Instructions
Practical 2A Use of online data resources for molecular biology

In the computer practicals related to online data resources, several resources will be addressed: genomic databases (Ensembl, NCBI), SNP and genetic variation databases, sources on genetic diversity (populations) and personalised genomics, databases related to genetic regulation (miRNAs and their targets, epigenetics), metabolite databases, and the links of all these molecular players to disease.

In this session, part A, we will look into all of those except for the sources on genetic variation and diversity and their consequences, which will be introduced in computer practical 2B.

In case 3, you are studying neuroplasticity, a complex process in which many molecular interactions are involved. In this practical we are going to explore information in molecular biological databases, taking examples from the neuroplasticity pathway. The main goal of this practical is to get you more familiar with these databases and the information you can find. The assignments of this short practical cannot target detailed biological interpretation, but we strongly advise you to take these resources into account when preparing for this and the next cases (and furthermore throughout your study). These databases are a reliable source of up-to-date information and can complement knowledge in text books and other literature sources. When trying to answer the questions, use the *Reference guide online resources* that was provided to you. Instructors are around to help you, but consult the guide first!

Assignment 1: Genetic and functional information on neurotrophic factors in Ensembl *(20 minutes)*

Note: if Ensembl responds slowly, click 'Mirrors' at the top of the page and try another mirror.

Look up the human Brain-Derived Neurotrophic Factor (BDNF) signalling pathway in WikiPathways.

a) The neurotrophins, including NGF, are located at the top of the pathway. Which interaction does NGF have with BDNF dimers according to this pathway?

   Protein-protein interaction.

b) Which is the receptor that binds NGF and which other protein does the receptor target?

   It binds to the NTRK1 receptor, which targets the CTNNB1 protein.

Click on the NGF box, to display its identifiers in several databases.

c) What are the Entrez Gene (NCBI Gene), Ensembl, RefSeq, and HGNC symbols of the human NGF gene and the Uniprot and RefSeq identifiers of the human NGF protein?

   4803, ENSG00000134259, NM_002506, NGF and P01138, NP_002497.

Now, open the Ensembl page for the human NTRK1 receptor encoding gene.

d) What is the chromosomal position (chromosome, location, and strand) of this gene?

Chromosome 1: 156,815,640-156,881,850 forward strand.

e) What was the position on chromosome in the previous genome build (CRCh37)?

Scroll down to 'other assemblies'.
156,785,432-156,851,642(so more than 30kb different).

f) How many transcripts (splice variants) are known for this gene? How many are protein coding?

10 known, 4 protein coding.

g) Which transcripts is the longest in base pairs? And which produces the longest protein? Explain your findings.

Longest in base pairs: NTRK1-202; ENST00000368196
Longest protein: NTRK1-206; ENST00000524377
Finding: longest mRNA is from different splicing variant as longest protein
Explanation: not all exons are (completely) protein-coding; we also have 3'and 5' UTRs.

Now open the Ensembl page for the human NGF gene.

h) What is the genomic position of this gene?

Chromosome 1: 115,285,918-115,338,236 reverse strand.

i) Give the name of one Gene Ontology (GO) annotation for each of the three GO domains (biological process, molecular function, cellular component). Pick annotations that are specifically related to neuro molecular functions.

Examples of terms (others are possible, it has many annotations):
Biological process: regulation of neuron differentiation, negative regulation of neuron apoptotic process, neuron projection morphogenesis, positive regulation of axonogenesis
Molecular function: nerve growth factor receptor binding
Cellular component: extracellular region.

j) What is the Ensembl identifier of the single transcript known for NGF?

ENST00000369512 (or ENST00000369512.2 including version number).

Open the transcript's tab by clicking its name.

k) How many coding and non-coding exons does this transcript have? How are these displayed? Which is the coding exon?

It has 3 exons, of which 1 is coding. Coding (parts of) exons are shown by filled boxes, non-coding (parts of) exons as open boxes. The first exon (note that the gene is on the

reverse strand, so starts from the right side) is the coding one. Note that the first exon also has a small non-coding part (there is a 3'UTR, as usual), but this cannot be seen at this scale.

Now open the location tab for the NGF gene.

l) Which is the closest gene to the NGF gene? Which type of gene is this? Which is the closest protein coding gene on the same strand?

AL512638.2 is the closest gene, an RNA gene; the closest protein coding gene on the same strand (reverse, indicated by a < symbol in the graphical overview) is TSPAN2.

Leave your browser window open; you will continue using it in the next assignment.

Assignment 2: Regulatory information on neurotrophic factors in Ensembl *(15 minutes)*

In the 'Regulatory build' track, you can find information on regulatory features in the gene region.

a) Where is the promotor located? Does this make sense?

At the right side of the screen. Yes, this surrounds the first exon (the gene is located on the reverse strand).

Configure the page to add the CpG Islands track to the display on the location tab for the NGF gene.

b) Is there a CpG Island in this gene region? If so, at which position? Does this make sense?

Yes, there is one CpG Island. It is located at position: 115337547-115338711. This makes sense, it is located right at the first exon (it is a gene on the reverse strand) and in its promoter region.

Also add DNA methylation information, select all available cell types.

c) What can you say about the methylation status of the CpGs in the CpG Island?

They can be methylated or not, dependent on cell type. This makes sense as the NGF protein will not be needed in all cell types.

d) What can you say about the methylation status of the other CpGs in the gene region?

Almost all are methylated (note that whether or not a CpG has information, depends on the measurement technology used, if there is no coloured mark that CpG has not been included).

To keep the number of tracks limited, disable all DNA methylation information, except for H1ESC cells. Add information on histone modification: add the default selection of histone modification marks for H1ESC cells.

e) What can you say about the presence of H3K4me3 marks in the gene? What does this mean? How does this relate to the methylation information? And to the Pol-II binding?

Marks are present in the promoter region, which means that the gene could be active in H1ESC cells. The methylation status is unclear, showing half methylated parts detected in the promoter flank (not measured in the promoter itself). Polymerase II binding is present in the promoter area, strengthening evidence for expression.

Finally, enable the default Open chromatin & TFBS tracks (DNase1 and CTCF binding), for H1ESC cells.

f) What can you say about the DNase1 mark for open chromatin, does this confirm the answer to the previous question?

It is present in the promoter region, indicating open chromatin over there. Yes, this does confirm likely expression of the gene in H1ESC cells.

Assignment 3: Genetic and functional information on neurotrophic factors in NCBI *(15 minutes)*

Look up the human BDNF gene in NCBI.

a) What is the NCBI Gene identifier of this gene?

627

b) What are other names of this gene, besides to official HGNC name BDNF?

ANON2; BULN2.

c) Give the RefSeq identifier of the mRNA and protein for one of its splice variants.

NM_1707314 and NP_733927 (other possible).

d) What is the chromosomal position (chromosome, location, and strand) of this gene?

Chromosome 11: 27654893..27722058, reverse strand.

e) How many exons does the gene have (total regardless of transcript)?

12 (see 'exon count').

f) In which tissue is this gene expressed most?

Brain (scroll down to 'Expression').

g) To which neurological Phenotype is it associated?

Obsessive-compulsive disorder (other possible, scroll down to 'Phenotypes').

h) Name one KEGG pathway in which the gene is involved.

Neurotrophin signalling pathway (others possible, scroll down to 'Pathways from BioSystems).

i) Give the name and UniProt identifier of one protein with which BDNF interacts.

NTF3, P20783 (others possible; scroll down to 'Interactions')

Note: it also interacts with itself, as we already saw in the WikiPathways process diagram, where it formed a dimer

j) Give one GO biological process annotation that has been based on a direct assay (experiment).

Positive regulation of synapse assembly (others possible). Note that the Evidence code for 'Inferred from Direct Assay' is IDA. Scroll down to 'General gene information'.

Open the OMIM page for this gene, for example by clicking OMIM in the right menu on this page.

k) What is the OMIM identifier of the gene?

113505

l) To which conditions has the gene been associated?

Central hypoventilation syndrome, congenital
Anorexia nervosa
Bulimia nervosa
Memory impairment
Obsessive-compulsive disorder

m) Name one paper that is mentioned in the expert-written text in OMIM and has studied the association with Obsessive-compulsive disorder

Hall, D., Dhilla, A., Charalambous, A., Gogos, J. A., Karayiorgou, M. Sequence variants of the brain-derived neurotrophic factor (BDNF) gene are strongly associated with obsessive-compulsive disorder. Am. J. Hum. Genet. 73: 370-376, 2003.

Look up Obsessive-compulsive disorder (OCD) in OMIM.

n) What is the OMIM identifier of OCD?

164230

o) Are other genes described in OMIM that are related to Obsessive-compulsive disorder?

Yes, HTR2A and SLC6A4.

Assignment 4: miRNAs targeting BDNF *(15 minutes)*

Look up the human BDNF gene in miRTarBase, a database of validated miRNA targets.

a) By which experimental measurements has hsa-miR-210-3p been demonstrated to be able to target the BDNF mRNA?

Reporter assay, western blot, qPCR, microarray, and other methods.

b) Why are there also miRNAs from other species in the list of results?

They may still found in experiments to be able to target the human BDNF mRNA; it is not said this necessarily biologically also happens (as for the others).

Look up the hsa-miR-210-3p mature miRNA in miRBase

c) What is the accession number of hsa-miR-210-3p in miRBase?

MIMAT0000267

Open the hsa-mir-210 stem-loop sequence (precursor) page in miRBase for this miRNA

d) What is the accession number of hsa-mir-210?

MI00000286

e) From the deep sequencing reads, name a tissue in which a high read count (expression) has been observed. Give the study (literature reference) in which this has been found.

Uterine cervix, 399 reads (others possible)
Found by: "Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls"
Witten D, Tibshirani R, Gu SG, Fire A, Lui WO; BMC Biol. 8:58(2010).

f) How many mature miRNA are produced from this stem-loop sequence?

Two, a 3p and a 5p miRNA.

Open the validated targets for hsa-miR-210-3p as determined by TarBase, for example by clicking the link from within miRBase

g) Do these targets include BDNF? How can you explain your findings?

No it does not. TarBase and miRTarBase use other experimental resources to collect their information from. As such, reported interactions may be different.

h) Also links to databases with predicted targets are given. How are these determined?

By computational prediction based on the seed sequence (sequence complementarity between miRNA and mRNA) and possibly conformational information (like self-looping of the mRNA, which may block the target sequence).

Find target genes for hsa-miR-210-3p in TargetScan (for example by clicking the 'TARGETSCAN-VERT' link in miRBase.

i) Do the predicted targets by TargetScan include BDNF?

Yes they do.

Look up hsa-mir-210 in miR2Disease.

j) Has this miRNA been associated to any diseases? If so, do these include neurological conditions?

Yes it has been associated to many diseases, mostly cancers and myopathies. And, yes, a reference for Alzheimer's disease is also in the list.

Assignment 5: metabolites *(15 minutes)*

One of the metabolites present in the human Brain-Derived Neurotrophic Factor (BDNF) signaling pathway on WikiPathways is cAMP. Open the HMDN page for this metabolite, either directly or by clicking the link when selecting this element using the pathway diagram.

a) What is the HMDB identifier of cAMP?

HMDB0000058

b) What is the chemical formula of cAMP and its average molecular weight?

C10H12N5O6P and 329.2059.

c) To which health effects has cAMP been associated? Also give the literature references provided.

Meningitis

Lerche A, Svenson M, Wiik A: Cerebrospinal fluid levels of cyclic nucleotides in meningitis and idiopathic polyneuritis. Acta Neurol Scand. 1984 Mar;69(3):168-75.

Headache

Kruuse C, Frandsen E, Schifter S, Thomsen LL, Birk S, Olesen J: Plasma levels of cAMP, cGMP and CGRP in sildenafil-induced headache. Cephalalgia. 2004 Jul;24(7):547-53.

(Scroll down to 'Ontology' – Physiological effect; references given at bottom of the page).

d) Name a process in which cAMP is involved according to HMDB, pick a neurological one.

One of:

Dopamine Activation of Neurological Reward System

Excitatory Neural Signalling Through 5-HTR 4 and Serotonin

Excitatory Neural Signalling Through 5-HTR 6 and Serotonin

Excitatory Neural Signalling Through 5-HTR 7 and Serotonin

Scroll down to the "Ontology" section

e) In which bodily fluids has cAMP been determined? What is its normal concentration in adult blood?

Blood, CSF, Feces, Urine. Normal concentration in adult blood is 0.0085 +/- 0.0005 µM.

Open the page for cAMP in ChEBI, either directly or by following the link from the HMDB page.

f) What is the ChEBI identifier of cAMP?

17489

g) Name one derivative molecule of cAMP given by ChEBI.

Bucladesine (many others possible). Look at the ChEBI Ontology for molecules of which cAMP is reported to be a functional parent.

Open the entry for cAMP in DrugBank, either directly or by following the link from the HMDB page.

h) What is the DrugBank identifier of cAMP

DB02527

Now look up BDNF in DrugBank.

i) Name three drugs targeting the receptor NTRK1 (note to search for it as a 'target'; to verify: NTRK1 is called 'High affinity nerve growth factor receptor' in DrugBank) and their status (experimental or approved)

Imatinib, Amitriptyline, Regorafenib – all approved.

j) In what form and for what indications is Regorafenib prescribed according to DrugBank?

As a tablet; for colorectal and hepatocellular cancer.

k) Does it stimulate or inhibit NTRK1?

Inhibit (see Mechanism of action under 'Pharmacology'.