.

Collaborate. Translate. Change lives.

## Assembly is a job for Bioinformaticians!

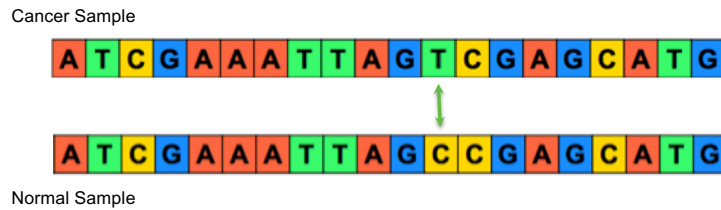


Collaborate. Translate. Change lives.

## Using a sequence assembly

Once the sequences from the tumor and normal DNA are assembled, the two assemblies are compared

We can use the assemblies to find mutations



Collaborate. Translate. Change lives.

## How do we find our mutated gene in the human genome?

A genome is the complete set of an organism's DNA including all the genes

The human genome is over 3 billion bases long!

GCTACCTTTATCCCTAGCCCCCTGCGCCCCGCGCCCTTGGCA

?



Collaborate. Translate. Change lives.

# Which gene is mutated resulting in cancer?

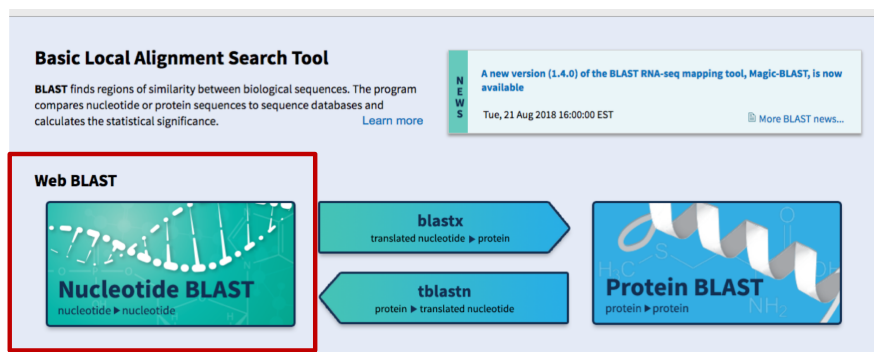
## Exercise:

The cancer sample from our patient has been sequenced and we need to find what gene the mutated sequence belongs to so that we can personalize their cancer treatment. What gene is likely driving their cancer? Is this a common mutation in cancer?

**Link to Cancer Bioinformatics Exercise:** <https://tinyurl.com/y8gsopz4>

Collaborate. Translate. Change lives.

## Using BLAST to identify the mutated gene



<https://blast.ncbi.nlm.nih.gov>

# Using BLAST to identify the mutated gene

**BLAST** → blastn suite Home Recent Results Saved Strategies Help

Standard Nucleotide BLAST

Enter Query Sequence:  Reset page Bookmarks

Enter accession number(s), g(i)s, or FASTA sequence(s)  Query subrange

From:  To:

Job Title:

Align two or more sequences

Choose Search Parameters

Database:  Human genomic + transcript  Mouse genomic + transcript  Others (nr etc.):

Human genomic plus transcript (Human G+T)

Exclude Optional:  Models (XM/XP)  Uncultured/environmental sample sequences

Limit to Optional:  Sequences from type material

Entrez Query Optional:

Program Selection

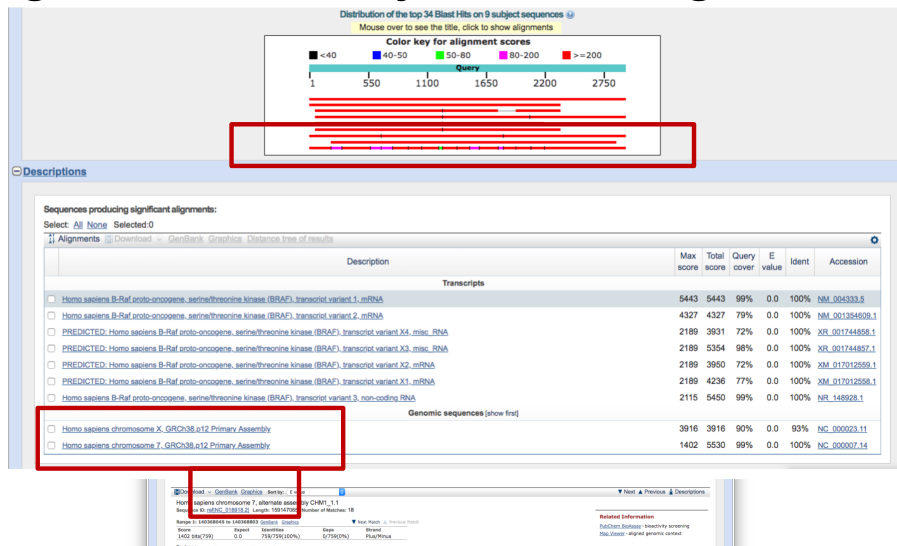
Optimize for:  Highly similar sequences (megablast)  More dissimilar sequences (discontiguous megablast)  Somewhat similar sequences (blastn)

Choose a BLAST algorithm:

Search database Human G+T using Megablast (Optimize for highly similar sequences)  Show results in a new window

15

# Using BLAST to identify the mutated gene



16

# Answers to BLAST

1. The sequence is located on Chromosome X and 7.
2. The sequences matches *BRAF*.
3. Gene title: B-raf proto-oncogene, serine/threonine kinase; location: complement(140,368,045-140,559,148); length: 191,104

17

Collaborate. Translate. Change lives.

## Using COSMIC to find common mutations

The screenshot shows the COSMIC website interface. At the top, there is a navigation bar with links for Projects, Data, Tools, News, Help, About, and Genome Version. A search bar is located in the top right corner. Below the navigation bar, the main content area features a section for 'COSMIC v86, released 14-AUG-18' and a 'COSMIC News' section. A search bar is highlighted with a red box, containing the text 'eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell' and a 'SEARCH' button. The 'COSMIC News' section includes a 'Follow @cosmic\_sanger' button and a 'Have your say!' section. The 'Projects' section lists 'COSMIC' and 'Cell Lines Project'.

<https://cancer.sanger.ac.uk/cosmic>

18

ONTARIO INSTITUTE FOR CANCER RESEARCH

Collaborate. Translate. Change lives.

# Using COSMIC to find common mutations

**Gene**  
BRAF

**Gene view**

The gene view histogram is a graphical view of mutations across BRAF. These mutations are displayed at the amino acid level across the full length of the gene by default. Restrict the view to a region of the gene by dragging across the histogram to highlight the region of interest, or by using the sliders in the filters panel to the left. [Show more](#)

**Filters**  
Show advanced filters

**Range** Show input fields  
531 - 621

**Coordinate system**  
 Amino-acid  
 cDNA

Apply filters | Reset filters

GRCh38 · COSMIC v86

Substitutions

Amino acid

Plam

Complex

SS

Insertions

Deletions

CNV Gain

# Using COSMIC to find common mutations

**Gene**  
BRAF

**Gene view**

The gene view histogram is a graphical view of mutations across BRAF. These mutations are displayed at the amino acid level across the full length of the gene by default. Restrict the view to a region of the gene by dragging across the histogram to highlight the region of interest, or by using the sliders in the filters panel to the left. [Show more](#)

**Filters**  
Show advanced filters

**Range** Show input fields  
600 - 621

**Coordinate system**  
 Amino-acid  
 cDNA

Apply filters | Reset filters

GRCh38 · COSMIC v86

Substitutions

Amino acid

Plam

Complex

SS

Insertions

Deletions

CNV Gain



## Using COSMIC to find common mutations

**Mutation**  
COSM476

- Overview
- Tissue distribution
- Samples
- Pathways affected
- References

Reset page

**HomoloGene** 3197, view the [multiple sequence alignment](#)

**Ever confirmed somatic?** Yes

**FATHMM prediction** Pathogenic (score 0.99)

**Remark** n/a

**Recurrent** n/a

**Drug resistance** In the samples curated with this mutation, resistance has been observed for the following drugs. **Note** that the same resistance pattern may not apply to all samples. For more details, look at the [Samples](#) section.  
Imatinib

**Tissue distribution**

This section displays the distribution of mutated samples and tissue types (top 5). You can see more information on our [help pages](#).

**Tissue Distribution**

Tissue	Total number of samples
Thyroid	~1400
Skin	~600
Large intestine	~500
NS	~100
Haematopoietic and lymphoid	~50

21

## Using COSMIC to find common mutations

Stomach (3863 / 27166)  
Testis (334 / 2294)  
Thymus (157 / 1116)  
Thyroid (2993 / 80019)  
Upper aerodigestive tract (4966 / 20625)  
Urinary tract (8830 / 23258)

Ankle (11)  
Anorectal (3)  
Anus (2)  
Arm (156)  
Axilla (13)  
Back (251)  
Breast (13)

Filter by screen type:

**Cancer browser**

**Results**

- Genes
- Genome browser
- Mutation matrix
- Distribution
- Variants
- Samples

Reset page

**Your selections:**  
Skin (18504 / 46578)

**Genes**

This tab shows genes that have mutations for the current tissue/histology selections. Read more on our [help pages](#)>

Show 10 entries

Gene	Mutated samples	Samples tested
BRAF	11114	22774
NRAS	2184	14541
TP53	1213	4608
TERT	1122	4744
CDKN2A	778	4562
HRAS	613	5827
KIT	523	7307
TTN_ENST00000356127	521	1277
TTN_ENST00000342992	492	1277

22

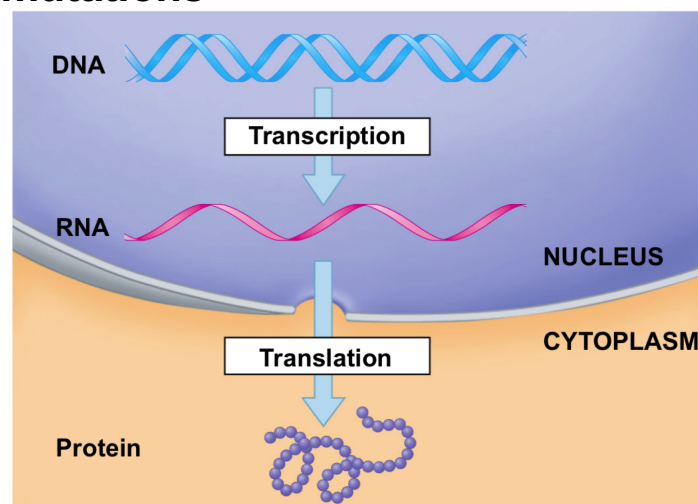
## Answers to COSMIC

1. The mutation is a T > A causing an amino acid change from V > E. 27800 samples have this mutation. It is common.
2. BRAF mutations are commonly found in Thyroid, Skin, and Large Intestine.
3. BRAF, NRAS, and TP53 are the 3 most commonly mutated genes.

23

Collaborate. Translate. Change lives.

## Impact of mutations

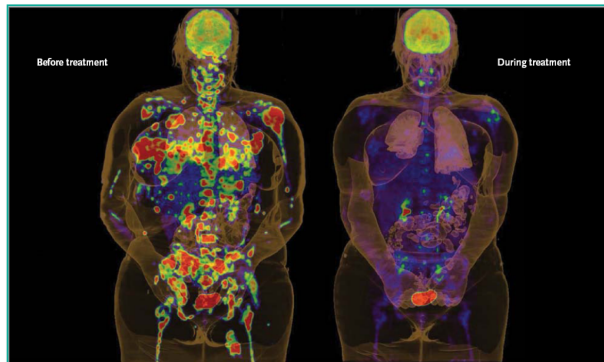


© 2012 Pearson Education, Inc.

Collaborate. Translate. Change lives.

## Common mutations provide potential new drug targets

V600E mutation discovered in 2002 at the Sanger Institute and the Institute for Cancer Research



Images courtesy of Grant McArthur, Jason Callahan, and Rod Hicks of the Peter MacCallum Cancer Centre.  
*McDermott, Downing and Stratton. N Engl J Med 2011;364:340-50.*

Collaborate. Translate. Change lives.

## Challenges to treatment



Link to video: <https://www.yourgenome.org/video/braf-from-gene-to-cancer-therapy-video>

Collaborate. Translate. Change lives.

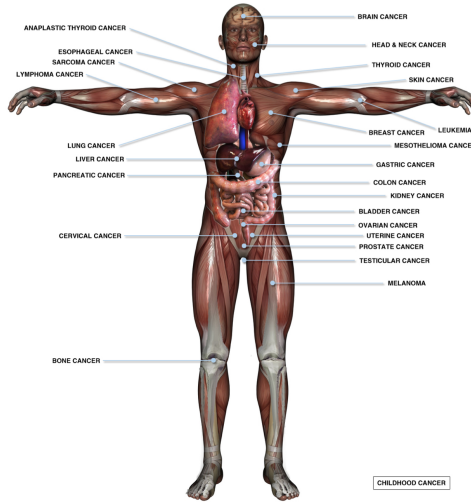


Funding for the Ontario Institute for Cancer Research  
is provided by the Government of Ontario



## Optional background slides

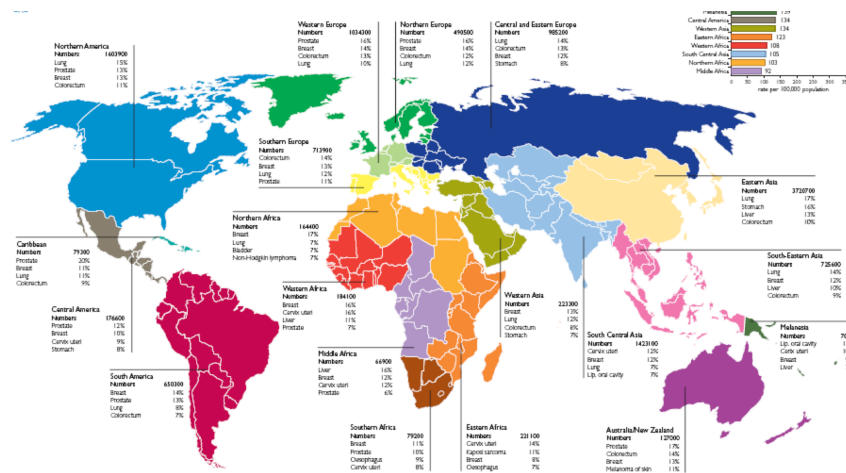
# There are more than 200 types of cancer



<http://kadev2.com/sites/all/themes/nfortheme/css/images/body.jpg>

Collaborate. Translate. Change lives.

# Cancer incidence worldwide



Source: Cancer Research UK.  
<http://info.cancerresearchuk.org/cancerstats/world/the-global-picture/>

Collaborate. Translate. Change lives.

# 2 in 5 Canadians will develop cancer in their lifetimes



Source: Canadian Cancer Society

Collaborate. Translate. Change lives.

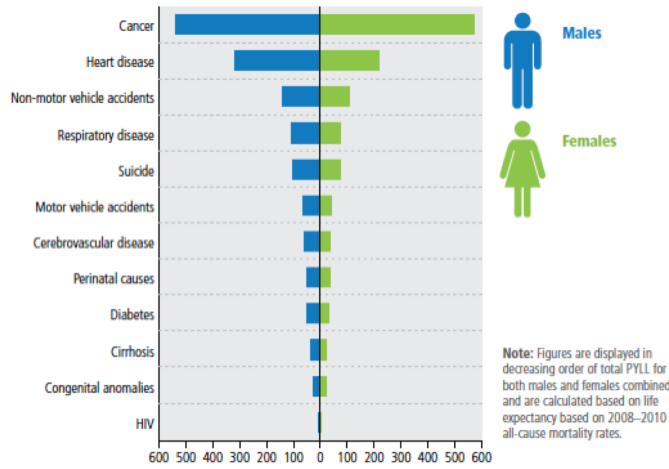
# 1 in 4 Canadians will die from cancer



Source: Canadian Cancer Society

Collaborate. Translate. Change lives.

# Cancer is the leading cause of death in Canada

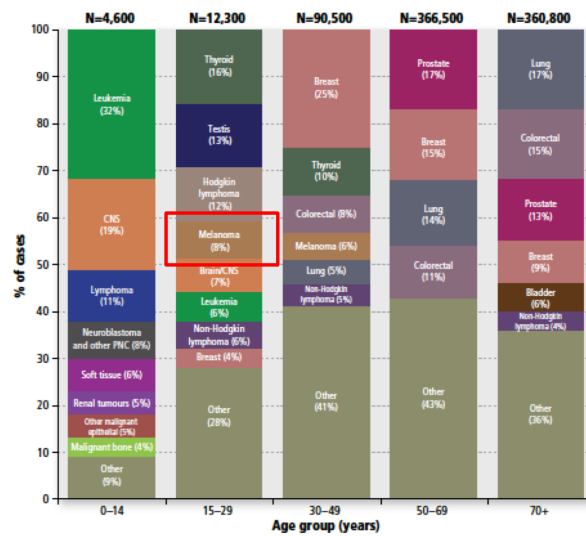


Source: Canadian Cancer Society

33

Collaborate. Translate. Change lives.

# Distribution of new cancer cases by age group

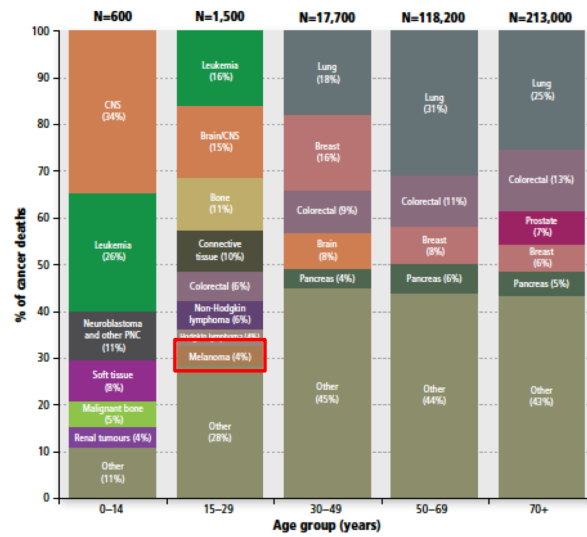


Source: Canadian Cancer Society

34

Collaborate. Translate. Change lives.

# Distribution of cancer deaths by age group



Source: Canadian Cancer Society 35

Collaborate. Translate. Change lives.





## Cancer Bioinformatics for High Schools

---

**Developed by:** Ontario Institute for Cancer Research and [bioinformatics.ca](http://bioinformatics.ca) with acknowledgements for original work from Wellcome Genome Campus.

**Text adaptation:** Ann Meyer

**License:** All the included material is protected by the Creative Commons ShareAlike 2.5 Canada license CC BY-SA 2.5 CA (<https://creativecommons.org/licenses/by-sa/2.5/ca/>)

For any questions or comments, please contact the [bioinformatics.ca](http://bioinformatics.ca) at [course\\_info@bioinformatics.ca](mailto:course_info@bioinformatics.ca)

---

### INTRODUCTION

#### **Cancer research in the personalized genomics era**

Cancer research has benefited greatly from recent advances in DNA sequencing technologies and the availability of large cancer datasets and databases. This has spurred our ability to personalize cancer diagnosis and treatments based on a patient's cancer DNA mutations. Ongoing research projects in Canada and worldwide aim to find more treatment targets to increase the success of cancer treatments.

We will be using 2 bioinformatics resources in this exercise. The first, the Basic Local Alignment Search Tool (BLAST) from the National Center for Biotechnology Information (NCBI) will let us identify the gene that is mutated in our patient. It works by comparing DNA sequences from unknown, sample genes to a database of DNA sequences of known genes. The top hit it finds is very likely to be our mystery gene. The second tool we will be using, Catalogue of Somatic Mutations in Cancer (COSMIC) will let us explore common cancer mutations.

## CLASSROOM HANDS-ON EXERCISE

Requirements: A computer with internet access

### Using BLAST to Identify the Mutated Gene

1. To start, go to the NCBI BLAST page: <https://blast.ncbi.nlm.nih.gov>
2. Under “Web BLAST”, click the “Nucleotide BLAST” button. This will let us compare DNA sequences to DNA sequences.
3. In the “Query Sequence” section, in the text box under "Enter accession number(s), gi(s), or FASTA sequence(s)", copy and paste all lines in the file "Cancer Gene Sequence for Cancer Bioinformatics Exercise". The DNA sequence can be found here <https://tinyurl.com/y76jbecm>.
4. In the “Choose Search Set” box, beside “Database”, select “Human genomic and transcript.” This will let us compare human DNA to human DNA.
5. Hit BLAST to start the comparison of your DNA sequence to the whole Human genome and transcriptome.
6. Wait until your results page appears. BLAST will return locations in the human genome and transcriptome that match our input sequence.
7. On which chromosome (or chromosomes) is our sequence located? Hover over each of the red lines to determine the chromosome number. Hint: the bottom 2 red lines will show the genomic hits with chromosome numbers.
8. In the ‘Descriptions’ box under ‘Genomic Sequences’, select the result with the highest percent identity (column labelled ‘Ident’). This jumps down the page to the result.
9. Download the ‘Graphics’ for this result. This opens a new ‘Graphics’ tab.
10. In the ‘Genes’ section, which gene does your sequence match with? Hover over the gene: what is the gene title, location, and length?
11. Select ‘View MIM’ from the pop up window to learn about the function of your gene. Learn about mutations in your gene in the ‘Molecular Genetics’ section.

### Using COSMIC to find common mutations

1. Go to <http://cancer.sanger.ac.uk/cosmic/>
2. In the search box, type “BRAF”, the gene we identified with BLAST.
3. In the table at the bottom of the screen, in the Gene column, click on BRAF.
4. Move the triangle sliders to focus in on the amino acid range for the BRAF mutation (581-621) and click apply filter.
5. In the histogram, there is a tall divided bar which represents the number of samples with a mutation at this position. Hover over the largest segment. What is the mutation? How many samples have this mutation? Is it common?
6. Click on the largest portion of the bar. Scroll to Tissue Distribution. In which tissues are BRAF mutations commonly found? Hover over the bar to see numbers.
7. Click on Skin. Click the Genes with Mutations tab. Sort the table by mutated samples from highest number to lowest number. What are the 3 most commonly mutated genes?

# Lesson Plan #3 – Dutch Techcentre for Life Sciences

Charlotte Zwetsloot



## Bioinformatics: a bit of life



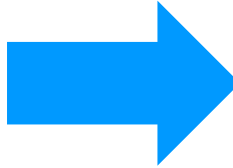
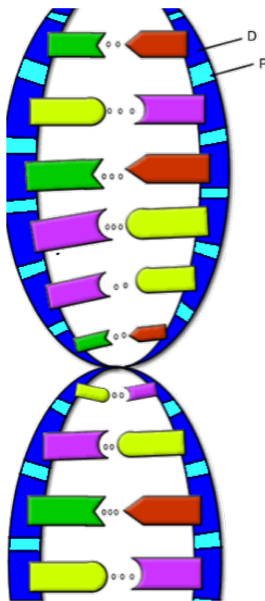
## Bioinformatics: a bit of life

Two units:

1. Murder at the airport  
(35 minutes)
2. Designing an antivenom  
(65 minutes)



# DNA is a code

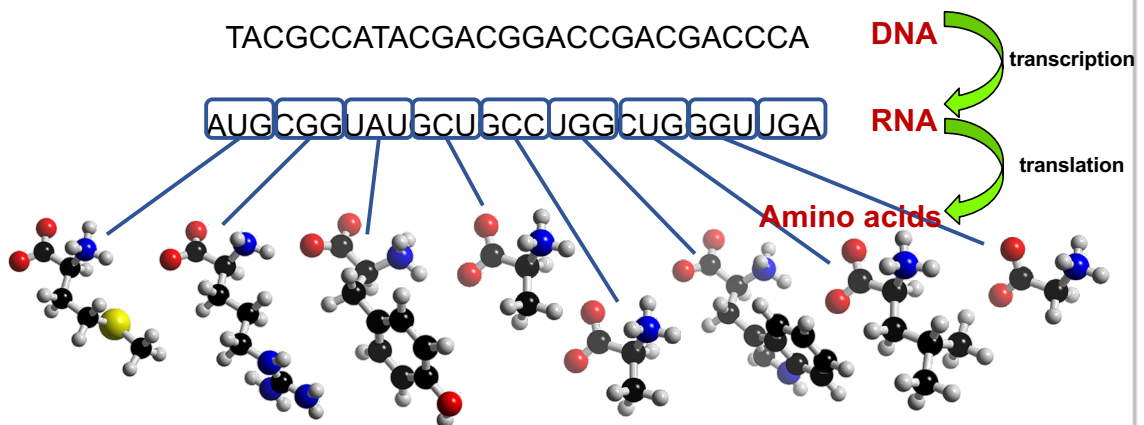


```

ATTATACGAAGCCAAAAAGTCTACCCGACTCAAAACTAAATCCTG
AGTATTTGCTATTTCAGTCAAGACATACAGCTCTCGAAACAGGCAG
ACCTCAAACCAAAGAGATTTACTATTGAGAGTGATTTGCAAAA
TTCATCATAACAATTATAACCCGGAAGAGTTCCCGAAAGCACAT
ACTACGTCGCCAGAGGCTCAAATGTTATGCATATTGTCCCCTTG
ATTTAGACCCCAATGTTTGAACGAAATGATTTAGATACTACAT
FCGATCAAGTTTTTAAAGATAAAAATGCTGTGAGCAAAAAGTTC
AATTTTTAAAGAACAACCTTTAAGAATTGACTCAAATAAAACATTCG
AAAACCTGGAACATATCCATGTTTACAACACTTTAAAAACTACA
ACCTTAATGTGGTAAATCAACTTTGATCAATACTCTACTACAGA
AGGTTAAAATTGATTCTACTGGGAAAATTAACCTCCCCTCTGAAG
TTTTTACGAACCCGAAGAATTTTTTCAAGATTCAGCTGCTGGTG
ATTTAACTAGATCCGTACAAGCCTACCAAGTTGGAGGTAAAATAC
ECTATTCAACTTCCACCTCCAGATTACGCCTAGAAGAATAATCG
AAGGCTGCGAAAAACCGATTTATTCAATCGGAAGCATATAAAAC
TATGAAGGGCACATCACAGGGAGGCTGTTATACCGTGGGAGGGA
TCCGAAAAGGATCCATAAATCAGATTGTGAAATACATACCGAGG
AGAACATTGAAAAGGTATTGATGTATTCAACTCGTCAATTCAT
TGTATCACGGTATTGTGGAATTAAGCGTTATATGCGAAAAAA
ACGCAATACCTCCATTATTTGGAAATAGAGATCGTCTGAAAAG
PATTAAGGACAACCTGGGAGGTATGAGTTCAAAGGACTACATGAGA
TATCCAAGTAGGTATAAGGGAACCGCTAGAAAACCTTGATTGAGT
ACATTGAAAACAATGGTAAAGAGTCAAGCTGTCCAGAGATAGAC
TGTATGAGATGGCCCCGGATGAAGCTGATACTTTGAACGCAGTGA
AAAAACCGAGAAAGATTTATCTGCTAGAAGATTTGTGACGATG
    
```



# From DNA to protein



Methionine (start), arginine, tyrosine, alanine, alanine, tryptophan, leucine, glycine

One letter code: M R Y A A W L G



# Proteins



**AMGEN** Biotech Experience  
Scientific Discovery for the Classroom  
Nederland

DNA  
LABS

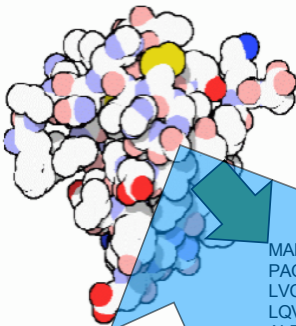


**DTL**  
DUTCH TECHCENTRE FOR LIFE SCIENCES

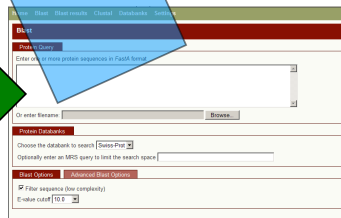


**Radboudumc**

# BLAST: identifying proteins



```
MALWMRLPL LALLALWGPD  
PAQAFVNQHL CGSHLVEALY  
LVCGERGFY TPKTRREAED  
LQVGQVELGG GPGAGSLQPL  
ALEGSLQKRG IVEQCCTSLC  
SLYQLENYCN
```



1. Scientists **sequence proteins**
2. The amino acid sequence is **stored in databases**
3. **BLAST compares** amino acid sequences
4. When a 'match' occurs, the protein is **identified**

**AMGEN** Biotech Experience  
Scientific Discovery for the Classroom  
Nederland

DNA  
LABS



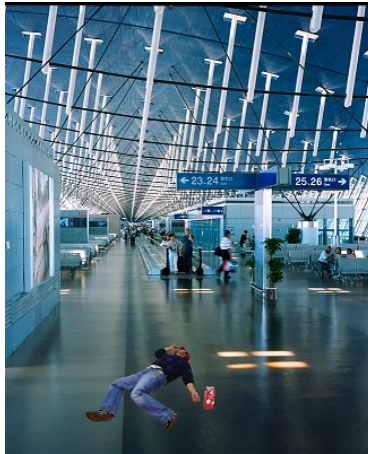
**DTL**  
DUTCH TECHCENTRE FOR LIFE SCIENCES



**Radboudumc**



# Murder at the airport



→ Identify the four proteins

→ <http://www.bioinformaticaindeklas.nl/en/education/murder-at-the-airport/>

**AMGEN** Biotech Experience  
Scientific Discovery for the Classroom  
Nederland

DNA  
LABS

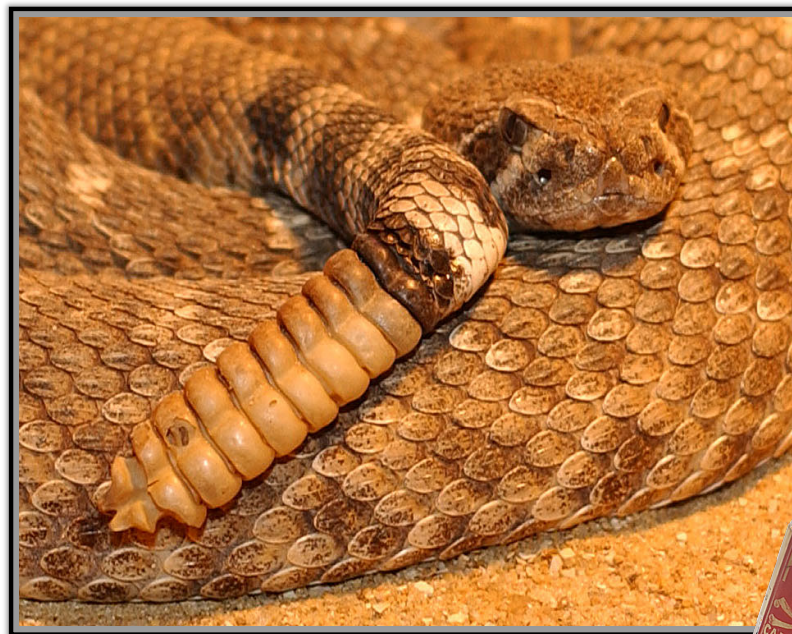


**DTL**  
DUTCH TECHCENTRE FOR LIFE SCIENCES



**Radboudumc**

# The killer!



**AMGEN** Biotech Experience  
Scientific Discovery for the Classroom  
Nederland

DNA  
LABS



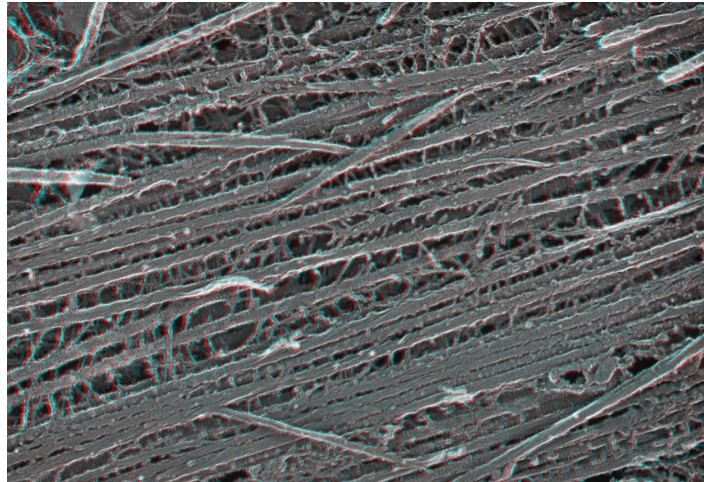
**DTL**  
DUTCH TECHCENTRE FOR LIFE SCIENCES



**Radboudumc**

# 3D Drug Design

## Collagen



**AMGEN** Biotech Experience  
Scientific Discovery for the Classroom  
Nederland

DNA  
LABS



**DTL**  
DUTCH TECHCENTRE FOR LIFE SCIENCES

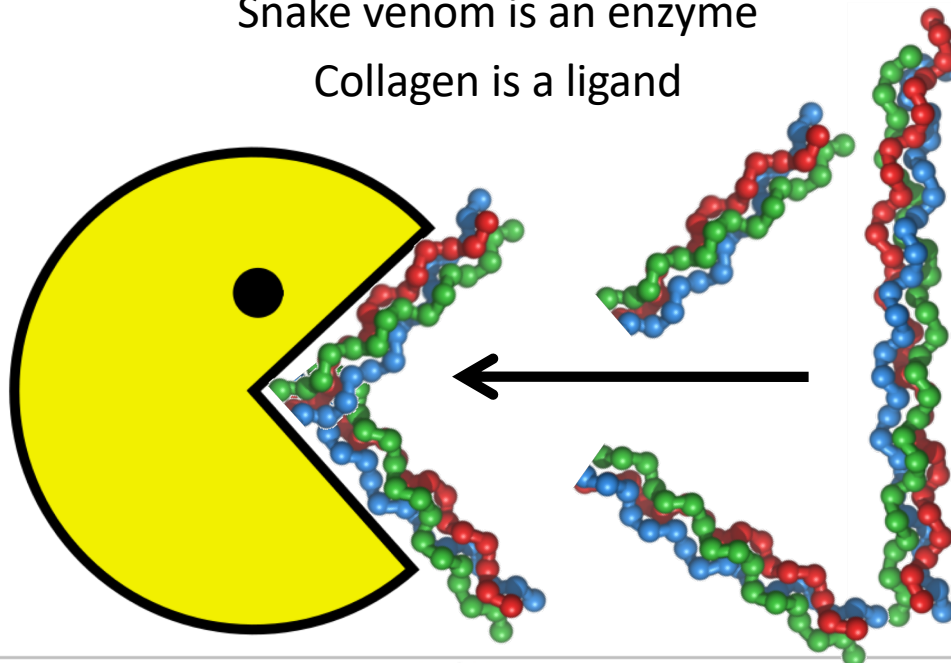


**Radboudumc**

# Snake venom

Snake venom is an enzyme

Collagen is a ligand



**AMGEN** Biotech Experience  
Scientific Discovery for the Classroom  
Nederland

DNA  
LABS



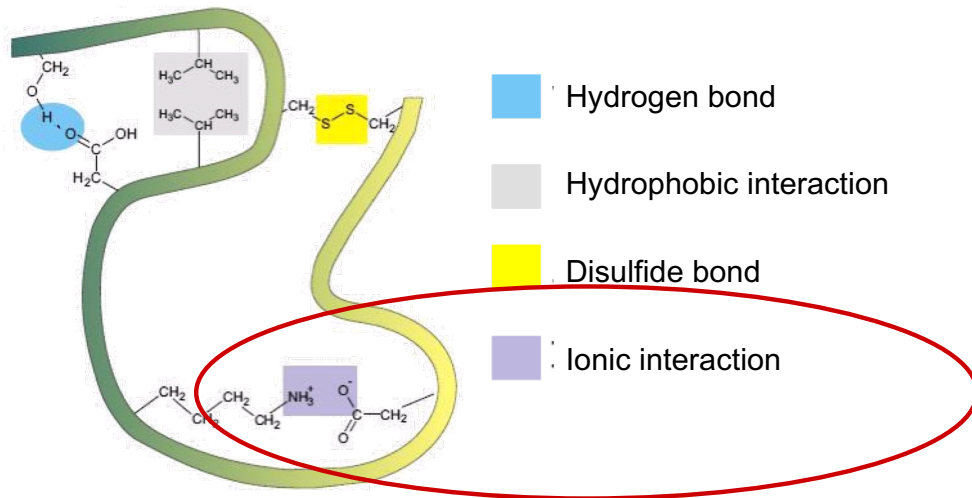
**DTL**  
DUTCH TECHCENTRE FOR LIFE SCIENCES



**Radboudumc**

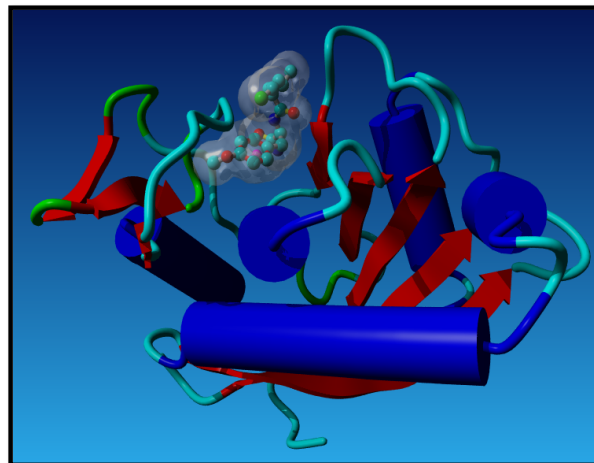


# Interactions in proteins



The strongest interaction is the ionic interaction

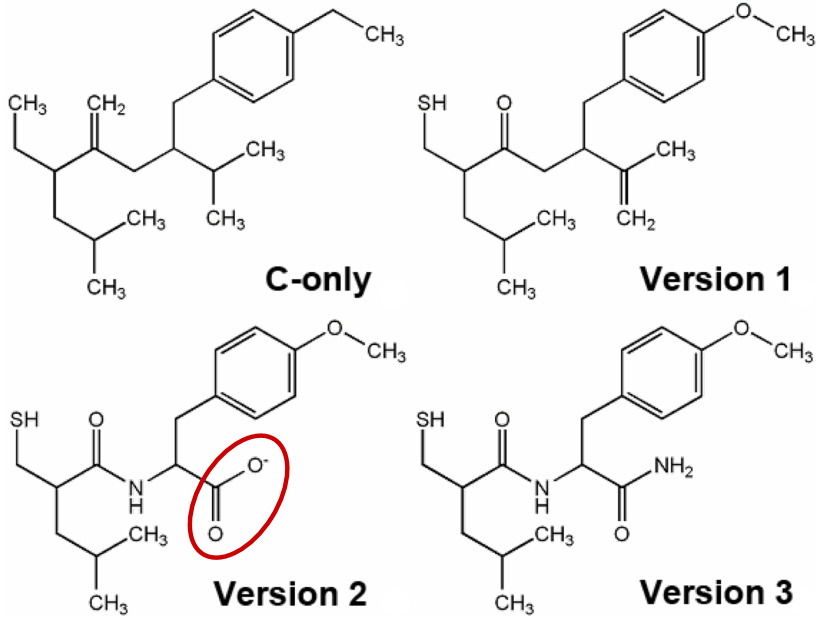
# 3D Drug Design



→ <http://www.bioinformaticaindeklas.nl/en/education/designing-an-antivenom-in-3d/>

→ 3D Drug design

# The best antidote?



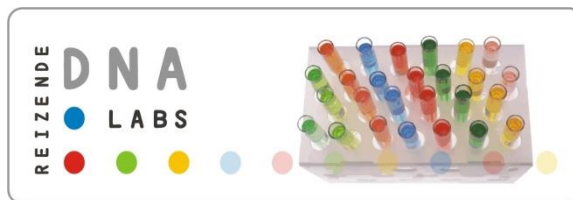
**AMGEN** Biotech Experience  
Scientific Discovery for the Classroom  
Nederland

DNA  
LABS



**DTL**  
DUTCH TECHCENTRE FOR LIFE SCIENCES

 **Radboudumc**



# *Teacher's Guide*

Level: expert

***Bioinformatics:  
a bit of life  
Do-It-Yourself***

# bioinform@tica in de klas



Radboudumc

## Amgen Biotech Experience

### Scientific Discovery for the Classroom

Developed by the Netherlands Bioinformatics Centre in cooperation with the Centre for Molecular and Biomolecular Informatics of the Radboud University Nijmegen Medical Centre

#### Text

Celia van Gelder, Robbie Joosten and Hienke Sminia

#### Illustrations

[www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)

#### Design

Identim, Wageningen

All the included material is protected by the Creative Commons Naamsvermelding-Niet-commercieel-Gelijk delen 3.0 Nederland license (<http://creativecommons.org/licenses/by-nc-sa/3.0/nl/>).

#### CC BY-NC-SA 2017 – NBIC

For any questions or comments, please contact the Netherlands Bioinformatics Centre ([nijmegen@dnlabs.nl](mailto:nijmegen@dnlabs.nl)).

Disclaimer: Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Amgen Foundation or Education Development Center, Inc.

## Table of contents

<b>DNA Labs on the Road</b>	<b>5</b>
Structure of the module 'Bioinformatics: a bit of life'	5
Target group	5
Required knowledge	5
Learning objectives	6
Grading	6
Practical information	6
Computer guide	6
Supplementary material	7
<b>Lesson 1: Introduction</b>	<b>8</b>
Required materials	<b>Fout! Bladwijzer niet gedefinieerd.</b>
Planning introductory lesson	<b>Fout! Bladwijzer niet gedefinieerd.</b>
<b>Lessons 2 and 3: Practicum</b>	<b>8</b>
Structure of the practicum	10
Required materials	10
Answers	10
<b>Lesson 4: Concluding lesson</b>	<b>14</b>
Concluding modules	14
Suggestions	14

## **DNA Labs on the Road**

The DNA Labs on the Road integrate biology and chemistry and introduce the newest developments in genomics by having students work with advanced techniques and current subjects in scientific research. The DNA Labs were developed by Dutch universities and the Genomics Centres of Excellence.

There are six different DNA Labs, each of which introduce a different aspect of the modern-day DNA research. These labs show that knowledge of genes and molecules in the cell plays an important part in matters that are important to every human being: nutrition, health and environment. Furthermore, the practical lessons demonstrate that scientific progression may produce ethical and social questions.

The DNA Labs are meant for students in classes havo/vwo classes 4 and up. The practicum takes 100 minutes.

### **Structure of the module 'Bioinformatics: a bit of life'**

The DNA Lab on the Road module 'Bioinformatics: a bit of life' introduces students to modern-day computer research of DNA and proteins. During the practical lesson, students will do their own research as a bioinformatician to investigate an eye-related disease. Studying 3D protein structures and targeted searching through online databases are important components of this DNA Lab. Several biological concepts will be discussed in context of this research, giving students a better visual image and understanding. For example, they will see that proteins are three dimensional molecules of which the shape and structure is essential for its function, and that this structure is dependent on the nucleotide sequence in the DNA.

This module consists of four lessons: the introductory lesson, the practical lesson (lesson 2 and 3) and the concluding lesson. Materials for the lessons are given in this Teacher's Guide, to prepare the students optimally for the DNA Lab on the Road. The evaluation and feedback from participating teachers has shown that the introductory and concluding lessons are very important for achieving the optimal learning experience. Without proper introduction, the practical lesson is too difficult to understand properly, and without proper conclusion, the students often have difficulty to effectively process what they've learned and how to place this in context. Moreover, the introductory lesson is very suitable to refresh the students' memory on DNA, proteins and cells. The concluding lesson can be used to elaborate on a biological or chemical aspect, or to expand on social aspects of the research.

### **Target group**

The module is meant for students grade 10-12, to be given in the Biology or Chemistry. Due to the difference in level of knowledge and skills of different classes, we decided to provide this module at two levels. The basic version is meant for students that are not yet fully acquainted with subjects like DNA and proteins. The expert version offers more challenging concepts for classes that already master the subject.

### **Required knowledge**

Specific knowledge about bioinformatics is not necessary. However, the students will need to possess a certain level of knowledge of molecular biology in order to be able to participate in this DNA Lab. The students should at least know the terms and concepts concerning DNA, RNA, proteins, protein synthesis, amino acids and enzymes.



## Learning objectives

This module gives students more insight in the new developments in DNA research that is conducted more and more in computers nowadays. Due to the immense amounts of biological data obtained by Sequencers, Micro-arrays and Mass Spectrometers, computers are now required to process these data. Using smart software and online databases, the modern-day scientist analyses the whole genome. The students will learn how to apply bioinformatics in a practical fashion in the context of an eye-related disease.

### Subject specific objectives

*At the end of the module, students will be able to*

- explain how the amino acid sequence determines the structure of a protein
- explain the role of atomic interactions in proteins
- explain how an enzyme works
- explain how a toxin exerts its harmful effects on the body
- understand that proteins have a three dimensional structure

### Practical learning objectives

*At the end of the module, students will be able to*

- identify amino acid sequences using BLAST
- operate 3D modelling software YASARA
- explain in what way bioinformatics can be used in solving a crime
- explain in what way bioinformatics can be used in developing a medicine

## Grading

To give your students a grade for their work, you can use exam questions and a practical assignment which can be given after the concluding lesson. This grading tool can be found on the website [www.dnalabs.nl](http://www.dnalabs.nl).

## Practical information

'Bioinformatics: a bit of life' is brought to you by the Netherlands bioinformatics Centre in cooperation with the Radboud University Medical Centre.

For more information about this module, please contact us via [nijmegen@dnalabs.nl](mailto:nijmegen@dnalabs.nl).

*The conditions for doing this DNA lab on the Road practical are:*

- The computers in the classroom meet the requirements as stated in 'computer guide'
- A beamer is present in the classroom
- Students are prepared with the introductory lesson (lesson 1)
- The teacher will give a concluding lesson to elucidate or expand on the DNA Lab lesson
- The teacher prints a sufficient amount of students' guides (at least one per student duo)

## Computer guide

To make the practicum take place as smoothly as possible, these requirements should be met

**Computer** – Enough computers should be present: at least one computer per student duo.

**Internet connection** – The students will conduct their practicum through the website [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu). Moreover, they will search an online database for homologous proteins. The computers therefore need to be connected to the internet.

**Yasara** – For designing a medicine, the students will use 3D modelling software Yasara. Below, you will find instructions for installing this software on your computers.

**Yasara files** – Specific files are needed for this practicum, which the students will load into Yasara. These files can be downloaded from this website (at the bottom of the web page):

<http://www.bioinformaticsatschool.eu/lesmateriaal/3d/3d.html>

## Website

*Follow these steps to have a look at the practicum:*

- Go to: [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)
- Click 'Education' in the menu

- Then, click 'Material' in the menu. You are now directed to an overview of all the bioinformatics lessons material. This *DNA lab on the Road* module is named 'Losing Sight and vision'.
- Click the practicum subject you want to view. A new window will open, containing a short description of the practicum.
- At the left side of the menubar, click 'start the lab' button in the menu.
- The practicum has now started. At the bottom of the page, you can click to continue to the next section.

### Yasara

Our new version of Yasara can be used with Windows as well as Linux. Minimal system requirements are often not able to indicate if Yasara will work properly. Therefore, we advise you test Yasara on one (or more, preferably) of the computers first. Systems that are three years old or newer usually work fine. If possible, we recommend the use of optic mice.

On the website [www.yasara.org](http://www.yasara.org), you can download Yasara-view. You don't need any license or admin rights on the computer, and the software can even be run from a CD-ROM or a USB stick.

During the practicum, we use several files in Yasara. These can be downloaded from our website: <http://www.bioinformaticsatschool.eu/lesmateriaal/rp/rples.html>.

For a more detailed instruction of Yasara, you can visit website made by a bioinformatics research assistant: <http://www.cmbi.ru.nl/~hvensela/yasara/>.

### Supplementary material

If you appreciate this practicum, we recommend you take a look at our website [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu). We are currently very busy developing lesson material about DNA and proteins which is easily used in class, not only for biology, but also for chemistry, crafts, mathematics, NLT and elementary school level. Moreover, we organize several activities and trainings. Want to know more? Contact [nijmegen@dnalabs.nl](mailto:nijmegen@dnalabs.nl).



## Lesson 1: Introduction

The introductory lesson consists of a classical PowerPoint presentation with several exercises. The exercises are integrated in the presentation and the students can write the answers in their Student's Guide. The exercises can be discussed using the following slides. It is advised to study the slides beforehand, so that you know when an exercise starts and when the answer slides will appear.

### Required materials

Documents<sup>1</sup>

- Presentation (*introduction.ppt*)
- Student Guide

Miscellaneous

- Beamer and computer
- Blackboard and chalk or whiteboard
- Handbook

### Planning introductory lesson

The PowerPoint presentation serves as a guideline for the introductory lesson.

2 minutes – Slide 1	Introduction of this lesson and the practicum that will follow
8 minutes – Slide 2 to 4	<ul style="list-style-type: none"> <li>- Explain: What is DNA? What are proteins? And how do proteins play a role in an organism?</li> <li>- Start the movie: What does protein synthesis in the cell look like?</li> <li>- How is a protein made from DNA and RNA?</li> </ul> Exercise: Which amino acids do the codons encode?
5 minutes – Slide 5	Explain: A process may influence an organism on different levels (i.e. cell, tissue, whole organism). For example, UV radiation. Exercise: complete the table on the answer sheet.
15 minutes – Slide 6 to 8	Briefly discuss the answers. The answers on the slides are not necessarily complete. Students are encouraged to come up with other answers.
5 minutes – Slide 9	What is bioinformatics? Give the students some time to think about this question. Write some of the students' ideas on the blackboard. At the end of the presentation, this will be used as a recap. <i>Bioinformatics is solving biological and biomedical problems using computers and databases.</i>
5 minutes – Slide 10 to 13	Example of bioinformatics: the platypus. The platypus possesses a wide variety of characteristics found in birds, mammals and reptiles. These genes responsible for these characteristics have indeed been proven to come from different ancestors. The gene for its venom is very similar to the reptile venom gene. Based on these novel discoveries, the platypus is placed in another place than mammals in the evolutionary pedigree.
5 minutes – Slide 14 to 16	Example of bioinformatics: Friedrich's Ataxia. A cure is yet to be found for the heritable disease called Friedrich's Ataxia. Nevertheless, bioinformatics has allowed major progression in unravelling this disease. Patients suffering from this disease have many more repetitions of the GAA bases in the <i>Frataxine</i> gene (a protein with a protective function for muscle and nerve cells) than healthy people.

<sup>1</sup> You can either find the documents on the Travelling DNA Labs website ([www.dnalabs.nl](http://www.dnalabs.nl)) or request them from [nijmegen@dnalabs.nl](mailto:nijmegen@dnalabs.nl).

5 minutes – Slide 17 and 18	<ul style="list-style-type: none"><li>- There are many more applications of bioinformatics. Some examples are briefly mentioned.</li><li>- Recap on the thoughts the students came up with on the blackboard. Are there any new thoughts about bioinformatics that come to mind?</li></ul>
-----------------------------	--

## Lessons 2 and 3: Practical

### Structure of the practicum

The practicum consists of two parts:

Practicum hour 1 – Murder at the airport

In the first hour, the students will investigate 4 proteins which were found in the drink of a murdered tourist. They will do this by comparing the amino acid sequences of these proteins to those in an online database, and fill out their findings in a “CSI-form”. In the 2<sup>nd</sup> practicum hour, they will continue with these results.

Practicum hour 2 – Designing an antivenom in 3D

The students will continue their research on the protein that killed this victim. They will try to design an antivenom based on the structure of the protein and its possible interactions with a ligand. To do this, the students will use the 3D modelling software Yasara.

The students will be guided through the lesson using the website [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)<sup>2</sup>.

### Required materials

Documents<sup>3</sup>

- Answer sheet for the students (*Expert-Bioinformatics\_Student-Guide*)

Miscellaneous

- Beamer
- Computers with internet connection and Yasara software installed (see computer guide, page 6) for every student duo.

### The practical

*Follow these steps to start the practicum:*

- Go to: [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)
- Click ‘Education’ in the menu
- Then, click ‘Material’ in the menu. You are now directed to an overview of all the bioinformatics lessons material. This *DNA lab on the Road* practical consist of two modules ‘Murder on the airport’ and ‘3D Drug Design’.
- Click the practicum subject you want to view. A new window will open, containing a short description of the practicum.
- At the left side of the menubar, click ‘start the lab’ button in the menu.
- The practicum has now started. At the bottom of the page, you can click to continue to the next section.
- The students write down their answers in the Student Guide.

### Answers

#### Practicum hour 1 – Murder at the airport

For each protein, the answers are given below

#### Candidate 1

1. CASA1\_BOVIN. This is amino acid 16-214 of alpha-s1-caseine from a cow (1-15 is the signaling peptide). This protein is one of the main ingredients of milk. Keywords: Milk, Comments: Secreted in milk.
2. Bos Taurus (cattle/cow)
3. From the comments: Important role in the capacity of milk to transport calcium phosphate.
4. The protein is not suspicious, it is found in milk.

<sup>2</sup> On page 6 of this guide, you will find practical information, including where to find the guide on the webpage. Please feel free to use other materials from the [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu) website.

<sup>3</sup> You can either find the documents on the DNA Labs on the road website ([www.dnalabs.nl](http://www.dnalabs.nl)) or request them from [nijmegen@dnalabs.nl](mailto:nijmegen@dnalabs.nl).

5. This is amino acid 16-214 of alpha-s1-caseine of the cow (1-15 is the signaling peptide).

**Candidate 2**

1. AMYS\_HUMAN, the human alpha-amylase found in saliva.
2. Human
3. Endohydrolysis of 1,4-alpha-D-glucosidic linkages in oligosaccharides and polysaccharides.
4. The protein originates from the victim and is therefore not responsible for his death.
5. The sequence given here starts at amino acid 16 of the sequence in the database (with QYSSN). MKLFWLLFTI GFCWA(1-15) is the signaling peptide of amys\_human.

**Candidate 3**

1. HRTD\_CROAT. Hemorrhagic metalloproteinase (EC 3.4.24.42) (Atrolysin D/C). Hemorrhagic means that it causes internal bleeding.
2. Crotalus atrox (Western diamondback rattlesnake).
3. It is a hydrolase, a metalloproteinase with zinc as cofactor. It is a toxin of a very venomous rattle snake.
4. The protein is suspicious, since it causes internal bleeding. In Synonyms it reads toxin, in Comments it says snake venom.
5. Zinc protease. The sequence in the database is much longer, but you can deduce that the above mentioned part is the functional protein. Amino acid 191-393.

**Candidate 4**

1. LACB\_BOVIN. This is the precursor of beta-lactoglobulin from a cow, amino acids 17-178 (1-16 is the signaling peptide), 162 amino acids. This is a major component of whey, a milk protein.
2. Bos Taurus (cattle/cow)
3. From the function in the Comments section: Primary component of whey, it binds retinol and is probably involved in the transport of that molecule. Keywords: Allergen; Milk; Retinol-binding; Signal; Transport
4. The protein is not suspicious. The keywords tell us: Milk. In Comments, the tissue specificity reads: Synthesized in mammary gland and secreted in milk.
5. From the comments: Causes an allergic reaction in human. Is one of the causes of cow's milk allergy.

**Final conclusion**

The victim died from poisoning by milk containing snake venom.

**Practicum hour 2 – Designing an antivenom in 3D****Exercise 1**

Usually, in mutation studies, individual amino acids are being mutated (mostly to Alanin). The effect of the mutation on the function of the protein is then investigated: does the protein's function improve, does it stop exerting its function or is there no change at all? Most of the mutations will have no or little effect. However, mutating the most important amino acids will have a significant impact on the function of the protein. Usually, the protein will completely fail to work when an important amino acid is mutated.

**Exercise 3**

The elements in proteins are carbon, nitrogen, oxygen, sulfur and hydrogen.

a. Atoms are colored in Yasara as follows:

- Red: oxygen
- Dark blue: nitrogen
- Green: sulfur
- Light blue: carbon

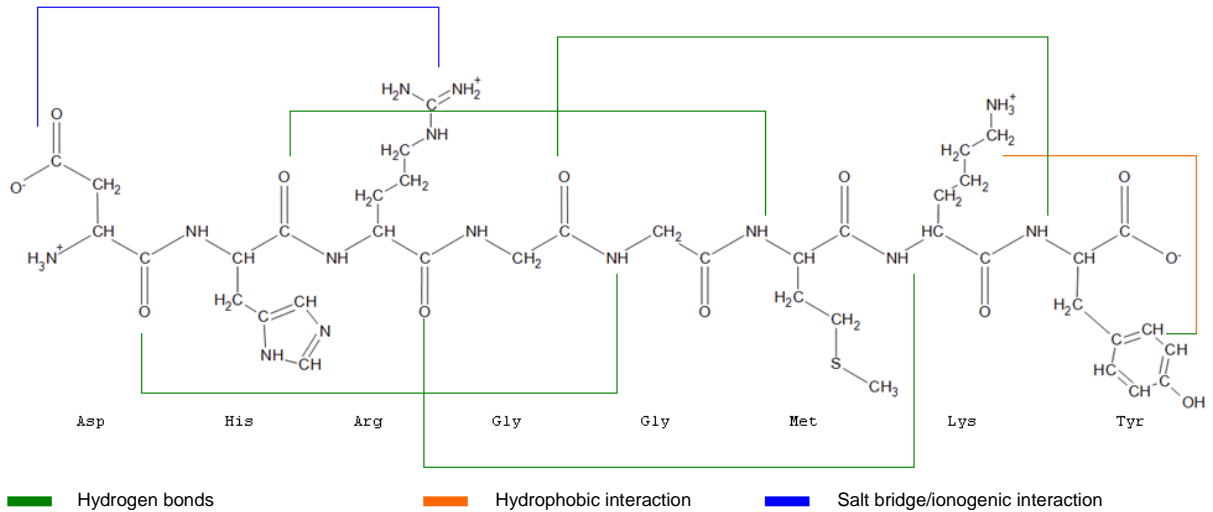
**Exercise 4**

White bonds: single (covalent) bonds; yellow bonds: double (covalent) bonds.

**Exercise 5**

- a. The three types of interactions are:
- Green: hydrogen bonds
  - Blue: salt bridge/ionogenic interaction
  - Orange: hydrophobic interactions

b.



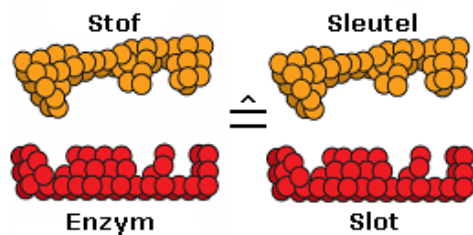
**Exercise 7**

Blue components:  $\alpha$ -helix, red components:  $\beta$ -strand (multiple strands make a  $\beta$ -sheet)

**Exercise 8**

- The double bonded oxygen is best visualized in the Ball-and-stick display.
- The Histidine side chain can be seen using any view in which the side chains are visible.
- The alpha helix can best be found using the backbone-trace or cartoon display mode.

These answers depend on your own personal preference. Only the display modes in which the object is absolutely not visible, can be accounted as wrong. The object of this exercise is to show that the ball-display, although most realistic, is not always the most informative.



**Exercise 9**

The lock-and-key principle states that an enzyme (the lock) only works if the right ligand (the key) is being bound. Another assumption of this model is that the enzyme only changes its shape slightly.

**Exercise 10**

A protein can have a cavity. This cavity has a relatively large contact surface, making many different interactions (hydrogen bonds, ionogenic interactions, hydrophobic interactions) possible.

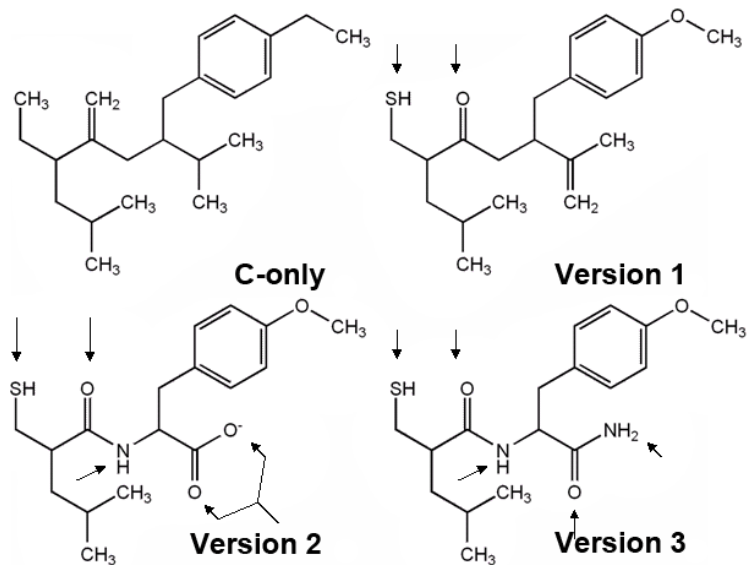
**Exercise 14**

There are different ways to disable a venom protein:

1. Destroy the harmful enzyme.
2. Block the active site with a ligand that won't be cut.

**Exercise 16**

Most of the indicated atoms can undergo hydrogen bonding. The –SH (thiol) group can only undergo weak hydrogen bonds. Such a group can also make a disulfide bond. The negatively charged oxygen in version 2 can undergo ionogenic interactions with a positively charged group in the enzyme. Note that the charge is being distributed over the whole carboxyl group. The salt bridge is thus also established by the whole group (the charged oxygen and the double-bonded oxygen).

**Final question**

Ligand version 2 is the best candidate to serve as an antivenom. In this ligand, the peptide bond creates two hydrogen bonds and the carboxyl group undergoes an ionogenic interaction with the positively charged zinc ion in the protein. The other ligands have fewer interactions that are energetically favorable and therefore won't bind as well.

You can view the ligands in Yasara. Load (File > Load > YASARA scene) the files `int1.sce`, `int2.sce` and `int3.sce`.

## Lesson 4: Concluding lesson

The practicum of the *DNA lab on the Road* 'Bioinformatics: a bit of life' is meant to be concluded by one or more concluding modules. These modules are meant to offer the students a more in-depth look in the subject, or to place the subject in a wider context.

Using the concluding modules, you can build your own concluding lesson. To help you make a choice from our wide variety of possibilities, we suggest two possible conclusions under the 'Suggestions' header.

### Concluding modules

The modules you can choose for concluding this practicum are:

- **Biobanking**  
Biological data can be stored in databases. But are you allowed to link the data from these databases to other data? For example, think of the Albert Heijn supermarket Bonus Card. Genetic susceptibility to become obese combined with buying unhealthy food products can be useful information for health insurance companies... - Suitable for havo/vwo classes 4, 5 and 6, Biology.
- **Protein Next Top Model**  
What is your favorite protein? Which protein is the most important one, has the most important function? Which protein has the nicest shape? In other words, which protein is the 'Protein Next Top Model'? - Suitable for havo class 5/vwo classes 4, 5 and 6, Biology and Chemistry.
- **Toothpickase**  
Want to experience how an enzyme works? With simple experiments, students will get a better image of how enzymes work. They will cooperate, discuss, calculate and make drawings in teams. - Suitable for havo/vwo classes 4, 5 and 6, Biology and Chemistry.
- **Visualizing proteins**  
Visualize proteins on the computer. Using special software, proteins can be viewed from all angles using different kinds of display. Students learn the structural characteristics of a protein and how the structure is related to the protein function. - Suitable for havo/vwo classes 4, 5 and 6, Biology and Chemistry.

Not only can these modules be used as a conclusion to the practicum, but they can also be used separately.

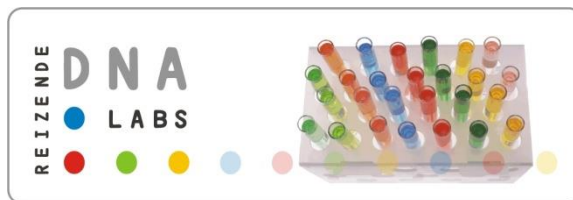
Aside from these specific modules, you can also choose from a wide range of concluding modules that the other *DNA labs on the Road* have developed. More information about these modules can be found on [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu).

### Suggestions

Here, you can find two suggestions for concluding the practicum.

1. Subject-specific conclusion: to take your students more in-depth into the subject (research and proteins), you can use the concluding module 'Toothpickase'. Using simple experiments, the students will experience what it's like to be an enzyme.
2. Societal conclusion: to help your students develop an opinion on the consequences of modern-day DNA research, you can use the concluding module 'Biobanking'. The students will find out what new information can be generated from combining databases. More importantly they will think about who benefits from these new data.





# *Student Guide*

Level: expert

***Bioinformatics:  
a bit of life***



# bioinform@tica in de klas



Radboudumc

## Amgen Biotech Experience

### Scientific Discovery for the Classroom

Developed by the Netherlands Bioinformatics Centre in cooperation with the Centre for Molecular and Biomolecular Informatics of the Radboud University Nijmegen Medical Centre

#### Text

Celia van Gelder, Robbie Joosten and Hienke Sminia

#### Illustrations

[www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)

#### Design

Identim, Wageningen

All the included material is protected by the Creative Commons Naamsvermelding-Niet-commercieel-Gelijk delen 3.0 Nederland license

(<http://creativecommons.org/licenses/by-nc-sa/3.0/nl/>).

#### CC BY-NC-SA 2009 – NBIC

For any questions or comments, please contact the Netherlands Bioinformatics Centre

([nijmegen@dnalabs.nl](mailto:nijmegen@dnalabs.nl)).

Disclaimer: Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Amgen Foundation or Education Development Center, Inc.

## Bioinformatics: a bit of life

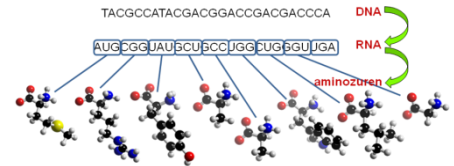
DNA, genes, proteins and genomes. That is what these lessons are about. You will learn the modern techniques of DNA-research during a two-hour practical. Before the practical, you need good preparation. And afterwards, you will even go a step further in genomics. You will work with scientific research in depth or map the relation between genomics and society. Have fun while working with these small molecules with big consequences!

### Introduction

What are the names of the processes given by the arrows?

From DNA to RNA: .....

From RNA to amino acids: .....



Which amino acids are encoded by the codons given below?

AUG: .....  
 CGG: .....  
 UAU: .....  
 GCU: .....

GCC: .....  
 UGG: .....  
 CUG: .....  
 GGU: .....

What is the short code (1-letter) for these amino acids? : . . . . .

UGA is a stopcodon. How does a stopcodon work?

.....

Complete the scheme below to show what happens at each level when a certain change occurs

	An organism is exposed to a lot of UV-radiation	A livercell divides	The DNA mutates in a non-coding part of the sequence	One protein is folded in a wrong way
Organism				
Cell				
DNA				
Protein				

## Practical part 1

### Screening Suspects

Follow these steps to start the practicum:

- Go to: [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)
- Click 'Education' in the menu
- Then, click 'Material' in the menu. You are now directed to an overview of all the bioinformatics lessons material. This *DNA lab on the Road* practical consist of two modules 'Murder on the airport' and '3D Drug Design'.
- Click the practical 'Murder on the airport' as first. A new window will open, containing a short description of the practicum.
- At the left side of the menubar, click 'start the lab' button in the menu.
- The practicum has now started. At the bottom of the page, you can click to continue to the next section.
- Fill in your findings and conclude which protein may have killed the American tourist.

### Questions for each protein:

1. Which protein is it?
2. From which organism does it originate?
3. What is de function of the protein?
4. Is this protein guilty? Could it be responsible for the death of the tourist? Why (not)?
5. Other comments on your findings

<b>Suspect 1</b>	1
	2
	3
	4
	5
<b>Suspect 2</b>	1
	2
	3

	4
	5
<b>Suspect 3</b>	1
	2
	3
	4
	5
<b>Suspect 4</b>	1
	2
	3
	4
	5
<b>Final conclusion: What is the cause of death?</b>	

## Practical part 2

### 3D Drug Design

Follow these steps to start the practicum:

- Go to: [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)
- Click 'Education' in the menu
- Then, click 'Material' in the menu. You are now directed to an overview of all the bioinformatics lessons material. This *DNA lab on the Road* practical consist of two modules 'Murder on the airport' and '3D Drug Design'.
- Click now the practical '3D Drug Design'. A new window will open, containing a short description of the practicum.
- At the left side of the menubar, click 'start the lab' button in the menu.
- The practicum has now started. At the bottom of the page, you can click to continue to the next section.
- Write down your answers below.

#### Exercise 1

.....

.....

.....

.....

#### Exercise 2

Start Yasara and load: introduction.sce.

#### Exercise 3

A. *Atoms in amino acids*:.....

B. *Atoms in Yasara*:

Red: ..... Dark blue: ..... Green: ..... Light blue: .....

#### Exercise 4

*White bonds*: .....

*Yellow bonds*: .....

#### Exercise 5

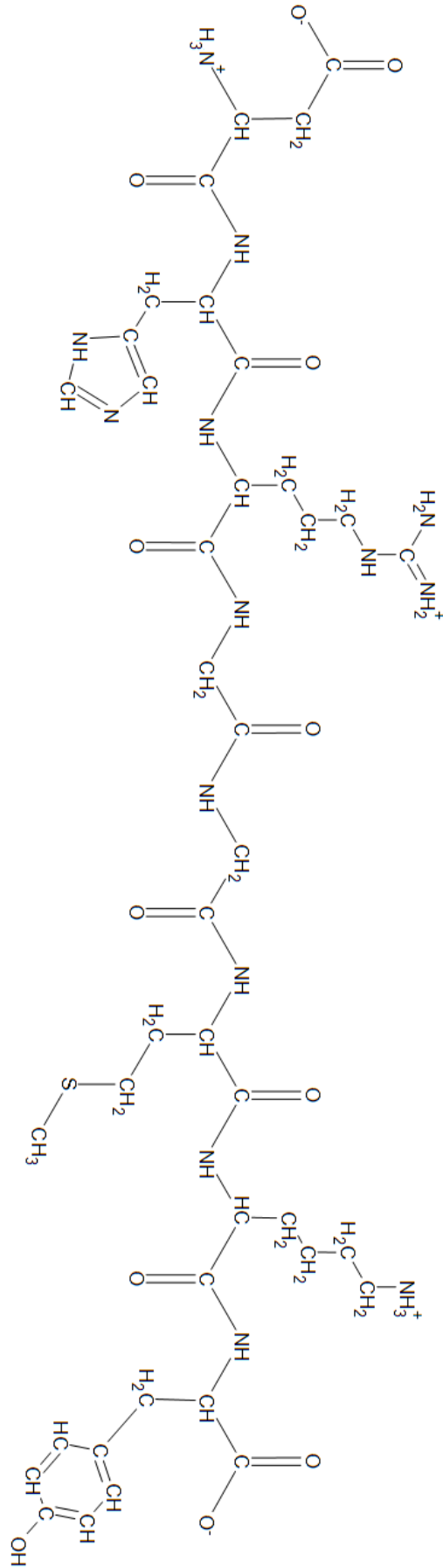
a. *Green*: .....

*Blue*: .....

*Orange*: .....

b. Draw the bonds in the 2D representation on the next page.

Figure: A polypeptide of eighth amino acids.



**Exercise 6**

Load the file slangengif.pdb

**Exercise 7**

*Blue:* .....

*Red:* .....

**Exercise 8**

*Double bonded oxygen atom:* .....

*A histidine sidechain:* .....

*An alpha helix:* .....

**Exercise 9**

.....  
.....  
.....

**Exercise 10**

.....  
.....  
.....

**Exercise 11**

Find the active site of the snake venom

**Exercise 12**

Load the file slangengif2.pdb

**Exercise 13**

Find zinc in the active site

**Exercise 14**

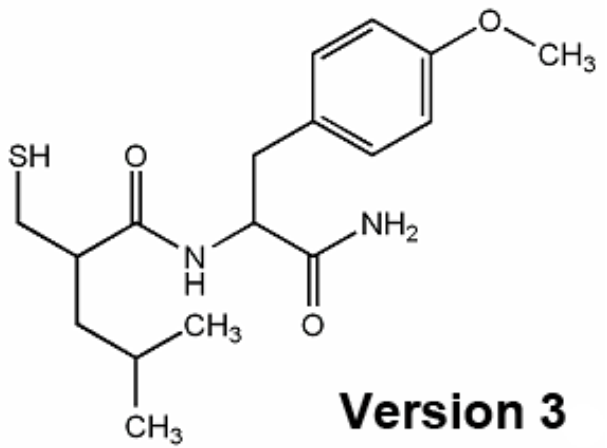
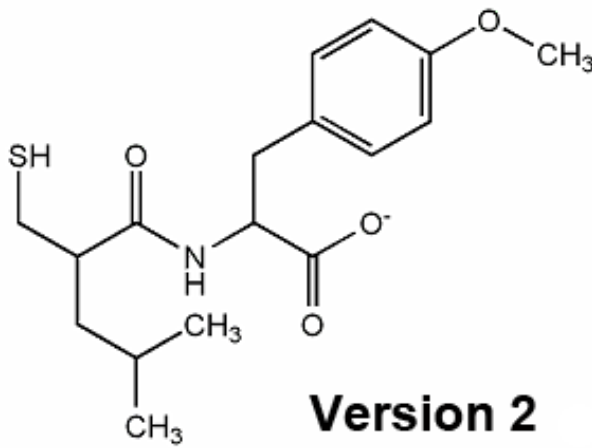
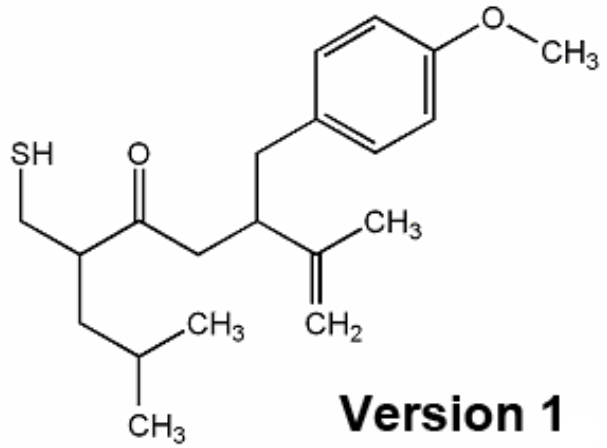
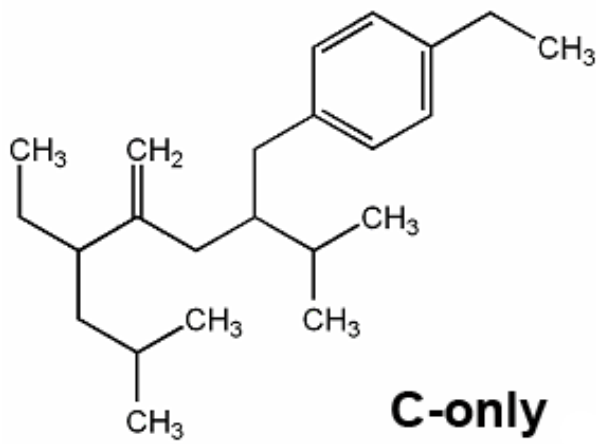
.....  
.....  
.....

**Exercise 15**

Look at 6.8 Rattle snake poison help movie

**Exercise 16**

Mark the atoms that may be involved in hydrogen bonds and ionic interaction with the protein:



**Exercise 17**

Examine ligand\_version1.sce, ligand\_version2.sce en ligand\_version3.sce

**Final conclusion**

.....

.....

.....

.....

More information on bioinformatics?

[www.bioinformaticaindeklas.nl](http://www.bioinformaticaindeklas.nl) and [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)



# **Appendix - Additional High School Teacher Resources**

### **Useful Links for further explorations**

This page lists a number of activities that were prepared by the Bioinformatics and Research Computing Department at Whitehead Institute for Biomedical Research.

<http://jura.wi.mit.edu/bio/education/>

This page lists articles about using bioinformatics in the secondary school setting.

[www.ploscollections.org/cbstartingearly](http://www.ploscollections.org/cbstartingearly)

This web site contains a compilation of materials used at the secondary school level.

<https://www.iscb.org/bioinformatics-resources-for-high-schools>

# Netherlands

Charlotte Zwetsloot, MSc  
Dr. Celia van Gelder

DTL Dutch Techcentre for Life Sciences  
October 2018

## *Dear high school teacher,*

---

Please find enclosed a selection of materials we have developed for high school teachers and pupils in the Netherlands in the context of the Bioinformatics@school programme.

### **Bioinformatics@school**

Since 2006, we organize a DNA Lab on the Road about bioinformatics called Bioinformatica in de klas/ Bioinformatics@school ([www.bioinformatica-in-de-klas.nl](http://www.bioinformatica-in-de-klas.nl), [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu)). The project has been implemented by DTL, Dutch Techcentre for Life Sciences.

Since 2017, the DNA Labs on the Roads (<http://www.dnalabs.nl/english/>) is the Dutch site of the Amgen Biotech Experience. The Amgen Biotech Experience is a science education programme that empowers teachers to bring biotechnology into their classrooms by providing professional development for teachers and the equipment.

Since the start of the project over 24000 high school pupils have participated in one of our Bioinformatics@school practicals in their own classroom. These pupils gain interest in and knowledge about new scientific subjects like genomics and can use real research technology at their school. Our lab is free of charge for high schools and is taught at the high schools by science students of the Radboud University Nijmegen.

The mission of Bioinformatics@school is to get bioinformatics elements embedded in the high school curriculum by educating pupils and teachers and also to show the relevance of bioinformatics and genomics to a broader audience (for example we use a 3D-beamer to visualize proteins for the general public).

During the years we have developed a large portfolio of activities and materials. An overview of our materials and an example of the Navigene are given in this booklet. The Navigene is a tool to help find your way in bioinformatics and design your own bioinformatics lesson materials. A summary is given here in this booklet, including the Navigene scheme; the complete guide (20 pages) can be downloaded from our website.

The lessons that we do in the Dutch high schools can be accessed at [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu).

***We wish you a nice journey through the world of Bioinformatics!***

### **The Bioinformatics@school team:**

Charlotte Zwetsloot ([nijmegen@dnalabs.nl](mailto:nijmegen@dnalabs.nl)); Celia van Gelder ([celia.van.gelder@dtls.nl](mailto:celia.van.gelder@dtls.nl))

October 2018

All Bioinformatics@school materials are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Licence

## Overview of our materials

Materials available via [www.bioinformaticsatschool.eu](http://www.bioinformaticsatschool.eu). Teacher manuals can be obtained by sending an email to [nijmegen@dnalabs.nl](mailto:nijmegen@dnalabs.nl).

Name	Requires computer	Information
<b>Introduction lesson for DNA Lab on the Road</b>		Introductory lesson for the DNA Lab on the Road. 50 min. Level: high school
<b>DNA Lab on the Road Basic: Losing sight and vision</b>	X	The link between DNA-mutations and an eye disease. 100 min. Level: high school
<b>DNA Lab on the Road Expert: Murder at the airport (lesson 1 of 2)</b>	X	Solve a crime by identifying proteins. 35 min. Level: high school
<b>DNA Lab on the Road Expert: 3D drug design (lesson 2 of 2)</b>	X	Design an antidote using 3D modelling. 50 min. Level: high school
<b>Visualisation of proteins</b>	X	Visualize proteins and determine their function. 50 min. Level: high school
<b>Ion channel</b>	X	About ion channels and how they work. 35 min. Level: high school
<b>Nutrition and digestion</b>	X	About lock-key concept. 35 min. Level: high school
<b>Evolution</b>		About homologous proteins. 35 min. Level: high school
<b>Protein Next Top Model</b>	X	Which protein becomes “Protein Next Top Model”? 100 min. Level: high school
<b>Biobanking</b>		Gain more knowledge by integrating databases. 30 min. Level: high school
<b>Tandestokerase +</b>		Work as a protease. 40 min. Level: high school
<b>Crossword puzzle</b>		What do you remember? Did you pay attention during the bioinformatics practical? 15 min. Level: high school
<b>DNA-scale</b>	X	Suitable for all ages. 10 min. Determine your DNA!
<b>NaviGene</b>	X	Instruction tool for high school teachers to develop their own lessons <a href="https://www.dtls.nl/wp-content/uploads/2017/02/NaviGene_EN_versie_2014.pdf">https://www.dtls.nl/wp-content/uploads/2017/02/NaviGene_EN_versie_2014.pdf</a>

## ***The NAVIGENE: a tool to help you find your way in bioinformatics***

---

Within the Dutch Bioinformatics@school project an unique instruction tool, the Navigene, has been developed to help teachers and students navigate through online bioinformatics tools and software and enable them to design their own bioinformatics lesson materials.

You can download the latest version at <https://www.nbic.nl/education/high-school-programmes/bioinformaticsschool/teacher-training/navigene/>.

### **Why bioinformatics in the classroom?**

The recent flood of data from genome sequences and functional genomics had given rise to a new field, bioinformatics, which combines elements of biology and computer science. Bioinformatics is nowadays an inherent part of research in molecular biology. Gelbart and Yarden<sup>1</sup> write that a bioinformatics learning environment promotes the construction of new knowledge structures of the genetics domain and therefore influences students' acquisition of a deeper, multidimensional understanding of the domain.

We think that databases and software used in bioinformatics can contribute to several challenges in biology education:

#### **1. *Students understanding of abstract concepts like protein, genome and evolutionary relationship***

Proteins and genes cannot be observed by the human eye. Expensive equipment is needed to visualize these molecules. And even then it remains to be seen whether students would gain a better understanding of the processes and functions. Cheaper and probably more helpful is a computer-based approach. Using 3D-software, you will be able to see a certain protein from all different angles. You can zoom in, turn the protein around and select specific amino acids. A protein structure can be downloaded from the Protein Data Bank. Other databases make it possible to show the structure of genes in a scientific way. You can simply zoom in on a gene and distinguish the exons, introns and regulating domains. You can even make simplistic phylogenetic trees or look directly at proteins that are related to your protein of interest.

**We think that when students can work with these tools, abstract genomic concepts become more tangible and therefore easier to understand.**

---

<sup>1</sup> Gelbart H and Yarden A (2006) Learning genetics through an authentic research simulation in bioinformatics. Authentic research simulation 40-3: 107-112

## **2. The coherence between DNA, protein and traits, and other themes in biology**

Schoolbooks often discuss the relation between DNA, genes and heredity in the context of visible traits like the colour of the eyes or hair. The fact that humans have 99,9% of (mostly non-visible) heritable characteristics in common is hardly ever taught to students. One way of giving attention to the relationships between DNA and traits outside the chapter on heredity is by making a link to proteins, which are discussed as part of other themes within the biology curriculum. For example: when discussing digestion, you can simply look up on what chromosome the gene for amylase is and/or show the 3D-structure of amylase. These links can be packaged as small assignments (max. ten minutes) directly connected to proteins in the biology curriculum.

**We think that making more links from different chapters throughout the biology curriculum to genes and proteins helps students' understanding of the genome.**

## **3. Insight in current research methods**

Almost every discipline in life science employs bioinformatics. Moreover, bachelor and university programmes in life sciences also use bioinformatics.

**We think that high school education that aims to provide insight in current research methods, cannot ignore bioinformatics.**

## **What is Navigene?**

The NaviGene is a guide that helps you to find your way in online databases and software that are used by bioinformaticians and link it to your biology knowledge and to what you would like to discuss in the classroom with your students. Our experience is that when you are not a bioinformatics expert, it is very difficult to find any useful information in online sources. That is why we set up an understandable instruction guide to make it feasible to get real and authentic research into your classroom.

## **Who can use the NaviGene?**

The NaviGene is initially developed for high school biology teachers. It is our experience that teachers use the Navigene each in their own way. Here are some examples:

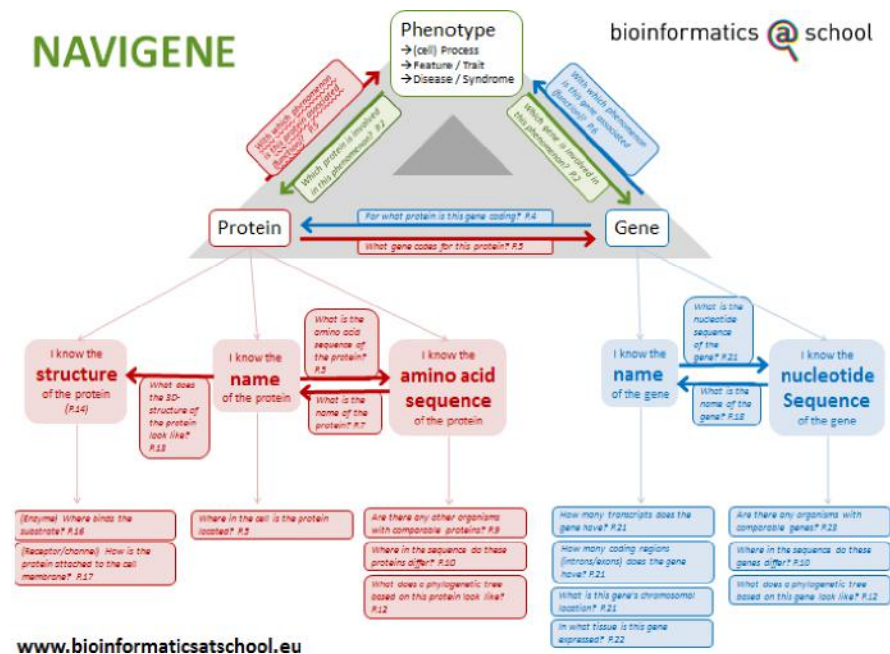
- “I use the NaviGene to plenary show my students 3D-proteins when we come across a protein in the text book. This gives them better insight in what these molecules look like.”
- “I made a few assignments for my students with help from the NaviGene. I let students browse into Ensembl and let them make a phylogenetic tree. I couldn't have made these assignments without the NaviGene.”
- “I used the NaviGene to find background information on blood groups. This blood group system was far more difficult than I expected and Wikipedia couldn't give me the information that I wanted. So I looked into protein databases to find the right information.”
- “I have the NaviGene printed out in the back of my class. When excellent students want to do something extra in my biology lesson, I let them work from the NaviGene on a subject we just treated in class. Students also used the NaviGene on own initiative for school projects.”

## How can I use the NaviGene?

The NaviGene consists of two parts: a scheme and an instruction booklet. In the papers you are holding right now, you will only find the scheme (on the next page).

The rest of the booklet can be downloaded at:

<https://www.nbic.nl/education/high-school-programmes/bioinformaticschool/teacher-training/navigene/>



You read about the BRCA1-gene in a news article and wonder for what protein this gene codes. Or you find the protein Amylase in the chapter 'Digestion' in your biology book. These are excellent starting points for further research with help of the NaviGene.

You start with the Navigene in the grey triangle at the top of the scheme. Let's take Amylase as an example. Amylase is a protein, so you start at the red box *Protein* on the left side of the grey triangle. From there you can follow a red arrow to *Gene*. There is a question linked to that arrow: *What gene codes for this protein? P.5*. If you want to know the answer on this question for Amylase, than go to page 5 of the instruction booklet. There you will find extended and comprehensible instructions on how to find the answer with help of online tools.

You cannot only 'move' around in the grey triangle, but also follow the lighter coloured arrows down. Depending on the information you already have, you go to either *structure*, *name* or *amino acid sequence*. In the case of Amylase you know the name, so you will have to start at *I know the name of*



*the protein*. From there you can hunt down the structure, the amino acid sequence or follow the arrow downward to find out where the protein is located in the cell. All questions in the scheme are followed by *P* and a number. This refers to a page in the (online) instruction booklet.

The instructions are given in this format:

→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://www.ncbi.nlm.nih.gov>
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://www.ncbi.nlm.nih.gov> 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 (*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 has not the highest relevance when searching with *titin*, so it may not appear on top in the list.

In coloured bold letters the question from the scheme is repeated.

In the grey text box you will find short instructions to find the answer to your question. These instructions convenient when you are an experienced user of the NaviGene.

The extended instructions are given underneath the grey text box. Just read them trough and use them for your specific question.

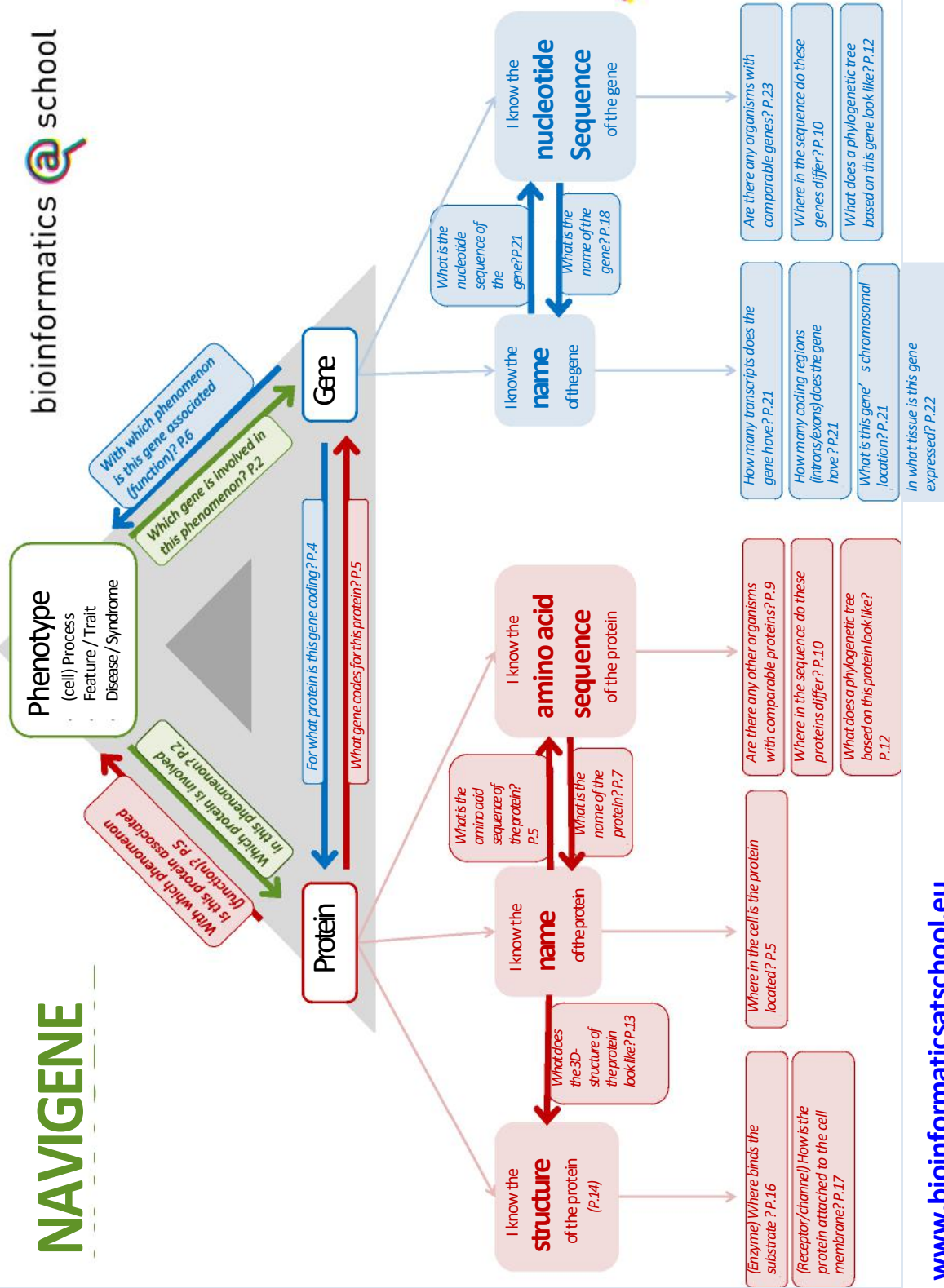
Each instruction ends with a coloured text box with a small assignment to get you acquainted with the instructions.

At the bottom of the page you can find the page number.

We wish you many useful discoveries and valuable surprises when using the NaviGene!

## Finally,

- NAVIGENE is available for you to use under under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Licence
- We welcome all your feedback. If you have used it and created a student exercise we would be happy to post in on the [bioinformaticsatschool.eu](http://bioinformaticsatschool.eu).
- If you would like to edit the NAVIGENE guide (translate to your own language, add information or improve otherwise), we are glad to help you. Just let us know!
- NAVIGENE is, and will always be, under development due to updates from tools, websites and new features in bioinformatics resources. Please let us know when you find dead or wrong links. Than we can correct it!
- The original version of NAVIGENE is in Dutch. Updating the English translation is in full progress, but is lagging behind a bit. We trust you can understand that.



## Dear teacher/student

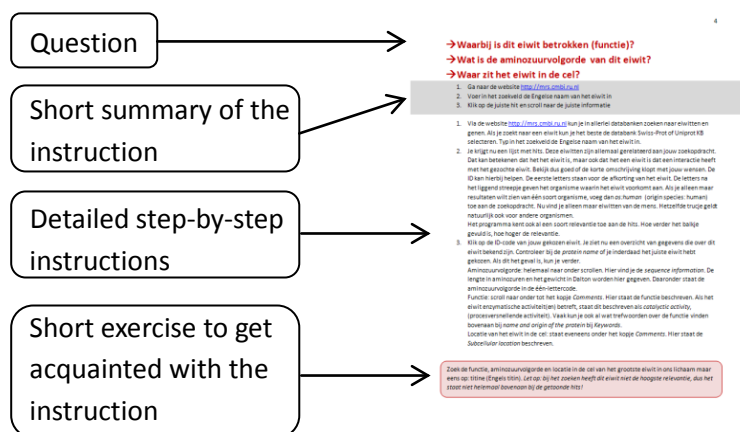
This booklet is the NAVIGENE. It is a navigation tool developed for biology teachers at secondary schools, developed by the Netherlands Bioinformatics Centre and the Freudenthal Institute for Science and Mathematics Education. Bioinformatics is a field of research that focusses on DNA and proteins. You 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.



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](mailto:onderwijs@nbic.nl).

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

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

1. Use Google ([www.google.com](http://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 abbreviation 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: PDBs](#) [RCSB](#)

[List of PDB id codes](#) [show]

**Identifiers**

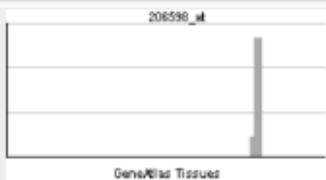
**Symbols** [INS](#); [IDDM2](#); [ILPR](#); [IRDN](#); [MODY10](#)

**External IDs** [OMIM: 176730](#) [MG: 96573](#) [HomoloGene: 173](#)  
[ChEMBL: 5881](#) [GeneCards: INS Gene](#)

[Gene Ontology](#) [show]

**RNA expression pattern**

206598\_at



GeneAtlas Tissues

[More reference expression data](#)

**Orthologs**

Species	Human	Mouse
Entrez	3630	16334
Ensembl	ENSG00000254647	ENSMUSG00000000215
UniProt	P01308	P01326
RefSeq (mRNA)	NM_002207.2	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](http://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

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq	Flags
BRCA2-001	ENST00000380152	11986	3418 aa	Protein coding	CCDS9344	-	TSL:5   GENCODE basic (P)
BRCA2-201	ENST00000544455	10984	3418 aa	Protein coding	CCDS9344	NM_000059 NP_000050	TSL:1   GENCODE basic (P)
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*.

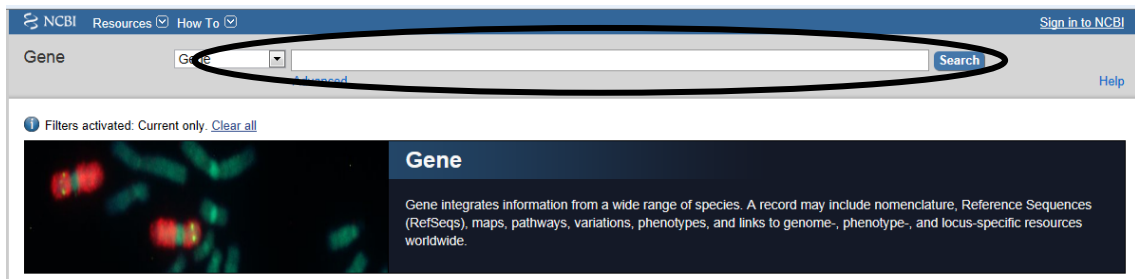
Find the function, amino acid sequence and location of the protein in a cell of the largest protein in our body: titin.

*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.



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

The screenshot shows the BLAST web interface. At the top, there are navigation links: Home, Blast, Align, Status. Below is a search bar with 'All Databases' selected. A text area contains a protein sequence in FASTA format, starting with '>Protein1'. Below the search bar, there are dropdown menus for 'Database to search' (set to 'SwissProt') and 'Filter low complexity' (checked). An 'E-value cutoff' is set to '10.0'. A 'Blast' button is visible. Below the search area, a 'Blast results' summary shows '1 Protein1' in the 'sprot' database with '243 hits found'. At the bottom, a table lists the results for 'Protein1' with columns for 'Nr', 'ID', 'Accession', 'Description', 'Hits', 'BitScore', and 'E-value'. The first few results are: 1 AMY1\_HUMAN (0.00), 2 AMY2\_HUMAN (0.00), 3 AMYP\_HUMAN (0.00), 4 AMY1\_MOUSE (1.29e-270), 5 AMYP\_PIG (2.59e-270), 6 AMYP\_MOUSE (3.31e-268), 7 AMYP\_RAT (2.12e-266), 8 AMYP\_STRCA (5.43e-264), 9 AMY\_PECMA (1.54e-172), 10 AMY1\_MOUSE (2.49e-158), 11 AMY1\_DROAN (3.98e-157), 12 AMY2\_DROAN (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.

- 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 QVSPNENVA IHNPFRPWWE
RYQPVSYKLC TRSGNEDEFR NMVTRCANNV VRIYVDAVIN HMCNAVSAG 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 IRNMVNFNRV VDGQPF'TNWY DNGSNQVAFG
RGNRGFIVFN NDDWTFSLTL QTGLPAGTYC DVISGDKING NCTGIKIYVS DDGKAHFSIS
NSAEDPFIAI HAESKL
```

## → Are there any organisms with similar proteins?

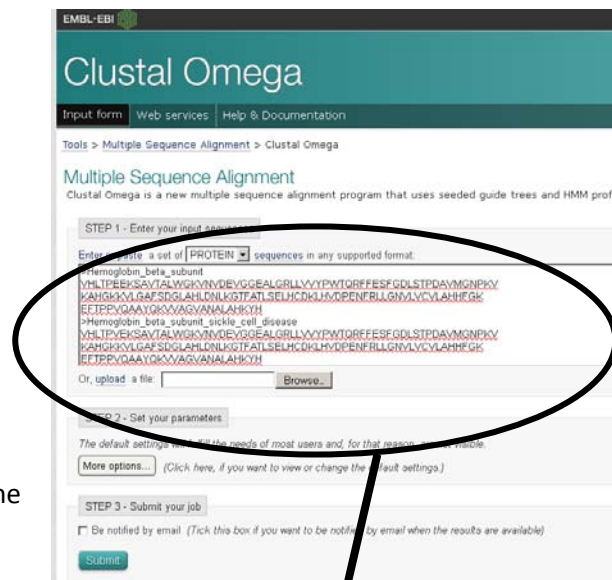
1. Visit <http://mrs.cmbi.ru.nl>
  2. Search for the protein that you would like to use
  - 3a. Click '*Find similar*'
  - 3b. Click '*Blast*'
- 
1. The website <http://mrs.cmbi.ru.nl> enables you to search for genes in proteins in numerous databases. When looking for proteins the databases Swiss-Prot 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?

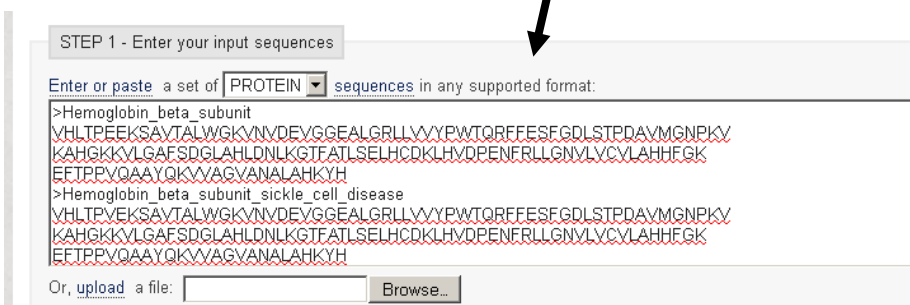
1. Visit <http://www.ebi.ac.uk/Tools/clustalw2/index.html>
2. Copy both amino acid sequences into the text box.
3. Click 'Submit'

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?*

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
VHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLNLDKGTFFATLSEIHCDKLVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
>Hemoglobin_beta_subunit_sickle_cell_disease
VHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLNLDKGTFFATLSEIHCDKLVDPENFRLLGNVLVCVLAHHFGK
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 DIPGHGQEV L IRLFKGHPET LEKFDKFKHL KSEDEMKASE
DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI PVKYLEFISE CIIQVLQSKH
PGDFGADAQG AMNKALELFR KDMASNYKEL GFQG
>rabbit
MGLSDAEWQL VLNVWGKVEA DLAGHGQEV L IRLFHTHPET LEKFDKFKHL KSEDEMKASE
DLKKHGNTVL TALGAILKKK GHHEAEIKPL AQSHATKHKI PVKYLEFISE AIIHVLHSHK
PGDFGADAQA AMSKALELFR NDIAAQYKEL GFQG
>shark
MABWDKVN SV WSAVEQNITA IGQNILLRLF EQYPESEDYF PKLKNKSLGE LKDTADIKAQ
ADTVLRALGN IVKKKGDHSQ PVKALAATHI TTHKIPPHYF TKITTIAVGV LSEMYPSEM N
AQAQAAFSGA FKNICS DIEK EYKAANFQG
>tuna
MADFDAVLKC WGPVEADYTT MGGLVLTRLF KEHPETQKLF PKFAGIAQAD IAGNAAISAH
GATVLKKLGE LLKAKGSHAA ILKPLANSHA TKHKIPINNF KLISEVLVKV MHEKAGLDAG
GQTALRNVMG IIIADLEANY KELGFSG
>gibbon
MGLSDGEWQL VLNVWGKVEA DIPSHGQEV L IRLFKGHPET LEKFDKFKHL KSEDEMKASE
DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI PVKYLEFISE CIIQVLQSKH
PGDFGADAQG AMNKALELFR KDMASNYKEL GFQG
>baboon
MGLSDGEWQL VLNVWGKVEA DIPSHGQEV L IRLFKGHPET LEKFDKFKHL KSEDEMKASE
DLKKHGATVL TALGGILKKK GHHEAEIKPL AQSHATKHKI PVKYLELISE SIIQVLQSKH
PGDFGADAQG AMNKALELFR NDMAAKYKEL GFQG
>common carp
MHDAELVLKC WGGVEADFEG TGGEVLTRLF KQHPETQKLF PKFVGIASNE LAGNAAVKAH
GATVLKKLGE LLKARGDHAA ILKPLATTHA NTHKIALN NF RLITEVLVKV MAEKAGLDAG
GQSALRRVMD VVIGDIDTTY KEIGFAG
>zebra
MGLSDGEWQ VLNVWGKVEA DIAGHGQEV L IRLFTGHPET LEKFDKFKHL KTEAEMKASE
DLKKHGTVL TALGGILKKK GHHEAELKPL AQSHATKHKI PIKYLEFISD AIIHVLHSHK
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](http://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 transferrin, 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](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 software isn't available on your computer, you can download and use it for free.

- 1 Visit [www.yasara.org](http://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, .job 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	<ul style="list-style-type: none"> <li>- Display hydrogen atoms (<i>Edit &gt; Add &gt; Hydrogens to All</i>).</li> <li>- Select <i>view &gt; color &gt; residue</i></li> <li>- In the third column, select (<i>belongs to or has</i>) <i>Charge &lt; 0</i> and click <i>Ok</i>.</li> <li>- Choose your color and hit <i>Apply Unique color</i>.</li> </ul>
Color positively charged residues	<ul style="list-style-type: none"> <li>- Display hydrogen atoms (<i>Edit &gt; Add &gt; Hydrogens to All</i>).</li> <li>- Select <i>view &gt; color &gt; residue</i></li> <li>- In the third column, select (<i>belongs to or has</i>) <i>Charge &gt; 0</i> and click <i>Ok</i>.</li> <li>- Choose your color and hit <i>Apply Unique color</i>.</li> </ul>
Color hydrophylic residues	<ul style="list-style-type: none"> <li>- Select <i>view &gt; color &gt; residue</i></li> <li>- In the second column, select Arg, Asp, Asn, Glu, Gln, His, Lys, Ser and Thr (while holding Ctrl) and click <i>Ok</i>.</li> <li>- Choose your color and hit <i>Apply Unique color</i>.</li> </ul>
Color hydrophobic residues	<ul style="list-style-type: none"> <li>- Select <i>view &gt; color &gt; residue</i></li> <li>- In the second column, select Ile, Leu, Met, Phe en Val (while holding Ctrl) and click <i>Ok</i>.</li> <li>- Choose your color and hit <i>Apply Unique color</i>.</li> </ul>
Show hydrogen atoms	Select <i>Edit &gt; Add &gt; Hydrogens to All</i>
Show hydrogen bonds	<ul style="list-style-type: none"> <li>- Select <i>Edit &gt; Add &gt; Hydrogens to All</i></li> <li>- Select <i>View &gt; Show hydrogen bonds &gt; All</i></li> </ul>
Show secondary structures	F6
Show sidechains	F8
Delete water molecules	Select <i>Edit &gt; Delete &gt; Waters</i>

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

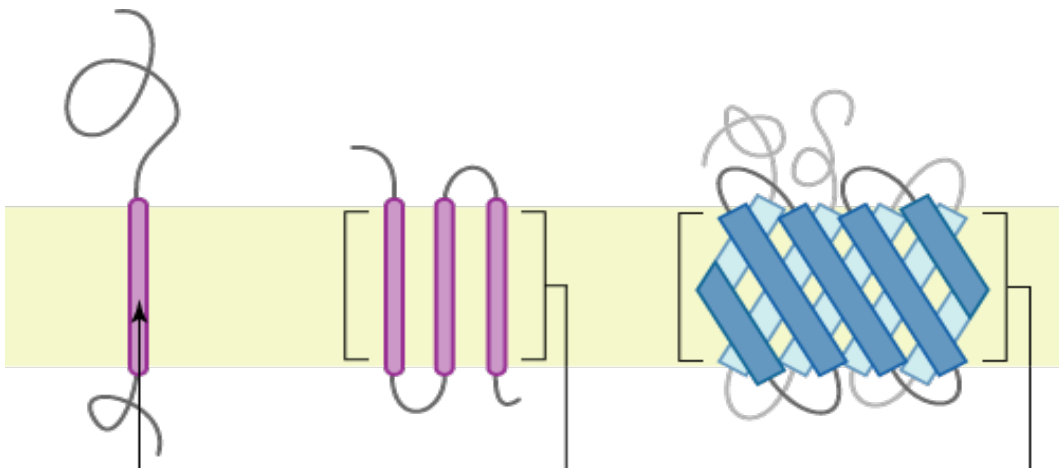
## → (receptor/channel) How is the protein bound in the membrane?

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  $\alpha$ -helices and  $\beta$ -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  $\beta$ -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  $\beta$ -barrel consists of multiple  $\beta$ -strands, which are interwoven to form a pore in the membrane.



Schematic display of a single helix, a helix bundle and a  $\beta$ -barrel

Source: [http://en.wikipedia.org/wiki/Transmembrane\\_protein](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 optimisation 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

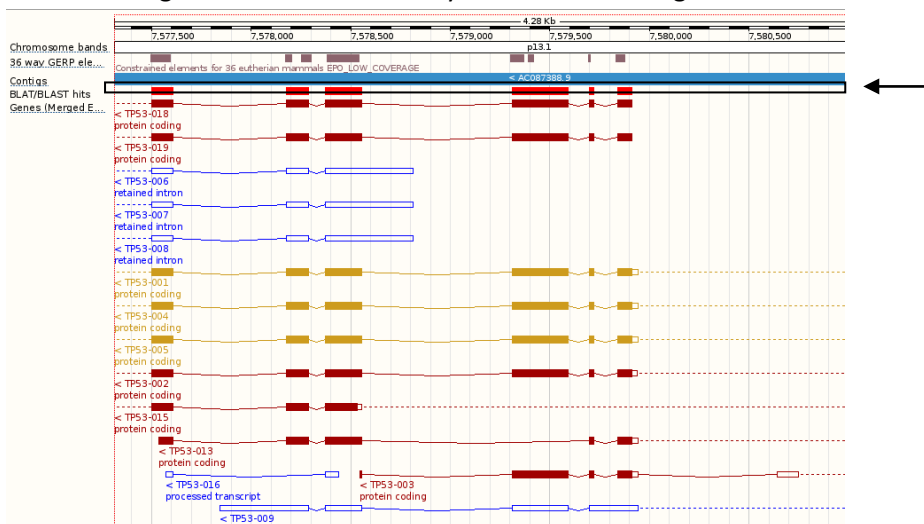
purpose, this information is irrelevant and can be discarded if desired.

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.

- 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*.

- 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.

**Try this for:**

>Nucleotidesequence1

```

ATGGAGGAGCCGCGAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAGGAAACATTTTCAGACCTATGGA
AACTACTTCCGAAAACAACGTTCTGTCCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCC
GGACGATATTGAACAATGGTTCACCTGAAGACCCAGGTCCAGATGAAGCTCCCAGAATGCCAGAGGCTGCT
CCCCCGTGGCCCCGTCACCAGCAGCTCCTACACCGCGCGCCCCGTCACCAGCCCCCTCCTGGCCCCCTGT
CATCTTCTGTCCCTTCCAGAAAACCTACCAGGGCAGCTACGGTTTCCGTCTGGGCTTCTTGCATTCTGG
GACAGCCAAGTCTGTGACTTGCACGTACTCCCCGCCCCCAACAAGATGTTTTGCCAACTGGCCAAGACC
TGCCCTGTGCAGCTGTGGGTTGATTCCACACCCCCGCCCCGCGACCCGCGTCCGCGCCATGGCCATCTACA
AGCAGTCACAGCACATGACGGAGTTGTGAGGCGCTGCCCCACCATGAGCGCTGCTCAGATAGCGATGG
TCTGGCCCCCTCCTCAGCATCTTATCCGAGTGAAGGAAATTTGCGTGTGGAGTATTTGGATGACAGAAAC
ACTTTTCGACATAGTGTGGTGGTGCCTATGAGCCGCCTGAGGTTGGCTCTGACTGTACCACCATCCACT
ACAACCTACATGTGTAACAGTTCTTGCATGGGCGGCATGAACCGGAGGCCATCCTCACCATCATCACACT
GGAAGACTCCAGTGGTAATCTACTGGGACGGAACAGCTTTGAGGTGCGTGTGTTGTGCCTGTCTGGGAGA
GACCGGCGCACAGAGGAAGAGAATCTCCGCAAGAAAGGGGAGCCTCACCACGAGCTGCCCCAGGGAGCA
CTAAGCGAGCAGTCCCAACAACACCAGCTCCTCTCCCCAGCCAAAGAAGAAACCCTGGATGGAGAATA
TTTACCCTTTCAGATCCGTGGGCGTGAGCGCTTCGAGATGTTCCGAGAGCTGAATGAGGCCTTGGAACTC
AAGGATGCCAGGCTGGGAAGGAGCCAGGGGGGAGCAGGGCTCACTCCAGCCACCTGAAGTCCAAAAGG
GTCAGTCTACCTCCCGCCATAAAAACTCATGTTCAAGACAGAAGGGCCTGACTCAGACTGA
    
```

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?

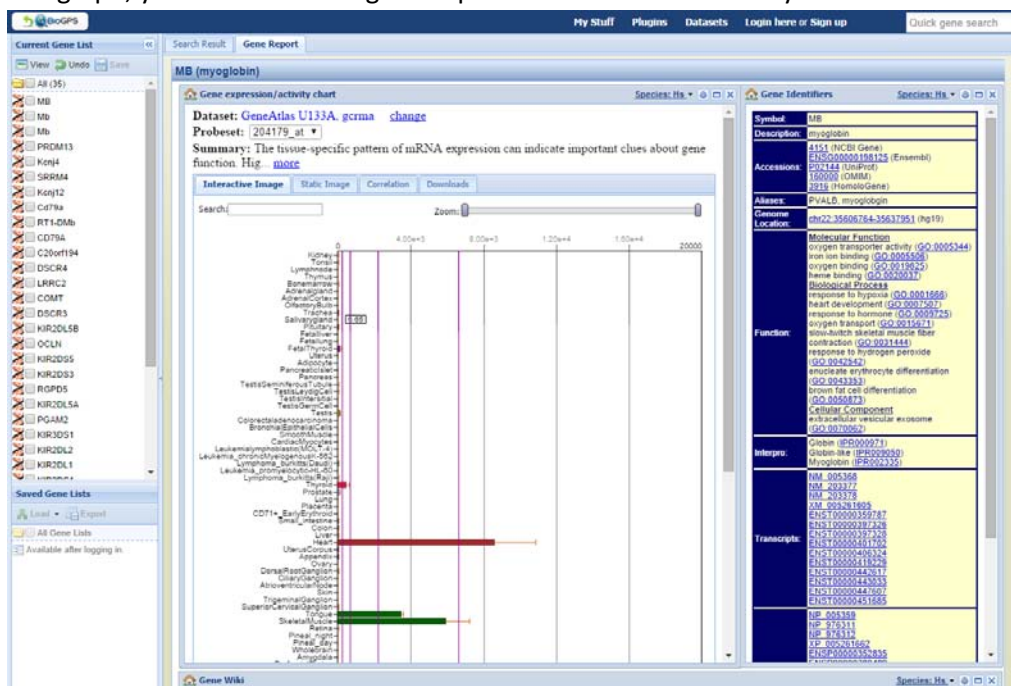
1. Visit the Ensembl genome database: <http://www.ensembl.org/index.html>
2. Enter the name of the gene in the searchfield
3. Select 'gene' of the best hit
4. For exons/introns: click the desired transcript

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](http://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.



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).





**mygoblet.org**

Sponsored by:



Contributions from:

