Homework:

https://www.youtube.com/watch?v=-uBY0rnkf8U
Basic bioinformatics concepts, databases and tools

Module 3

Sequence analysis
We can do a lot more with sequences than just checking if they’re similar.

Module 3: we will analyze sequences to infer biological properties.

Relation between modules 1, 2 and 3:
Biological properties that can be inferred from sequences

AGCTACTACGGACTACTAGCAGCTACCTCTCTCTG

For this nucleotide sequence you can answer questions like:

Is this a coding sequence?

Can this sequence bind a certain transcription factor?

What is the melting temperature of the sequence?

What is the GC content of the sequence?

Does it fold into a stable secondary structure?

...
Inferring biological function based on motifs/domains

Method is straightforward

You calculate a score (bits score or E-value) to measure similarity between a sequence and the motif/domain model

Overview of tools for predicting protein motifs, active or functional sites, and function

How are most biological properties inferred?

Biological properties are predicted based on experience

= Thousands available experimentally verified sequences with a certain property

= Training set

Tools that do these predictions are trained to recognize sequences with the property

This approach is called **classification**

Done by machine learning algorithms: neural networks, support vector machines...

Black box methods: they don’t use any biological rationale, just the training set
Never trust a tool’s output blindly!

Interpretation of results depends on the kind of output

When a prediction result is obtained, the question arises 'Is it true?'

Some tools give a 'binary' result: 1 or 0, a hit or a miss

Compare results of different prediction tools for higher confidence

e.g. Prediction of secretion: a protein is either secreted or not

Prediction of TM region: a region is a TM region or not

Some tools give a score/p-value/E-value: the chance that the 'result' is 'not true'

Asses the p-value: lower score -> more trust worthy result

e.g. ScanProsite for a motif
Neural network use training set to fine-tune parameters for prediction

Neural network consists of layers of nodes
Each node gets one / more inputs
Performs a mathematical operation on inputs
To calculate an output
Output is passed on to the next layer

**Training:**
You specify global input and output
Network adjust parameters of mathematical operation so that input results in output

**Prediction:**
Same parameter settings are used to predict output of new input sequence
Support Vector Machine finds plane that optimally separates classes

**Training:**

Training set consists of two groups:

*Sequences that are part of a helix*

*Sequences that are not part of a helix*

Project sequences as points in space

Transform projection into a space where the two groups are perfectly separated

Calculate the plane that separates the two groups

**Prediction:**

New sequences are projected into the feature space and you can calculate on which side of the plane they are located

*SVM introduction* [https://www.statsoft.com/textbook/support-vector-machines](https://www.statsoft.com/textbook/support-vector-machines)
Neural networks are the hipsters of machine learning

Originally nobody believed in them

Google expanded the size of these networks to make them more powerful

⇒ Deep learning

⇒ Used to infer protein function directly from sequence

protein-protein interaction networks

proteome from peptide MS profiles
You can **predict** modification sites in peptide sequences

Overview of tools: [https://omictools.com/ptms-category](https://omictools.com/ptms-category)

Most tools predict one type of modification

Tools that can find multiple modifications:

  - one protein at a time
  - **experimentally verified** PTMs for all UniProtKB protein entries

- **Findmod**: [https://web.expasy.org/findmod/](https://web.expasy.org/findmod/)
  - one protein at a time
  - **prediction** based on MS data of the protein
  - for many proteins you need advanced software e.g. MaxQuant
Plant specific PTM databases

http://phosphat.uni-hohenheim.de
Arabidopsis phosphorylation sites identified by large-scale MS

http://www.p3db.org/
9 plants mainly Arabidopsis, medicago, rice: http://www.p3db.org/organism.php
Phosphorylation sites identified by MS

https://bioinformatics.psb.ugent.be/webtools/ptm_viewer/
Arabidopss, wheat, maize
Phosphorylation and ubiquitination sites identified by MS and COFRADIC
You can **predict** protein’s **structural** properties

Done by

**Classification**

**Comparison with models** of structure components

Structural components you can search for

- secondary structures see next slides
- transmembrane regions e.g. Phobius (comparison to HMMs)
- coiled coil regions e.g. COILS, Marcoils (comparison to models)
- globular regions
- disorder regions e.g. MeDor, Poodle (SVM)

PredictProtein:  [https://www.predictprotein.org/](https://www.predictprotein.org/)

combined predictions of all listed structural components
Secondary structures via interactions between amino acids

Parallel β pleated sheet

Antiparallel β pleated sheet

Right-handed α helix

Secondary structure resources: [http://www.geneinfinity.org/sp/sp_proteinsecondstruct.html](http://www.geneinfinity.org/sp/sp_proteinsecondstruct.html)  
[http://www.molbiol-tools.ca/Protein_secondary_structure.htm](http://www.molbiol-tools.ca/Protein_secondary_structure.htm)
Tools for predicting proteins’ secondary structure

http://bioinf.cs.ucl.ac.uk/psipred/

PSIPRED aggregates several structure prediction methods into one location

Users can select the prediction methods they want to apply to their sequence:

You can select multiple tools and immediately compare the results

Good overview of sec structure prediction tools: http://toolkit.tuebingen.mpg.de/sections/secstruct
You do not always need classification for sequence analysis

Examples for protein sequence analysis

\( pI \) (isoelectric point) = \textbf{sum} of all charges in a protein
Composition metrics = \textbf{number} of aromatic amino acids, A’s...
Hydrophobicity calculation = \textbf{sum} of hydrophobicity of amino acids in window
Translation (DNA → protein) = \textbf{translate} codons into amino acids
Occurrence of simple patterns (e.g. does KDL occurs and how many times)
...
You can simply \textbf{calculate} these properties, you don’t need classification
In nucleotide sequences you can also search for motifs

DNA motifs: short DNA patterns that are believed to have a biological function

They bind to proteins:

- **TFs**
- Nucleases
- Ribosomes
- mRNA processing factors
- hnRNP proteins
- ...


*DNA motifs course*  http://www.slideshare.net/Stewbacca/dna-motif-finding-2010
Regulatory DNA **motifs** regulate gene expression

Motifs = sites in the **promoter** that are recognized by transcription factors (TFs)

= Transcription Factor Binding Sites

Promoter = regulatory region in the DNA in front of gene required for expression

TFs bind to motifs in promoter and regulate where and when the gene is expressed

TFs can activate or repress expression

Motifs often occur in **modules**: combinations of motifs bound by interacting TFs

*From DNA motifs course [http://www.slideshare.net/Stewbacca/dna-motif-finding-2010](http://www.slideshare.net/Stewbacca/dna-motif-finding-2010)*
Gene expression is not only regulated by TFs!

but also by

**Chromatin remodelling:**

Open parts of chromatin are accessible for TFs/RNAP


RNA processing:

e.g. degradation of mRNA mediated by miRNAs and siRNAs

[http://wiki.utep.edu/display/siRNA/Home](http://wiki.utep.edu/display/siRNA/Home)
Expression regulation via chromatin remodeling

Histone acetylation/methylation/phosphorylation

DNA is wrapped around histones and as such not accessible

Modifications alter interaction between histone and DNA

DNA methylation

C is methylated, typically in CpG islands
Shuts off or reduces expression

Figure taken from http://www.nature.com/nrc/journal/v1/n3/full/nrc1201-194a.html
Data on chromatin accessibility

CpG islands can be predicted but mostly based on experimental data.

Most data available for human and mouse.

Large scale projects: ENCODE
Epigenome Roadmap

Overview of resources: http://generegulation.info/
The ENCODE project

ENCyclopedia Of DNA Elements

Mission: annotate **functional elements** in human genome in various **lab cell lines** regions that encode a product e.g. Protein transcriptional regulatory regions

**Limited clinical applications**

Findings of ENCODE pilot project:  [http://www.genome.gov/Pages/Research/ENCODE/nature05874.pdf](http://www.genome.gov/Pages/Research/ENCODE/nature05874.pdf)

ENCODE experiments are based on NGS

RNA bind proteins
Purify DNA with histones still attached
TF
Fragment DNA
RNA
Selection with antibodies directed to modified histone
TF
Methyl groups
RNA and perform RT
RNA bind proteins
Release DNA
RNA and perform RT
Sequence
ENCODE data form tracks in Ensembl (human+mouse)

Regulatory features:

Open chromatin & TFBS:
Antibodies targeting DNase (only fragments open chromatin) + TFs

Histones & polymerases:
Antibodies targeting modified histones / RNA polymerases

DNA methylation:
Bisulphite sequencing: bisulphite C -> U but methyl groups protect C
Methylated C -> C
Unmethylated C -> U
The Epigenome roadmap project

Mission: annotate **epigenomic modifications** in human genome in various **cell types**

e.g. stem cells

mature cells from different tissues from healthy people

mature cells from patients with cancer, neurodegenerative, autoimmune disease

**Great clinical relevance**

Similar experiments as ENCODE

Results: [http://epigenomemegateway.wustl.edu/browser/](http://epigenomemegateway.wustl.edu/browser/)

*Findings of Epigenomics roadmap pilot project:* [https://www.nature.com/articles/518314a](https://www.nature.com/articles/518314a)

Overview of all large-scale epigenetics projects

http://epigenomesportal.ca/ihec/
Download data
Visualize in genome browser

https://epigenie.com/epigenetic-tools-and-databases/
List of epigenetics tools / databases
Also non-coding RNA resources

Findings of Epigenomics roadmap pilot project: https://www.nature.com/articles/518314a
Cell types and data download http://www.genboree.org/epigenomeatlas/multiGridViewerPublic.rhtml
Ensembl Genome Browser has track hubs for epigenetics

Collections of genomic data that can be viewed on the Ensembl Genome Browser

Load via Location page

Most tracks available for previous version of human genome (GRCh37)

[Image of Ensembl Genome Browser interface with track hubs highlighted]

http://ensemblgenomes.org/info/access/public-track-hubs
Gene expression is not only regulated by TFs!

but also by

Chromatin remodelling:

Open parts of chromatin are accessible for TFs/RNAP


RNA processing:

e.g. degradation of mRNA mediated by miRNAs and siRNAs

[http://wiki.utep.edu/display/siRNA/Home](http://wiki.utep.edu/display/siRNA/Home)
Databases of published miRNA sequences

**miRBase**  [http://www.mirbase.org/](http://www.mirbase.org/)

Published miRNA sequences

Analysis: You want to know details (e.g. sequence) and targets for published miRNA


*Predicted* and *validated* target sites of miRNAs from miRBase

Analysis: You have a gene and want to know if it contains miRNA binding sites

Human, mouse, rat
miRWalk contains predicted and validated target sites

**Validated sites:** obtained by text mining of scientific literature from databases of validated miRNAs e.g. miRTarBase

**Predicted sites:** miRBase + RefSeq + 5 prediction algorithms

Predictions of other websites

Tips for miRWalk searches:

! Do not only search for target sites in 3’UTRs !

! miRNAs can also target other miRNAs !

Overview of miRNA target prediction tools  http://www.exiqon.com/microrna-target-prediction
Most miRNA target prediction tools only search for published miRNAs

In the following tools you can submit

- Your own miRNA sequence
- Potential target(s)

They will predict if the miRNA can bind the target

RNAhybrid: http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/
miRTar: http://mirtar.mbc.nctu.edu.tw/
Regulatory DNA **motifs** regulate gene expression

**Motifs** = sites in the promoter that are **recognized by transcription factors** (TFs)

= Transcription Factor Binding Sites

Promoter = regulatory region in the DNA in front of gene required for expression

TFs bind to motifs in promoter and regulate where and when the gene is expressed

TFs can activate or repress expression

Motifs often occur in modules: combinations of motifs bound by interacting TFs

*From DNA motifs course*  [http://www.slideshare.net/Stewbacca/dna-motif-finding-2010](http://www.slideshare.net/Stewbacca/dna-motif-finding-2010)*32*
How to represent TFBS

Same representations as for protein motifs:

**Consensus sequence**

**Position Weight Matrix**

**Logos**

HMMs are not used for DNA motifs!
Consensus sequences are not accurate

TFBS are degenerate: bind to variable but similar motifs

If you do not allow mismatches only 2/6 TFBS will be found
This motif will occur by chance every 4096 bp

If you allow 2 mismatches you find all TFBS
You will find one match every 30 bp

=> You have to use IUPAC codes in consensus sequences

<table>
<thead>
<tr>
<th>Y</th>
<th>Pyrimidine (C or T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>Purine (A or G)</td>
</tr>
<tr>
<td>W</td>
<td>weak (A or T)</td>
</tr>
<tr>
<td>S</td>
<td>strong (G or C)</td>
</tr>
<tr>
<td>K</td>
<td>keto (T or G)</td>
</tr>
<tr>
<td>M</td>
<td>amino (C or A)</td>
</tr>
</tbody>
</table>
PWMs for representing DNA motifs

counts

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>1</th>
<th>6</th>
<th>1</th>
<th>0</th>
<th>7</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

frequencies

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0.125</th>
<th>0.75</th>
<th>0.125</th>
<th>0</th>
<th>0.875</th>
<th>0.125</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
<td>0.125</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0.875</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.125</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0.125</td>
<td>0.875</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
</tbody>
</table>

log odds scores = weights

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>1.2</th>
<th>0</th>
<th>-1.6</th>
<th>0</th>
<th>-1.6</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-1.6</td>
<td>-1.6</td>
<td>-1.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>-1.6</td>
<td>0.96</td>
<td>0.96</td>
<td>-1.6</td>
<td>0.59</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-1.6</td>
<td>-1.6</td>
<td>0</td>
<td>0.59</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Bit scores on the Y-axis of sequence logos

Y-scale range: maxBits = $\log_2(4) = 2$

Base occurring in all sequences: 2 bits

Base occurring in half of the sequences: 1 bit

Base occurring in a quarter of the sequences (randomly): 0 bits

Review on DNA motifs  http://www.nature.com/nbt/journal/v24/n4/full/nbt0406-423.html
Databases containing **known** TF binding motifs

In eukaryotes 2000 TFs are predicted, 900 are **known**

Overview of TF binding databases: [http://generegulation.info/](http://generegulation.info/)


Motifs represented as PSMs

One of few databases with a lot of motifs from plants !

**JASPAR** : [http://jaspar.binf.ku.dk/](http://jaspar.binf.ku.dk/)

Motifs represented as PSMs, regularly updated
cisBP is a composite TFBS database

http://cisbp.ccb.r.utoronto.ca/

Composite database + tools for finding known motifs

Motifs from 25 databases + publications = validated motifs

They add predicted motifs by cross species comparisons

Represented as IUPAC consensus or PWMs

> 160000 TFs from > 300 species
Tools for finding **known** TF binding motifs


Set of tools that work together all with similar user interface

**dna-pattern**: Search for similarity to *regular expressions* of user defined motifs

**matrix-scan**: search for similarity to user defined **PSM**

**Cscan** - [http://159.149.160.51/cscan/](http://159.149.160.51/cscan/)

Large collection of ChIP-Seq data on TFs, histone modifications, RNA polymerases...

Enriched peak regions are crossed with the genomic coordinates of input genes
Tools for predicting **novel** TF binding motifs from coregulated genes

RSAT info-gibbs - [http://www.rsat.eu/](http://www.rsat.eu/)


Discover novel motifs in sets of unaligned DNA or protein sequences

First choice for finding novel motifs

**Weeder** - [http://159.149.160.51/modtools/](http://159.149.160.51/modtools/)

Based on simple word counts

But OK for yeast / bacteria

Mask repeats in promoter sequences

Run tool multiple times with different parameters

Let tool return multiple motifs
Low specificity in motif searches if no measures are taken

Motif searches (known + unknown) return many false positives
e.g. if you submit a random set of genes, you get many highly 'significant' motifs

known motifs are found everywhere in the genome

Less problematic in simple organisms: yeast, bacteria

In higher organisms, you need additional info to improve the searches:

- **Phylogenetic footprinting**
- **Search for modules** of motifs
- **Comparison to genomic background**
Modules is best method but not always feasible

Phylogenetic footprinting:
Only search for motifs in conserved regions in promoters
Allows to eliminate 90% false positives
Only thing you need are promoters of orthologs: no further knowledge required

Search for modules of motifs
Allows to eliminate 99% of false positives
You need to know upfront that transcription factors interact

Comparison to genomic background:
Motif occurs more frequently in your set than in genome (= expected by chance)?
Tools for phylogenetic footprinting

Function and DNA-binding preferences of TFs are well conserved between species

-> TF binding motifs will also be conserved while remainder of promoter is not

1. Search orthologs for gene of interest
2. Select promoter regions
3. Align promoters
4. Use conserved regions for motif searches

ConTra: http://bioit.dmbr.ugent.be/contrav2/
Tools for binding motif overrepresentation

Compare your results with those obtained on a set of random promoters
Calculate statistics to see if the motif frequency that you see in your set is significantly higher than in the set of random promoters

**Unknown motifs:**
RSAT-info-gibbs

**Known motifs:**
RSAT Matrix-Scan
MotifLab can do all three techniques simultaneously

Download + local installation: [http://tare.medisin.ntnu.no/motiflab/](http://tare.medisin.ntnu.no/motiflab/)

Can perform phylogenetic footprinting + overrepresentation


Gene regulation training: on demand

Exercises:
DNA sequence analysis: gene regulation
Approaches for structural gene annotation

Gene prediction: where are the genes and what do they look like?

Identifying signals surrounding genes: TFBS, promoters, CpG islands
splice sites
polyA sites ...

Motifs available for these signals $\rightarrow$ combined **motif searching**

Difference in composition between coding and noncoding DNA
$\rightarrow$ **compositional analyses**

Alignment with annotated homologs / transcripts + proteins (like Ensembl does)
$\rightarrow$ **similarity searches**
Software for gene prediction using motifs and composition

**GENSCAN**: commercial but free for academics
Models for exons, TATA box, polyA signal... for vertebrates/Arabidopsis/maize
Used at EBI and NCBI for genome annotation
http://genes.mit.edu/GENSCAN.html

**GeneMark**: commercial but free for academics
Models for many prokaryotic and eukaryotic organisms
http://exon.gatech.edu

**Augustus**: free, many organisms
http://bioinf.uni-greifswald.de/webaugustus/about.gsp

http://www.embl.de/~seqanal/courses/spring00/GenePred.00.html
Software for gene prediction using motifs and composition

**fgenesb**: free

Models of CDS, operons, promoters, terminators for bacteria


**fgenesh**: free

Models for animals and plants


Accuracy dependent on quality of models

Models specific for organism

=> So you need a set of structurally well annotated genes for your organism
Software for gene prediction using similarity searches

BLAST algorithms

**RATT:** [http://ratt.sourceforge.net/](http://ratt.sourceforge.net/)

- transfer annotation from an annotated to an unannotated genome
- Used at Sanger institute

More universally applicable

Accuracy depends on quality of the alignment
Often a combination of the 3 methods is used

Like VEGA/HAVANNA does
Software that combine 2 approaches gives best result

**Eugene:** [http://eugene.toulouse.inra.fr/index.html#more](http://eugene.toulouse.inra.fr/index.html#more)

- HMM for prediction of CDS, splice sites
- Integration of predictions of other software
- Similarity with existing sequences
- All organisms
- Used for various plant genomes and C. elegans genome

**Maker:** [http://www.yandell-lab.org/software/maker.html](http://www.yandell-lab.org/software/maker.html)

- Repeat identification
- Prediction of genes
- Similarity with existing sequences
- Prokaryotes and small eukaryotes
Portals for structural gene annotation

Overview of tools for small-scale analysis:
https://omictools.com/genome-annotation2-category

Good practices for genome assembly and annotation:
https://f1000research.com/articles/7-148/v1

For newbies, provides answers to beginners questions
Software for non-coding RNA detection

**INFERNAL**: [http://eddylab.org/infernal/](http://eddylab.org/infernal/)

Searching DNA sequences for RNA structure and sequence similarities

**tRNAscan-se**: [http://lowelab.ucsc.edu/tRNAscan-SE/](http://lowelab.ucsc.edu/tRNAscan-SE/)

Searching DNA sequences for tRNA sequences
Databases of repeat sequences

RepBase: https://www.girinst.org/server/RepBase

Dfam: http://www.dfam.org/

HMM model library derived from Repbase sequences
Software for detection of repeats

**RepeatMasker:** [http://repeatmasker.org/](http://repeatmasker.org/)

- Screens DNA sequences for known repeats
- Detection: HMMER-based algorithm (using Dfam)
  - BLAST-based algorithms (using RepBase)

**REPET:** [https://urgi.versailles.inra.fr/Tools/REPET](https://urgi.versailles.inra.fr/Tools/REPET)

- Can detect, classify, and analyze novel repeats in genomic sequences
- Detection: BLAST sequence to itself
- Classification: Cluster similar repeats into families
  - MSA (using MAFFT) + consensus for each family
- Check sequence for extra instances of repeats (using RepeatMasker)
- Used by international consortia for > 50 genomes
Bioinformatics portals: EBI

http://www.ebi.ac.uk/services

Bioinformatics services

We maintain the world’s most comprehensive range of **freely available** and up-to-date **molecular databases**. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our **web services** to access our resources programmatically.

### DNA & RNA
- genes, genomes & variation

### Gene expression
- RNA, protein & metabolite expression

### Proteins
- sequences, families & motifs

### Structures
- Molecular & cellular structures

### Systems
- reactions, interactions & pathways

### Chemical biology
- chemogenomics & metabolomics

### Ontologies
- taxonomies & controlled vocabularies

### Literature
- Scientific publications & patents

### Other software
- cross-domain tools & resources

Sequence analysis tools
Other bioinformatics portals: Expasy

http://www.expasy.org

Sequence analysis tools
Other bioinformatics portals: EMBOSS

http://emboss.bioinformatics.nl/
Primer design guidelines

18-28 nucleotides long

G+C content: 45-60%

**Melting T:** 60-75°C, max 5°C apart

3’-end: G, C, CG or GC to promote binding

Middle: have the mismatched bases in the middle of the primer for mutagenesis

5’-end: add 3 to 4 nucleotides 5’ of the restriction site to promote efficient cutting

No runs of > 3 identical nucleotides e.g. CCCC

No runs of > 3 identical dinucleotides e.g. ATATATAT

No runs of > 3 bases that complement within primer: primer dimers

No complementarity between primers: primer dimers

No secondary structures such as hairpins

Take into account location of SNPs

Specific
3’ ends of primers are crucial

This is where DNA polymerase adds nucleotides
qPCR primer design guidelines

**intron or exon-exon junction spanning** to avoid measuring genomic DNA
length: 18-28 bp (ideally: 20 bp)
melting temperature Tm: 58-60°C, **max 2°C apart**
5 nucleotides at 3’ end < 3 G or C
avoid runs > 3 identical nucleotides

**Specific**

PCR Product: Take into accounts the existence of **splice variants**

70-200 bp

120-200 bp when using **SYBRgreen** to discern primer dimers – product
Primer design to reduce false positives from genomic DNA

1. Primer spans an exon-intron boundary:

- Genomic DNA:
  - exon 1
  - intron
  - exon 2
  - NO product

- cDNA:
  - exon 1
  - exon 2

2. Primers flank an intron:

- Genomic DNA:
  - exon 1
  - long intron
  - exon 2
  - NO product

- cDNA:
  - exon 1
  - exon 2
Primer design software

Checking splice variants: Ensembl

Masking SNPs
SNPMasker: http://bioinfo.ebc.ee/snpmasker/
Highlight all SNPs that occur in target sequence

Masking repeats
RepeatMasker: http://www.repeatmasker.org
Highlight runs of identical nucleotides

Primer Design
Primer3Plus: http://primer3plus.com/cgi-bin/dev/primer3plus.cgi
PrimerQuest: https://eu.idtdna.com/PrimerQuest/Home/Index
Primer design software

Checking primer specificity
BiSearch: http://binearch.enzim.hu/

Checking primer characteristics
OligoAnalyzer: http://eu.idtdna.com/calc/selector

Checking secondary structures in the PCR product
UNAfold: http://mfold.rna.albany.edu/
Commercial -> not free

Checking location of SNPs
UCSC in silico PCR + SNPs track: http://genome.ucsc.edu/cgi-bin/hgPcr
Primer3 does not check specificity of your primers!

You have to BLAST the suggested primers against genome of organism you work in to see if the primers are uniquely targeting the region you are interested in.

Often you will see that the primers will also bind other regions in the genome => You have to start your search all over again.

Especially 3’ end of primers is critical.
Primer BLAST: defining where to select primers

**Primers cannot lie in region** e.g. 100-1000bp

Impossible unless excluded region is at the start/end of the input sequence

**Primers must surround region closely**

<table>
<thead>
<tr>
<th>Region</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward primer</td>
<td>477</td>
<td>577</td>
</tr>
<tr>
<td>Reverse primer</td>
<td>745</td>
<td>845</td>
</tr>
</tbody>
</table>

**Also see intron parameters below**

<table>
<thead>
<tr>
<th>Region</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward primer</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Reverse primer</td>
<td>1000</td>
<td>1100</td>
</tr>
</tbody>
</table>

**Primers must lie in region**

<table>
<thead>
<tr>
<th>Region</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward primer</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>Reverse primer</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>
PrimerBLAST offers less but better parameters than Primer3

- **Exon junction containing primers**
  - Exon junction span
  - Exon junction match: Exon at 5' side 7, Exon at 3' side 4

- **Intron spanning primers**
  - Intron inclusion
  - Intron length range
PrimerBLAST performs BLAST during primer design so that it returns only specific primers

Parameters of the BLAST search

It is important that the 3’ end of a primer is specific since extension is done here
SNPs can have large impact on Tm

<table>
<thead>
<tr>
<th>Results</th>
<th>5' mods</th>
<th>Internal Mods</th>
<th>3' mods</th>
<th>Mixed Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EXACT AND SINGLE BASE MISMATCH DNA THERMODYNAMICS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5' ACTTGAATGATGCCCTGGAGA 3'</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3' TGAACTTACTACGGGACC CCT 5'</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SNP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PRIMER</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TEMPLATE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MELTING TEMPERATURES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EXACT MATCH T&lt;sub&gt;M&lt;/sub&gt;:</strong></td>
<td>56.2°C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MISMATCH T&lt;sub&gt;M&lt;/sub&gt;:</strong></td>
<td>51.5°C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DELTA T&lt;sub&gt;M&lt;/sub&gt;:</strong></td>
<td>4.7°C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PERCENT BOUND AT 56.2°C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EXACT MATCH:</strong></td>
<td>50.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MISMATCH:</strong></td>
<td>4.7%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the SNP occurs at the final base at the 3’ end: no PCR product will be formed.
OligoAnalyzer to check primer characteristics

https://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/

Have large impact on Tm (± 10°C)
Similar settings as Primer3 uses

General characteristics e.g. Tm, GC content...

Can primers bind themselves?
Can primers bind each other?

Exercises:
DNA sequence analysis: primer design
Summary on classification

Biological properties can be inferred from sequences
This is done by training prediction tools to recognize sequence patterns
Training is based on a set of sequences with known properties
This process is called classification

Classification of protein modifications
protein secondary structure

You don’t always need classification: properties can be simply calculated
Some characteristics are inferred by comparison to models
Protein sequence analysis tools are gathered on Expasy, EBI, EMBOSS
Never blindly trust a tool’s output !!
Summary on properties based on DNA/RNA sequences

Simple nucleotide sequence analysis based on calculations

Different ways to regulate gene expression
  chromatin accessibility: experimental data
  RNA processing: alignment of miRNA to target
  TFs: models of TFBS

Databases / tools for finding known TF binding sites based on model alignment

Tools for detecting novel TF binding sites in co-regulated genes

Many false positives!

Phylogenetic footprinting

Comparison to background