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 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'**

* 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):

```plaintext
1
2
3
4
5
6
7
8

Diabetes type II (40 years)

Diabetes type I (13 years)
```

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.

```
cagccgcagccctttgtgaaccaacacctgtgctacacaccttgatgtaaagctctctacc
tagtgccacagccaaagagctctcttctcTacacccacagctgcccggacagagggacc
>2

cagccgcagccctttgtgaaccaacacctgtgctacacaccttgatgtaaagctctctacc
tagtgccacagccaaagagctctcttctcTacacccacagctgcccggacagagggacc
>3

cagccgcagccctttgtgaaccaacacctgtgctacacaccttgatgtaaagctctctacc
tagtgccacagccaaagagctctcttctcTacacccacagctgcccggacagagggacc
>4

cagccgcagccctttgtgaaccaacacctgtgctacacaccttgatgtaaagctctctacc
tagtgccacagccaaagagctctcttctcTacacccacagctgcccggacagagggacc
>5

cagccgcagccctttgtgaaccaacacctgtgctacacaccttgatgtaaagctctctacc
tagtgccacagccaaagagctctcttctcTacacccacagctgcccggacagagggacc
>6

cagccgcagccctttgtgaaccaacacctgtgctacacaccttgatgtaaagctctctacc
tagtgccacagccaaagagctctcttctcTacacccacagctgcccggacagagggacc
>7

cagccgcagccctttgtgaaccaacacctgtgctacacaccttgatgtaaagctctctacc
tagtgccacagccaaagagctctcttctcTacacccacagctgcccggacagagggacc
>8

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.

**Amino acid sequence of the 'normal' insulin**

```
MALNWELL FLLAIAL MCIFPD AAA PQHGWLSHE VEAK FLPCHEG FTFENV RIMCAELGQFQVELQDGPCAGS LGFLAEIGLQRNLVTVRGCCTIG C331QQEWHY
```

**Amino acid sequence of the 'mutated' insulin (R 55 C)**

```
MALNWELL FLLAIAL MCIFPD AAA PQHGWLSHE VEAK FLPCHEG FTFENV RIMCAELGQFQVELQDGPCAGS LGFLAEIGLQRNLVTVRGCCTIG C331QQEWHY
```

The mutation R -> C prevents insulin from being cut and thus from being biologically 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):
MALWMRLPLPLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLGQVQVQELGPGAGSLQPLALEGSLQKRGIV
EQCTSCSLYQLENYC

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](https://en.wikipedia.org/wiki/Insulin), 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](http://www.philophylo.fr) (in French)

**Activity 6: www.chromosomewalk.ch**

[www.chromosomewalk.ch](http://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](http://www.chromosomewalk.ch))

Find out if you are real expert: [quiz expert](http://www.chromosomewalk.ch)!

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