Bringing Bioinformatics to the Biology Classroom

Presented by the Global Organization for Bioinformatics Learning, Education and Training (GOBLET)

Global Organisation for Bioinformatics Learning, Education & Training

Organizations working together

on Education and Training

GOBLET

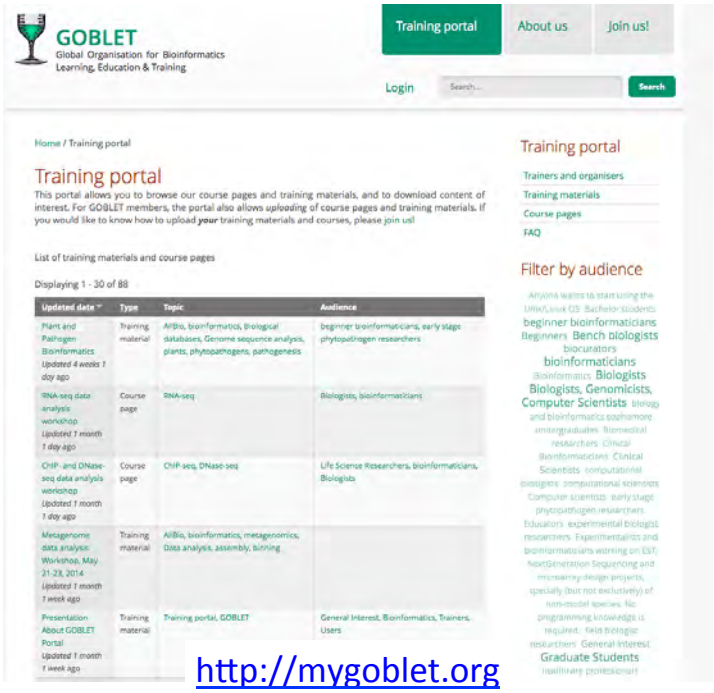
Global Organization for Bioinformatics Learning, Education and Training

http://mygoblet.org

ISCB

International Society for Computational Biology

http://www.iscb.org



http://mygoblet.org

Bioinformatics Resources for High School Teacher

hosted by ISCB Education Committee

- Bioinformatics is an integral part of biology
- Held a workshop in July 2014 for local Boston teachers
- 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

Lesson Plan #1 – Swiss Institute of Bioinformatics

Dr. Marie-Claude Blatter

Swiss Institute of Bioinformatics

http://education.expasy.org/bioinformatique/Diabetes.html

Understanding a genetic disease thanks to Bioinformatics

proposed by the SIB Swiss Institute of Bioinformatics

Context:

Insulin is a protein that allows sugar (glucose) to enter the body's cells (mainly liver, adipose tissue and skeletal muscle). This hormone plays a key role in the regulation of glucose levels in the blood ('hypoglycemic' effect). It is produced by the beta cells in the pancreas.

Type I diabetes (insulino dependent; IDDM) is more often than not due to the absence of insulin: for various poorly understood reasons (virus, autoimmune aetiology, ...), the pancreas is no longer able to produce the protein.
Type II diabetes (non insulino dependent; NIDDM) is a metabolic disease (insulin resistance). Obesity is thought to be the primary cause of type II diabetes in people who are genetically predisposed to the disease.

A very rare genetic variation - rs121908261 - leads to the the production of a non functional insulin and is the cause of type I diabetes in a Norwegian family, (Molven et al., 2008).

This workshop will explore how bioinformatics can help to better understand the causes of this rare genetic disorder ... and also to learn more about insulin.

Activity 1: The insulin gene and the human genome

Bellow is a piece of the gene sequence that encodes for the insulin protein ('*wild sequence*')...

cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc

Question:

- **On which of our 23 chromosomes is this gene located?**

Bioinformatics approach:

Use the tool 'BLAT'
*Technical information: 'BLAT' is a bioinformatics tool for comparing a DNA sequence against the whole genome sequence (the human genome has 3 billion nucleotides). If the sequence exists, BLAT finds the sequence that is the most similar in just a few seconds. It's a bit like a small **'google map'** of the human genome.*

* Copy the DNA sequence and paste it in the tool **'BLAT'**
* Click on 'submit'
* In the page 'BLAT Search Result': choose the best score and click 'browser'

- **On which chromosome is located the gene for insulin?**
- **What are the beginning and end positions of the sequence on the chromosome (nucleotide 'numbers')?**

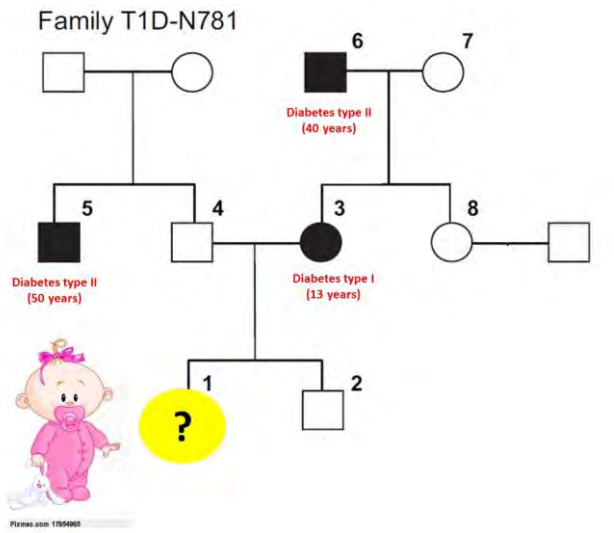
- **For fun: write a random sequence (about 30 letters), always using the 4-letter alphabet (a, t, g, c) into 'BLAT': can you find it in the genome?**

Activity 2: Comparing DNA sequences - Diagnosing a rare genetic disease

About 1 nucleotide in 1000 differs from one person to another, and from one genome to another. These differences are called variations or mutations. Some have no effect on a person, while others may be associated with genetic diseases. In 2008, scientists studied a Norwegian family in which several members had diabetes (type I or type II) (Molven et al., 2008).

All diabetic type I members of the family carry the same rare variation in the gene which encodes for insulin.

Here is the family's pedigree (phenotype and family relationship):



Question:

- **Is this baby diabetic?**
- To answer this question, researchers extracted DNA from 8 members of the Norwegian family and sequenced part of the gene that encodes for insulin.

```
>1
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacctgccgggaggcagaggacc
>2
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>3
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacctgccgggaggcagaggacc
>4
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>5
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>6
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>7
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggagaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>8
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
```

Compare these sequences, and locate the common variation for diabetes.

'Paper and pencil' approach:
... You can do it **manually which will help you better understand the principle of sequence comparison and alignment.**
Take into account all the given clues and play with our strips of DNA sequences...

- 8 family members - 4 DNA sequences - one allele
- 8 family members - 1 DNA sequence - two alleles
- 8 family members - 2 DNA sequences - two alleles (not easy) (see printed document)

Bioinformatics approach:
Build an alignment of these 8 sequences using a bioinformatics tool and look out for the common variation among those with diabetes

* Copy these 8 sequences (including the lines starting with '>1') and paste them into the align tool
* Click on the *Run Align* button.
* On the results page, on the lefthand column 'Highlight': select 'Similarity'

For those who are curious:

.... additional information on the family (Molven et al., 2008):

The subject **(1)** with the c->t R55C mutation (heterozygous mutation) is a girl who presented frank diabetes at the early age of 10. Her blood glucose level was of 17.6 mmol/l – which is very high. The girl's mother **(3)** has type I diabetes that was diagnosed when she was 13. Currently, she is being treated with insulin. She also carries the heterozygous mutation. The girl's maternal grandfather **(6)** has type 2 diabetes, which was diagnosed at the age of 40. He is currently being treated with insulin. Neither he nor the healthy maternal grandmother carry mutations. Thus, the girl's mother is carrying a *de novo* mutation, which must be a germline mutation since it has been inherited by her daughter.
C-peptides, or connecting peptides, are the peptides which connect the insulin's A chain to the B chain. Both carriers (mother and daughter) of the c-> t R55C mutation have C-peptide levels in the normal range, which suggests that some insulin is being processed and secreted. Currently, no one really knows why the mother and daughter have severe insulin deficiency despite evidence of insulin secretion.

.... Sequence of the gene for insulin (with the c->t R55C mutation site highlighted)
.... The list of known variants (in red) in the insulin gene; note that many variations are neither pathogenic, nor associated with diabetes.

Activity 3: DNA translation -> protein

Check the effect of the mutation c-> t ('R55C')...

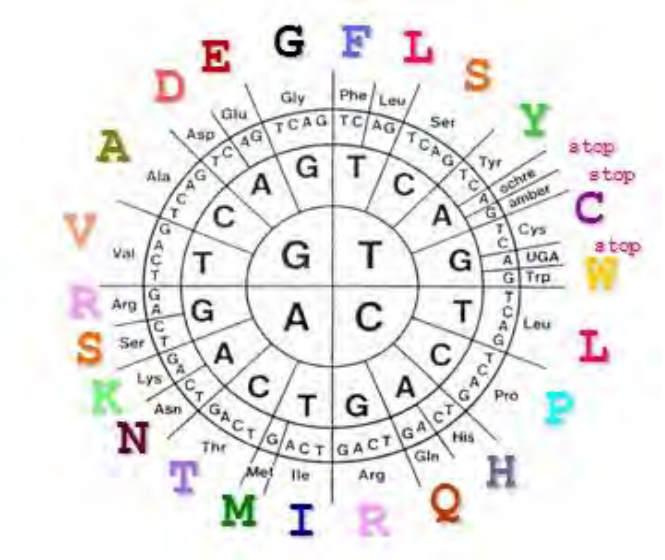
Like all proteins, insulin is composed of a sequence of amino acids. The order of the amino acids is determined by the nucleic acid sequence of the insulin gene.
3 letters of DNA (codon) correspond to one amino acid (symbolized by letters: K for lysine, M for methionine, etc.).

This is a piece of the DNA sequence of the normal insulin gene.
aag acc cgc cgg gag
This is a piece of the DNA sequence of the insulin gene with the c -> t variation, associated with type I diabetes.
aag acc tgc cgg gag

Question:

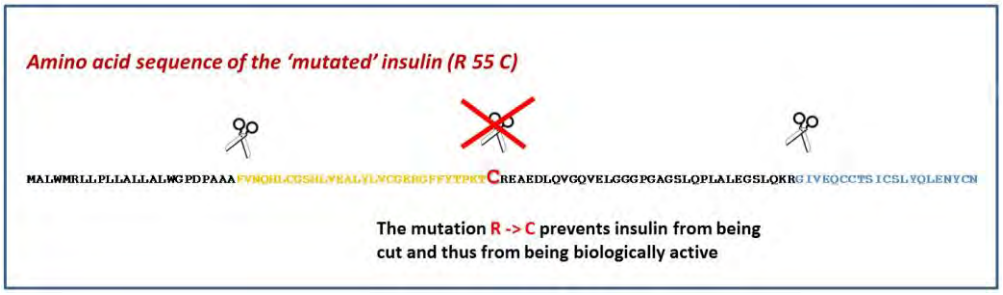
- **Does the c->t mutation change the amino acid sequence of insulin?**
- **Does the aag -> aaa mutation change the amino acid sequence of insulin?**

You could manually translate the nucleic acid sequences into amino acid sequences ('1 'letter code) using the genetic code below: :

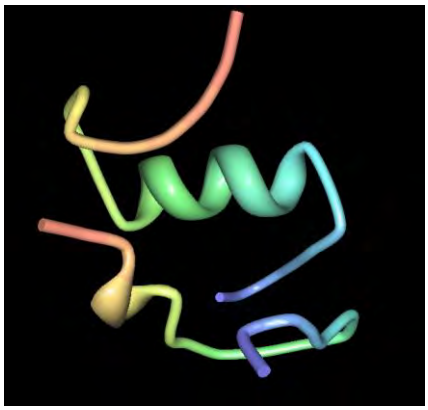


You can also use the bioinformatics tool [Translate]

Answer: The c -> t mutation in the insulin gene led to the replacement of the amino acid R (arginine, cgc codon) by the amino acid C (cysteine, cgt codon) at position 55.
This change prevents the insulin protein from being 'cut', a process which is essential for insulin to be functional (Molven et al., 2008).
Insulin is cleaved by an enzyme called 'Protease' (insulin protease or insulinase). The cleavage site recognized by insulinase is very specific: a change in the amino acid sequence of the cleavage site (such as the R55C mutation), will prevent the protease from being active.



Activity 4: 3D structure of insulin



Since 1958, researchers have been able to crystallize proteins and then 'take a picture' of them by using X-rays. The results of these experiments are then analyzed using bioinformatic programs which make it possible to **view** the 3D structure of proteins such as insulin.

View the 3D structure of insulin

* Go to the PDB entry 2LWZ
* Select the 3D viewer 'Protein Workshop'.
    A Jmol application will be launched and you will be asked to accept it. Jmol is a viewer for chemical structures in 3D. The Jmol application requires Java to be installed in your computer. Both programs are free.
* In Shortcuts: Recolor the backbone 'By compound' - and then look at the positions of the different amino acids (mouse over)
* In Tools: 'Surfaces' play with the Transparency slider
* In Tools: 'Visibility', 'atoms and bonds', click on 'Chain A: Insulin" and see the atoms (balls and sticks) that are displayed
* In Option: Reset - to go back to the original image

For fun, here are the raw experimental data, the spatial coordinates(X, Y, Z) of every atom in each amino of insulin (search ATOM in the page)

Note: There is no 3D structure data for insulin with the R55C mutation.

The amino acid sequence of a protein determines its shape and its function.
Here is a gallery of pictures, which will give you an idea of the relative sizes and shapes of different proteins (enlarged x 3,000,000) (pdf (5Mb)).
* Find the insulin among the different proteins and compare its size with the others.

Activity 5: Is insulin specific to humans?

BLAST

This is the full sequence of human insulin amino acid (in UniProtKB):
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIV
EQCCTSICSLYQLENYCN

Question:

- **Is this protein specific to humans?**

Bioinformatics approach:
Do a 'BLAST' against a database of proteins called UniProtKB

Technical information: BLAST is a bioinformatics tool that compares the sequence of a protein with millions of other sequences contained in a database. If they exist, it finds those that resemble a given sequence the most within a few seconds. We can thus find out quickly whether a protein exists in a given species, or not.

* Copy the sequence and paste it into the tool 'BLAST'
* Select '**Target Database = UniProtKB/Swiss-Prot**'
* Click on 'Run BLAST'
* Check the conservation of amino acids ('View alignment') and the conservation of the disulfide bonds ('Highlight' 'disulfide bond', when available)
* Search on Google for images corresponding to the different Latin names of the species (example 'Octodon degus')

According to wikipedia, insulin is a very old protein that may have originated one billion years ago.
Apart from animals, insulin-like proteins are also known to exist in Fungi and Protista kingdoms.

* Select '**Target Database = ...Nematoda**' or '**Target Database = ...Arthropoda**'

Multiple alignment

Starting from the following set of insulins from different species (in UniProtKB/Swiss-Prot)
* Select different species (mammals, birds, fish; *include human*)
* Do a multiple alignment (*Align*)
* Result page: '*Highlight Annotation*' '*Disulfide bond*' and '*Natural Variant*':
  - look at the cystein conservation (involved in disulfide bonds).
  - look at the conservation of the R55 amino acid .
  - look at the 2 major conserved regions which correspond to the A and B chains of mature insulin, respectively.

Introduction to phylogeny

You can also compare the insulin sequences of different species and sketch a phylogenetic tree with PhiloPhylo (in French)

Activity 6: www.chromosomewalk.ch

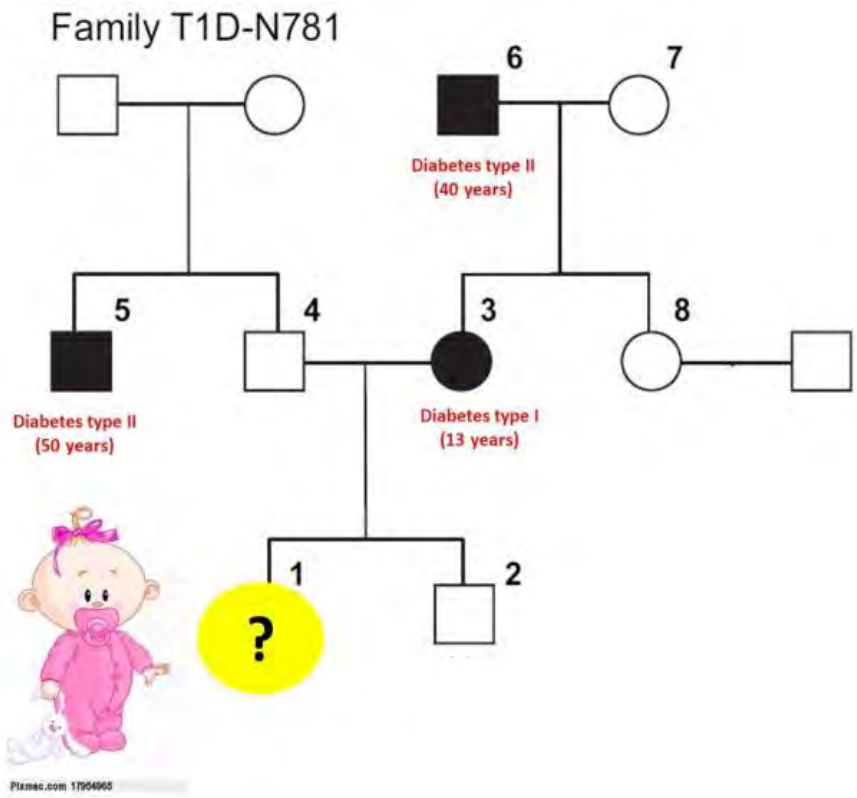
www.chromosomewalk.ch is a virtual exhibition to (re) discover the world of genes, proteins and bioinformatics ....

From the list of human chromosomes: search for 'insulin'

- **On which chromosome is located the insulin gene?**
- **What is the size of the chromosome (number of nucleotides and length in centimeters)?**
- **How many genes are there on this chromosome?**
- **Some additional information on insulin *(information)***

Find out if you are real expert: quiz expert !

To contact us... sp-com@isb-sib.ch



All diabetic type I members of the family are carriers of the same rare variation in the gene which encodes for insulin.
8 family members
2 alleles (maternal and paternal)
2 different DNA sequences (= 2 different regions of the insulin gene)

>1.1
tagtgtgcggggaacgaggcttcttctaca

>1.2
tagtgtgcggggaacgaggcttcttctaca

>1.3
cacccaagacccgccgggaggcagaggacc

>1.4
cacccaagacctgccgggaggcagaggacc

>2.1
tagtgtgcggggaacgaggcttcttctaca

>2.2
tagtgtgcggggaacgaggcttcttctaca

>2.3
cacccaagacccgccgggaggcagaggacc

>2.4
cacccaagacccgccgggaggcagaggacc

>3.1
tagtgtgcggggaacgaggcttcttctaca

>3.2
tagtgtgcggggaacgaggcttcttctaca

>3.3
cacccaagacccgccgggaggcagaggacc

>3.4
cacccaagacctgccgggaggcagaggacc

>4.1
tagtgtgcggggaacgaggcttcttctaca

>4.2
tagtgtgcggggaacgaggcttcttctaca

>4.3
cacccaagacccgccgggaggcagaggacc

>4.4
cacccaagacccgccgggaggcagaggacc

>5.1
tagtgtgcggggaacgaggcttcttctaca

>5.2
tagtgtgcggggaacgaggcttcttctaca

>5.3
cacccaagacccgccgggaggcagaggacc

>5.4
cacccaagacccgccgggaggcagaggacc

>6.1
tagtgtgcggggaacgaggcttcttctaca

>6.2
tagtgtgcggggaacgaggcttcttctaca

>6.3
cacccaagacccgccgggaggcagaggacc

>6.4
cacccaagacccgccgggaggcagaggacc

>7.1
tagtgtgcggagaacgaggcttcttctaca

>7.2
tagtgtgcggagaacgaggcttcttctaca

>7.3
cacccaagacccgccgggaggcagaggacc

>7.4
cacccaagacccgccgggaggcagaggacc

>8.1
tagtgtgcggggaacgaggcttcttctaca

>8.2
tagtgtgcggggaacgaggcttcttctaca

>8.3
cacccaagacccgccgggaggcagaggacc

>8.4
cacccaagacccgccgggaggcagaggacc

"Bringing Bioinformatics into the Biology Classroom"

Marie-Claude Blatter & Patricia Palagi

Marie-Claude.Blatter@isb-sib.ch

SIB Swiss Institute of Bioinformatics

Global Organisation for Bioinformatics Learning, Education & Training

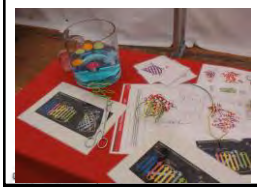
SIB Swiss Institute of Bioinformatics

- academic, non-profit foundation established in 1998
- coordinates **research** and **education** in bioinformatics throughout Switzerland
- provides high quality **bioinformatics services** to the national and international research community.
- helps shape the future of life sciences
- 52 groups, more than 600 scientists
- GOBLET member from the beginning

SIB outreach activities

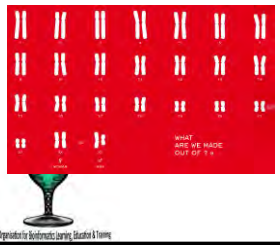
education & public at large

- Since 2000 (science fairs, electronic publication, exhibitions, hands on workshops, high school (HS) teacher continuing education training, etc.)
- Collaboration with public laboratories, didacticians and HS teachers
- **www.chromosomewalk.ch** (EN, FR, DE)
- **Protein Spotlight** (EN): http://web.expasy.org/spotlight/
- **'Ateliers de bioinformatique'** (FR): http://education.expasy.org/bioinformatique/
- *New project*: Drug Design and personalized medicine



www.ChromosomeWalk.ch

- **a saunter along the human genome**
- …take a walk and discover the world of **genes**, **proteins** and **bioinformatics**.
- quizzes, videos, links to databases and bioinformatics tools



proteinspotlight

> ONE MONTH, ONE PROTEIN <

- http://web.expasy.org/spotlight/
- above 160 articles, informal tone (V. Gerritsen)

«The German inventor Nikolaus Otto is credited with having invented the first automobile engine that ran on alcohol.»

Moving Forward

September 2014

Nature's imagination seems endless, and so is Man's. For as long as humans have existed, they have twisted Nature to meet their own needs. Wood has been used to keep them warm. Whale oil has been used to make light. Water has been harnessed to make electricity. And when the era of bio-engineering developed, it was not long before scientists found ways to tinker with an organism's genome for the benefits of mankind...

Swiss Institute of Bioinformatics

'Ateliers de Bioinformatique'

http://education.expasy.org/bioinformatique/ (FR)

Understanding a genetic disease thanks to Bioinformatics

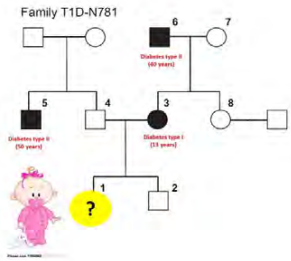
http://education.expasy.org/bioinformatique/Diabetes.html
(Atelier 7: L'insuline de A à Z ; English version)

additional documents are available here:
http://education.expasy.org/cours/Toronto

Swiss Institute of Bioinformatics

Context

In a **special case** of type I diabetes in a Norwegian family, a genetic variation has been found, leading to the production of inactive insulin



http://www.ncbi.nlm.nih.gov/pubmed/18192540

Diabetes. 2008 Apr;57(4):1131-5. doi: 10.2337/db07-1467. Epub 2008 Jan 11.

Mutations in the insulin gene can cause MODY and autoantibody-negative type 1 diabetes.

Molven A[1], Ringdal M, Nordbø AM, Raeder H, Støy J, Lipkind GM, Steiner DF, Philipson LH, Bergmann I, Aarskog D, Undlien DE, Joner G, Søvik O; Norwegian Childhood Diabetes Study Group, Bell GI, Njølstad PR

⊕ **Collaborators (27)**

⊕ **Author information**

Abstract

OBJECTIVE: Mutations in the insulin (INS) gene can cause neonatal diabetes. We hypothesized that mutations in INS could also cause maturity-onset diabetes of the young (MODY) and autoantibody-negative type 1 diabetes.

RESEARCH DESIGN AND METHODS: We screened INS in 62 probands with MODY, 30 probands with suspected MODY, and 223 subjects from the Norwegian Childhood Diabetes Registry selected on the basis of autoantibody negativity or family history of diabetes.

RESULTS: Among the MODY patients, we identified the INS mutation c.137G>A (R46Q) in a proband, his diabetic father, and a paternal aunt. They were diagnosed with diabetes at 20, 18, and 17 years of age, respectively, and are treated with small doses of insulin or diet only. In type 1 diabetic patients, we found the INS mutation c.163C>T (R55C) in a girl who at 10 years of age presented with ketoacidosis and insulin-dependent, GAD, and insulinoma-associated antigen-2 (IA-2) antibody-negative diabetes. Her mother had a de novo R55C mutation and was diagnosed with ketoacidosis and insulin-dependent diabetes at 13 years of age. Both had residual beta-cell function. The R46Q substitution changes an invariant arginine residue in position B22, which forms a hydrogen bond with the glutamate at A17, stabilizing the insulin molecule. The R55C substitution involves the first of the two arginine residues localized at the site of proteolytic processing between the B-chain and the C-peptide.

CONCLUSIONS: Our findings extend the phenotype of INS mutation carriers and suggest that INS screening is warranted not only in neonatal diabetes, but also in MODY and in selected cases of type 1 diabetes.

Comment in

Insulin mutations in diabetes: the clinical spectrum. [Diabetes. 2008]

PMID: 18192540 [PubMed - indexed for MEDLINE] **Free full text**

This publication is not available as free 'full text' in PubMed Central (PMC).
For full text:
http://education.expasy.org/cours/Toronto/

http://education.expasy.org/bioinformatique/Diabetes.html

Activity 1: The insulin gene and the human genome
(Genome browser (USCS), BLAT)

Activity 2: Comparing DNA sequences - Diagnosing a rare genetic disease
(alignment tool, database dbSNP)

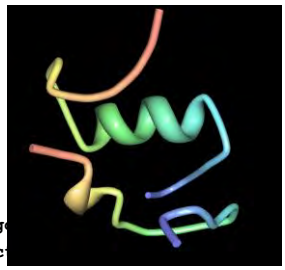
Activity 3: DNA translation -> protein
(translate tool)

Activity 4: 3D structure of insulin
(database PDB, 3D visualization tool)

Activity 5: Is insulin specific to humans?
(similarity search (BLAST), dabatase UniProtKB, alignment tool)

&

http://thumbs.dreamstime.com/z/cartoon-man-working-computer-13780903.jpg



protein (amino acid)

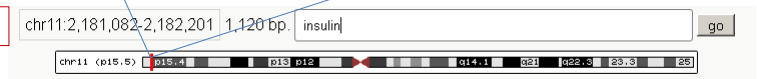
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVE
ALYLVCGERGFFYTPKTCREAEDLQVGQVELGGGPGA
GSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

gene (DNA; nucleic acid)

atggccctgtggatgcgcctcct[illegible]gtgctggcg[illegible]
ccagccgcagcctttgtga[illegible]gtgcggc[illegible]
tagtgtgcggggaacgag[illegible]acacacccaagacccgtcgggaggcagaggac
tgcaggtggggcaggtggag[illegible]gggcgggggccctggtgcaggcagcctgcagcccttg
gccctggaggggtccctgcagaagcgtggcattgtggaacaatgctgtaccagcatctgc
tccctctaccagctggagaactactgcaactag

mutation

genome



Activity 1

Activity 1: The insulin gene and the human genome

Bellow is a piece of the gene sequence that encodes for the insulin protein ('*wild sequence*')...

```
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
```

Question:

- **On which of our 23 chromosomes is this gene located?**

Bioinformatics approach:

Use the tool 'BLAT'
*Technical information: 'BLAT' is a bioinformatics tool for comparing a DNA sequence against the whole genome sequence (the human genome has 3 billion nucleotides). If the sequence exists, BLAT finds the sequence that is the most similar in just a few seconds. It's a bit like a small '**google map**' of the human genome.*

* Copy the DNA sequence and paste it in the tool **'BLAT'**
* Click on 'submit'
* In the page 'BLAT Search Result': choose the best score and click 'browser'

- **On which chromosome is located the gene for insulin?**
- **What are the beginning and end positions of the sequence on the chromosome (nucleotide 'numbers')?**

- **For fun: write a random sequence (about 30 letters), always using the 4-letter alphabet (a, t, g, c) into 'BLAT': can you find it in the genome?**

http://education.expasy.org/bioinformatique/Diabetes.html

Bioinformatics approach:

Use the tool 'BLAT' @ USCS

Technical information: 'BLAT' is a bioinformatics tool for comparing a DNA sequence against a whole genome sequence.

*If the sequence exists, BLAT finds the sequence that is the most similar in just a few seconds. It's a bit like a small **'google map'** of the human genome.*

1. Google: look for 'BLAT UCSC'

2. Choose the latest release of the human genome (GRCh38)

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human BLAT Search

BLAT Search Genome

Genome:	Assembly:	Query type:	Sort output:	Output type:
Human	Dec. 2013 (GRCh38/hg38)	BLAT's guess	query,score	hyperlink

cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc

3. Click on submit

submit | I'm feeling lucky | clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be

Bioinformatics – Genome Browser (Blat USCS)
http://genome.ucsc.edu/cgi-bin/hgBlat

Swiss Institute of Bioinformatics

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human BLAT Results

BLAT Search Results *human genome (GRCh37)*

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	60	1	60	60	100.0%	11	-	2182081	2182140	60
browser details	YourSeq	20	26	45	60	100.0%	9	+	138958442	138953461	20

At each genome release the positions may change

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human BLAT Results

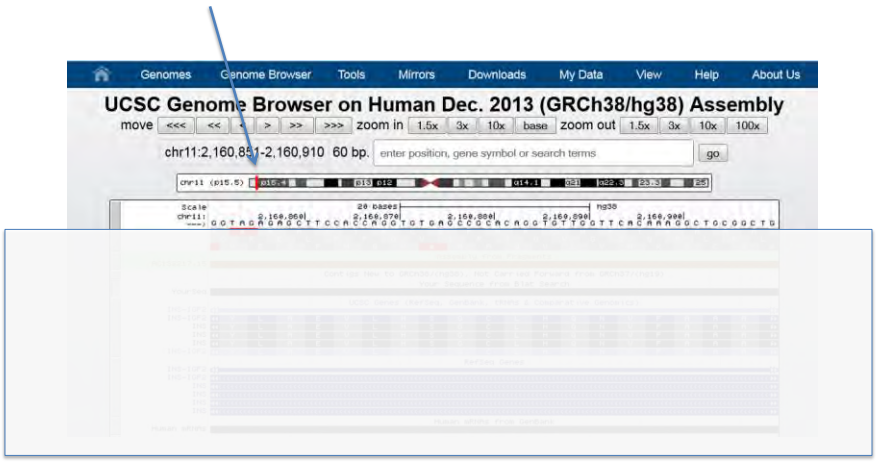
BLAT Search Results *human genome (GRCh38)*

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	60	1	60	60	100.0%	11	-	2160851	2160910	60
browser details	YourSeq	20	26	45	60	100.0%	9	+	136061596	136061615	20

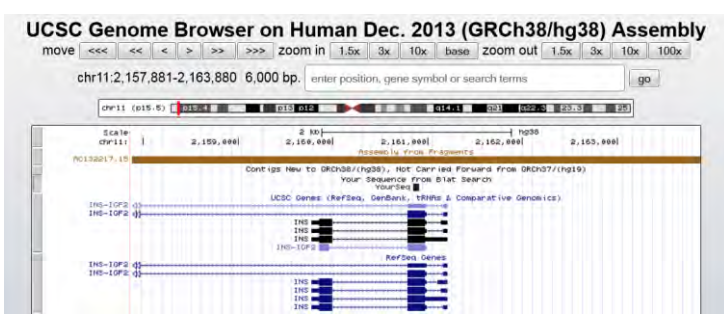
By default, choose the best score
Click on 'browser'

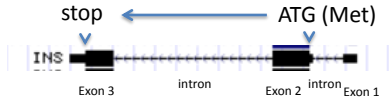
Swiss Institute of Bioinformatics

The insulin DNA sequence is located on **chromosome 11 (11p15.5)**
(positions: 2,160,851-2'160,910 (GRCh38))



Zoom out 100 x





The insulin gene consists of 3 exons and 2 introns

A readthrough transcript INS-IGF2 involves INS and IGF2 genes (neighboring genes)

Write a random sequence (about 30 letters), always using the 4-letter alphabet (a, t, g, c)

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human BLAT Search

BLAT Search Genome

Genome: Human | Assembly: Dec. 2013 (GRCh38/hg38) | Query type: BLAT's guess | Sort output: query,score | Output type: hyperlink

atgctagatcatgctagctagtcgatgctattgatcgatttagatcgat

Click on 'submit'

submit | I'm feeling lucky | clear

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help

Human BLAT Results

Sorry, no matches found

It is virtually impossible to match a randomly typed sequence (ATGC, n=30) on the human genome sequence, even on «junk» DNA regions (Application: PCR and primer selection)

Randomly selected letters (i.e. n=5) rarely create a correct word….

Write a random sequence (about 30 letters), always using the 4-letter alphabet (a, t, g, c)

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

White rhinoceros BLAT Search

BLAT Search Genome

Genome: White rhinoceros | Assembly: May 2012 (CerSim/CerSim1.0/cerSim1) | Query type: BLAT's guess | Sort output: query,score | Output type: hyperlink

atgctagatcatgctagctagtcgatgctacttgatcgatttagatcgat

Select another genome

submit | I'm feeling lucky | clear

http://jon-atkinson.com/Large%20Images%207/White%20Rhino.jpg

Swiss Institute of Bioinformatics

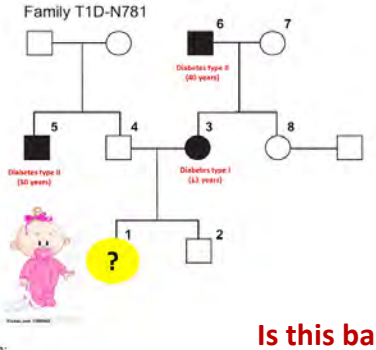
Activity 2

Activity 2: Comparing DNA sequences - Diagnosing a rare genetic disease

In 2008, scientists studied a Norwegian family in which several members had diabetes (type I or type II) (Molven et al., 2008).

All diabetic type I members of the family carry the same rare variation in the gene which encodes for insulin.

Here is the family's pedigree (phenotype and family relationship):



- affected male
- male
- affected female
- female

Is this baby diabetic (type I) ?

Question:

- Is this baby diabetic?

http://education.expasy.org/bioinformatique/Diabetes.html

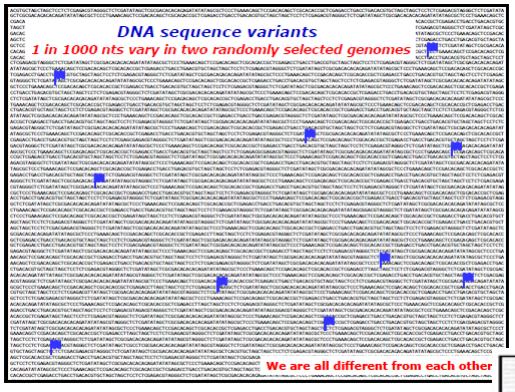
Biological context: Human genome & variations

The human genome = a text of 3’000’000’000 pb

= a reference sequence

All the differences (also called *variations*, *variants*, *Single Nucleotide Polymorphisms (SNPs)*, ~*mutations*, …) between human subjects are described on the basis of this ‘text’

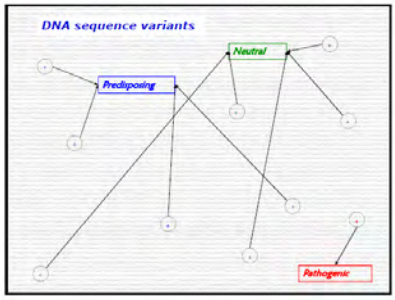
Swiss Institute of Bioinformatics



~ 3.3 millions of SNPs between 2 individuals

~10 millions of SNPs in the human population
- neutral (the majority)
- associated with a particular phenotype…
- associated with a predisposition
- associated with a genetic disease

~10 – 30 new mutation at each new generation



Pr S.E. Antonarakis (UNIGE)

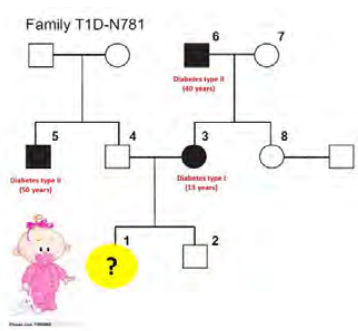
Biological context: Human genome & variations

1. Comprehensive characterization of **human genome** variation by high coverage whole-**genome** sequencing of **forty four Caucasians**.
   Shen H, Li J, Zhang J, Xu C, Jiang Y, Wu Z, Zhao F, Liao L, Chen J, Lin Y, Tian Q, Papasian CJ, Deng HW.
   PLoS One. 2013;8(4):e59494. doi: 10.1371/journal.pone.0059494. Epub 2013 Apr 5.
   PMID: 23577066 [PubMed - indexed for MEDLINE] **Free PMC Article**
   Related citations

Publication (free full text) in PubMed Central (PMC @NCBI) are freely available for everyone

"On average, each individual genome carried ~3.3 million SNPs and ~492,000 indels/block substitutions, including approximately 179 variants that were predicted to cause loss of function of the gene products. "

Swiss Institute of Bioinformatics



Question: is the baby diabetic ?

To answer this question, researchers extracted DNA from 8 members of the Norwegian family and sequenced part of the gene that encodes for insulin.

http://www2.grifil.com/album.html

Compare these sequences, and locate the common variation for diabetes.

'Paper and pencil' approach:
... You can do it **manually which will help you better understand the principle of sequence comparison and alignment.**
Take into account all the given clues and play with our strips of DNA sequences...

- 8 family members - 4 DNA sequences - one allele
- 8 family members - 1 DNA sequence - two alleles
- 8 family members - 2 DNA sequences - two alleles (not easy)

Bioinformatics approach:
Build an alignment of these 8 sequences using a bioinformatics tool and look out for the common variation among those with diabetes

* Copy these 8 sequences (including the lines starting with '>1') and paste them into the align tool
* Click on the *Run Align* button.
* On the results page, on the lefthand column 'Highlight': select 'Similarity'

Global Organisation for Bioinformatics Learning, Education & Training

SIB Swiss Institute of Bioinformatics

'Paper and pencil' approach:
... You can do it **manually which will help you better understand the principle of sequence comparison and alignment.**
Take into account all the given clues and play with our strips of DNA sequences...

- 8 family members - 4 DNA sequences - one allele
- 8 family members - 1 DNA sequence - two alleles
- 8 family members - 2 DNA sequences - two alleles (not easy)

Global Organisation for Bioinformatics Learning, Education & Training

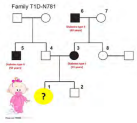
2 different DNA sequences (INS gene)

8 subjects (same family)

```
>1.1
tagtgtgcggggaacgaggcttcttctaca

>1.2
Tagtgtgcggggaacgaggcttcttctaca

>1.3
cacccaagacccgccgggaggcagag

>1.4
cacccaagacctgccgggaggcagaggacc

>2.1
tagtgtgcggggaacgaggcttcttctaca

>2.2
tagtgtgcggggaacgaggcttcttctaca

>2.3
cacccaagacccgccgggaggcagaggacc

cacccaagacccgccgggaggcagaggacc

>3.1
tagtgtgcggggaacgaggcttcttctaca

>3.2
tagtgtgcggggaacgaggcttcttctaca

>3.3
cacccaagacccgccgggaggcagaggacc

>3.4
cacccaagacctgccgggaggcagaggacc

>4.1
tagtgtgcggggaacgaggcttcttctaca

>4.2
tagtgtgcggggaacgaggcttcttctaca

>4.3
cacccaagacccgccgggaggcagaggacc

>4.4
cacccaagacccgccgggaggcagaggacc

>5.1
tagtgtgcggggaacgaggcttcttctaca

>5.2
tagtgtgcggggaacgaggcttcttctaca

>5.3
cacccaagacccgccgggaggcagaggacc

>5.4
cacccaagacccgccgggaggcagaggacc
```

2 alleles (maternal / paternal)

Sequence 1

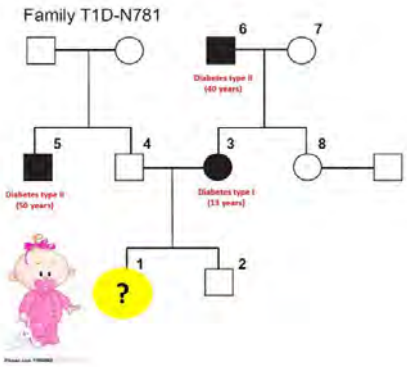


```
1.1   tagtgtgcggggaacgaggcttcttctaca
1.2   tagtgtgcggggaacgaggcttcttctaca
2.1   tagtgtgcggggaacgaggcttcttctaca
2.2   tagtgtgcggggaacgaggcttcttctaca
3.1   tagtgtgcggggaacgaggcttcttctaca
3.2   tagtgtgcggggaacgaggcttcttctaca
4.1   tagtgtgcggggaacgaggcttcttctaca
4.2   tagtgtgcggggaacgaggcttcttctaca
5.1   tagtgtgcggggaacgaggcttcttctaca
5.2   tagtgtgcggggaacgaggcttcttctaca
6.1   tagtgtgcggggaacgaggcttcttctaca
6.2   tagtgtgcggggaacgaggcttcttctaca
7.1   tagtgtgcggagaacgaggcttcttctaca
7.2   tagtgtgcggagaacgaggcttcttctaca
8.1   tagtgtgcggggaacgaggcttcttctaca
8.2   tagtgtgcggggaacgaggcttcttctaca
```

Where are the differences ?

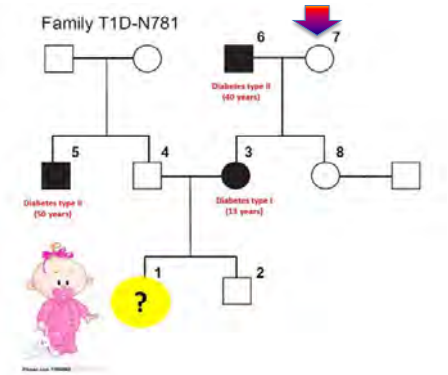
http://www2.grifil.com/album.html

Sequence 1




```
1.1   tagtgtgcggggaa
1.2   tagtgtgcggggaa
2.1   tagtgtgcggggaa
2.2   tagtgtgcggggaa
3.1   tagtgtgcggggaa
3.2   tagtgtgcggggaa
4.1   tagtgtgcggggaa
4.2   tagtgtgcggggaa
5.1   tagtgtgcggggaa
5.2   tagtgtgcggggaa
6.1   tagtgtgcggggaacgaggcttcttctaca
6.2   tagtgtgcggggaacgaggcttcttctaca
7.1   tagtgtgcggagaacgaggcttcttctaca
7.2   tagtgtgcggagaacgaggcttcttctaca
8.1   tagtgtgcggggaacgaggcttcttctaca
8.2   tagtgtgcggggaacgaggcttcttctaca
```

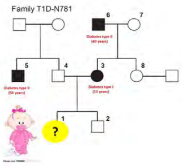
Sequence 1




```
1.1   tagtgtgcgggga
1.2   tagtgtgcgggga
2.1   tagtgtgcgggga
2.2   tagtgtgcgggga
3.1   tagtgtgcgggga
3.2   tagtgtgcgggga
4.1   tagtgtgcgggga
4.2   tagtgtgcgggga
5.1   tagtgtgcgggga
5.2   tagtgtgcgggga
6.1   tagtgtgcggggaacgaggcttcttctaca
6.2   tagtgtgcggggaacgaggcttcttctaca
7.1   tagtgtgcgg agaacgaggcttcttctaca
7.2   tagtgtgcgg agaacgaggcttcttctaca
8.1   tagtgtgcggggaacgaggcttcttctaca
8.2   tagtgtgcggggaacgaggcttcttctaca
```

The SNP g -> a (homozygous; subject 7) is not associated with diabetes (neutral)

Sequence 2

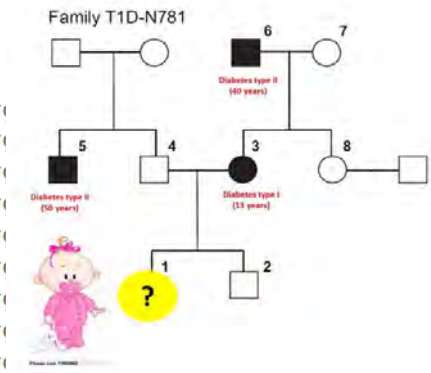



```
11   cacccaagacccgccgggaggcagaggacc
12   cacccaagacctgccgggaggcagaggacc
21   cacccaagacccgccgggaggcagaggacc
22   cacccaagacccgccgggaggcagaggacc
31   cacccaagacccgccgggaggcagaggacc
32   cacccaagacctgccgggaggcagaggacc
41   cacccaagacccgccgggaggcagaggacc
42   cacccaagacccgccgggaggcagaggacc
51   cacccaagacccgccgggaggcagaggacc
52   cacccaagacccgccgggaggcagaggacc
61   cacccaagacccgccgggaggcagaggacc
62   cacccaagacccgccgggaggcagaggacc
71   cacccaagacccgccgggaggcagaggacc
72   cacccaagacccgccgggaggcagaggacc
81   cacccaagacccgccgggaggcagaggacc
82   cacccaagacccgccgggaggcagaggacc
```

Where are the differences ?

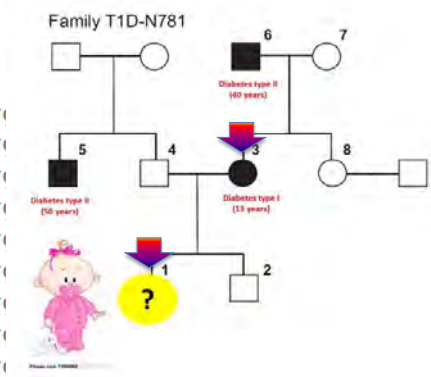
Sequence 2

```
11  cacccaagacccgccgg
12  cacccaagacctgccgg
21  cacccaagacccgccgg
22  cacccaagacccgccgg
31  cacccaagacccgccgg
32  cacccaagacctgccgg
41  cacccaagacccgccgg
42  cacccaagacccgccgg
51  cacccaagacccgccgg
52  cacccaagacccgccgggaggcagaggacc
61  cacccaagacccgccgggaggcagaggacc
62  cacccaagacccgccgggaggcagaggacc
71  cacccaagacccgccgggaggcagaggacc
72  cacccaagacccgccgggaggcagaggacc
81  cacccaagacccgccgggaggcagaggacc
82  cacccaagacccgccgggaggcagaggacc
```



Sequence 2

```
11  cacccaagacccgccgg
12  cacccaagacctgccgg
21  cacccaagacccgccgg
22  cacccaagacccgccgg
31  cacccaagacccgccgg
32  cacccaagacctgccgg
41  cacccaagacccgccgg
42  cacccaagacccgccgg
51  cacccaagacccgccgg
52  cacccaagacccgccgggaggcagaggacc
61  cacccaagacccgccgggaggcagaggacc
62  cacccaagacccgccgggaggcagaggacc
71  cacccaagacccgccgggaggcagaggacc
72  cacccaagacccgccgggaggcagaggacc
81  cacccaagacccgccgggaggcagaggacc
82  cacccaagacccgccgggaggcagaggacc
```



ANSWER: The SNP c -> t is present in subjects 3 and 1 (heterozygous) and is associated with Type I Diabetes ('all type I diabetic members carry the same variation in the INS gene')

Bioinformatics approach:
Build an alignment of these 8 sequences using a bioinformatics tool and look out for the common variation among those with diabetes

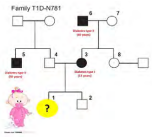
* Copy these 8 sequences (including the lines starting with '>1') and paste them into the align tool
* Click on the *Run Align* button.
* On the results page, on the lefthand column 'Highlight': select 'Similarity'

Bioinformatics – Alignment tool (UniProt)
http://www.uniprot.org/align/

This tool is used to align protein sequences, but it can also properly align short DNA sequences

Swiss Institute of Bioinformatics

Global Organisation for Bioinformatics Learning, Education & Training



Question: is the baby diabetic ?

To answer this question, researchers extracted DNA from 8 members of the Norwegian family and sequenced part of the gene that encodes for insulin.

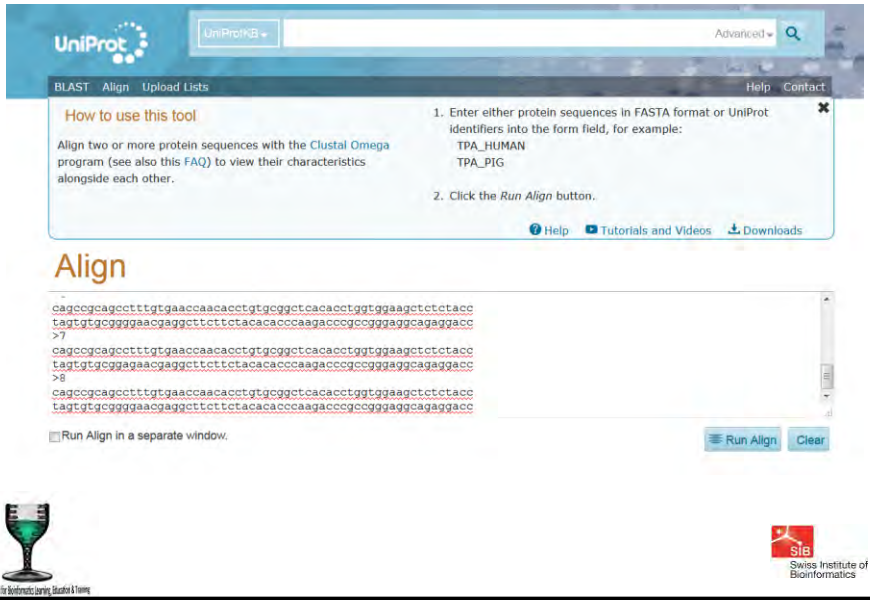
```
>1
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacctgccgggaggcagaggacc
>2
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>3
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacctgccgggaggcagaggacc
>4
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>5
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>6
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>7
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggagaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
>8
cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc
tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc
```

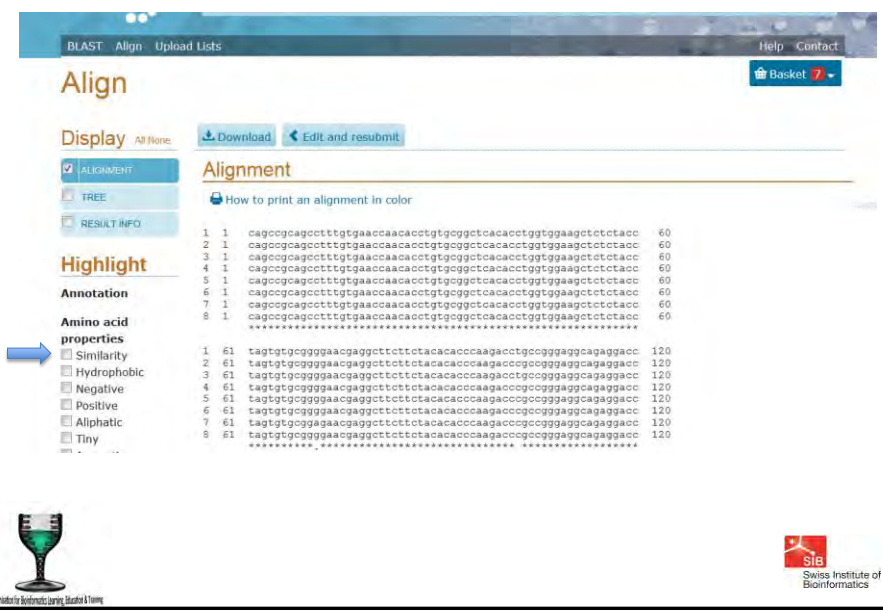
Where are the differences ?

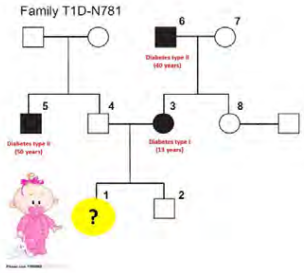
http://www2.grifil.com/album.html

Global Organisation for Bioinformatics Learning, Education & Training

http://www.uniprot.org/align/



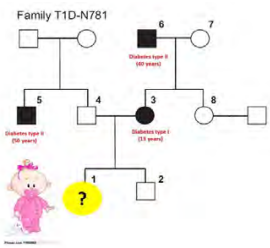




Bioinformatics – Alignment tool (UniProt)
http://www.uniprot.org/align/

(only one DNA sequence (allele) per subject)

```
1  1   cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc   60
2  1   cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc   60
3  1   cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc   60
4  1   cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc   60
5  1   cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc   60
6  1   cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc   60
7  1   cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc   60
8  1   cagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctacc   60
       ************************************************************

1  61  tagtgtgcggggaacgaggcttcttctacacacccaagacctgccgggaggcagaggacc  120
2  61  tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc  120
3  61  tagtgtgcggggaacgaggcttcttctacacacccaagacctgccgggaggcagaggacc  120
4  61  tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc  120
5  61  tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc  120
6  61  tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc  120
7  61  tagtgtgcggagaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc  120
8  61  tagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggacc  120
       **********.****************************** ******************
```



- The subject **(1)** with the **c -> t** mutation (heterozygous mutation) is a girl who presented type I diabetes at the early age of 10.
- The girl's mother **(3)** has type I diabetes that was diagnosed when she was 13. Currently, she is being treated with insulin. She also carries the heterozygous mutation **c -> t** .
- The girl's maternal grandfather **(6)** has type 2 diabetes, which was diagnosed at the age of 40. He is currently being treated with insulin. Neither he nor the healthy maternal grandmother carry mutations.
- **-> Thus, the girl's mother is carrying a *de novo* c -> t mutation, which must be a germline mutation since it has been inherited by her daughter.**

(Molven et al., 2008)

Biology / Statistics / Bioinformatics

Molven et al., 2008

"We screened INS (*gene coding for insulin*) in 92 probands, and 223 subjects from the Norwegian Childhood **Diabetes** Registry selected on the basis of autoantibody negativity or family history of diabetes."

Identification of a rare genetic variation - rs121908261 – in the human INS gene which is the cause of type I diabetes in a Norwegian family

Swiss Institute of Bioinformatics

Insulin gene
(4992 pb ; chromosome 11)

In red, part of the DNA sequence translated into the protein sequence

rs121908261
c -> t

Bioinformatics – Biological Database: dbSNP @ NCBI
http://www.ncbi.nlm.nih.gov/SNP/

Variant accession number in dbSNP

NCBI Resources How To Sign in to NCBI

dbSNP SNP rs121908261 Search

Save search Advanced Help

Display Settings: Summary Send to:

rs121908261 *[Homo sapiens]*

1.

AGGCTTCTTCTACACACCCAAGACC [C/T] GCCGGGAGGCAGAGGACCTGCAGGG

Chromosome: 11:2160809

Gene: INS-IGF2 (GeneView) INS (GeneView)

Functional Consequence: missense,nc transcript variant

Clinical significance: Pathogenic

Validated: no info

HGVS: NC_000011.10:g.2160809G>A, NC_000011.9:g.2182039G>A, NG_007114.1:g.5386C>T, NM_000207.2:c.163C>T, NM_001042376.2:c.163C>T, NM_001185097.1:c.163C>T, NM_001185098.1:c.163C>T, NM_001291897.1:c.163C>T, NP_000198.1:p.Arg55Cys, NP_001035835.1:p.Arg55Cys, NP_001172026.1:p.Arg55Cys, NP_001172027.1:p.Arg55Cys, NP_001278826.1:p.Arg55Cys, NR_003512.3:n.222C>T

Pubmed Varview Protein3D OMIM

Search details

rs121908261[All Fields]

Search See more...

Recent activity

Turn Off Clear

rs121908261 (1)

SNP

Mutations in the insulin gene can cause MODY and autoantibodies PubMed

See more...

Link to the publication of Molven et al (2008)

Global Organization for Bioinformatics Learning, Education & Training

Bioinformatics – Biological Database: dbSNP @ NCBI
http://www.ncbi.nlm.nih.gov/SNP/

Look for all the SNPs located within human INS gene

NCBI Resources How To

dbSNP SNP ins

Save search Advanced

Show additional filters

Clear all

Organism

Homo sapiens

Bos taurus

More ...

Variation Class

in del

snp

Clinical Significance

Clinical/LSDB Submissions

benign

likely benign

other

pathogenic

uncertain significance

untested

Annotation

Cited in PubMed

OMIM

PubMed

nucleotide

protein

structure

Function Class

3' utr

5' utr

coding synonymous

intron

Display Settings: Summary, 20 per page, Sorted by SNP_ID Send to:

Results: 1 to 20 of 314 Page 1 of 16 Next > Last >>

Filters activated: Homo sapiens. Clear all to show 3415 items.

rs5505 *[Homo sapiens]*

1

CCCTGCCTGTCTCCCAGATCACTGT [C/T] CTTCTGCCATGGCCCTGTGGATGCG

Chromosome: 11:2160980

Gene: INS-IGF2 (GeneView) INS (GeneView)

Functional Consequence: nc transcript variant,utr variant 5 prime

Allele Origin: C(germline)/T(germline)

Clinical significance: Benign

Validated: no info

Global MAF: A=0.0058/28

HGVS: NC_000011.10:g.2160880G>A, NC_000011.9:g.2182210G>A, NG_007114.1:g.5215C>T, NM_000207.2:c.-9C>T, NM_001042376.2:c.-9C>T, NM_001185097.1:c.-9C>T, NM_001185098.1:c.-9C>T, NM_001291897.1:c.-9C>T, NR_003512.3:n.51C>T

Varview

rs5507 *[Homo sapiens]*

2

CCGCTGTTCCGGAACCTGCTCTGCG [C/T] GGCACGTCCTGGCAGTGGGGCAGGT

Chromosome: 11:2160013

Gene: INS-IGF2 (GeneView) INS (GeneView)

Functional Consequence: intron variant

Allele Origin: C(germline)/T(germline)

Clinical significance: Benign

Validated: no info

Global MAF: A=0.0082/40

HGVS: NC_000011.10:g.2160013G>A, NC_000011.9:g.2181243G>A, NG_007114.1:g.6182C>T, NM_000207.2:c.188-16C>T, NM_001042376.2:c.187+772C>T.

Global Organization for Bioinformatics Learning, Education & Training

Activity 3

Activity 3: DNA translation -> protein

Check the effect of the mutation 'R55C'...

Like all proteins,insulin is composed of a sequence of amino acids. The order of the amino acids is determined by the nucleic acid sequence of the insulin gene. 3 letters of DNA (codon) correspond to one amino acid (symbolized by letters: K for lysine, M for methionine, etc.).

This is a piece of the DNA sequence of the normal insulin gene.

```
aag acc cgc cgg gag
```

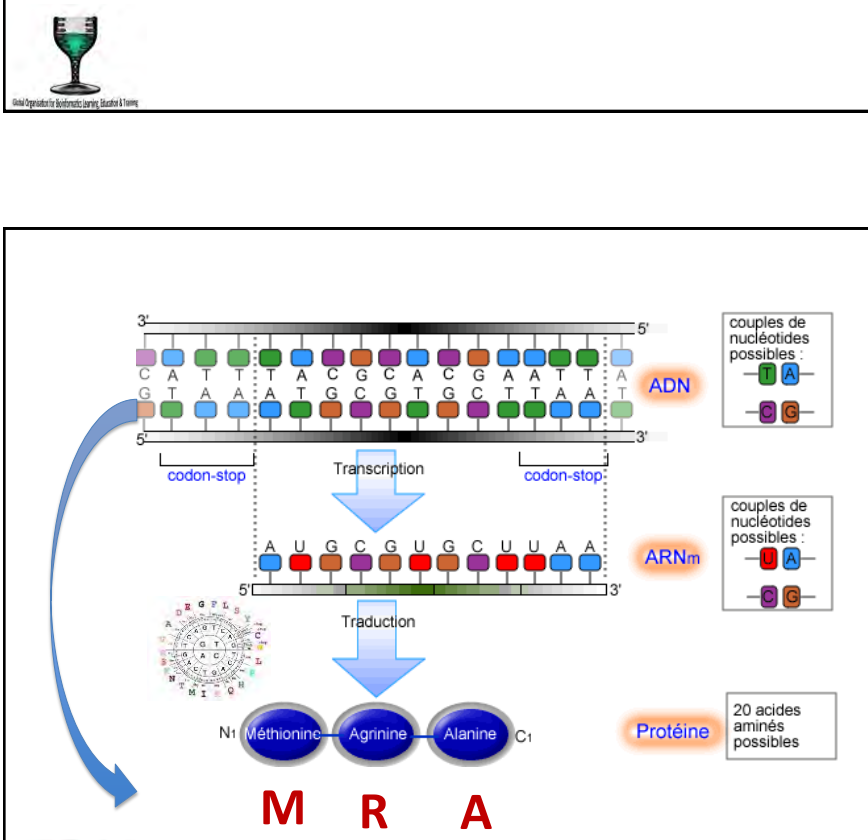
This is a piece of the DNA sequence of the insulin gene with the c -> t variation, associated with type I diabetes.

```
aag acc tgc cgg gag
```

Question:

- **Does the c->t mutation change the amino acid sequence of insulin?**
- **Does the aag -> aaa mutation change the amino acid sequence of insulin?**

http://education.expasy.org/bioinformatique/Diabetes.html



http://fr.wikipedia.org/wiki/Synth%C3%A8se_des_prot%C3%A9ines#mediaviewer/File:Proteines.png

Activity 3: DNA translation -> protein

Check the effect of the mutation 'R55C'

Like all proteins,insulin is composed of a sequence of amino acids. The order of the amino acids is determined by the nucleic acid sequence of the insulin gene. 3 letters of DNA (codon) correspond to one amino acid (symbolized by letters: K for lysine, M for methionine, etc.).

This is a piece of the DNA sequence of the normal insulin gene

```
aag acc cgc cgg gag
```

This is a piece of the DNA sequence of the insulin gene with the c-->t variation, associated with type I diabetes...

```
aag acc tgc cgg gag
```

Question:

- **Does the c->t mutation change the amino acid sequence of insulin?**
- **Does the aag -> aaa mutation change the amino acid sequence of insulin?**

You could manually translate the nucleic acid sequences into amino acid sequences ('1 'letter code) using the genetic code below: :

Swiss Institute of Bioinformatics

Global Organisation for Bioinformatics Learning, Education & Training

Bioinformatics – Translate tool
http://www.bioinformatics.org/sms2/translate.html

www.bioinformatics.org/sms2/translate.html

traduction | Linguee | SPmanual | Google Drive | Dico | sp interne | UniProt | SIBintranet | Curation Manual | confluence | UNIL2013

SMS — Sequence Manipulation Suite:

Translate

Translate accepts a DNA sequence and converts it into a protein in the reading frame you specify. Translate supports the entire IUPAC alphabet and several genetic codes.

Paste a raw sequence or one or more FASTA sequences into the text area below. Input limit is 200000 characters.

```
>normal
aag acc cgc cgg gag

>mutated
aag acc tgc cgg gag
```

Fasta format

Please check the browser compatibility page before using this program.

Submit | Clear | Reset

- Translate in reading frame 1 on the direct strand.
- Use the standard (1) genetic code.

Translate results

```
>rf 1 normal
KTRRE

>rf 1 mutated
KTCRE
```

Swiss Institute of Bioinformatics

Global Organisation for Bioinformatics Learning, Education & Training

Amino acid sequence of the 'normal' insulin

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

Chain B: FVNQHLCGSHLVEALYLVCGERGFFYTPKT

Chain A: GIVEQCCTSICSLYQLENYCN

Amino acid sequence of the 'mutated' insulin (variant c-> t; R 55 C)

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTCREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

The mutation R -> C prevents insulin from being cut and thus from being biologically active

Swiss Institute of Bioinformatics

http://www.ncbi.nlm.nih.gov/pubmed/18192540

Diabetes. 2008 Apr;57(4):1131-5. doi: 10.2337/db07-1467. Epub 2008 Jan 11.

Mutations in the insulin gene can cause MODY and autoantibody-negative type 1 diabetes.

Molven A[1], Ringdal M, Nordbø AM, Raeder H, Støy J, Lipkind GM, Steiner DF, Philipson LH, Bergmann I, Aarskog D, Undlien DE, Joner G, Søvik O; Norwegian Childhood Diabetes Study Group, Bell GI, Njølstad PR

Collaborators (27)

Author information

Abstract

OBJECTIVE: Mutations in the insulin (INS) gene can cause neonatal diabetes. We hypothesized that mutations in INS could also cause maturity-onset diabetes of the young (MODY) and autoantibody-negative type 1 diabetes.

RESEARCH DESIGN AND METHODS: We screened INS in 62 probands with MODY, 30 probands with suspected MODY, and 223 subjects from the Norwegian Childhood Diabetes Registry selected on the basis of autoantibody negativity or family history of diabetes.

RESULTS: Among the MODY patients, we identified the INS mutation c.137G>A (R46Q) in a proband, his diabetic father, and a paternal aunt. They were diagnosed with diabetes at 20, 18, and 17 years of age, respectively, and are treated with small doses of insulin or diet only. In type 1 diabetic patients, we found the INS mutation c.163C>T (R55C) in a girl who at 10 years of age presented with ketoacidosis and insulin-dependent, GAD, and insulinoma-associated antigen-2 (IA-2) antibody-negative diabetes. Her mother had a de novo R55C mutation and was diagnosed with ketoacidosis and insulin-dependent diabetes at 13 years of age. Both had residual beta-cell function. The R46Q substitution changes an invariant arginine residue in position B22, which forms a hydrogen bond with the glutamate at A17, stabilizing the insulin molecule. The R55C substitution involves the first of the two arginine residues localized at the site of proteolytic processing between the B-chain and the C-peptide.

CONCLUSIONS: Our findings extend the phenotype of INS mutation carriers and suggest that INS screening is warranted not only in neonatal diabetes, but also in MODY and in selected cases of type 1 diabetes.

Comment in

Insulin mutations in diabetes: the clinical spectrum. [Diabetes. 2008]

PMID: 18192540 [PubMed - indexed for MEDLINE] **Free full text**

SORRY

This publication is not available as free 'full text' in PubMed Central (PMC).
For full text:
http://education.expasy.org/cours/Toronto/

Activity 4

Activity 4: 3D structure of insulin

Since 1958, researchers have been able to crystallize proteins and then 'take a picture' of them by using X-rays. The results of these experiments are then analyzed using bioinformatic programs which make it possible to view the 3D structure of proteins such as insulin.

View the 3D structure of insulin

* Go to the PDB entry **2LWZ**
* Select the 3D viewer 'Protein Workshop'
  A Jmol application will be launched and you will be asked to accept it. Jmol is a viewer for chemical structures in 3D.
  The Jmol application requires Java to be installed in your computer. Both programs are free.
* In Shortcuts: Recolor the backbone 'By compound' - and then look at the positions of the different amino acids (mouse over)
* In Tools: 'Surfaces' play with the Transparency slider
* In Tools: 'Visibility', "atoms and bonds", click on 'Chain A: Insulin" and see the atoms (balls and sticks) that are displayed
* In Option: Reset - to go back to the original image

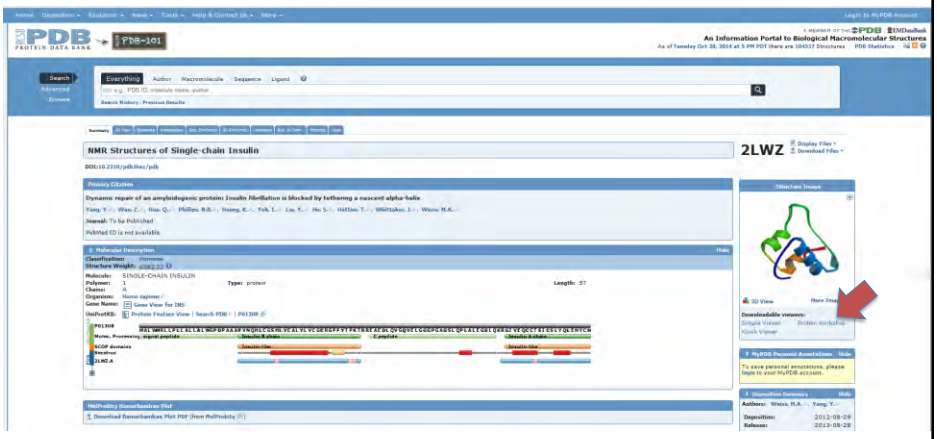
For fun, here are the raw experimental data, the spatial coordinates(X, Y, Z) of every atom in each amino of insulin (search ATOM in the page)

Note: There is no 3D structure data for insulin with the R55C mutation.

http://education.expasy.org/bioinformatique/Diabetes.html

Global Organization for Bioinformatics Learning, Education & Training

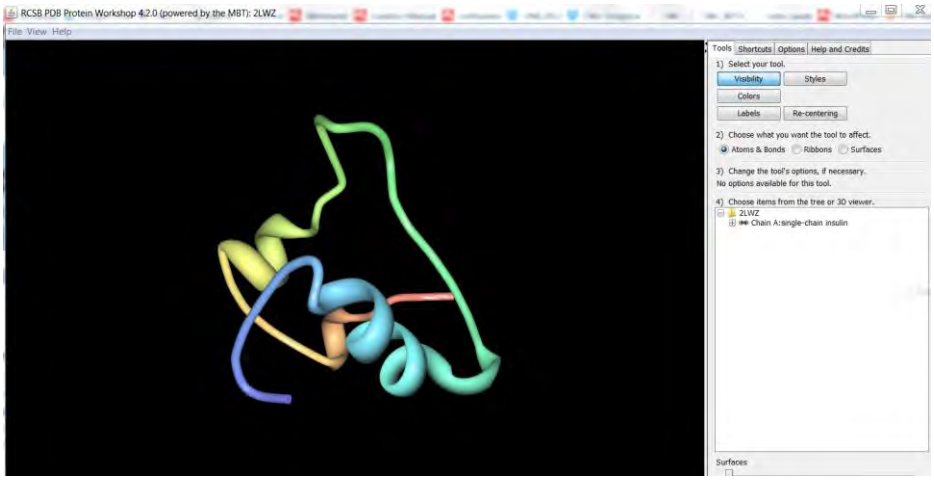
http://www.pdb.org/pdb/explore/explore.do?structureId=2LWZ



….. requires Java to be installed in your computer.

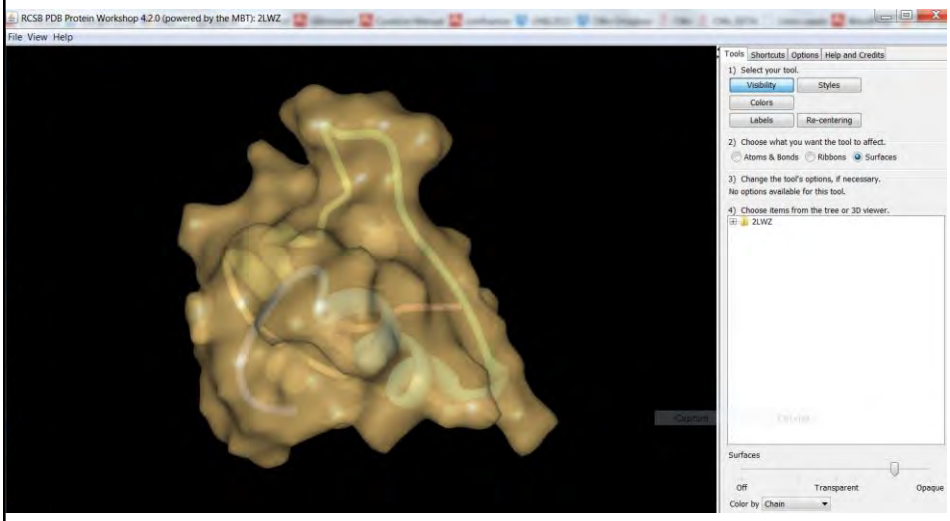
Global Organization for Bioinformatics Learning, Education & Training

Bioinformatics – Insulin 3D structure in PDB database (2LWZ) (Protein Workshop)
http://www.pdb.org/pdb/explore/explore.do?structureId=2LWZ



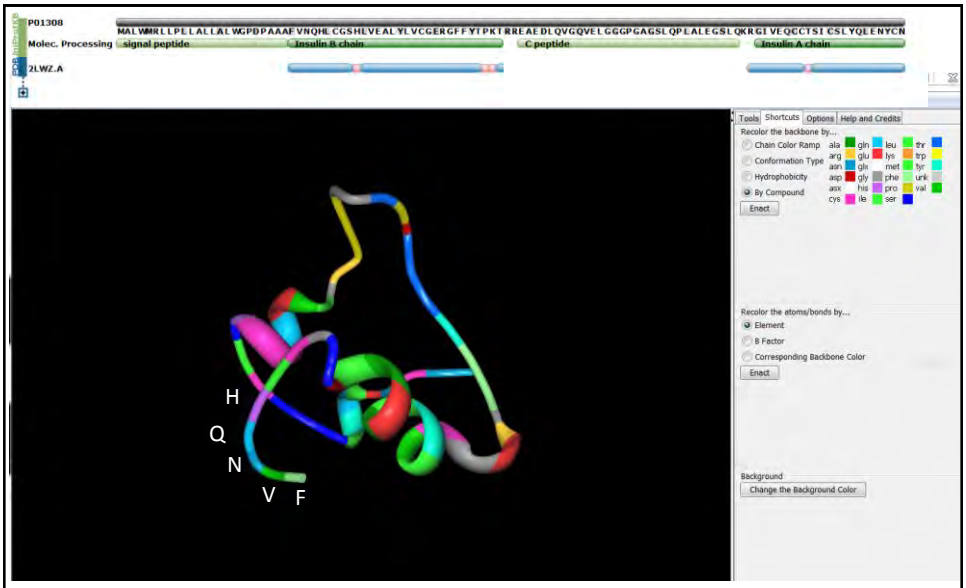
Visualization tool: Protein Workshop

Swiss Institute of Bioinformatics

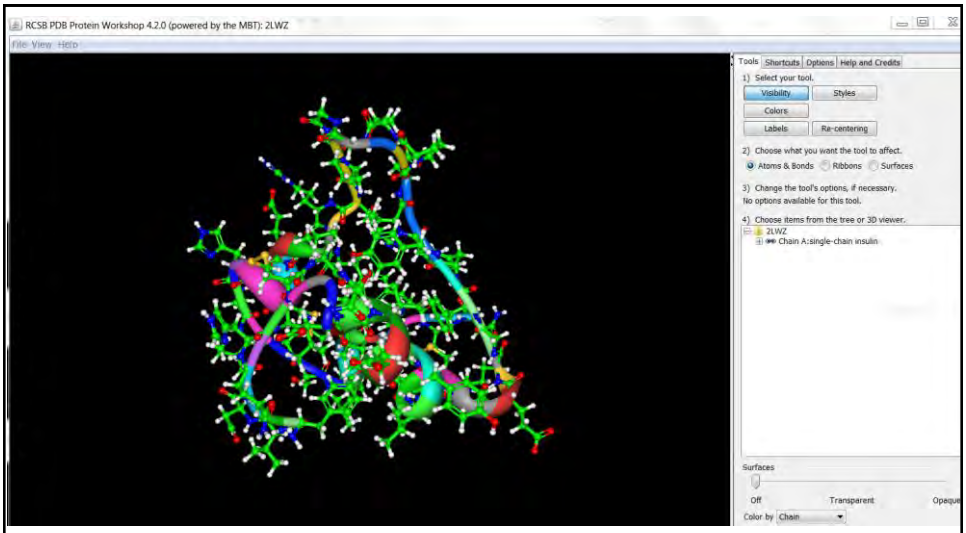


Protein Workshop: in Tools: 'Surfaces' play with the Transparency slider

Swiss Institute of Bioinformatics



Protein workshop:
In Shortcuts: Recolor the backbone 'By compound' - and then look at the positions of the different amino acids (mouse over)



Protein workshop: In Tools: 'Visibility', 'atoms and bonds', click on 'Chain A: Insulin" and see the atoms (balls and sticks) that are displayed

Activity 5

Activity 5: Is insulin specific to humans?

BLAST

This is the full sequence of human insulin amino acid (in UniProtKB):

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

Question:

- **Is this protein specific to humans?**

Bioinformatics approach:
Do a 'BLAST' against a database of proteins called UniProtKB

Technical information: BLAST is a bioinformatics tool that compares the sequence of a protein with millions of other sequences contained in a database. If they exist, it finds those that resemble a given sequence the most within a few seconds. We can thus find out quickly whether a protein exists in a given species, or not.

* Copy the sequence and paste it into the tool 'BLAST'
* Select '**Target Database = UniProtKB/Swiss-Prot**'
* Click on 'Run BLAST'
* Check the conservation of amino acids ('View alignment') and the conservation of the disulfide bonds ('Highlight' 'disulfide bond', when available)
* Search on Google for images corresponding to the different Latin names of the species (example 'Octodon degus')

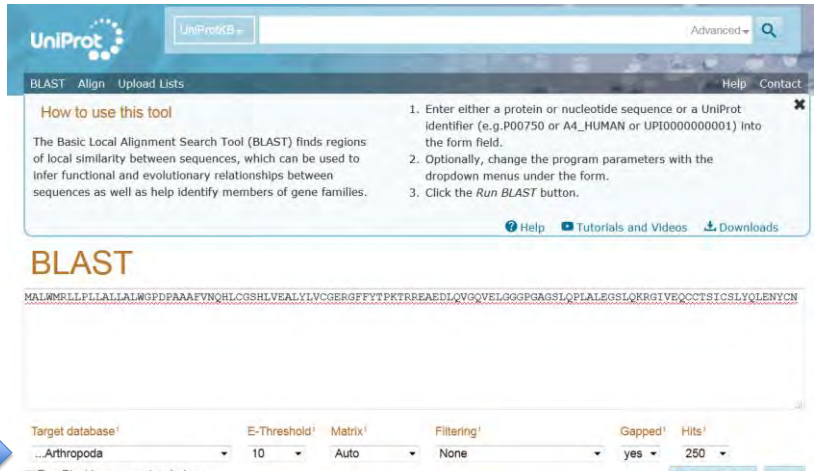
According to wikipedia, insulin is a very old protein that may have originated one billion years ago.
Apart from animals, insulin-like proteins are also known to exist in Fungi and Protista kingdoms.

* Select '**Target Database = ...Nematoda**' or '**Target Database = ...Arthropoda**'

http://education.expasy.org/bioinformatique/Diabetes.html

Global Organization for Bioinformatics Learning, Education & Training

Bioinformatics – BLAST Similarity search tool
www.uniprot.org/blast/



Swiss Institute of Bioinformatics

Global Organization for Bioinformatics Learning, Education & Training

UniProt

BLAST Align Upload Lists

Help Contact

Basket 7

BLAST

100 80 60 40 20 0
Identity %

Filter by

Reviewed (33)
Swiss-Prot

Unreviewed (127)
TrEMBL

With 3D structure (3)
Proteomes (112)

Organisms
Fruit fly (11)
MANSE (1)
STRMM (1)
CAMFO (3)
AMBVA (1)
Other organisms
Go

Edit and resubmit
Order by: Score Limit to sequences from organism: All

Overview

Show all 160

Entry	Protein names	Match hit (500 1k 1.5k)	Identity
Q4JJX8	Bombyxin (Manduca sexta)		35.0%
T1JF91	Uncharacterized protein (Strigamia maritima)		33.0%
E2AZ92	Insulin (Camponotus floridanus)		38.0%
F0J9V1	Insulin (Amblyomma variegatum)		35.0%

Alignments

Columns BLAST Align Download

1 to 25 of 160 Show 25

'reviewed' entries (UniProtKB/Swiss-Prot section) are manually reviewed
'unreviewed' entries (UniProtKB/TrEMBL section) are automatically annotated

Global Organisation for Bioinformatics Learning, Education & Training

UniProt

UniProtKB uniprot:(job:B201411012E154U7QF6) Advanced

BLAST Align Upload Lists

Help Contact

Basket 7

BLAST

100 80 60 40 20 0
Identity %

Filter by

Reviewed (33)
Swiss-Prot

With 3D structure (3)
Proteomes (27)

Organisms
Fruit fly (3)
BOMMO (24)
AGRCO (2)
SAMCY (3)
LOCMI (1)

Map To
UniProtKB
UniRef
UniParc

Edit and resubmit Order by: Score Limit to sequences from organism: All

Overview

Show all 33

P33721	Bombyxin B-1 homolog (Samia cynthia)	43.0%
P33722	Bombyxin B-2 homolog (Samia cynthia)	45.0%
Q9VT52	Probable insulin-like peptide 3 (Drosophila melanogaster)	26.0%
P26733	Bombyxin B-1 (Bombyx mori)	30.0%
P26741	Bombyxin B-7 (Bombyx mori)	30.0%
P26729	Bombyxin A-6 (Bombyx mori)	30.0%

Alignments

Columns BLAST Align Download

1 to 25 of 33 Show 25

Swiss Institute of Bioinformatics

Global Organisation for Bioinformatics Learning, Education & Training

How similar are the human and drosophila sequences ?

>sp|Q9VT52|INSL3_DROME Probable insulin-like peptide 3 OS=Drosophila melanogaster GN=Ilp3 PE=2 SV=2
MGIEMRCQDRRILLPSLLLLILMIGGVQATMKLCGRKLPETLSKLCVYGFNAMTKRTLDP
VNFNQIDGFEDRSLLERLLSDSSVQMLKTRRLRDGVFDECCLKSCTMDEVLRYCAAKPRT

>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

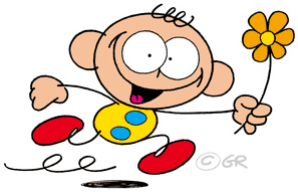
Swiss Institute of Bioinformatics

Bioinformatics – alignment tool
www.uniprot.org/align/

Swiss Institute of Bioinformatics

Many thanks to all of you

and
to Michelle Brazas



http://www2.grifil.com/album.html

Global Organization for Bioinformatics Learning, Education & Training

SIB
Swiss Institute of Bioinformatics

List of URLs - workshop – 14 November 2014

GOBLET
http://mygoblet.org/

SIB Swiss Institute of Bioinformatics
www.isb-sib.ch

'Ateliers de Bioinformatique' (FR)
http://education.expasy.org/bioinformatique/

Understanding a genetic disease thanks to Bioinformatics
http://education.expasy.org/bioinformatique/Diabetes.html

Additional documents for this workshop
http://education.expasy.org/cours/Toronto/

Bioinformatics tools

Genome browser – BLAT @ UCSC
http://genome.ucsc.edu/cgi-bin/hgBlat

Similarity search tool - BLAST (protein) @ UniProt
http://www.uniprot.org/blast/

Alignment tool (protein) – Clustal @ UniProt
http://www.uniprot.org/align/

Translate tool (DNA -> protein)
http://www.bioinformatics.org/sms2/translate.html

Biological databases

Database protein – UniProtKB
http://www.uniprot.org/

Database Single Nucleotide Polymorphism (SNP) – db SNP @ NCBI
http://www.ncbi.nlm.nih.gov/snp

Database 3D structure – PDB @ RSCB
http://www.rcsb.org/pdb/home/home.do

SIB Outreach

ChromosomeWalk.ch
www.chromosomewalk.ch

An introduction to Phylogeny – PhiloPhylo (FR)
http://education.expasy.org/cgi-bin/philophylo/philophylo.cgi

Electronic publication - Protein spotlight
http://web.expasy.org/spotlight/

Mutations in the Insulin Gene Can Cause MODY and Autoantibody-Negative Type 1 Diabetes

Anders Molven,[1,2] Monika Ringdal,[3,4] Anita M. Nordbø,[3,4] Helge Ræder,[5] Julie Støy,[6] Gregory M. Lipkind,[7] Donald F. Steiner,[6,7] Louis H. Philipson,[6] Ines Bergmann,[8] Dagfinn Aarskog,[9] Dag E. Undlien,[10,11] Geir Joner,[12,13] Oddmund Søvik,[3] the Norwegian Childhood Diabetes Study Group,* Graeme I. Bell,[6,14] and Pål R. Njølstad[3,5]

OBJECTIVE—Mutations in the insulin (*INS*) gene can cause neonatal diabetes. We hypothesized that mutations in *INS* could also cause maturity-onset diabetes of the young (MODY) and autoantibody-negative type 1 diabetes.

RESEARCH DESIGN AND METHODS—We screened *INS* in 62 probands with MODY, 30 probands with suspected MODY, and 223 subjects from the Norwegian Childhood Diabetes Registry selected on the basis of autoantibody negativity or family history of diabetes.

RESULTS—Among the MODY patients, we identified the *INS* mutation c.137G>A (R46Q) in a proband, his diabetic father, and a paternal aunt. They were diagnosed with diabetes at 20, 18, and 17 years of age, respectively, and are treated with small doses of insulin or diet only. In type 1 diabetic patients, we found the *INS* mutation c.163C>T (R55C) in a girl who at 10 years of age presented with ketoacidosis and insulin-dependent, GAD, and insulinoma-associated antigen-2 (IA-2) antibody-negative diabetes. Her mother had a de novo R55C mutation and was diagnosed with ketoacidosis and insulin-dependent diabetes at 13 years of age. Both had residual β-cell function. The R46Q substitution changes an invariant arginine residue in position B22, which forms a hydrogen bond with the glutamate at A17, stabilizing the insulin molecule. The R55C substitution involves the first of the two arginine residues localized at the site of proteolytic processing between the B-chain and the C-peptide.

CONCLUSIONS—Our findings extend the phenotype of *INS* mutation carriers and suggest that *INS* screening is warranted not only in neonatal diabetes, but also in MODY and in selected cases of type 1 diabetes. ***Diabetes* 57:1131–1135, 2008**

From the [1]Gade Institute, University of Bergen, Norway; the [2]Department of Pathology, Haukeland University Hospital, Bergen, Norway; the [3]Department of Clinical Medicine, University of Bergen, Bergen, Norway; the [4]Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway; the [5]Department of Pediatrics, Haukeland University Hospital, Bergen, Norway; the [6]Department of Medicine, The University of Chicago, Chicago, Illinois; the [7]Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, Illinois; [8]Kristiansund Hospital, Kristiansund, Norway; [9]Buskerud Hospital, Drammen, Norway; the [10]Institute of Medical Genetics, Faculty Division, Ullevål University Hospital, University of Oslo, Oslo, Norway; the [11]Department of Medical Genetics, Ullevål University Hospital, Oslo, Norway; the [12]Department of Pediatrics, Ullevål University Hospital, Oslo, Norway; the [13]Faculty of Medicine, University of Oslo, Oslo, Norway; and the [14]Department of Human Genetics, The University of Chicago, Chicago, Illinois.

Address correspondence and reprint requests to Dr. Pål R. Njølstad, Department of Pediatrics, Haukeland University Hospital, N-5021 Bergen, Norway. E-mail: pal.njolstad@uib.no.

Received for publication 14 October 2007 and accepted in revised form 6 January 2008.

Published ahead of print at http://diabetes.diabetesjournals.org on 11 January 2008. DOI: 10.2337/db07-1467.

*Other members of the Norwegian Childhood Diabetes Study Group are listed in the APPENDIX.

IA-2, insulinoma-associated antigen-2; MODY, maturity-onset diabetes of the young.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

See accompanying original articles on pgs. 1034 and 1115 and commentary on p. 799.

Molecular genetic studies of monogenic forms of diabetes such as maturity-onset diabetes of the young (MODY) and neonatal diabetes have provided important insight into the pathophysiology and have led to improved diagnosis and treatment (1–7). In type 1 diabetes, immune-mediated destruction of the pancreatic β-cells plays an important role in the pathogenesis (8). However, some type 1 diabetic children do not present with signs of autoimmunity and are classified as having autoantibody-negative type 1 diabetes, also denoted idiopathic or type 1b diabetes (9–11). Recently, we observed that heterozygous missense mutations in the insulin gene (*INS*) can cause permanent neonatal diabetes (12). The majority of these mutations occurred de novo. Moreover, this phenomenon has been noted in previous studies of *KCNJ11* and *ABCC8* in patients with neonatal diabetes and is in accordance with the sporadic nature of permanent neonatal diabetes.

We hypothesized that *INS* mutations might also cause MODY and could explain some cases of apparent type 1 diabetes. The aim of the present study was therefore to search for *INS* mutations in patients with MODY of unknown etiology as well as in selected patients from the Norwegian Childhood Diabetes Registry.

RESEARCH DESIGN AND METHODS

Physicians refer subjects to the Norwegian MODY Registry based on at least two of the following criteria: first-degree relative with diabetes, onset of diabetes before 25 years of age in at least one family member, insulin level <0.5 units $\cdot$ kg^{-1} $\cdot$ day^{-1}, diabetes diagnosed between age 25 and 40 years of age, or unusual type 1 diabetes (low-dose insulin requirement, no antibodies, or atypical history). The conventional criteria of MODY (13) are therefore not met in all cases. Still, inclusion of subjects based strictly on the conventional criteria would exclude some true MODY patients, e.g., those with de novo mutations, age at diagnosis >25 years, or limited clinical data on the family history of diabetes. We screened DNA samples from 92 probands of the Norwegian MODY Registry for mutations in *INS*; 62 fulfilled conventional MODY criteria, while 30 were categorized as "suspected MODY." None of the probands had mutations in *HNF1A* (14). Moreover, 57 of the probands had a phenotype clinically evaluated as MODY2-like. *GCK* mutations had therefore been excluded in them. Standard oral glucose tolerance testing was performed, and World Health Organization criteria for diabetes were applied.

In addition, we investigated samples from the population-based Norwegian Childhood Diabetes Registry (15). From June 2002 to June 2007, 1,373 subjects were eligible and enrolled. We excluded subjects with mutations in *HNF1A* or *KCNJ11* and one subject with diabetes secondary to pancreatectomy. We then chose to screen two subsets of subjects in the present study. The first consisted of patients who were GAD and insulinoma-associated antigen-2

(IA-2) antibody negative, with or without a family history of diabetes (n = 124). The second subset consisted of GAD and/or IA-2 antibody-positive patients with at least one parent with diabetes (n = 99). Antibodies were measured the day after diagnosis. We used the following cut offs to define antibody status as negative: GAD <0.08 units and IA-2 <0.1 units. Diabetes in the parents included all types of diabetes. Thus, we sequenced *INS* in a total of 223 subjects regarded as having type 1 diabetes. The reference ranges for fasting C-peptide were 220–1,400 pmol/l for subjects in the MODY family N580 and 400–1,700 pmol/l for subjects in the type 1 diabetes family N781.

We obtained written informed consent from all participants or their parents. The study was approved by the Regional Committee for Research Ethics and the Norwegian Data Inspectorate and performed according to the Helsinki Declaration.

Genotyping. DNA was purified from EDTA blood samples by standard methods. Human *INS* was amplified in two segments using PCR and the primers 5′-CAAGGGCCTTTGCGTCA-3′ together with 5′-GAAGCCAACAC-CGTCCTCA-3′ (exon 2) and 5′-CCCTGACTGTGTCCTCCTGT-3′ together with 5′-AGAGAGCGTGGAGAGAGCTG-3′ (exon 3). The exon and flanking noncoding regions of *INS* were sequenced in both directions using an Applied Biosystems 3730 DNA Analyzer (Applied Biosystems, Foster City, CA). We imported all sequence sample files into the SeqScape Software (Applied Biosystems) and analyzed them for variation in *INS*. Template sequence applied for *INS* was NM_000207 (NCBI database).

RESULTS

INS mutations and MODY. We did not find any pathogenic mutations in the 30 subjects with suspected MODY. We did, however, find a heterozygous mutation in 1 of the 62 families fulfilling conventional MODY criteria (Fig. 1*A* and Table 1). The mutation c.137G>A is predicted to alter arginine to glutamine at residue number 46 (R46Q) of the preproinsulin molecule (Fig. 2). The proband (N580-1) was diagnosed with diabetes at 20 years of age. Initially he was treated with diet only. After 1 year, he needed psychopharmacologic treatment, his BMI increased to 29.6 kg/m^2, and he required insulin. Subsequently, his psychopharmacologic treatment was changed, he lost weight, and he is now on diet only. N580-1 was GAD and IA-2 antibody negative; his nonfasting C-peptide was undetectable, and his most recent A1C was 5.9% (normal range 4.0–6.0%). His father (N580-3) was diagnosed with diabetes at 18 years of age; he was initially treated with diet only, and after ~20 years a sulfonylurea was introduced. In later years, he has been treated with small doses of insulin. The proband's paternal aunt (N580-4) was diagnosed with diabetes at 17 years of age and has been on a strict diet without need for pharmacological treatment.

INS mutations and type 1 diabetes. Nearly all Norwegian subjects diagnosed with diabetes at age 18 years or under are included in the Norwegian Childhood Diabetes Registry. We first screened *INS* in a group of 124 antibody-negative cases and found a heterozygous mutation in 1 proband. The mutation, c.163C>T, is predicted to cause an arginine to cysteine substitution at residue 55 (R55C) of the preproinsulin molecule (Fig. 2). We subsequently sequenced a further 99 subjects with antibody-positive type 1 diabetes and a parental history of diabetes but identified no further mutations.

The subject N781-1 with the R55C mutation presented with frank diabetes at 10 years of age (Fig. 1*B* and Table 1). She had a blood glucose of 17.6 mmol/l and ketoacidosis, and her A1C was 9.1% (normal range 4.0–6.2%). Autoantibodies against insulin were 5.8 units/ml (normal range <1.0), while fasting C-peptide was detectable (500 pmol/l). She was insulin dependent from the time of diagnosis. Her most recent insulin requirement and A1C were 0.72 units · kg^{-1} · day^{-1} and 8.0%, respectively. A recent meal-stimulated C-peptide was detectable (1,050 pmol/l, paired glucose 9.6 mmol/l). Recent autoantibodies against insulin were positive (6.3 units/ml), while GAD and IA-2 were negative. Her mother (N781-3) had type 1 diabetes diagnosed at 13 years of age (Table 1). She is currently treated with insulin (0.96 units · kg^{-1} · day^{-1}). Recent meal-stimulated C-peptide was barely detectable (420 pmol/l, paired glucose 11.1 mmol/l). Recent autoantibodies against insulin were positive (5.8 units/ml), while GAD and IA-2 were negative. She also carries the heterozygous mutation. The proband's maternal grandfather (N781-6) had type 2 diabetes diagnosed at 40 years of age. He is treated with insulin (0.47 units · kg^{-1} · day^{-1}), and his most recent A1C was 6.4%. His current BMI is 42.7 kg/m^2,

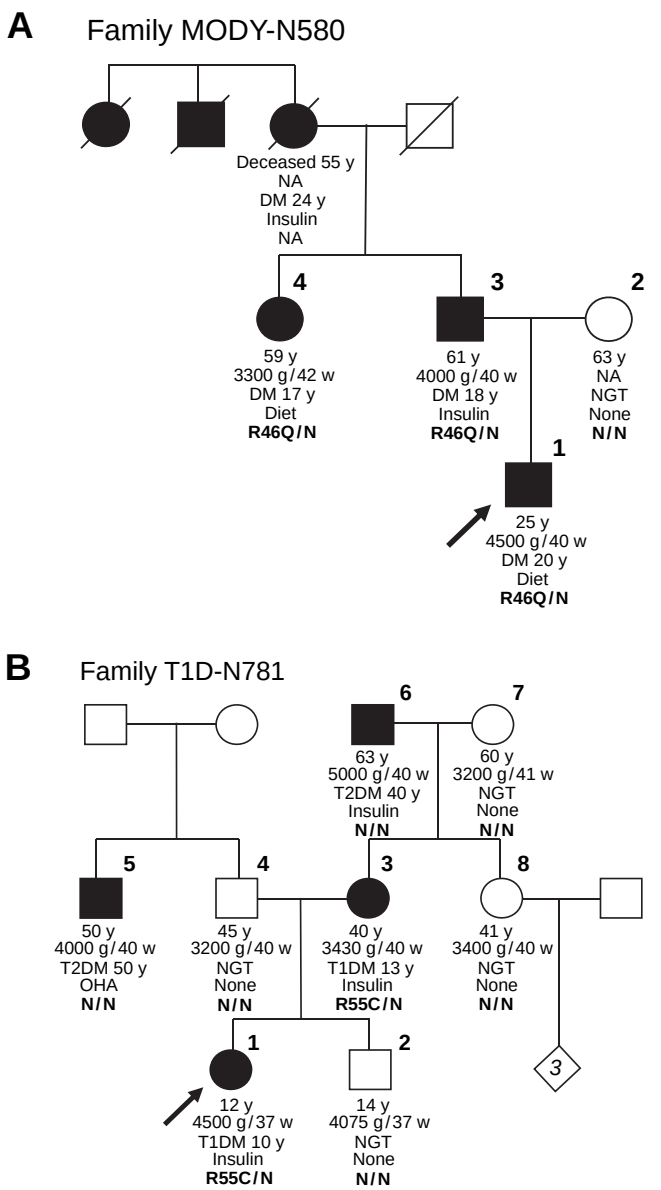


FIG. 1. Mutations in *INS* can cause MODY and type 1 diabetes. *A*: Pedigree of a family with MODY due to the mutation R46Q. The three cases of diabetes in the first generation were unavailable for genetic analysis, but limited clinical information could be obtained for one of them. *B*: Pedigree of a family with antibody-negative type 1 diabetes and the mutation R55C. For both pedigrees, current age, birth weight/gestational age, age of diagnosis, current treatment, and mutation status are listed. Subjects with diabetes are shown in black. Females are represented by circles and males by squares. The probands are marked by arrows. DM, diabetes mellitus; NA, not available; NGT, normal glucose tolerance; OHA, oral hypoglycemic agents; T1DM, type 1 diabetes; T2DM, type 2 diabetes.

TABLE 1
Clinical characteristics of the subjects with the *INS* mutations R46Q and R55C

	Family				
	MODY-N580			T1D-N781	
Subject	N580–1	N580–3	N580–4	N781–1	N781–3
General characteristics					
INS mutation	R46Q	R46Q	R46Q	R55C	R55C
Sex	M	M	F	F	F
Current age (years)	25	61	59	12	40
Centile for birth weight (SD score)	+1.5	+1	−1.5	>+2	+0
Onset of diabetes					
Age (years)	20	18	17	10	13
Clinical manifestation	Hyperglycemia	Glucosuria	Glucosuria	Hyperglycemia, ketoacidosis	Hyperglycemia, ketoacidosis
Recent status					
BMI (kg/m^2)	23.9	25.0	18.1	24.6	37.6
A1C (%)	5.9	6.0	6.0	8.0	8.9
Insulin dose (units · kg^{-1} · day^{-1})	None (diet treated)	0.25	None (diet treated)	0.72	0.96
Other	Bipolar disorder	Neuropathy Hypertension	—	—	—

and he has nephropathy, retinopathy, and neuropathy. Neither he nor the healthy maternal grandmother are mutation carriers. Thus, the proband's mother has a de novo mutation. The paternal uncle (N781-5) was diagnosed with type 2 diabetes at 50 years of age. He is treated with glimepiride (2 mg/day); his most recent A1C was 7.3%, and BMI was 21.4 kg/m^2. He is not carrying the mutation.

The pathogenic role of the INS mutations R46Q and R55C. We did not detect either mutation among 100 healthy blood donors. Neither mutation has been described previously (12,16). The mutation R46Q alters an invariant arginine at residue 22 of the B-chain. The guanidino group of arginine forms a hydrogen bond with the glutamate at residue 17 of the A-chain and participates in a network of electrostatic interactions with surrounding carbonyl and carboxyl oxygens, which stabilizes the structure of the insulin molecule (Fig. 3). The substitution of arginine B22 by glutamine will disrupt this critical hydrogen bond. The mutation R55C affects the first of the two arginines at the B-chain—C-peptide junction, i.e., the first site of proteolytic processing of proinsulin to insulin. The

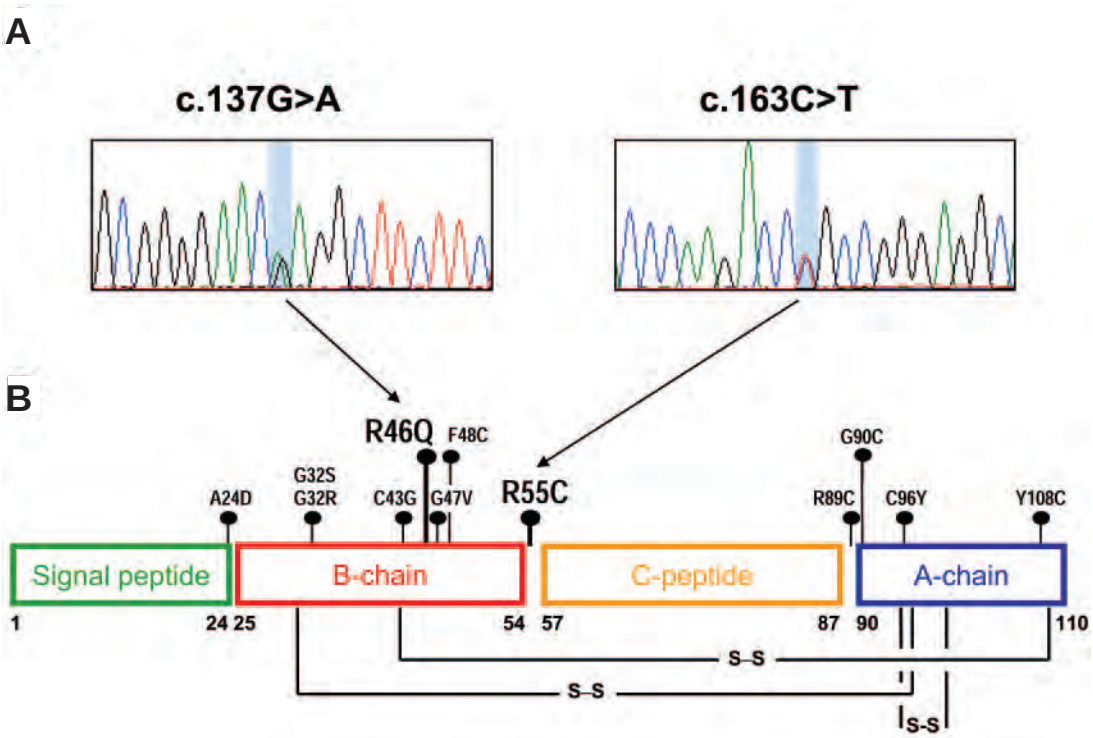


FIG. 2. *A*: DNA sequences of the *INS* mutations c.137G>A (R46Q) and c.163C>T (R55C) found in the Norwegian MODY Registry and the Norwegian Childhood Diabetes Registry, respectively. *B*: Location of the two corresponding amino acid substitutions in the preproinsulin molecule. The 10 mutations identified by Støy et al. (12) are shown in smaller font. Amino acid numbers below the bars show the extension of each peptide fragment in preproinsulin. Note that the amino acids 55/56 and 88/89 form the recognition sites for the proteolytic removal of the C-peptide but are not part of the mature insulin molecule. "S-S" indicates disulfide bridge.

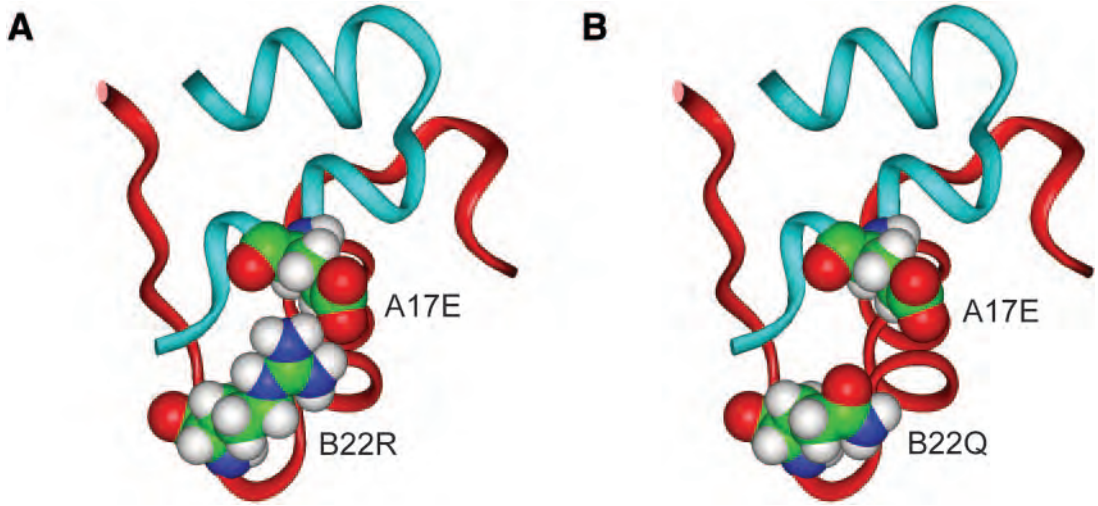


FIG. 3. Predicted effect of the R46Q mutation on structural stability of the insulin molecule. *A*: The native structure of insulin, shown by space-filled image, where the side chain of arginine B22 (B22R) forms a hydrogen bond with the side chain of glutamate A17 (A17E). This hydrogen bond stabilizes the COOH-terminal ends of the A- and B-chains (shown by red and blue ribbons, respectively). *B*: Effect of mutating the arginine to glutamine at B22 (B22Q). Substitution of the long side chain of arginine by the 4.5 Å shorter side chain of glutamine disrupts the formation of a hydrogen bond between residues B22 and A17. B22R is invariant, while A17 tolerates only two stereochemically equivalent amino acid residues, glutamate and glutamine, both of which allow the hydrogen bond between B22 and A17.

substitution of arginine with a neutral residue (in this case cysteine) is not predicted to interfere with the proteolytic processing by proinsulin endoprotease PC1/3. It is thus more likely that the introduction of an unpaired cysteine may affect insulin biosynthesis, as noted for C96Y, the mutant insulin in the Akita mouse, by introducing a defect in folding of the preproinsulin molecule (12,17). Both carriers of the R55C mutation have C-peptide levels in the normal range, thus suggesting that some insulin is being processed and secreted. It is currently not fully understood why these patients, despite evidence of insulin secretion, have severe insulin deficiency, as indicated by ketoacidosis at diagnosis and subsequent requirement for insulin in full replacement doses.

DISCUSSION

We have found that mutations in the gene encoding insulin can cause MODY and antibody-negative type 1 diabetes. Our findings add *INS* to the list of causes of MODY, which currently includes *HNF4A*, *GCK*, *HNF1A*, *IPF1*, *HNF1B*, *NEUROD1*, and *CEL*. The relatively mild phenotype of the three family members with the R46Q mutation suggests that a spectrum of phenotypes may exist in patients with *INS* mutations, ranging from mild diabetes and hyperinsulinemia in patients with the previously described mutations that cause reduced biological activity of the insulin molecule (i.e., B24 Ser, B25 Leu, and A3 Leu) (18,19), to MODY in patients with mutations that are predicted to reduce the structural stability of the insulin molecule (R46Q) and ultimately to neonatal diabetes in patients with mutations that cause severe defects in the biosynthesis of the insulin molecule (for example B8 Ser and B19 Gly) (12).

One could argue that the case with apparent type 1 diabetes (R55C) was MODY that was misclassified. The presentation, however, was like classical type 1 diabetes, with ketoacidosis and frank diabetes. Hence, we believe that most pediatricians on a clinical basis will classify such a patient as having type 1 diabetes. Not all clinics are routinely screening children with newly developed diabetes for antibodies. Although rare, we nevertheless think it is interesting that patients with a monogenic form of diabetes can be found among those with a diagnosis of type 1 diabetes, an observation that has important implications for diagnosis, genetic counseling, and possibly treatment.

Generally, subjects with neonatal diabetes and *INS* mutations are small for gestational age (12,16). None of our five mutation-positive subjects had low birth weights (Fig. 1). The R46Q mutation of family N580 appears to be functionally mild compared with the *INS* mutations causing neonatal diabetes, as suggested by the much later age of onset, a low A1C, and less-intensive treatment needed. The effect of R46Q on fetal insulin secretion may therefore be negligible, explaining the lack of effect on birth weight. In family N781, the diabetic mother with a de novo mutation had a birth weight in the lower normal range. The relatively high birth weight of her R55C-carrying child can be explained by the mother being diabetic during pregnancy and a near-normal insulin secretion capacity in fetal life. As for R46Q, the age of onset suggests that the phenotype of R55C is milder than that of *INS* mutations causing neonatal diabetes.

Although 80% of the *INS* cases found in patients with neonatal diabetes are de novo, both the probands described here inherited the mutation from a diabetic parent. Thus, our findings as well as those of Edghill et al. (16) indicate that de novo mutations in the *INS* gene are possible when diabetes presents after the neonatal period.

In summary, our results suggest that patients with MODY and autoantibody-negative type 1 diabetes should be screened for mutations in *INS*. The presence of residual β-cell function in the subjects with apparent type 1 diabetes indicates that new approaches for treatment should be considered in such cases with *INS* mutations.

APPENDIX

Other members of the Norwegian Childhood Diabetes Study Group

The following physicians also contributed to the study: Henning Aabech and Sven Simonsen, Fredrikstad; Helge Vogt, Lørenskog; Kolbeinn Gudmundsson, Anne Grethe

Myhre, and Knut Dahl-Jørgensen, Oslo; Jon Grøtta, Elverum; Ola Tallerås and Dag Helge Frøisland, Lillehammer; Halvor Bævre, Gjøvik; Kjell Stensvold, Drammen; Bjørn Halvorsen, Tønsberg; Kristin Hodnekvam, Skien; Ole Kr. Danielsen, Arendal; Jorunn Ulriksen and Unni Mette Köpp, Kristiansand; Jon Bland, Stavanger; Dag Roness, Haugesund; Per Helge Kvistad, Førde; Steinar Spangen, Ålesund; Per Erik Hæreid, Trondheim; Sigurd Børsting, Levanger; Dag Veimo, Bodø; Harald Dramsdahl, Harstad; Bård Forsdahl, Tromsø; and Kersti Elisabeth Thodenius and Ane Kokkvoll, Hammerfest.

ACKNOWLEDGMENTS

This study was supported by the University of Bergen, Haukeland University Hospital, Helse Vest, Innovest, and the Functional Genomics Programme (FUGE) of the Research Council of Norway. Research carried out in Chicago was supported by U.S. Public Health Service Grants DK-13914, DK-20595, DK-44752, DK-73541, and DK-77489 and a gift from the Kovler Family Foundation.

REFERENCES

1. Yamagata K, Furuta H, Oda N, Kaisaki PJ, Menzel S, Cox NJ, Fajans SS, Signorini S, Stoffel M, Bell GI: Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). *Nature* 384:458–460, 1996
2. Yamagata K, Oda N, Kaisaki PJ, Menzel S, Furuta H, Vaxillaire M, Southam L, Cox RD, Lathrop GM, Boriraj VV, Chen X, Cox NJ, Oda Y, Yano H, Le Beau MM, Yamada S, Nishigori H, Takeda J, Fajans SS, Hattersley AT, Iwasaki N, Hansen T, Pedersen O, Polonsky KS, Turner R, Velho G, Chevre J-C, Froguel P, Bell GI: Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature* 384:455–458, 1996
3. Froguel P, Zouali H, Vionnet N, Velho G, Vaxillaire M, Sun F, Lesage S, Stoffel M, Takeda J, Passa P, Permutt MA, Beckmann JS, Bell GI, Cohen D: Familial hyperglycemia due to mutations in glucokinase: definition of a subtype of diabetes mellitus. *N Engl J Med* 328:697–702, 1993
4. Njølstad PR, Søvik O, Cuesta-Munoz A, Bjørkhaug L, Massa O, Barbetti F, Undlien DE, Shiota C, Magnuson MA, Molven A, Matschinsky FM, Bell GI: Neonatal diabetes mellitus due to complete glucokinase deficiency. *N Engl J Med* 344:1588–1592, 2001
5. Gloyn AL, Pearson ER, Antcliff JF, Proks P, Bruining GJ, Slingerland AS, Howard N, Srinivasan S, Silva JM, Molnes J, Edghill EL, Frayling TM, Temple IK, Mackay D, Shield JP, Sumnik Z, van Rhijn A, Wales JK, Clark P, Gorman S, Aisenberg J, Ellard S, Njølstad PR, Ashcroft FM, Hattersley AT: Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N Engl J Med* 350:1838–1849, 2004
6. Sagen JV, Ræder H, Hathout E, Shehadeh N, Gudmundsson K, Bævre H, Abuelo D, Phornphutkul C, Molnes J, Bell GI, Gloyn AL, Hattersley AT, Molven A, Søvik O, Njølstad PR: Permanent neonatal diabetes due to mutations in KCNJ11 encoding Kir6.2: patient characteristics and initial response to sulfonylurea therapy. *Diabetes* 53:2713–2718, 2004
7. Pearson ER, Flechtner I, Njølstad PR, Malecki MT, Flanagan SE, Larkin B, Ashcroft FM, Klimes I, Codner E, Iotova V, Slingerland AS, Shield J, Robert JJ, Holst JJ, Clark PM, Ellard S, Søvik O, Polak M, Hattersley AT: Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6.2 mutations. *N Engl J Med* 355:467–477, 2006
8. Pihoker C, Gilliam LK, Hampe CS, Lernmark A: Autoantibodies in diabetes. *Diabetes* 54 (Suppl. 2):S52–S61, 2005
9. Expert Committee on the Diagnosis and Classification of Diabetes Mellitus: Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* 20:1183–1197, 1997
10. American Diabetes Association: Diagnosis and classification of diabetes mellitus (Position Statement). *Diabetes Care* 27 (Suppl. 1):S5–S10, 2004
11. Edghill EL, Dix RJ, Flanagan SE, Bingley PJ, Hattersley AT, Ellard S, Gillespie KM: HLA genotyping supports a nonautoimmune etiology in patients diagnosed with diabetes under the age of 6 months. *Diabetes* 55:1895–1898, 2006
12. Støy J, Edghill EL, Flanagan SE, Ye H, Paz VP, Pluzhnikov A, Below JE, Hayes MG, Cox NJ, Lipkind GM, Lipton RB, Greeley SA, Patch AM, Ellard S, Steiner DF, Hattersley AT, Philipson LH, Bell GI: Insulin gene mutations as a cause of permanent neonatal diabetes. *Proc Natl Acad Sci U S A* 104:15040–15044, 2007
13. Fajans SS, Bell GI, Polonsky KS: Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *N Engl J Med* 345:971–980, 2001
14. Bjørkhaug L, Sagen JV, Thorsby P, Søvik O, Molven A, Njølstad PR: Hepatocyte nuclear factor-1 alpha gene mutations and diabetes in Norway. *J Clin Endocrinol Metab* 88:920–931, 2003
15. Bjørnvold M, Amundsen SS, Stene LC, Joner G, Dahl-Jørgensen K, Njølstad PR, Ek J, Ascher H, Gudjonsdottir AH, Lie BA, Skinningsrud B, Akselsen HE, Rønningen KS, Sollid LM, Undlien DE: FOXP3 polymorphisms in type 1 diabetes and coeliac disease. *J Autoimmun* 27:140–144, 2006
16. Edghill EL, Flanagan SE, Patch AM, Boustred C, Parrish A, Shields B, Shepherd MH, Hussain K, Kapoor RR, Malecki M, Macdonald MJ, Støy J, Steiner DF, Philipson LH, Bell GI, the Neonatal Diabetes International Collaborative Group, Hattersley AT, Ellard S: Insulin mutation screening in 1044 patients with diabetes: mutations in the INS gene are a common cause of neonatal diabetes but a rare cause of diabetes diagnosed in childhood or adulthood. *Diabetes* 57:1034–1042, 2007
17. Liu M, Hodish I, Rhodes CJ, Arvan P: Proinsulin maturation, misfolding, and proteotoxicity. *Proc Natl Acad Sci U S A* 104:15841–15846, 2007
18. Tager H, Given B, Baldwin D, Mako M, Markese J, Rubenstein A, Olefsky J, Kobayashi M, Kolterman O, Poucher R: A structurally abnormal insulin causing human diabetes. *Nature* 281:122–125, 1979
19. Steiner DF, Tager HS, Nanjo K, Chan SJ, Rubenstein AH: Familial syndromes of hyperproinsulinemia and hyperinsulinemia with mild diabetes. In *The Metabolic Basis of Inherited Disease.* 7th ed. Scriver CR, Beaudet AL, Sly WS, Valle D, Eds. New York, McGraw-Hill, 1995, p. 897–904

Lesson Plan #2 – NBIC Netherlands Bioinformatics Centre

Dr. Celia van Gelder

netherlands
bioinformatics
centre

bioinformatics @ school

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 travelling DNAlab about bioinformatics called Bioinformatica in de klas/ Bioinformatics@school (www.bioinformatica-in-de-klas.nl, www.bioinformaticsatschool.eu). The project has been implemented by NBIC, the Netherlands Bioinformatics Centre, and CMBI, the department of bioinformatics of Radboudumc, Nijmegen. Since the start of the project over 17000 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.
Two examples are given here in this booklet:

- A fun classroom activity: Bioinformatics Crossword (duration 15 min), this exercise relates to topics in our travelling DNAlab practical
- The Navigene: 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, where you can also find teacher materials belonging to this lessons.

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

The Bioinformatics@school team:
Judith Rotink & Hienke Sminia (onderwijs@nbic.nl)
Celia van Gelder (celia.vangelder@radboudumc.nl)

November 2014

bioinformatics @ school

Bioinformatics Crossword

Introduction

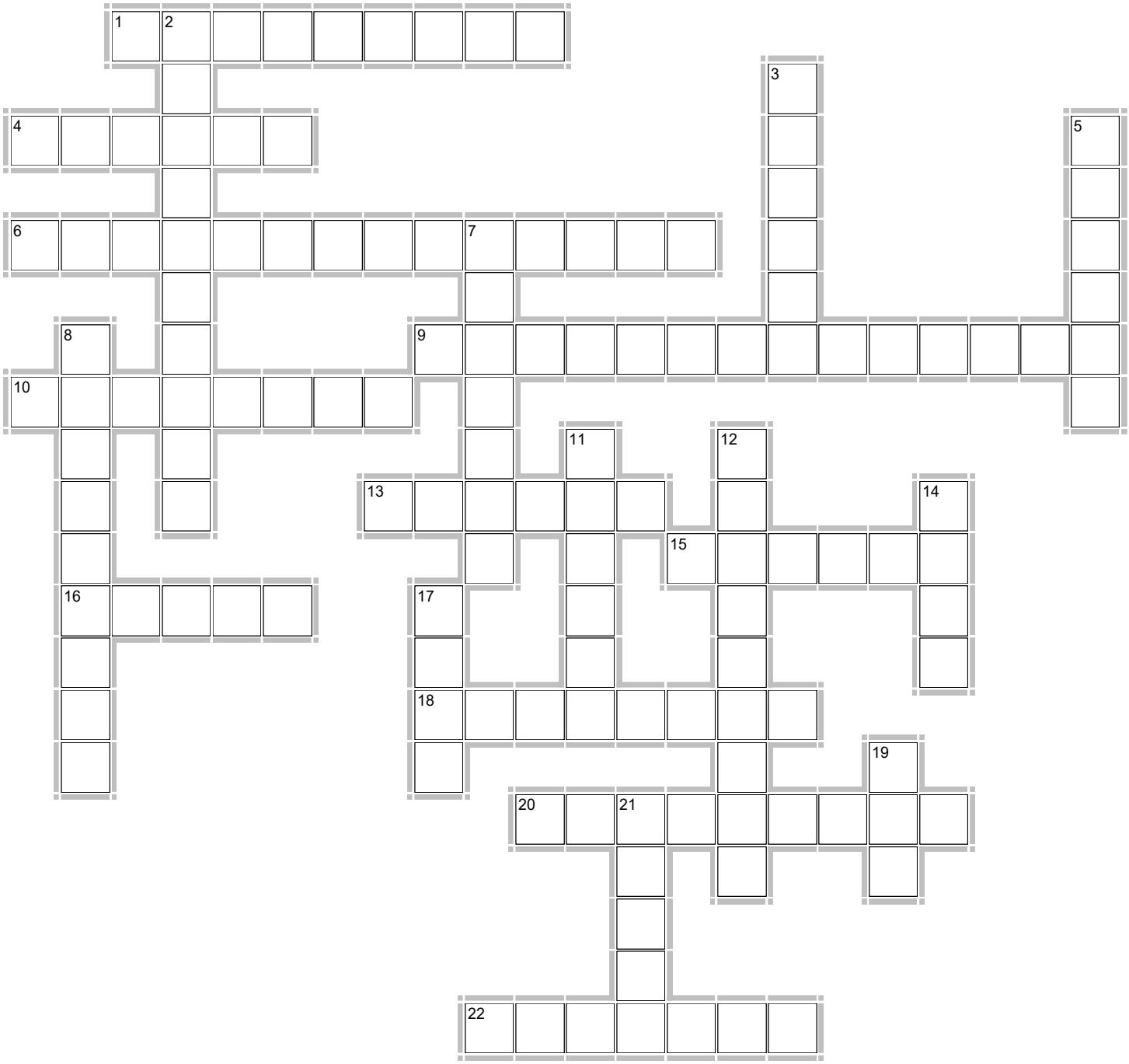
This crossword puzzle is being used as a follow up after Dutch pupils have done the *"Bioinformatics: a bit of life"* practical on their school (see www.dnalabs.eu, www.bioinformaticsatschool.eu and www.bioinformatica-in-de-klas.nl).

However, it can also be used separately in other classroom settings in case you like to do a small, fun exercise with your students about bioinformatics.

Students are requested to do the crossword individually. Afterwards they can consult with their fellow students.

This activity takes about 15 minutes.

bioinformatics @ school



EclipseCrossword.com

bioinformatics @ school

Across

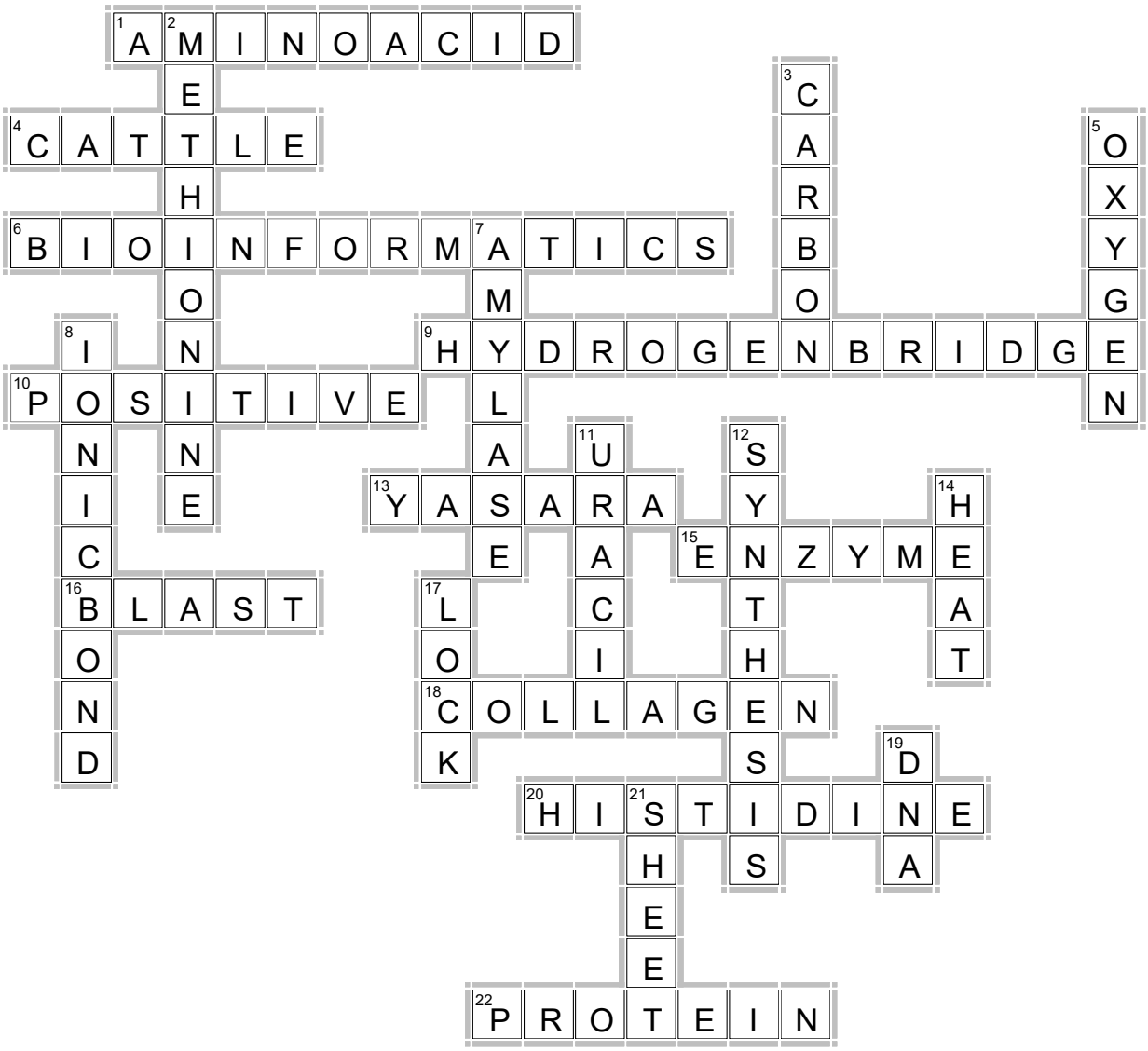
1. What does a set of three bases code for in RNA?
4. What is 'Bos taurus'?
6. What is the name of the research area that uses computers to solve biochemical problems?
9. In which inter-atomic interaction is hydrogen involved?
10. What is the charge of the zinc ion?
13. What is the name of the software that enables you to watch proteins in 3D?
15. Can the snake poison be classified as a structural protein, an enzyme or a substrate?
16. What is the name of the software tool that bioinformaticians use to search for proteins in databases?
18. What is the name of the structural protein that is cleaved by the poison of the Texas diamond-back rattlesnake?
20. Which amino acid is abbreviated with the word 'His'?
22. DNA -> RNA ->… What is the next step in this chain?

Down

2. Which amino acid is represented by the so-called start codon?
3. Which element from the periodic table is the most abundant one in proteins?
5. Which element is represented by a red sphere in the 3D-software?
7. What is the name of the enzyme that digests starch? Hint: It is also present in your mouth.
8. Which atomic interaction can one compare to the functioning of a magnet?
11. Which base can one find in RNA but not in DNA?
12. What is the biological term for the production of a protein?
14. How can one cause a protein to denature?
17. In the lock-and-key principle, is the enzyme the lock or the key?
19. What is abbreviation of deoxyribonucleic acid?
21. What is the name of the secondary structure that looks like a drapery?

bioinformatics @ school

Answer Guide



EclipseCrossword.com

bioinformatics @ school

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
http://www.bioinformaticsatschool.eu/docenten.php or at
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[1] 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.

[1] 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?

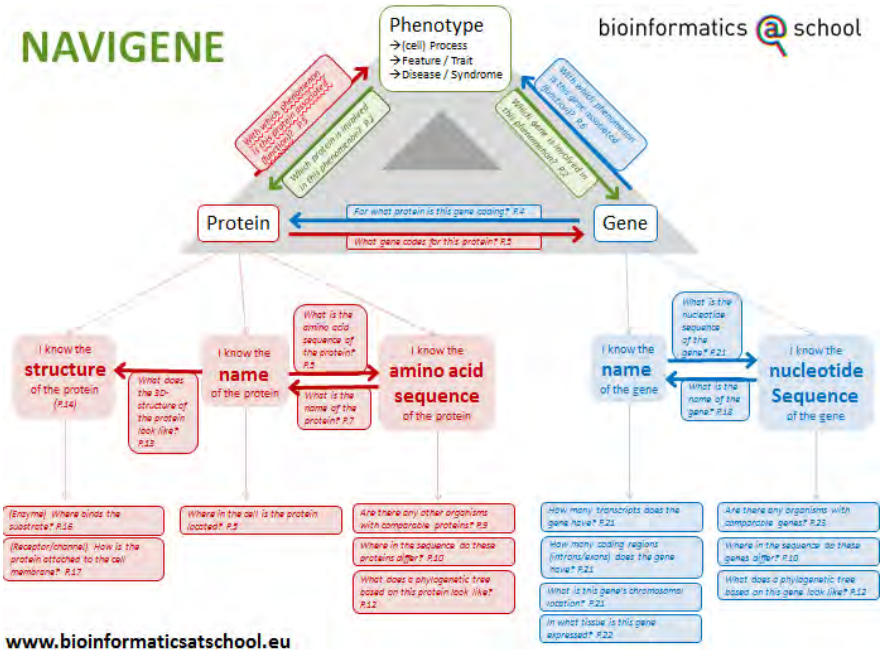
The NaviGene is initially developed for high school biology teachers. It is our experience that teachers use the Navigene each in their own way. . Here are some examples:

- “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.”

We have several examples of student assignments made by Dutch biology teachers. Please contact us via onderwijs@nbic.nl for more information.

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:
http://www.bioinformaticsatschool.eu/docenten.php or at:
www.nbic.nl/education/high-school-programmes/bioinformaticsschool/teacher-training/navigene

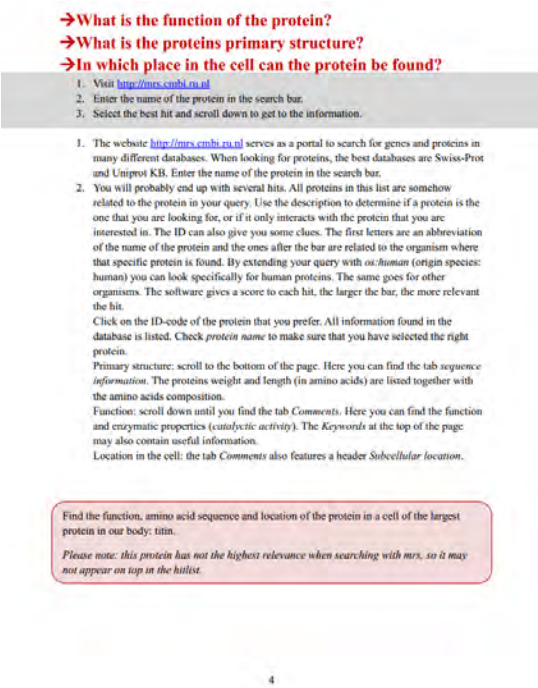


You read about the BRCA1-gene in a news article and wonder for what protein this gene codes. Or you find the protein Amylase in the chapter 'Digestion' in your biology book. These are excellent starting points for further research with help of the NaviGene.

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

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

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

Please note: this protein has not the highest relevance when searching with mrs, so it may not appear on top in the hitlist.

4

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.

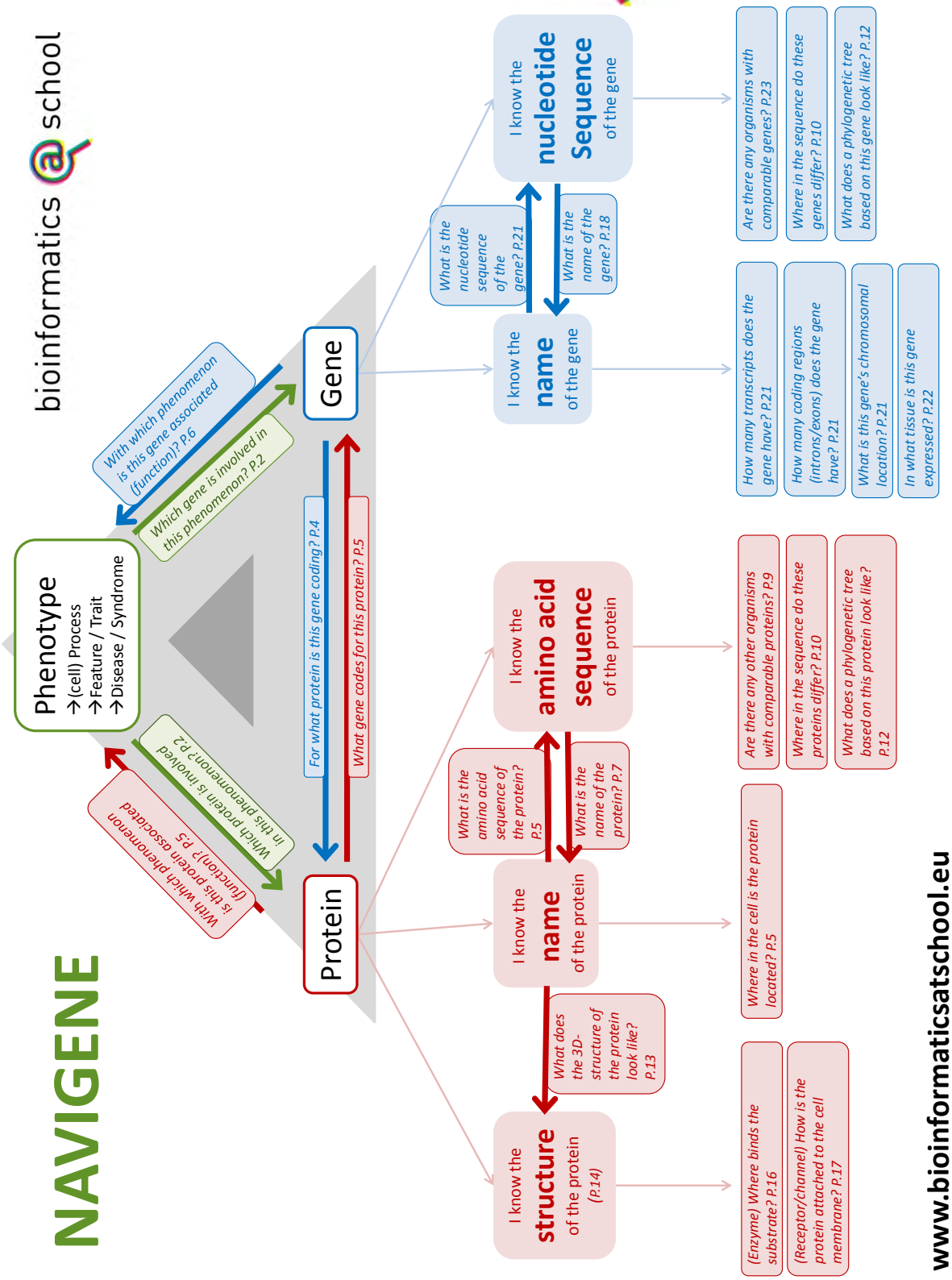
We wish you many useful discoveries and valuable surprises when using the NaviGene!

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.
- 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 Engish translation is in full progress, but is lagging behind a bit. We trust you can understand that.

The Bioinformatics@school team
Contact: onderwijs@nbic.nl
November 2014

bioinformatics @ school



www.bioinformaticsatschool.eu

Lesson Plan #3 – Bioinformatics.ca

Dr. Michelle Brazas

bioinformatics.ca

Be a Cancer Researcher for a Day

Michelle Brazas, PhD
Ontario Institute for Cancer Research

Acknowledgements

John McPherson
Geoffrey Fong
Genome Technologies Team

OICR
Ontario Institute
for Cancer Research

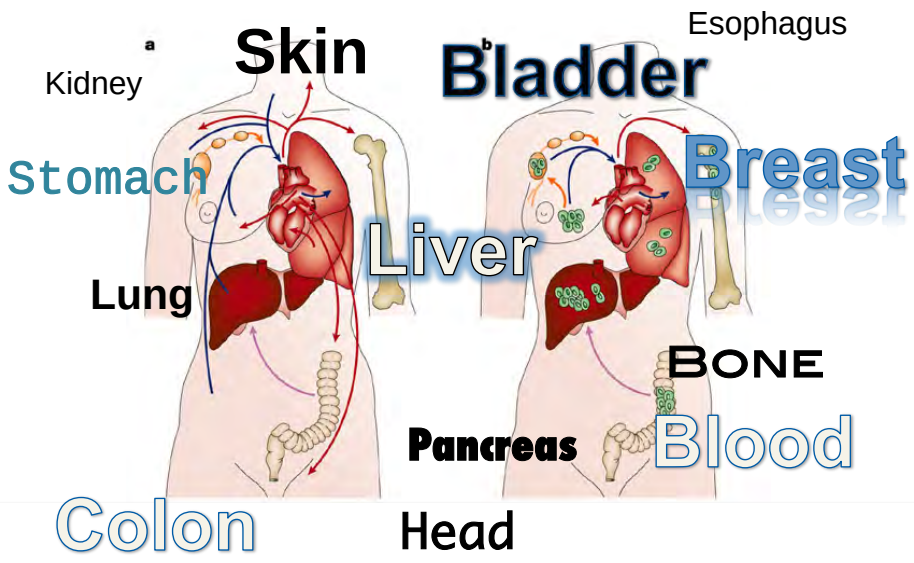
Cancer 101

What is Cancer?

Cancers are diseases of unstoppable cell growth

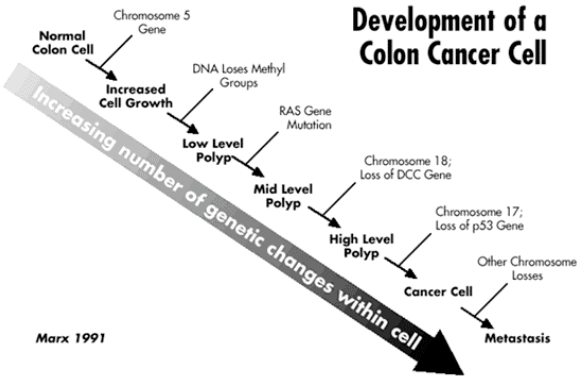
Over 200 different types of cancer!

Any normal cell can turn into a cancer cell



How does a normal cell change into a cancer cell?

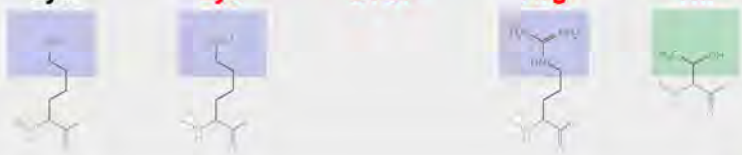
Mutations build up in your DNA



DNA mutations change a normal cell into one that grows unstoppably

What is a mutation?

	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr



basic
polar

Also insertions and deletions of bases.
Also translocations (different chromosomes joining together).

What causes mutations in DNA?

1. Random mutation events in DNA replication
2. Family Genetics
3. Life Style and Habits — You can change your life style
4. Environmental — You can change your environment

Mutations in DNA accumulate over time to cause cancer

What can you do to prevent cancer?

1. Don't smoke.
2. Maintain a healthy weight.
3. Exercise regularly.
4. Eat a healthy diet.
5. Drink alcohol only in moderation, if at all.
6. Protect yourself from the sun.
7. Protect yourself from infections.
8. Get screening tests regularly.

From Dart et al, Cancer Causes and Control, 2012

OICR Ontario Institute for Cancer Research

bioinformatics.ca

OICR Ontario Institute for Cancer Research

Doing Cancer Research (with Bioinformatics)

Be a Cancer Researcher

Research Problem: Bud has blood cancer. He has come to you for help. He wants a treatment that will get rid of his blood cancer, but won't make him more sick.

To help Bud, your **Research Question** is:

What is the **difference** between normal blood cells and Bud's cancer blood cells?

Why do we want to look at the difference?

How can you answer this research question?

What is the difference between normal blood cells and Bud's cancer blood cells?

Possible Experiments that Answer Question:

1. Compare the normal cells & cancer cells under the microscope

Making a Blood Smear on a Microscope Slide

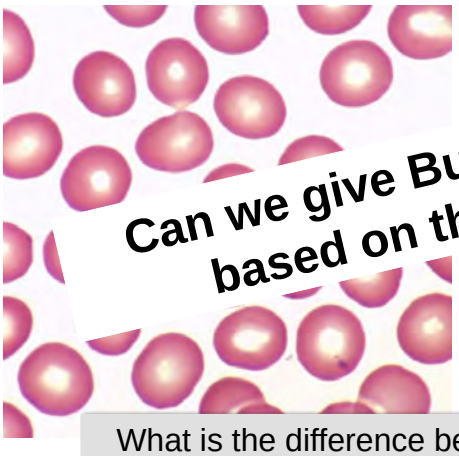
http://www.youtube.com/watch?v=O3d_4dkVVSE&feature=youtu.be

Staining a Blood Smear

http://www.youtube.com/watch?v=89VRmOJ10iA&feature=youtu.be

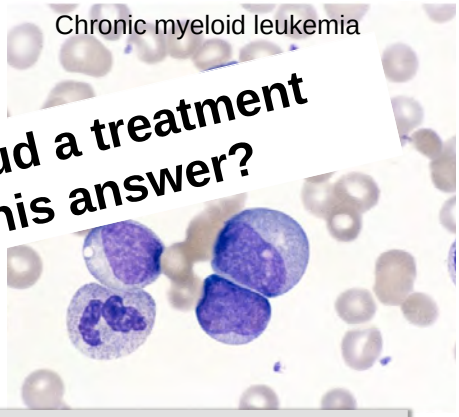
Microscope Comparison - Pathology

Normal Blood Cells



Cancer Blood Cells

Chronic myeloid leukemia



Can we give Bud a treatment based on this answer?

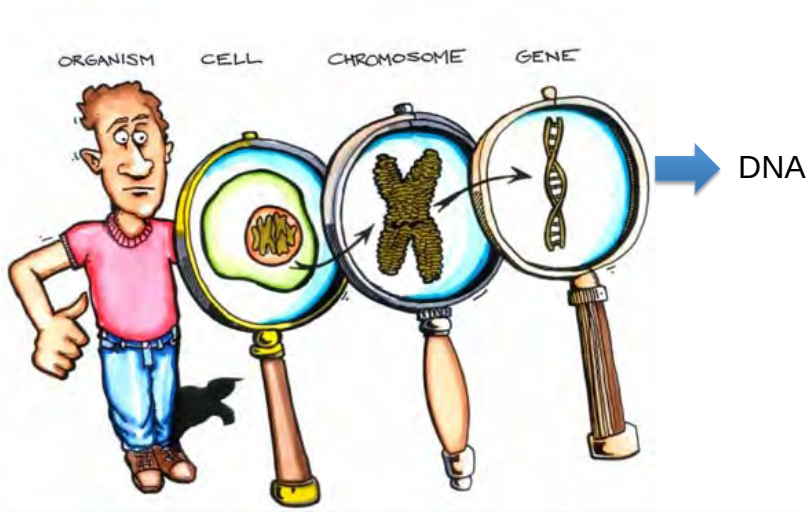
What is the difference between normal blood cells and Bud's cancer blood cells?

How can you answer this research question?

What is the difference between normal blood cells and Bud's blood cancer cells?

Possible Research Methods to Answer Question:

1. Compare the normal cells & cancer cells under the microscope
2. Compare the DNA from the normal & cancer cells



To compare the DNA between normal & cancer cells we must separate the DNA from the rest of the cell

DNA Extraction from Cheek Cells (20min)

Step 1

- Pour 3mL of TheraBreath mouthwash into your cup.

Step 2

- Swirl the mouthwash in your mouth for 30 seconds.
- Gently bite on your cheeks.
- Spit the mouthwash back into your cup.

Step 3

- Pour the mouthwash with cheek cells into the 15mL tube.
- Discard the cup into the yellow waste bag.

Step 4

- Using the bulb pipette, add 1mL of soap solution to the tube.
- Gently mix by inversion.

DNA Extraction from Cheek Cells

Step 5

- Layer cold isopropanol on top of the soap solution by slowly pouring down the side of the tube to the 15mL mark.

Step 6

- Let the solution sit for 5 minutes to separate.
- DNA will precipitate into the alcohol layer.

Step 7

- Use the stir stick to play with your DNA.

DNA dissolves in water but precipitates in alcohol.

To visualize our DNA, we need to run it on an agarose gel

http://www.youtube.com/watch?v=2UQIoYhOowM&feature=youtu.be

DNA Comparison – Molecular Biology

- Stain the agarose gel to view the DNA

Lane M – Lambda DNA/HindIII Ladder (for sizing)
Lane 1 – Normal Blood DNA
Lane 2 – Cancer Blood DNA

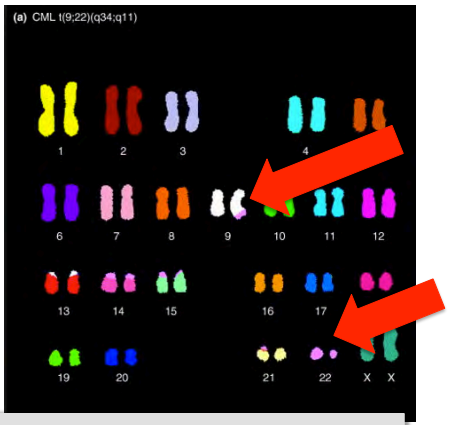
What is the difference between normal blood cells and Bud's cancer blood cells?

Can we give Bud a treatment based on this answer?

Can also visualize DNA by Spectral Karotype – Molecular Genetics

Normal Blood Cell Karotype

Cancer Blood Cell Karotype



What is the difference between normal blood cells and Bud's cancer blood cells?

How can you answer this research question?

What is the difference between normal blood cells and Bud's cancer blood cells?

Possible Research Methods to Answer Question:

1. Compare the normal cells & cancer cells under the microscope
2. Compare the DNA from the normal & cancer cells
3. Sequence the DNA bases & compare the DNA of the normal & cancer cells

OICR

bioinformatics.ca

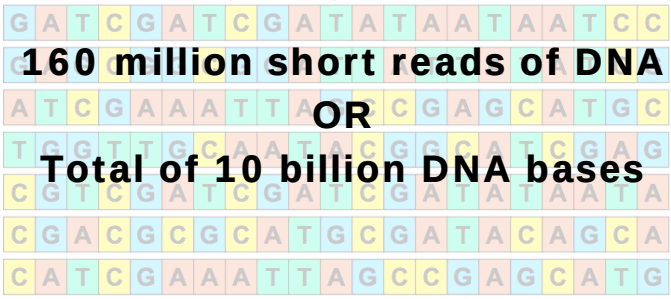
Sequencing DNA really fast

http://www.youtube.com/watch?v=HtuUFUnYB9Y&feature=youtu.be

OICR

bioinformatics.ca

Results from the sequencer…

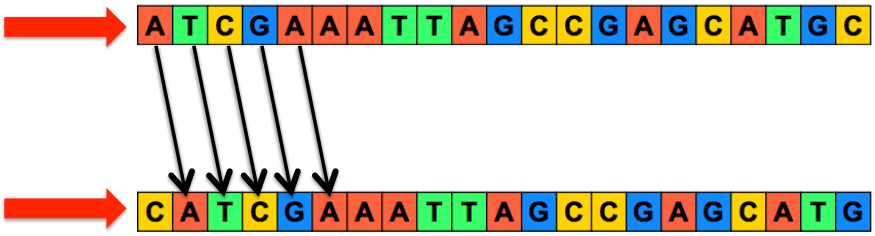


- To understand the information in the DNA sequence reads, we need to assemble them in the proper order
- Then we can compare normal DNA sequence to cancer DNA sequence

The Sequence Assembly Race (30min)

Work in your Team:

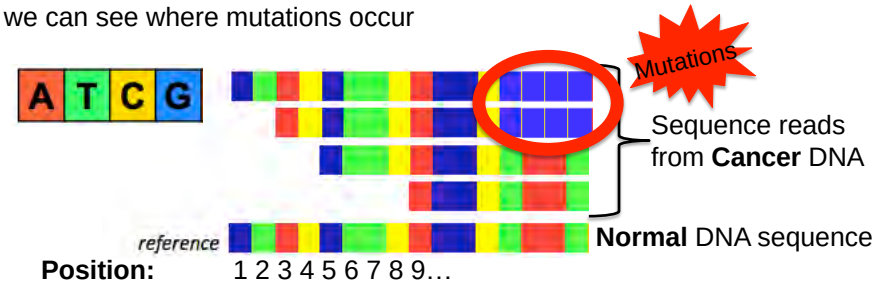
- One DNA sequence read will overlap with another DNA sequence read by a few bases.
- Assemble all of the sequence reads together.
- Tape them together to secure them.



(Computers are used to assemble 160 million sequence reads!)

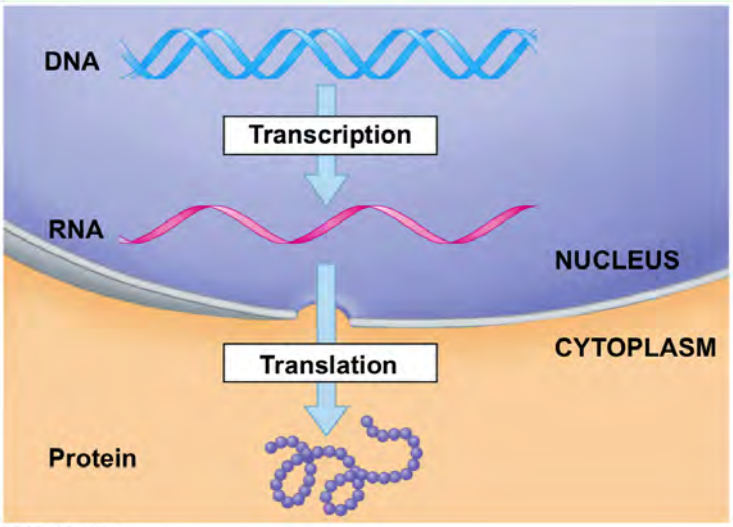
Getting Information from a Sequence Alignment

- After assembling the tumor DNA and assembling the normal DNA, the two assemblies are compared
- From a sequence alignment of cancer DNA to normal reference DNA, we can see where mutations occur



- 2 sequence reads have the same bases as the normal DNA sequence
- + 2 reads have different bases compared to the normal DNA sequence

What is the difference between normal blood cells and Bud's cancer blood cells?



DNA is the instruction manual for making proteins.
If DNA is mutated, then usually the protein is also mutated.

What is the impact of Bud's mutations?

Knowing that mutations exist is only useful information if we know what cellular function they change

1. Do Bud's mutations occur <u>outside</u> of important sections (protein coding sections) in the DNA?
   - Mutations here might not change anything in the cell
2. Do Bud's mutations occur <u>within</u> an important gene, like a gene that controls cell growth?
   - Mutations here might turn off the control, allowing the cell to keep growing without stopping

Determining Location of Mutations with BLAST

Step 1

- Go to http://goo.gl/6E7XoF

Step 2

- Select a sequence file of your choice. This sequence comes from our read assembly activity.
- Open the sequence file.
- Select (Ctrl+A) and copy (Ctrl+C) all of the sequence.

Step 3

- Go to http://blast.ncbi.nlm.nih.gov/
- Select **'Human'** under BLAST Assembled RefSeq Genomes.

Step 4

- Paste (Ctrl+V) the copied sequence into the **'Query Sequence'** box.
- Hit BLAST to start the comparison (alignment) of your DNA sequence to the whole Human genome sequence.
- BLAST will return locations in the Human genome that match (align to) your input sequence.

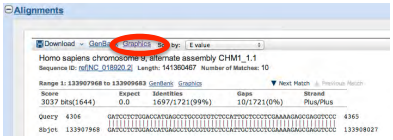
Determining Location of Mutations with BLAST

Step 5

- **Which chromosome (or chromosomes) is your sequence is located?**
- Hover over each of the red lines to determine the chromosome number.

Step 6

- In the '**Descriptions**' box, select the top result. This jumps down the page to the result.
- Download the **'Graphics'** for this result. This opens a new 'Graphics' tab.
- Note: Your sequence may match to more than one chromosome so you will need to repeat this step for each chromosome.



Step 7

- A new 'Graphics' window opened.
- In the **'Sequence'** section, which **Gene** does your sequence match with?

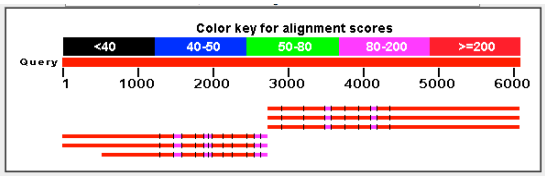
Step 8

- Hover over the gene: **What is the gene title, location, and length?**
- Select **'View MIM'** from the pop up window to learn about the function of your gene.
- Learn about the function of your gene in the '**Gene Function**' section.

Answers to BLAST

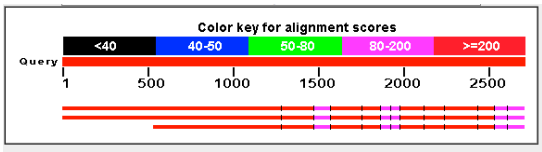
Sequence #1:

- Chromosome 22 = BCR gene
- Chromosome 9 = Abl1 gene



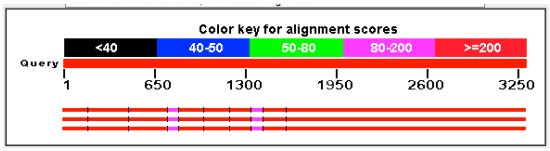
Sequence #2:

- Chromosome 22 = BCR gene



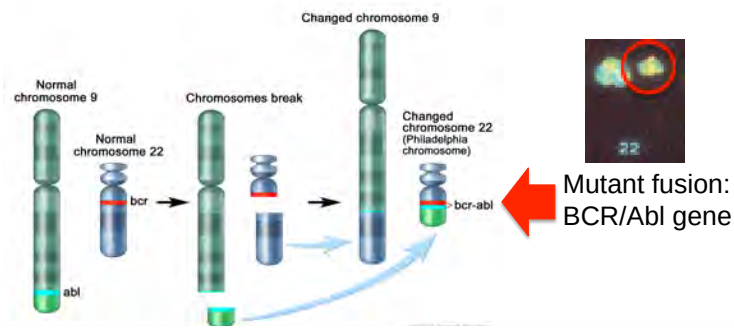
Sequence #3:

- Chromosome 9 = Abl1 gene



Impact of Mutations on Bud's Blood Cells

- Normal Abl1 protein (on chromosome 9) is a cell growth factor
- In chronic myeloid leukemia (CML), DNA mutates so that chromosome 9 + chromosome 22 exchange DNA segments
- A mutation fuses DNA to create BCR (Chr. 22) + Abl (Chr. 9) = BCR/Abl
- Mutant BCR/Abl protein never turns off
- The result is that the blood cell receives instructions to "Keep Growing!"



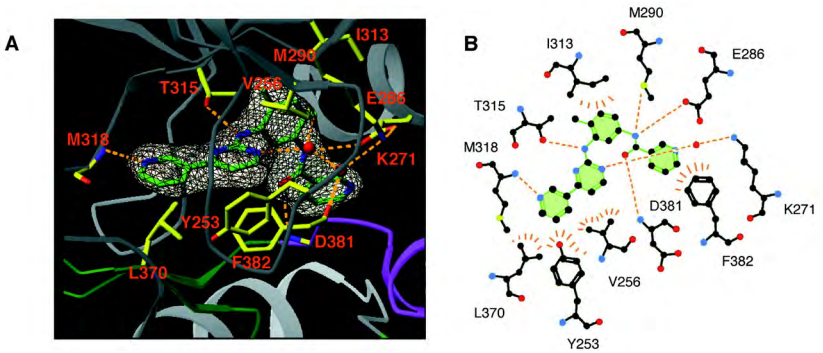
Common mutation in Chronic Myeloid Leukemia (CML)

Making a Drug to Stop the Mutant Protein

- Scientists could work with this information to model the normal Abl protein and compare it to the mutant BCR-Abl protein
- The goal is to design a drug that works against the BCR-Abl protein

- Need to look at the 3D structure of Abl protein
- Need to look at how a drug interacts with this protein in 3D

3D Model of Abl and Gleevec (STI-571)



Corbin A S et al. J. Biol. Chem. 2002;277:32214-32219

Using a 3D model of the Abl protein with Gleevec, we can determine which amino acids are important for Gleevec to work as a cancer drug

Using 3D Protein Models of Abl + Gleevec

Step 1

- Go to http://www.pdb.org/pdb/home/home.do

Step 2

- Search for **1IEP**
- Under the image, click on **3D View**

Step 3

- Select **Custom View**
  - Jmol mode = WebGL (beta)
  - Style = Backbone (or Ligands and Pocket)

Step 4

- Rotate the 3D model
- See how Gleevec fits into the pocket of the Abl protein

Using 3D Protein Models of Abl + Gleevec

Step 5

- Choose the Abl amino acids that you think are important for interaction with Gleevec (look at the figure for the chemical model of Abl and Gleevec)

Step 6

- Under the 3D image, expand the **Scripting Options** box
- Type in the **Input** box (the numbers are the amino acid numbers you chose to be important):

  select 271, 286, 290, 315, 381; spacefill; color yellow;

Step 7

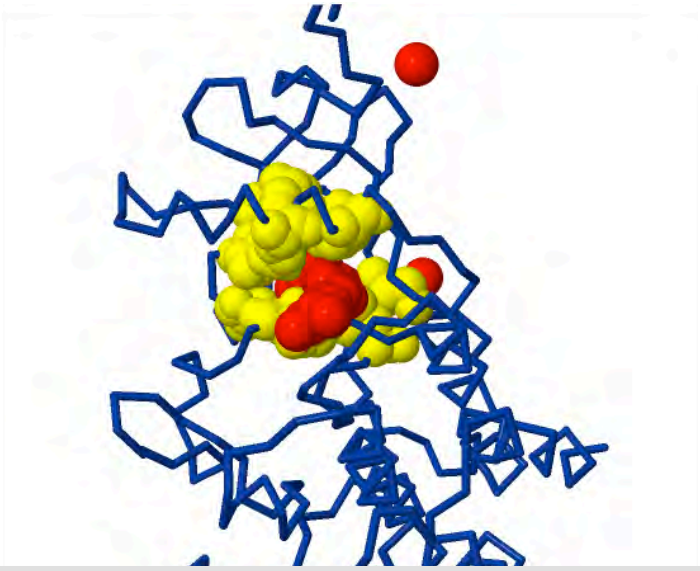
- To color Gleevec, type in the **Input** box:

  select ligand; spacefill; color red;

Step 8

- **Why is amino acid 310 not important to Gleevec function in Abl?**
- Type in the **Input** box:

  select 310; spacefill; color blue;



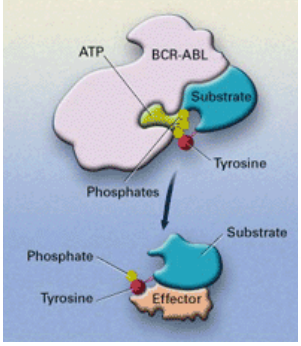
11 amino acids (yellow) within tyrosine kinase domain of Abl protein are important for interacting with Gleevec (red)

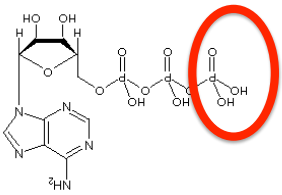
Can we give Bud a treatment based on this answer? YES

http://www.dnalc.org/view/15055-Using-DNA-science-to-control-CML-Brian-Druker.html

How Gleevec works on Mutant BCR/Abl

- Goal of cancer drug is to stop the activity of mutant BCR/Abl protein
  - Proteins have domains. The important domain in mutant BCR/Abl is tyrosine kinase.
- Tyrosine kinase domain moves phosphates (from ATP) around
  - The tyrosine kinase domain takes a phosphate from ATP and transfers it to the tyrosine amino acid on a substrate
  - The phosphated substrate then passes along the message to keep growing
- Mutant BCR/Abl is stuck on: Tyrosine kinase domain is always transferring a phosphate, so there is always a message to keep growing

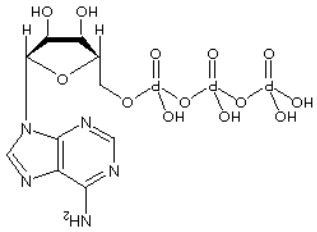




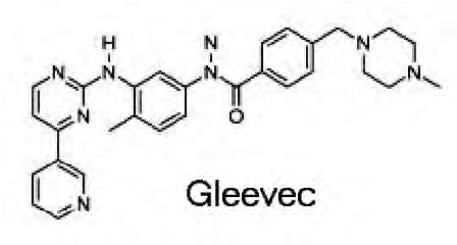
ATP – Adenosine Triphosphate

Stopping the Activity of Mutant BCR/Abl

- Need a molecule that can block ATP from entering the tyrosine kinase domain of the mutant BCR/Abl protein
  - No ATP = tyrosine kinase domain cannot transfer phosphates = No more growth message.

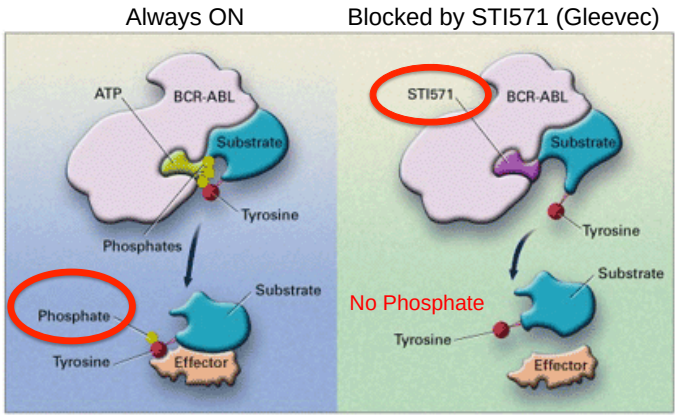


ATP – Adenosine Triphosphate



What is similar about these two molecules?

Gleevec Blocks the Mutant BCR/Abl Activity



We discovered that the DNA from Bud's blood cancer is mutated to create the BCR/Abl gene → BCR/Abl protein.

So we could give Bud this drug to block BCR/Abl activity.

Bud's Blood Counts Today

http://www.dnalc.org/view/15043-Bloodcount-returns-to-normal-with-Gleevec-Bud-and-Yvonne.html

THE END

Thank you!

DNA Extraction from Cheek Cells

Materials Needed:

- Cup
- 3mL TheraBreath mouthwash
- 1mL 16% soap mix
- 10mL cold isopropanol
- 15mL tube
- Pipette bulb
- Wooden stir stick
- Gloves
- Lab coat

(To make at home: Soap solution: ¾ teaspoon liquid soap in 2 tablespoons water)

Step 1
Pour 3mL of TheraBreath mouthwash into your cup.

Step 2
Swirl the mouthwash in your mouth for 30 seconds. Gently bite on your cheeks.

Step 3
Spit the mouthwash back into your cup.

Step 4
Pour the mouthwash with cheek cells into the 15mL tube. Discard the cup into the yellow waste bag.

Step 5
Using the bulb pipette, add 1mL of soap solution to the tube with your cheek cells.
Use the markers on the tube to guide you.
Gently mix together by inversion for 20 seconds.

Step 6
Layer cold isopropanol on top of the soap solution (up to the 15mL mark) by slowly pouring down the side of the tube.

Step 7
Let the solution sit for 5 minutes to separate.
DNA will precipitate into the alcohol layer.

Step 8
Use the stir stick to play with your DNA. Discard the stir stick and your tube into the yellow waste bag when you are finished.
DNA dissolves in water but precipitates in alcohol.

TAATCCGCCGCTCGATGCCGT

TCGATCGATCGATATAATAAT

CTAGCATCGAAATTAGCCGAG

TCGAGCGGCGGATTATTATAT

CGATCGATATAATAATCCGCC

TAGCCGAGCATGCTGTATCGC

CGGCATCGAGCGGCGGATTAT

CGACGCGCATGCGATACAGCA

ATATCGATCGATCGACGCGCA

CATCGAAATTAGCCGAGCATG

AAATTAGCCGAGCATGCTGTA

GCATCGAAATTAGCCGAGCAT

ATCGAAATTAGCCGAGCATGC

CGACGCGCATGCGATACAGCA

ACGGCATCGAGCGGCGGATTA

TCGGCTAATTTCGATGCTAGC

ATCGCATGCGCGTCGATCGAT

ATCGAGCGGCGGATTATTATA

TCGATCGATCGATATAATAAT

CGAAATTAGCCGAGCATGCTG

TCGAAATTAGCCGAGCATGCT

TGGTTGCAATACGGCATCGAG

CGATCGACGCGCATGCGATAC

TAATTTCGATGCTAGCTAGCT

GATTATTATATCGATCGATCG

CGAGCGGCGGATTATTATATC

TCGACGCGCATGCGATACAGC

TAGCTAGCTAGCATCGAAATT

CTAGCATCGAAATTAGCCGAG

CGTCGATCGATCGATATAATA

GATCGATCGATATAATAATCC

ACGGCATCGAGCGGCGGATTA

CATGCTCGGCTAATTTCGATG

GAGCGGCGGATTATTATATCG

GCTAGCATCGAAATTAGCCGA

ATTATATCGATCGATCGACGC

ATTAGCCGAGCATGCTGTATC

TAATTTCGATGCTAGCTAGCT

Determining Location of Mutations with BLAST

Step 1

- Go to http://goo.gl/6E7XoF

Step 2

- Select the sequence file of your choice. This sequence comes from our read alignment activity.
- Open the sequence file.
- Select and copy all of the sequence.

Step 3

- Go to http://blast.ncbi.nlm.nih.gov/
- Select **'Human'** under BLAST Assembled RefSeq Genomes.

Step 4

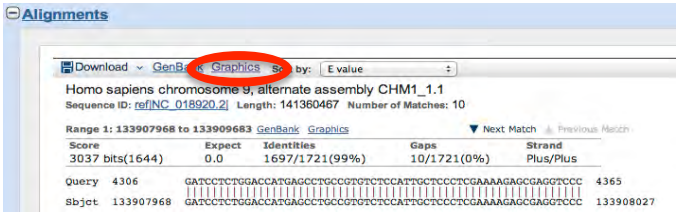
- Paste the copied sequence into the **'Query Sequence'** box
- Hit BLAST to start the comparison (alignment) of your DNA sequence to the whole Human genome sequence.
- BLAST will return locations in the Human genome that match (align to) your input sequence.

Step 5

- Note which chromosome (or chromosomes) your sequence is located.
- Hover over each of the red lines to determine the chromosome number.

Step 6

- In the '**Descriptions**' box, select the top result**.** This jumps down the page to the result.
- Download the '**Graphics'** for this result. This opens a new 'Graphics' tab.
- (Note: Your sequence may match to more than one chromosome so you may need to do this step for each chromosome.)



Step 7

- A new 'Graphics' window opened.
- In the '**Sequence'** section, which **Gene** does your sequence match with?

Step 8

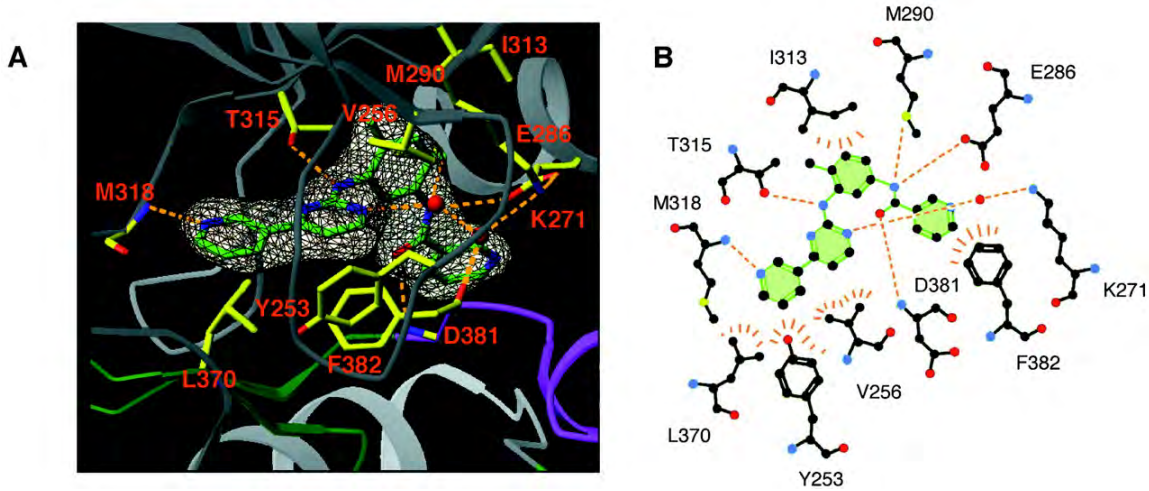
-
- Hover over the gene to learn its title, location, length, etc.
- Select **'View MIM'** from the pop up window to learn about the function of your gene.
- Learn about the function of your gene in the '**Gene Function**' section.

3D Structures of BCR/Abl + Gleevec
Content modified from DNAi.org

Instructions:

1. Identify the fit of Gleevec in Abl.
   - Go to Protein Data Bank (PDB) (http://www.pdb.org/pdb/home/home.do)
   - Search for 1IEP
   - Under the image, click '3D View'
   - Select 'Custom View' → Jmol mode = WebGL (beta)
   - Under 'Custom View':
     - Style = Backbone or Ligands and Pocket
   - Play around with the rotation of the molecule to see how Gleevec (STI-571) fits into a pocket within Abl

2. Which amino acids in Abl are essential for the interaction with Gleevec (STI-571)?
   - From the picture of Gleevec interacting with Abl, which amino acids do you think are important for this interaction?



   - Let's select these amino acids and color them on the 3D model.

   - To select amino acids, type the following commands in the 'Scripting Options' → 'Input' box located under the 3D image.
   - Be sure not to type mistakes. Spacing and commas and colons are very important:
     - select 271, 286, 290, 315, 381; spacefill; color yellow; (You can change the numbers to select for different amino acids. You can also change the color to another color)
     - select ligand; spacefill; color red;

   - Try selecting another amino acid like 310 and color it blue. Why is this amino acid not important for the interaction with Gleevec?

Lesson Plan #4 – International Society for Computational Biology

Dr. Fran Lewitter

High School Program - Bioinformatics Laboratory
(based on a lab developed at Whitehead Institute for Biomedical Research)
An Activity to Learn about the Spellchecker Gene and its Evolutionary History

Welcome to an introduction to Bioinformatics! Here we'll do some searches to get a feel for the kinds of biological information available on the web.

In today's lab, you will explore information about one type of human colon cancer - hereditary non-polyposis colon cancer (HNPCC) and the mismatch repair gene. This is one of the *"spellchecker*" genes for DNA replication. You will learn its relevance to yeast and bacteria, and see how tools available on the web can help keep researchers and the public informed.

To begin, we'll take a look at the Wikipedia entry for the gene to get a basic introduction. Next you will search the Online Mendelian Inheritance in Man (OMIM) database. This database is a catalog of human genes and genetic disorders and was developed and is maintained by scientists at Johns Hopkins University and elsewhere. You will then follow some links to explore other relevant information available to you. Finally, you will see how similar the gene responsible for HNPCC is in a variety of organisms.

What the formatting means:

- *Italics* indicates input you type into a web form
- **BOLD** indicates output from a web page

1. Type *msh2* into a browser search window. The first hit should be the entry in Wikipedia. Click on the link and read about this gene.

2. Next let's search OMIM (http://www.ncbi.nlm.nih.gov/omim). Here you will enter the words *mismatch repair* in the Search text box at the top of the page. Then click the **Search** Button.

3. You should see a page of results with many links. Click on the link (should be about halfway down the page) ***609309 MutS,E.coli,HOMOLOG of,2;MSH2**

4. First, read the paragraph under the subheading **"Description"** to get a quick summary. As stated, this gene, MSH2 is homologous (similar) to the E. coli MutS gene and is involved in DNA mismatch repair.

5. Scroll down the page and read additional information. One section of particular interest is **ALLELIC VARIANTS**. Take a look at the some of the more than 20 different mutations in this gene that cause this hereditary form of colon cancer. [Note: For example, people who have Allelic Variant .0001 in their MSH2 gene have a change in codon 622. Usually at this position in the protein, the amino

acid PROLINE is found. However, in this family with Colorectal Cancer, there is LEUCINE at position 632. A single DNA base change (C to T) causes this change in amino acids and causes the protein to behave differently.] When you are through, click on **Title** in the **Table of Contents** on the right hand side of the page to get back to the top of the article.

6. Notice the various **Links** on the right side of the page. Although you can click on any of these links, select **DNA** under **External Links for Entry**. Then click on **NCBI RefSeq** (RefSeq is a database of genetic sequences and has links to many resources.) Click on the second entry, the one with the Accession identifier **NM_000251.2**. Scroll down to the bottom of the document to see the DNA sequence of the MSH2 gene. Note that along the way, you will also see the protein translation of the DNA sequence represented in single letter amino acid codes. [Note: To see the single letter amino acid codes, visit http://tinyurl.com/3x2x4q]

7. Now let's look at the alignment for human, mouse and rat copies of the gene. (See page 4). The amino acids (identified by their single letter code) that are colored are identical in human and at least one other species. Notice how similar the sequences are from these related organisms. Now take a look at the alignment of more distantly related species known to have this gene (See Page 5.) Notice that now the only amino acids colored are those that are common to at least 5 organisms. Scroll through the alignment to find an area of the gene that is very similar in all organisms. This, so called "conserved" region may be important in the three-dimensional structure of the protein.

8. There's one more step to do before completing this lab. You will use a tool that scientists use on a daily basis - **BLAST.** This is a database search tool that takes as input a DNA or protein sequence and searches against a database of known sequences. The results are shown as alignments. The number of matches between the sequence you searched with and the hit in the database is used to predict has similar the two genes are.

9. Below is the human MSH2 gene. You can use this protein sequence to search against any of the available protein databases to see what sequences are similar to the human sequence. The human sequence is listed below:

MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYTAHGEDALLAAREVFKT
QGVIKYMGPAGAKNLQSVVLSKMNFESFVKDLLLVRQYRVEVYKNRAGNKASKENDWYLA
YKASPGNLSQFEDILFGNNDMSASIGVVGVKMSAVDGQRQVGVGYVDSIQRKLGLCEFPD
NDQFSNLEALLIQIGPKECVLPGGETAGDMGKLRQIIQRGGILITERKKADFSTKDIYQD
LNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIKFLELLSDDSNFGQFELTTFDFSQYM
KLDIAAVRALNLFQGSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERL
NLVEAFVEDAELRQTLQEDLLRRFPDLNRLAKKFQRQAANLQDCYRLYQGINQLPNVIQA
LEKHEGKHQKLLLAVFVTPLTDLRSDFSKFQEMIETTLDMDQVENHEFLVKPSFDPNLSE
LREIMNDLEKKMQSTLISAARDLGLDPGKQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFS
TVDIQKNGVKFTNSKLTSLNEEYTKNKTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLA

QLDAVVSFAHVSNGAPVPYVRPAILEKGQGRIILKASRHACVEVQDEIAFIPNDVYFEKD
KQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESAEVSIVDCILARVGAGDSQLK
GVSTFMAEMLETASILRSATKDSLIIIDELGRGTSTYDGFGLAWAISEYIATKIGAFCMF
ATHFHELTALANQIPTVNNLHVTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPKHV
IECAKQKALELEEFQYIGESQGYDIMEPAAKKCYLEREQGEKIIQEFLSKVKQMPFTEMS
EENITIKLKQLKAEVIAKNNSFVNEIISRIKVTT

10. BLAST searching can be slow if you search large databases, such as all organisms. Let's limit our search first to birds and then to budding yeast. These will tell us if these species have a gene similar to the human MSH2 gene.
    To use BLAST, go to http://blast.ncbi.nlm.nih.gov. Select **protein blast** under the **Basic** BLAST heading. Copy the sequence above and paste it into the large text box on the BLAST page. Alternatively, you can use the identifier for the protein sequence, *NP_000242*, and paste it into the large text box.

11. By default you'll search proteins from all organisms using the Non-redundant protein sequence (nr) database. This search can take a long time so let's limit the first search to birds. In the **Organism** text box, enter *birds (taxid:8782).* Click the box, **Show results in a new window**, and then click the **BLAST** button to start your search. You'll quickly see a "domain analysis" of your protein, but you may have to wait a couple of minutes to see the BLAST results. Once the results come back take a look at the alignments (you will have to scroll down past the graphic and the list of hits) and notice what parts of the sequence are conserved.

12. You can do another BLAST search limiting your search to just insects by repeating the steps above but entering *budding yeasts (taxid:4892)* in the **Organism** text box.

13. Look at the summary line for each BLAST search (just above the alignments) to see how many amino acids are "Identities" and how many are "Positives" in the alignment. Based on the BLAST searches, are fish or yeast more similar to humans? Why?

A Brief Summary of what we have done investigating the genetics of a disease:

1. We searched the OMIM database for the disease gene known as "mismatch repair" and examined some results.
2. We looked at the complete sequence of the mismatch repair gene.
3. We looked at mutations in the mismatch repair gene.
4. We looked at alignments of sequences in a variety of organisms.
5. We did database searches to find sequences in other organisms that are similar to the sequence in humans.

Alignment of the Protein sequences of MSH2 from Human, Mouse and Rat

```
HUMAN    1   MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYTAHGEDA
MOUSE    1   MAVQPKETLQLEGAAEAGFVRFFEGMPEKPSTTVRLFDRGDFYTAHGEDA
RAT      1   MAVQPKETLQLEGAAEVGFVRFFEGMPEKPSTTVGLFDRGDFYTAHGEDA

HUMAN   51   LLAAREVFKTQGVIKYMGPAGAKNLQSVVLSKMNFESFVKDLLLVRQYRV
MOUSE   51   LLAAREVFKTQGVIKYMGPAGSKTLQSVVLSKMNFESFVKDLLLVRQYRV
RAT     51   LLAAREVFKTQGVIKYMGPAGAKTLQTVVLSKMNFESFVKDLLLVRHYRV

HUMAN  101   EVYKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNNDMSASIGVVGV
MOUSE  101   EVYKNKAGNKASKENEWYLAFKASPGNLSQFEDILFGNNDMSASVGVMGI
RAT    101   EVYKNKAGNKASKENDWYLAYKASPGNLSQFEDILFGNNDMATSIGIMGI

HUMAN  151   KMSAVDGQRQVGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQIGPKECV
MOUSE  151   KMAYVDGQRHVGVGYVDSTQRKLGLCEFPENDQFSNLEALLIQIGPKECV
RAT    151   KLSTVDGQRQVGVGDVDSTQRKLGLCEFPDNDQFSNLEALLIQIGPKECI

HUMAN  201   LPGGETAGDMGKLRQIIQRGGILITERKKADFSTKDIYQDLNRLLKGKKG
MOUSE  201   LPGGETTGDMGKLRQVIQRGGILITERKRADFSTKDIYQDLNRLLKGKKG
RAT    201   LPGGETAGDMGKLRQVIQRGGILITERKRIDFSTKDIYQDLNRLLKGRKG

HUMAN  251   EQMNSAVLPEMENQVAVSSLSAVIKFLELLSDDSNFGQFELTTFDFSQYM
MOUSE  251   EQINSAALPEMENQVAVSSLSAVIKFLELLSDDSNFGQFELATFDFSQYM
RAT    251   EQMNSAVLPEMENQVAVSSLSAVIKFLELLSDDSNFGQFELATFDFSQYM

HUMAN  301   KLDIAAYRALNLFQGSVEDTTGSQSLAALLNKCKTPQGQRLYNQWIKQPL
MOUSE  301   KLDMAAYRALNLFQGSVEDTTGSQSLAALLNKCKTAQGQRLYNQWIKQPL
RAT    301   KLDMAAYRALNLFQGSVEDTTGSQSLAAFLNKCKTAQGQRLYSQWIKQPL

HUMAN  351   MDKNRIEERLNLVEAFVEDAELRQTLQEDLLRRFPDLNRLAKKFQRQAAN
MOUSE  351   MDRNRIEERLNLVEAFVEDSELRQSLQEDLLRRFPDLNRLAKKFQRQAAN
RAT    351   MDKNRIEERLNLVEAFVEDSELRRALQEDLLRRFPDLNRLAKKFQRQAAN

HUMAN  401   LQDCYRLYQGINQLPNVIQALEKHEGKHQKLLLAVFVTPLTDLRSDFSKF
MOUSE  401   LQDCYRLYQGINQLPSVIQALEKYEGRHQALLLAVFVTPLIDLRSDFSKF
RAT    401   LQDCYRLYQGVKQLPNVIQALEKYQGRHQALLLAVFVTPLTDLRSDFSKF

HUMAN  451   QEMIETTLDMDQVENHEFLVKPSFDPNLSELREIMNDLEKKMQSTLISAA
MOUSE  451   QEMIETTLDMDQVENHEFLVKPSFDPNLSELREYMDGLEKKMQSTLINAA
RAT    451   QEKIETTLDMDQVENHEFLVKPSFDPNLSELREYMDGLEKKMQSTLISAA

HUMAN  501   RDLGLDPGKQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVK
MOUSE  501   RGLGLDPGKQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVK
RAT    501   RGLGLDPGKQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVK

HUMAN  551   FTNSKLTSLNEEYTKNKTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLA
MOUSE  551   FTNSELSSLNEEYTKNKGEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLA
RAT    551   FTNSELSSLNEEYTKNKGEYEEAQDAIVKEIVNISSGYVEPMQTVNDVLA

HUMAN  601   QLDAVVSFAHVSNGAPVPYVRPAILEKGQGRIILKASRHACVEVQDEIAF
MOUSE  601   HLDAIVSFAHVSNAAPVPYVRPVILEKGKGRIILKASRHACVEVQDEVAF
RAT    601   HLDAVVSFAHVSNAAPVPYVRPVILEKGKGRIIVKASRHACVEVQHDVAF

HUMAN  651   IPNDVYFEKDKQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESA
MOUSE  651   IPNDVHFEKDKQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESA
RAT    651   IPNDVHFEKDKQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESA

HUMAN  701   EVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSATKDSLIIIDEL
MOUSE  701   EVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSATKDSLIIIDEL
RAT    701   EVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSATKDSLIIIDEL

HUMAN  751   GRGTSTYDGFGLAWAISEYIATKIGAFCMFATHFHELTALANQIPTVNNL
MOUSE  751   GRGTSTYDGFGLAWAISDYIATKIGAFCMFATHFHELTALANQIPTVNNL
RAT    751   GRGTSTYDGFGLAWAISEYIATNIGAFCMFATHFHELTALASQIPTVNNL

HUMAN  801   HVTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPKHVIECAKQKALE
MOUSE  801   HVTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPRHVIACAKQKALE
RAT    801   HVTALTTEETLTMLYQVKTGVCDQSFGIHVAELANFPRHVIECAKQKALE

HUMAN  851   LEEFQYIGESQGYDIMEPAAKKCYLEREQGEKIIQEFLSKVKQMPFTEMS
MOUSE  851   LEEFQNIGTSLGCDEAEPAAKRRCLEREQGEKIILEFLSKVKQVPFTAMS
RAT    851   LEEFQSIGTSQGHDETQPAAKRRCLEREQGEKIILEFLSKVKQVPFTDLS

HUMAN  901   EENITIKLKQLKAEYIAKNNSFVNEIISRIKVTT-
MOUSE  901   EESISAKLKQLKAEVYAKNNSFVNEIISRIKAPAP
RAT    901   EESVSVKLKQLKAEYLAKNNSFVNEIISRVKAP--
```

Alignment of the partial Protein sequences of MSH2 from 6 organisms

```
HUMAN     1   -MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTYRLFDRGDFYTAHG-EDALLAAREVF
MOUSE     1   -MAVQPKETLQLEGAAEAGFVRFFEGMPEKPSTTYRLFDRGDFYTAHG-EDALLAAREVF
RAT       1   -MAVQPKETLQLEGAAEVGFVRFFEGMPEKPSTTYGLFDRGDFYTAHG-EDALLAAREVF
FLIES     1   SNRQQAGTYPEYGHKCASNFIKFHAKLGEKPATTYRFFDHTDRYTVHGSDDCELYAKIVY
WORM      1   -MSSTRPELKFSDVSEERNFYKKYTGLPKKPLKTIRLYDKGDYYTYIG-SDAIFYADSYY
YEAST     1   ----------MSGGKDEASDKALLKILKSKSPNTIAIFSRGEYFSYYG-DDATFYATNIF
BACTERIA  1   -------MSAIENFDAHTPMMQQYLRLKAQHPEILLFYRMGDFYELFY-DDAKRASQLLD

HUMAN    59   KTQGYIK--YMGPAGAK----NLQSVYLSKMNFESFYKDLLLYRQYRYEYYKNRAGNKAS
MOUSE    59   KTQGYIK--YMGPAGSK----TLQSVYLSKMNFESFYKDLLLYRQYRYEYYKNKAGNKAS
RAT      59   KTQGYIK--YMGPAGAK----TLQTVYLSKMNFESFYKDLLLYRHYRYEYYKNKAGNKAS
FLIES    61   KSTAFIG--ALLPDDKK---ETLQFYSMSKGNFELAYRELLLYRNYRYEYYYKNS-----
WORM     59   HTQSYLKNCQLDPYTAKNFHEPTKYYTYSLQYLATLLKLCLLDLGYKYEIYDKG------
YEAST    50   KSDVCYKTFTLSTDNSQ----QMKYISYNRGQYEKYYRETIYLLRCSYELYSSEQ-----
BACTERIA 53   ISLTKRGASAGEPIPMAG--IPYHAYENYLAKLYNQGESYAICEQIGDPATSKGP-----

HUMAN   113   KENDWYLAYKASPGNLSQFEDILFGNNDMSASIGYYGYKMSAYDG-QRQYGYGYYDSIQR
MOUSE   113   KENEWYLAFKASPGNLSQFEDILFGNNDMSASYGYMGIKMAYYDG-QRHYGYGYYDSTQR
RAT     113   KENDWYLAYKASPGNLSQFEDILFGNNDMATSIGIMGIKLSTYDG-QRQYGYGDYDSTQR
FLIES   111   --SDWEIEYRGSPGNLLQFEDILFSNKEYLYGNSIISLLYKLDGGGQRRYGYASYEQNDC
WORM    113   ----WKLIKSASPGNIEQYNELMNMNIDSSIIIASLKYQWNSQDG-NCIIGYAFIDTTAY
YEAST   101   --GEWKMTKRGSPGNTYDFEQEIG----YSDQAPILAIYIHPGDD-DNRYTLCAWDAGNY
BACTERIA 106  --YERKYYRIYTPGTISDEALLQE------RQDNLLAAIWQDSKG----FGYATLDISSG

HUMAN   172   KLGLCEFPDNDQFSNLEALLIQIGPKECYLPGG----ETAGDMGKLRQIIQRGGILITER
MOUSE   172   KLGLCEFPENDQFSNLEALLIQIGPKECYLPGG----ETTGDMGKLRQYIQRGGILITER
RAT     172   KLGLCEFPDNDQFSNLEALLIQIGPKECILPGG----ETAGDMGKLRQYIQRGGILITER
FLIES   169   KFQLLEFLDDDFFTELEATYYLLGPKECLLP------SIEGEYSAYKTLLDRNGYMITMP
WORM    168   KYGMLDIYDNEYYSNLESFLIQLGYKECLYQDLTSNSNSNAEMQKYINYIDRCGCYYTLL
YEAST   154   RIYISEYIDTPSFSQTEQCIFGLCPTEYILYNE---GSYAPKAKKIASMFTRMEYHNKQQ
BACTERIA 154  RFRLSEPADR--ET-MAAELQRTNPAELLYA---------EDFAEMSLIEGRRGLRRRPL

HUMAN   228   KKADFSTKDIYQDLNRLLKGKKGEQMNSAYLPEMENQYAYSSLSAYIKFLELLSDDSNFG
MOUSE   228   KRADFSTKDIYQDLNRLLKGKKGEQINSAALPEMENQYAYSSLSAYIKFLELLSDDSNFG
RAT     228   KRIDFSTKDIYQDLNRLLKGRKGEQMNSAYLPEMENQYAYSSLSAYIKFLELLSDDSNFG
FLIES   223   KKS--GDNDLLQDLNRLLRFAKGQQEDATGLKELQLQLASNALKTAIKYLDLYNDAGNLG
WORM    228   KNSEFSEKDYELDLTKLLG-----DDLALSLPQKYSKLSMGACNALIGYLQLLSEQDQYG
YEAST   211   LKPKSQWSDYIESYHLDYK------------D-EAEKQNENIKECLQILHSNAADEYSISE
BACTERIA 202  WEFEIDTARQQLNLQFGTR-------DLYGFGYENAPRGLCAAGCLLQYAKDTQRTTLPH

HUMAN   288   QFELTTFDFSQYMKLDIAAYRALNLFQGSYEDTTGS------------------QSLAALL
MOUSE   288   QFELATFDFSQYMKLDMAAYRALNLFQGSYEDTTGS------------------QSLAALL
RAT     288   QFELATFDFSQYMKLDMAAYRALNLFQGSYEDTTGS------------------QSLAAFL
FLIES   281   HYEIKQLDLNRFYHLDSAAYAALNIMPKPGTHPSMPS--------------YRWQSYLGYL
WORM    283   KYELYEHKLKEFMKLDASAIKALNLFPQGPQNPFGSNNLAYSGFTSAGNSGKYTSLFQLL
YEAST   259   KYSIFNYGTHGNMLIDSCAYEALELFQLNYNYLEKSNN-----------------LTLYNYL
BACTERIA 255  IRSITMEREQDSIIMDAATRRNLEITQNLAGGAENT-------------------LASYL

HUMAN   331   NKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLYEAFYEDAELRQTLQEDLLRRFPDLNRL
MOUSE   331   NKCKTAQGQRLVNQWIKQPLMDRNRIEERLNLYEAFYEDSELRQSLQEDLLRRFPDLNRL
RAT     331   NKCKTAQGQRLVSQWIKQPLMDKNRIEERLNLYEAFYEDSELRRALQEDLLRRFPDLNRL
FLIES   328   DHCRTPQGHRLMGQWYKQPLRSRNILNDRHNIYQCLLESPDTMETLSLDYLKRIPDILML
WORM    343   NHCKTNAGYRLLNEWLKQPLTNIDEINKRHDLYDYLIDQIELRQMLTSEYLPMIPDIRRL
YEAST   304   NKCKTLPGEKLLRDWLSRPLCQIDHINERLDIYEALFENQTIRQKLRDSILARMPDCSQL
BACTERIA 296  DCTYTPMGSRMLKRWLHMPYRDTRYLLERQQTIGALQDFTAG----LQPYLRQYGDLERI

HUMAN   391   AKKFQRQAANLQDCYRLYQGINQLPNYIQALEKHEGKH------QKLLLAYFYTPLTDLR
MOUSE   391   AKKFQRQAANLQDCYRLYQGINQLPSYIQALEKYEGRH------QALLLAYFYTPLIDLR
RAT     391   AKKFQRQAANLQDCYRLYQGYKQLPNYIQALEKYQGRH------QALLLAYFYTPLTDLR
FLIES   388   TKKLMRRKANLQDLFRIYQYILRTPKILKYLHELDN---------STIESYICAPFKSFL
WORM    403   TKKLN-KRGNLEDYLKIYQFSKRIPEIYQYFTSFLEDDSPTEPYNELYRSYWLAPLSHHY
YEAST   364   ARRLMR-KCTLQDLNRFYQAATLLETYEMQLIQLSEAEQ----FAPSINRLLKSEITEIL
BACTERIA 352  LARLALRTARPRDLARMRHAFQQLPELRAQLETYDSAP---------------YQALREKM

HUMAN   445   SDFSKFQEMIE-TTLDMDQYE-NHEFLYKPSFDPNLSELREIMNDLEKKMQSTLISAARD
MOUSE   445   SDFSKFQEMIE-TTLDMDQYE-NHEFLYKPSFDPNLSELREYMDGLEKKMQSTLINAARG
RAT     445   SDFSKFQEKIE-TTLDMDQYE-NHEFLYKPSFDPNLSELREYMDGLEKKMQSTLISAARG
FLIES   439   KDLTGLKQMYE-QYYDFEAIE-RGEYLYKASFDSRLMELQQMMTELYSKMEELQFKCSQE
WORM    462   EPLSKFEEMYE-TTYDLDAYEENNEFMIKYEFNEELGKIRSKLDTLRDEIHSIHLDSAED
YEAST   419   KKYERFQYLCD-EFFDFDYEKENKEIRYRYDFYPEIQEISEKLEQMERYAEKLRKKYSAK
BACTERIA 398  GEFAELRDLLERAIIDTPPYLYRDGGYIASGYNEELDEWRALADGATDYLERLEYRERER

HUMAN   503   LGLDPGKQIKLDSSAQFGYYFRYTCKEEKYLRNNKNFSTYDIQKNGYKFTNSKLTSLNEE
MOUSE   503   LGLDPGKQIKLDSSAQFGYYFRYTCKEEKYLRNNKNFSTYDIQKNGYKFTNSELSSLNEE
RAT     503   LGLDPGKQIKLDSSAQFGYYFRYTCKEEKYLRNNKNFSTYDIQKNGYKFTNSELSSLNEE
FLIES   497   LNLDGKNQYKLESYAKLGHHFRITYKDDSYLRKNKNYRIYDYIKGGYRFTSDKLEGYADE
WORM    521   LGFDPDKKLKLENHHLHGWCMRLTRNDAKELRKHKKYIELSTYKAGIFFSTKQLKSIANE
YEAST   478   FECDN---LKLDKNSQYGFYFRYTLKEEKSIRKKDYHILETTKGSGYKFSYGELSDINDE
BACTERIA 458  TGLDT---LKYGFNAYHGYYIQISRGQSHLAPIN--YMRRQTLKNAERYIIPELKEYEDK

HUMAN   563   YTKNKTEYEEAQDAIYKEIYNISSGYYEPMQTLNDYLAQLDAYYSFAHYSNGAPYPYYRP
MOUSE   563   YTKNKGEYEEAQDAIYKEIYNISSGYYEPMQTLNDYLAHLDAIYSFAHYSNAAPYPYYRP
RAT     563   YTKNKGEYEEAQDAIYKEIYNISSGYYEPMQTYNDYLAHLDAYYSFAHYSNAAPYPYYRP
FLIES   557   FASCRTRYEEQQLSIYEEIIHYAYGYAAPLTLLNNELAQLDCLYSFAIAARSAPTPYYRP
WORM    581   TNILQKEYDKQQSALYREIINITLTYTPYFEKLSLYLAHLDYIASFAHTSSYAPIPYIRP
YEAST   535   FLEFHLKYTRAEEEYISMLCKKAEEFIPLIPAMAQLIATLDYFYSLSTFAATSSGIYTRP
BACTERIA 513  YLTSKGKALALEKQLYEELFDLLLPHLEALQQSASALAELDYLYNLAERAYTLN--YTCP

HUMAN   623   AILEK-GQGRIILKASRHACYEYQDEIAFIPNDYYFEKDKQMFHIITGPNMGGKSTYIRQ
MOUSE   623   YILEK-GKGRIILKASRHACYEYQDEYAFIPNDYHFEKDKQMFHIITGPNMGGKSTYIRQ
RAT     623   YILEK-GKGRIIYKASRHACYEYQHDYAFIPNDYHFEKDKQMFHIITGPNMGGKSTYIRQ
FLIES   617   KMLEE-GARELYLEDYRHPCLELQEHYNFIANSYDFKKEECNMFIITGPNMGGKSTYIRS
WORM    641   KLHPMDSERRTHLISSRHPYLEMQDDISFISNDYTLESGKGDFLIITGPNMGGKSTYIRQ
YEAST   595   NLLPL-GSKRLELKQCRHPYIEGNSEKPFIPNDYYLDKCR--LIILTGANMGGKSTYLRS
BACTERIA 571  TFIDK---PGIRITEGRHPYYEQYLNEPFIANPLNLSPQRR-MLIITGPNMGGKSTYMRQ

HUMAN   682   TGYIYLMAQIGCFYPCESAEYSIYDCILARYGAGDSQLKGYSTFMAEMLETASILRSATK
MOUSE   682   TGYIYLMAQIGCFYPCESAEYSIYDCILARYGAGDSQLKGYSTFMAEMLETASILRSATK
RAT     682   TGYIYLMAQIGCFYPCESAEYSIYDCILARYGAGDSQLKGYSTFMAEMLETASILRSATK
FLIES   676   YGTAYLMAHIGAFYPCSLATISMYDSILGRYGASDNIIKGLSTFMYEMIETSGIIRTATD
WORM    701   YGYISLMAQIGCFYPCEEAEIAIYDAILCRYGAGDSQLKGYSTFMYEILETASILKNASK
YEAST   652   AALSILLAQIGSFYPCSSATISYYDGIFTRYGASDKQSQGISTFMAEMLDCSAILQRATK
BACTERIA 627  TALIALMAYIGSYYPAQKYEIGPIDRIFTRYGAADDLASGRSTFMYEMTETANILHNATE
```

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

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

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