Bioinformatics Introductory Module

MED-BpM-2013-11-V: 1314

Molecular Mechanisms of Disease
MSc Programme

Period:
Summer 2013

Coordinators:
Dr. Wiljan J.A.J. Hendriks
Dr. Celia van Gelder

Contributors:
Prof.dr. Martijn Huijnen
Clive Michelo
Dr. Joost Martens
Dr. Edwin Lasonder
Prof.Dr. Gert Vriend

An introduction to Bioinformatics

MMD BIM: BIOINFORMATICS INTRODUCTORY MODULE

An introduction to Bioinformatics

Isis code	Course year	Scheduled
MED-BpM-2013-11-V	1st year MSc	Summer 2013

Coordinators
Dr. Wiljan J.A.J. Hendriks (w.hendriks@ncmls.ru.nl), Dr. Celia van Gelder (c.vangelder@cmbi.ru.nl)

This module is meant to prepare upcoming MMD students for Bioinformatics topics that are part of the MMD curriculum. The module is set up in such a way that, with a minimum on computer and internet resources, students will be able to familiarize themselves with the field of bioinformatics and to acquire some hands-on experience. The module is expected to require some 40 hours of reading and practising and *is advised to be completed before the MMD curriculum starts, in September 2013*.

Introduction

The past decades are characterized by a major expansion of biological and molecular data. Sources of information include genome sequences, protein structures, biomolecular interactions, functional pathway links, and 'omics'-type quantitative (expression) data collectively referred to as transcriptomes, proteomes or metabolomes. Understanding and exploiting these huge and expanding data sets is crucial for progress in medical biology and heavily depends on the application of computers that not only enable analysis and calculation but also visualisation and dissemination of the outcome. The field of bioinformatics is essentially covering this integration and exploitation of data resources, and as a result allows scientists to extract *in silico* hypotheses that subsequently can be put to the test experimentally. This module will provide a basic overview of various aspects of bioinformatics. In addition, practical skills will be trained by means of hands-on assignments. All this will prepare students for bioinformatics-based items that form part of the regular MMD curriculum.

Main objectives

Make students familiar with the basics of bioinformatics. Specifically the students will

- gain some basic hands-on experience with Blast searches and Domain databases;
- be introduced to human genome browsers and perform some basic exercises;
- learn about cluster analyses to classify genomic data;
- analyse retrieved protein and DNA sequences and learn about (multiple) sequence alignments;
- be familiarized with amino acid chemistry and principles of protein structure;
- read about state-of-the-art methods that generate data requiring bioinformatics analyses;
- realize that bioinformatics strategies can resolve biomedical questions.

Key words
Databases, Sequence homology, Phylogeny, Protein structure, Microarrays, Proteomics, Expression analysis, Protein-protein interaction, Cluster analysis, protein modelling, networks, bioinformatics.

Literature
The materials for this course consist of literature references that are all publically accessible on the web, via links in this Course Manual. Additional materials will be provided via Blackboard.

Contributors
Prof.dr. Martijn Huijnen, Clive Michelo, dr. Joost Martens, dr. Edwin Lasonder, Prof.dr. Gert Vriend

Table of contents

1 – Why Bioinformatics?

Teaching Bioinformatics – a must

The genome era is characterized by an avalanche of data which have, at first sight, only little biological meaning, such as the complete genomic sequence of a human being or bits of DNA that were detected in marine water. The research field of bioinformatics aims, by means of many different types of analyses, to interpret such 'genomics data' and connect biological and biomedical meaning to it, or to extract hypotheses concerning biological processes from the data which are relevant for the biomedical research that then subsequently can be put to the test experimentally.

We have noticed that students that embark on the MMD master's program have very different backgrounds with respect to Bioinformatics. To warrant a sufficient knowledge level for each student to participate in the various MMD courses and modules, this introductory module has been developed. This so-called Bioinformatics Introductory Module (BIM) thus precedes the MMD Master's program and as such cannot be obligatory or assessed. However, we strongly advise you to complete this module BEFORE SEPTEMBER 2013 to ensure yourself a swift start in Nijmegen.

The module provides an introduction to the types of data that are available, and where these can be retrieved. In addition, the module gives an introduction into the techniques that are used to obtain biomedical relevant information from the huge amount of data concerning e.g. sequences, gene expression and protein-protein interactions. To get a first impression of the field, start by reading the *'gentle overview'* indicated below).

Course Materials

- Achuthsankar S Nair (2007) Computational Biology & Bioinformatics - A gentle Overview. Communications of Computer Society of India.

Set-up of the Module & how to proceed

The complete module is covered in this Handbook. For your convenience some texts, hyperlinks and tools have also been uploaded onto the MMD digital learning environment called Blackboard. We anticipate that you will need some 40 hours to complete the module. And of course you do need internet connection to perform the hands-on exercises. There is no need, however, to install programs onto your computer and also low-speed data transfer should be sufficient. Slow internet data exchange may likely hamper the viewing of a few tutorial movies but as an alternative the corresponding powerpoint presentations will be indicated. All literature that is cited and used throughout the module was taken from 'open access' sources, and thus is fully accessible to all of you.

Next to an introductory and a closing chapter, this manual provides an overview of five main Bioinformatics topics (Chapters 2-6). Additional reading material will be indicated in the relevant sections, separately boxed under the headings "course materials" (as you have experienced above). After you read this additional literature you are well-prepared to embark on the questions part that follows each section on a given topic. This is to let you acquire some hands-on experience with essential bioinformatics tools on the web. Some of those questions are listed as "Additional practice (optional)"; feel free to embark on these extra questions, time permitting, but in essence these won't cover new grounds.

It is the struggle to find answers to the questions that actually provides the insight on how bioinformatics is used in biomedical research and it prepares you well for the (MMD) things to come. Therefore, **please document your answers** to the questions. Preferably collect them (via copy-paste) in a digital file (much like keeping a log or lab journal). After completing a chapter of the module, we hope that you will **upload your log file** onto Blackboard, in the BIM "**Assignment**" area: a dedicated button to upload each part has been prepared. As a response, we will then provide you with our pre-assembled answers to the questions in that chapter, for your comparison.

Since this is the third time the MMD Bioinformatics pre-Module is held we would highly welcome any **feedback** from your side to help us improve. What parts are obsolete, what is lacking, how about the required time to perform the various parts, is it doable with slow internet connections, etc. – any comments on these and other issues will be highly appreciated. For that purpose we have mounted a **Questionnaire** on Blackboard that we hope you will **fill out and subsequently upload** (again in the BIM "Assignment" area there is a dedicated spot to upload the form).

The MMD educational team wishes you a pleasant Bioinformatics training period and we look forward to your log files, feedback and most of all to meeting you in September, here in Nijmegen. Enjoy.

Course Materials

- Bioinformatics pre-Module Course Documents on Blackboard (blackboard.ru.nl).

Some background reading to refresh your memory

Many examples in this Handbook build on information that you may have acquired during your bachelor studies (organisms, diseases, mutations, genes, proteins etc.). In case you would like to look back at such information, it is good to know that you can find many textbooks freely available online; at the NCBI bookshelf (the 'Books' tab in PubMed).

To get 'fully charged' for the upcoming chapters it may be worthwhile to start off with reading some parts that introduce the diversity in genetic information and the tree of life. We chose the famous 'Molecular Biology of the Cell' textbook (Alberts et al, 4th ed.) for this because it is freely available online (see 'Course material' below).

By agreement with the publisher the book is accessible by the search feature but cannot be browsed. For your convenience, we have identified the **links to sections** of the book's chapters that cover relevant paragraphs for this module on Bioinformatics. Simply follow the hyperlinks to get to those pages. Depending on your computer system, you may subsequently be able to use the 'go to' drop-down menu to move to different paragraphs within a given Chapter's section. Alternatively, scroll down until the appropriate paragraph appears on your screen.

Course Material

Some background reading (optional) to refresh your memory:
Molecular Biology of the Cell (4th edition) Alberts et al.

Chapter 1: Cells and Genomes

Section: The diversity of genomes and the tree of life

The Tree of Life Has Three Primary Branches: Bacteria, Archaea, and Eucaryotes
Some Genes Evolve Rapidly; Others Are Highly Conserved
Most Bacteria and Archaea Have 1000–4000 Genes
New Genes Are Generated from Preexisting Genes
Gene Duplications Give Rise to Families of Related Genes Within a Single Cell
The Function of a Gene Can Often Be Deduced from Its Sequence
Mutations Reveal the Functions of Genes

Section: Genetic Information in Eucaryotes

Eukaryotes have hybrid genomes
Eukaryotic genomes are big
The expression levels of all the genes of an organism can be monitored simultaneously
The Mouse Serves as a Model for Mammals
Humans Report on Their Own Peculiarities
We Are All Different in Detail

2 – Sequences, homology, databases

Introduction

In this part the explosion of DNA sequence data and how this can contribute to biomedical research will be discussed. By means of several examples it will be shown how DNA sequence data can be exploited to elucidate the function of proteins that are implicated in diseases and how in that way knowledge can be obtained on the processes underlying diseases. Special emphasis will be on one of the most important parts in bioinformatics: homology prediction. It will be explained how homology is determined and how it can be used to predict protein function.
There are multiple "tools" available on the web to predict homologies, such as Blast and SMART, and to extract protein and other data (www.ncbi.nlm.nih.gov; www.ebi.ac.uk). These will be dealt with by means of questions that you will have to work on. This will provide hands-on experience with these tools and will give you some appreciation of what is 'out there' with respect to molecular life science databases and tools to work with these.

Course Materials

Some background reading (optional) to refresh your memory:
Molecular Biology of the Cell (4th edition) Alberts et al.

Chapter 4: DNA and Chromosomes
Section: Chromosomal DNA and Its Packaging in the Chromatin Fiber
The nucleotide sequence of the human genome shows how genes are arranged in humans
Comparisons between the DNAs of related organisms distinguish conserved and nonconserved Regions of DNA Sequence

Chapter 5: DNA replication, repair, and recombination
Section: The Maintenance of DNA Sequences
Mutation Rates Are Extremely Low
Many Mutations in Proteins Are Deleterious and Are Eliminated by Natural Selection
Low Mutation Rates Are Necessary for Life as We Know It.

Chapter 8: Manipulating proteins, DNA and RNA
Section: Analyzing Protein Structure and Function
Sequence Similarity Can Provide Clues About Protein Function
Section: Studying Gene Expression and Function
Genes Can Be Located by Linkage Analysis

Sequence comparison

The "genome era" was launched with the first endeavours towards the sequencing of complete genomes and it is this type of (DNA sequence) data that contributes a large part of genomic data. DNA sequences, whether derived from complete genomes or representing parts of genomic material from patients containing a suspect disease gene, are not obtained including the labels with information on where the genes are located and what their function is. In the pre-genomic era the DNA sequences and the corresponding predicted protein sequences often published together with the experimental data on the (protein) sequence. Whole genome sequences, however, are nowadays published without further experimental data, not even for a single gene. The annotation of a genome ("which DNA part corresponds to which gene?" and "where are the genes located and what do they do?") thus has to be done using other information sources. Gene prediction ("where do we find the genes in this genome?") is extremely difficult, especially for multicellular eukaryotes that make use of genes with a so-called exon-intron build-up (e.g. mammals), and consequently far from complete. In addition, phenomena like alternative splicing or alternative promoter use make that one no longer can assume that "one gene encodes for one protein". In fact, the majority of protein-coding genes in human actually encodes multiple different protein isoforms.

And even if we would know which genomic DNA parts are protein-coding then still we are left with the question what the function is for each of these proteins. Predicting protein functions is predominantly based on homology: by searching in databases for proteins that resemble the protein of interest – and therefore are expected to be evolutionary related – we may infer the type of function of our protein of interest from the functions ascribed to its lookalikes. Proteins that are related are called homologs. To determine whether proteins bear a resemblance that exceeds what one would expect by chance for two random proteins such search algorithms make use of randomization techniques. The protein in a database that resembles our protein of interest most is homologous to our protein only if the resemblance exceeds that what is to be expected from searching a database consisting of 'random sequence' proteins. The tool that is by far the most widely used for such searches is called BLAST (Basic Local Alignment Sequence comparison Tool). It compares the protein of interest that is used in the search (the "query") with all sequences in the databases and then reports the ones that bear the highest resemblance. This degree of resemblance is judged on the basis of several scores:

1) **identity**: the number and percentage of identical amino acid residues in the two sequences
2) **similarity**: how much the sequences would look alike if also "conservative" amino acid changes would be counted as identical. This is based on the observation that some amino acids are more equal than others. Leucine and Isoleucine are structurally much more related than for example Leucine and Cysteine. Thus, if in an alignment of two proteins at position 1 the one sequence has a Leucine and the other contains an Isoleucine, this would be scored as a higher similarity than in the case that alignment position 1 harbours a Leucine in the one and a Cysteine in the other protein.
3) **score**: this is a combination of the sequence similarity and the length of the piece of the query sequence that resembled the sequence that was found by BLAST. The higher the similarity and the longer the stretch, the higher the score will be.
   NB: the sequences do not have to match over the complete (query) length. During evolution genes do not remain the same; pieces may be added (gene fusion), other parts may be deleted or genes may split (gene fission). BLAST bears this is mind and therefore Blast is a "Local Alignment" method that searches for pieces of sequences that bear resemblance, instead of looking for complete overlap of resembling sequences.
4) **E-value** (Expectation value; the most important score). The E-value represents the number of sequences with that similarity score that would be expected (hence the term Expectation value) to result from a BLAST search in a 'random sequence' database. An E-value of 1 means that a search in a random database would have yielded one sequence hit with a similarity score equal to or higher than the one obtained in the actual database. Such an E-value of 1 is thus not very significant. An E-value can be higher than 1 (in contrast to the P-value). An E-value of 10 means that one would expect to find 10 sequences resembling the query from any database of that size. In general only E-values that are below 0.01-0.001 are regarded as significant: the chance to find such a resemblance by sheer luck is negligible.

Nuts and bolts of sequence comparison

Wrong annotations

By far not everything that is added to databases is correct. Sequences might be added with erroneous annotations, for example due to mistakes in the interpretation of experimental results. This type of errors may start living a life of their own: after all, new sequences that are homologous to the one with the erroneous annotation may well adopt this error. And sometimes such 'contagious wrong annotations' remain in the databases even long after the original error has been corrected. The only way to be sure that the annotation is correct is to trace back the source of the annotation. Some databases have a better quality (contain less erroneous annotations) than others. SWISSPROT (www.uniprot.org) is generally considered the best database for protein annotations. Such a "curated database", in which all data are checked manually, contain however much less data than a general database like Genbank, in which all DNA and protein sequences are entered directly. Sometimes it is unavoidable to use databases that have a lower quality, simply because the larger amount of sequence information is required.

False positives and False negatives
Although one may safely assume that a significant E-value (e.g. 0.001) implies that the sequences are homologous, a non-significant E-value does not automatically mean that there is no homology. A method like Blast detects about 30% of the homology relations between proteins. Profile-based methods such as domain databases (see below) are much more sensitive but even these cannot retrieve more than 70% of all homology relations with an error rate of 1 percent.

Low complexity regions
Sequences are homologous if they share a single common ancestor in a very distant past (divergent evolution), and not if they resemble each other by chance. Some sequences contain stretches of amino acids that are rather simple in composition, such as Proline- and Glutamine-rich sequences ("PPPPPPPPPQQQQQQQQQ"). Transmembrane parts of proteins, for example, are rich in hydrophobic amino acid residues. One would not like to mistakenly call proteins homologous just because the independent evolutionary pruning and grafting of their sequences resulted in an amino acid composition in certain regions of the proteins that has become similar (convergent evolution). Similarly, you would not like to call dolphins and fish evolutionary related just because they have a similar shape. The regions that are dominated by prevalent amino acids, displaying a simple composition, are called "low complexity regions". These can be filtered out in BLAST using a so-called "low complexity" filter that can be selected on the web tool.

Sequence comparison: domain databases

As stated above genes do not evolve as "units". Multiple genes can fuse into a single gene during the course of evolution and, likewise, genes can disassemble into smaller, but still functional, genes. The smallest functional units of proteins are the protein domains. These can be clearly discerned structurally (in the 3D structure of the protein they represent the individual 'building modules'), functionally (each domain has its own molecular function; e.g. catalyzing a certain reaction or mediating a particular protein interaction) and evolutionary (they can occur independently in other proteins and / or fused within one protein) as separate units. Bioinformaticians and structural biologists have put much effort in the identification of all those protein domains, but for sure there are still much more domains to be recognized.

Sequences that are composed of similar domains consequently are homologous. Domains can be recognized at the sequence level with the aid of protein domain databases. In such databases the sequence profiles of the various domains are stored. Sequence profiles are generalized descriptions of a number of homologous, aligned sequences. They contain the information on how often the different amino acids are found at the various positions in the domain. Such databases render the homology search at least twofold more sensitive than the pairwise comparison of plain sequences as done by BLAST. Consequently, it detects more homologous relationships. In the pairwise comparison there is no information available to discriminate which amino acid would be "typical" for a given spot in the sequence. Instead, such algorithms work with "general amino acid similarity matrices".

Examples of domain databases are "SMART" (Simple Modular Architecture Research Tool, smart.embl-heidelberg.de), PFAM (Protein FAMily, www.sanger.ac.uk/Software/Pfam), and Interpro (www.ebi.ac.uk/interpro). These databases contain also functional information on the protein domains ("what does it do"). When performing a Blast search one would get, in first instance, also results from a comparison with a particular domain database, the Conserved Domain Database (CDD). However, this domain database does not provide functional information on the protein domains.

Homology prediction and protein function

There are many ways to describe the function of a protein. For example, a protein is functioning in a certain metabolic route like glycolysis or the citric acid cycle, or in a signalling pathway like the MAP kinase cascade. One may also describe the specific subcellular localization of a protein, for example that it resides in the cytosol, the cell nucleus or the mitochondria. The description may also address its molecular function, e.g. a "glyceraldehyde-3-dehydrogenase" (from the glycolysis chain) or an aconitase (part of the citric acid cycle). And finally one may link a gene or protein to a disease or phenotype: f.e. the frataxin protein of which a too low an expression will lead to the disease known as Friedreich's ataxia, or the Presenilin protein that is involved in Alzheimer's disease.

With respect to the prediction of protein function on the basis of homology it is important to realize that proteins that are homologous will have the same molecular function. After all, the actions of a protein at the molecular level – catalyzing a reaction, phosphorylating another protein, binding to a specific phospholipid – is the part of its function that helped shaping it during evolution and is therefore best conserved. Much better than for example the protein's location in the cell (in trypanosomas, that cause a sleeping disease in humans, part of glycolysis takes place in a special peroxisome-like organelle whereas in humans and yeast all this is occurring in the cytosol) or the exact pathway it is in (protein kinases are all homologous to each other but they act in many different pathways and phosphorylate many different proteins). The different types of functions that can be attributed to proteins (molecular functions, biological processes and subcellular location) have been formalised in so-called Gene Ontologies (GO; www.geneontology.org).

Hands-on Exercises

Question 2.1

The aim of this task is to familiarise you with determining homologies using BLAST. By answering the questions you will learn 'where to click', how to find information and how that information should be interpreted.

In a terrorist lab the following DNA sequence has been PCR-amplified and sequenced:

```
gtgaaaaagactttaattacagggttattggttacagcggtatctacgagttgttttattcctgtaagcgc
ttacgctaaggaggggcaaacagaagtgaaaacagtatatgcacaaaatgtaattgctccaaatacattat
cgaattcaattagaatgttaggatcacaatcaccacttatacaagcatatggattagttattttacaacag
ccagacattaaggtaaacgcgatgagtagtttgacgaatcatcaaaaatttgcaaaggcaaatgtaagaga
gtggattgatgaatataatccgaagttaatcgacttaaatcaagagatgatgaggtatagtactagattta
atagctattatagtaagctttatgaactagcagggaacgtaaatgaggatgaacaagcaaaagcagatttt
acaaatgcatatggaaaattacaattgcaagtacaaagcatccaagagagtatggagcaagatttattaga
gttaaatcgattcaaaacggtattagataaagatagtagcaacttatcaattaaagctgatgaagcaataa
aaacactacaaggatcaagtggagatattgtgaaattaagagaagatat
```

2.1a) From which species does the sequence originate? Provide the alternative name that is commonly used for this species?

Hint: to answer this question you may investigate whether it matches a sequence that has already been determined for a given species and that has been deposited in the nucleotide database. Thus, perform a "Blast" search: go to the NCBI web page (www.ncbi.nlm.nih.gov), select "Blast", and pick the right "type" of Blast (mark that a DNA sequence must be compared to other DNA sequences). Copy and paste the sequence in the 'Query Sequence' window and press "Blast". If the resulting species name does not ring a bell then enter it as a search term in PubMed (www.ncbi.nlm.nih.gov/pubmed).

2.1b) Does the DNA sequence encode a protein? If yes, what protein?

Hint: To answer this question one actually would need to 'translate' the DNA sequence into six potential protein sequences, using all possible reading frames (3 potential reading frames if the given DNA strand itself is the coding one, and 3 potential reading frames that could be used on the complementary DNA strand that is not shown), and subsequently compare these six with all known proteins. There is a program, however, that performs these operations in one go: "BlastX". It translates all six potential reading frames and compares the obtained proteins against a protein database. BlastX can be found on the blast.ncbi.nlm.nih.gov/Blast.cgi site also.

2.1c) Does the DNA sequence encode the full-length protein?

Hint: Compare the length of the complete protein and the length of the alignment between the DNA and the protein that the program made. The size of the full-length protein is listed in the results output of the Blast search: next to the alignment it says: "length="

2.1d) Are there any other species that encode for protein sequences that are homologous to your query sequence? Provide at least five other species.

Hint: Homologies are best determined at the protein level because the functional capacity of the protein sequence is most directly assessed in evolutionary selection processes. The codons responsible for the amino acids in the protein are only amenable to selective pressure at a more indirect level (tRNA availability, di- or tri-nucleotide stabilities, CpG orders etc.). Thus, this time use "BlastP" and search directly at the protein level. In the BlastP search output one can click on "taxonomy browser", which provides an overview of the species in which significant hits were obtained.

Question 2.2
The aim of this question is to introduce the protein domain databases and to demonstrate that protein domains are functionally and evolutionary independent and that the domain organisation of proteins bears relevance for genetic diseases. We will work with the following sequence:

```
msqstqtnef lspevfqhiw dfleqpicsv qpidlnfvde psedgatnki eismdcirmq
dsdlsdpmwp qytnlgllns mdqqiqngss stspyntdha qnsvtapspy aqpsstfdal
spspaipsnt dypgphsfdv sfqqsstaks atwtystelk klycqiaktc piqikvmtpp
pqgaviramp vykkaehvte vvkrcpnhel srefnegqia ppshlirveg nshaqyvedp
itgrqsvlvp yeppqvgtef ttvlynfmcn sscvggmnrr piliivtlet rdgqvlgrrc
fearicacpg rdrkadedsi rkqqvsdstk ngdgtkrpfr qnthgiqmts ikkrrspdde
llylpvrgre tyemllkike slelmqylpq htietyrqqq qqhqhllqk qtsiqspssy
gnssplnkm nsmnklpsvs qlinpqqrna ltpttipdgm ganipmmgth mpmagdmngl
sptqalpppl smpstshctp pppyptdcsi vsflarlgcs scldyfttqg lttiyqiehy
smddlaslki peqfrhaiwk gildhrqlhe fsspshllrt pssastvsvg ssetrgervi
davrftlrqt isfpprdewn dfnfdmdarr nkqqrikeeg e
```

2.2a) Which protein domains are present in this sequence? Make a drawing – use the full page width. You will need this drawing later on.

Hint: to find this out one performs a sequence comparison using a protein domain database such as SMART (smart.embl-heidelberg.de). On the SMART page, please use the option "PFAM".

2.2b) What are the positions of the various domains in the protein sequence? Indicate these in your drawing.

Hint: Move with your cursor over the image displaying the domain organisation of the protein.

2.2c) What are the molecular functions of the distinct domains? For instance, do they interact with DNA or with other proteins?

Hint: Click on a segment in the display to get more information on that particular protein domain. Sometimes one needs to click more than once to reach the proper information: the description of the molecular function of the protein domain.

2.2d) For the current protein various mutations have been described that are connected to disease states like Ectrodactyly, ectodermal dysplasia, cleft lip/palate (EEC) syndrome, the Ankyloblepharon-ectodermal dysplasia-cleft lip/palate (AEC) syndrome and Acro-Dermato-Ungual-Lacrimal-Tooth (ADULT) syndrome.

Below you will find part of a table that has been published (Rinne et al, Am. J. of Med. Genet. 2006) that displays the mutations found in the protein and the corresponding phenotypes/syndromes they cause. (L162P means that the Leucine residue that is normally present at position 162 in the protein was mutated into a Proline; R204L/Q/W means that Arginine-204 is changed into an L, a Q (Glutamine) or a W (Tryptophan) residue).

Please check for the given mutations whether they reside within one of the identified protein domains. Redraw the cartoon provided by SMART in your journal and indicate the positions of the various syndrome-linked mutations. You don't need to indicate the type of mutation; just indicate the location and the corresponding syndrome type in the cartoon.

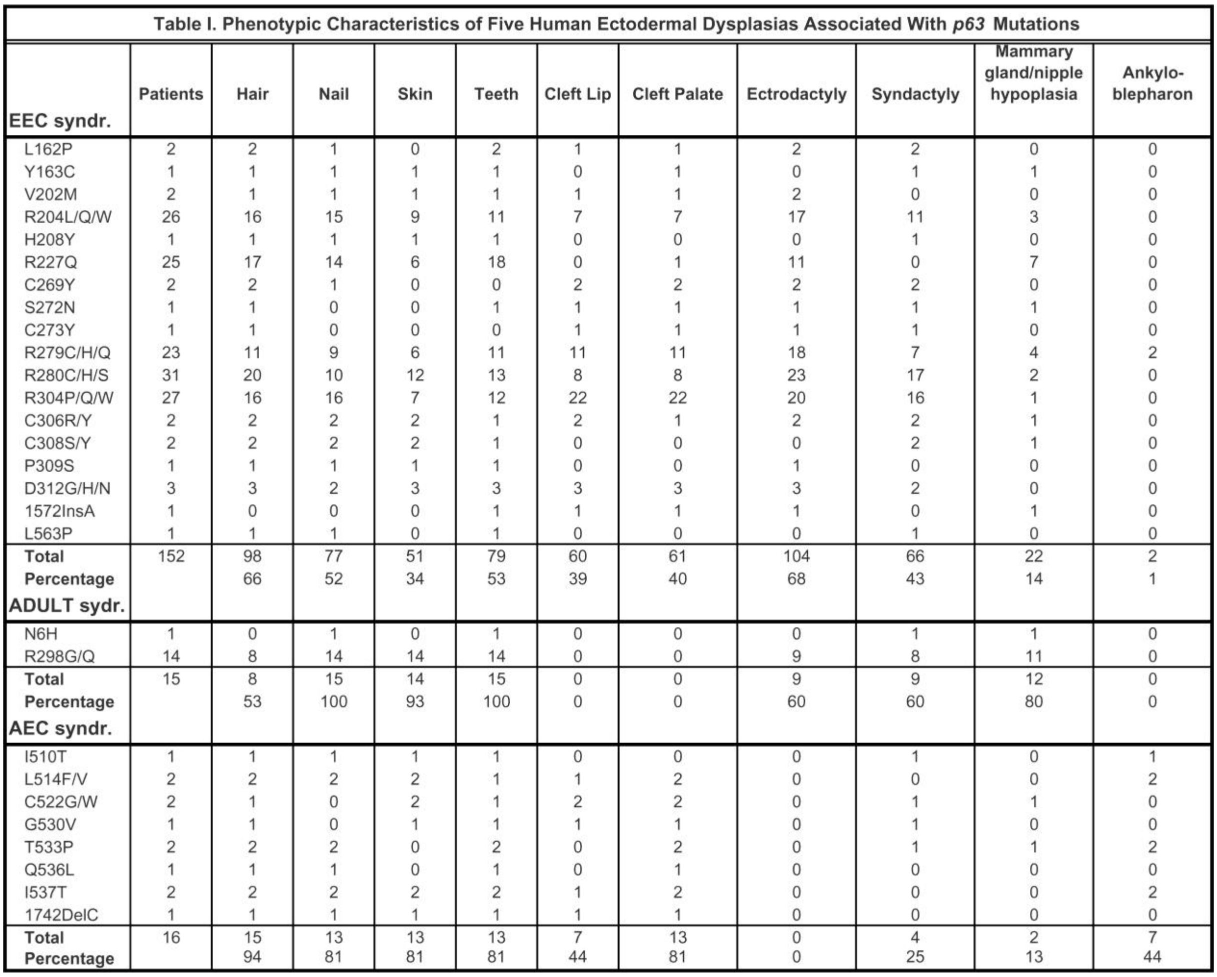


Table I. Phenotypic Characteristics of Five Human Ectodermal Dysplasias Associated With *p63* Mutations

EEC syndr.	Patients	Hair	Nail	Skin	Teeth	Cleft Lip	Cleft Palate	Ectrodactyly	Syndactyly	Mammary gland/nipple hypoplasia	Ankylo-blepharon
L162P	2	2	1	0	2	1	1	2	2	0	0
Y163C	1	1	1	1	1	0	1	0	1	1	0
V202M	2	1	1	1	1	1	1	2	0	0	0
R204L/Q/W	26	16	15	9	11	7	7	17	11	3	0
H208Y	1	1	1	1	1	0	0	0	1	0	0
R227Q	25	17	14	6	18	0	1	11	0	7	0
C269Y	2	2	1	0	0	2	2	2	2	0	0
S272N	1	1	0	0	1	1	1	1	1	1	0
C273Y	1	1	0	0	0	1	1	1	1	0	0
R279C/H/Q	23	11	9	6	11	11	11	18	7	4	2
R280C/H/S	31	20	10	12	13	8	8	23	17	2	0
R304P/Q/W	27	16	16	7	12	22	22	20	16	1	0
C306R/Y	2	2	2	2	1	2	1	2	2	1	0
C308S/Y	2	2	2	2	1	0	0	0	2	1	0
P309S	1	1	1	1	1	0	0	1	0	0	0
D312G/H/N	3	3	2	3	3	3	3	3	2	0	0
1572InsA	1	0	0	0	1	1	1	1	0	1	0
L563P	1	1	1	0	1	0	0	0	1	0	0
Total	152	98	77	51	79	60	61	104	66	22	2
Percentage		66	52	34	53	39	40	68	43	14	1
ADULT sydr.											
N6H	1	0	1	0	1	0	0	0	1	1	0
R298G/Q	14	8	14	14	14	0	0	9	8	11	0
Total	15	8	15	14	15	0	0	9	9	12	0
Percentage		53	100	93	100	0	0	60	60	80	0
AEC syndr.											
I510T	1	1	1	1	1	0	0	0	1	0	1
L514F/V	2	2	2	2	1	1	2	0	0	0	2
C522G/W	2	1	0	2	1	2	2	0	1	1	0
G530V	1	1	0	1	1	1	1	0	1	0	0
T533P	2	2	2	0	2	0	2	0	1	1	2
Q536L	1	1	1	0	1	0	1	0	0	0	0
I537T	2	2	2	2	2	1	2	0	0	0	2
1742DelC	1	1	1	1	1	1	1	0	0	0	0
Total	16	15	13	13	13	7	13	0	4	2	7
Percentage		94	81	81	81	44	81	0	25	13	44

2.2e) Is there an apparent correlation between a phenotype and the protein domain that harbours the mutation? Please specify for which syndrome this is the case (or not).

2.2f) Are there any homologous proteins present in the human genome?

Hint: The databases are filled with variants of this protein that result from the use of alternative transcription start sites or alternative splicing events. This is interesting on its own, but now let's focus on homologous proteins that are encoded by genes that differ from the current one. To help in this once could select, on the Blast page in the "Database" list, "Swissprot". This protein database does not contain all the splice variants.

2.2g) Do these homologs display a similar domain build-up?

2.2h) What is the level of sequence identity for the various protein domains when comparing the distinct homologs (i.e. the sequence identity when comparing the first domain in two homologs, when comparing the second domain, etc.) and what is the level of sequence identity in de regions in between the various domains (e.g. the 'spacer region' in between the first and the second domain, or the stretch in between the second and third domain)?

2.2i) What are the most conserved parts in the p63 protein: the domains that you identified or the sequences in between?

Hint: To determine the level of sequence identity per domain the SMART search is performed first. By clicking on the domain in the cartoon the sequence that corresponds to the protein domain is obtained. This then can be entered in a Blast search using again the Swissprot database and selecting for "Homo sapiens" in the "Organism" field, to filter for only the relevant output.

Question 2.3
The purpose of this question is to familiarize you with homology detection tools like Blast and domain databases, and to show how to search in protein databases. Hints on e.g. the available functional information for protein domains have been provided in question 2.

2.3a) There are at least four proteins (of which you may have heard of) that are involved in Alzheimers disease: Amyloid Precursor Protein (APP), Presenilin 1, Presenilin 2, and Apolipoprotein E. Are these proteins homologous? Use both Blast and a domain database to answer this question.

Hint: Protein sequences can be found at the NCBI site (www.ncbi.nlm.nih.gov). Select for "protein" as the "search" option, and provide a description of the protein as input (e.g. "Amyloid Precursor Protein (APP)" and "Alzheimer"). Please note that also incomplete protein sequences may be present in the database. When using Blast please make sure to only search the human genome, to limit the output to relevant items.
Hint: Instead of performing Blast searches you may also compare the proteins against a domain database. After all, proteins with similar domains are homologous.

2.3b) With respect to the results on question 3a, do you expect that the four proteins are involved in a similar manner in the mechanism of the disease?

2.3c) What is known about the function of the presenilin domain: where in the cell is it localized? Is there any information on what it is doing there?

Hint: the domain databases contain really a lot of information concerning the function of the protein domains.

Question 2.4
The purpose of this question is to introduce the phenomenon of "low complexity regions" that really complicate the detection of homology.

```
>Sequence A
matleklmka feslksfqqq qqqqqqqqqq qqqqqqqqqq pppppppppp pqlpqpppqa
qpllpqpqpp ppppppppgp avaeeplhrp kkelsatkkd rvnhcltice nivaqsvrns
```

2.4a) From which protein does this partial sequence originate, and from what species?

2.4b) What is the disease that is associated with this protein?

2.4c) Which type of mutation is causative of the disease? You may have heard about this type of mutation before, for example in the context of the Fragile X syndrome (where the mutation is present in the FMR1 gene).

2.4d) Compare in the output the alignment of your query sequence with that of its homolog in Gallus gallus (chicken). Where in the protein sequences are the main differences between the two species located? Compare this to what is happening in the case of Huntington's disease that is caused by an expansion of a trinucleotide CAG, encoding glutamine (Q) in the Huntingtin protein.

Hint: search in PubMed.

2.4e) The part of the sequence that contains many Qs (Glutamine) and Ps (Proline) is called a low-complexity region. Please use the "low-complexity filter" option by clicking on "Algorithm parameters" at the bottom of the Blast page and subsequently selecting "low complexity regions" by ticking the corresponding box. Now repeat the Blast search. The characters from the PPPPQQQQ regions will now look different in the alignment resulting from the Blast search. They have been filtered out while searching for homologs.
Do you now retrieve different sequences than with the "low-complexity filter" in the "off" position? If yes, which novel sequences did you obtain?

Additional practice (optional)

Question 2.5

This question partly resembles a previous one (1), so you may reuse the hints. The following DNA sequence was obtained from a patient suffering from an infection that did not respond to the administered antibiotics:

```
attctttcat tttttagtgt attaaatgaa atggttttaa atgtttcttt acctgatatt
gcaaatcatt ttaatactac tcctggaatt acaaactggg taaacactgc atatatgtta
actttttcga taggaacagc agtatatgga aaattatctg attatataaa tataaaaaaa
ttgttaatta ttggtattag tttgagctgt cttggttcat tgattgcttt tattggtcac
aatcactttt ttattttgat ttttggtagg ttagtacaag gagtaggatc tgctgcattc
ccttcactga ttatggtggt tgtagctaga aatattacaa gaaaaaaaca aggcaaagcc
tttggtttta taggatcaat tgtagcttta ggtgaagggt taggtccttc aataggggga
ataatagcac attatattca ttggtcttac ctacttatac ttcctatgat tacaatagta
```

2.5a) From which species does the sequence originate?

2.5b) Does the DNA sequence encode a protein? If yes, what protein?

2.5c) Does the DNA sequence encode the full-length protein?

2.5d) Is this protein involved in antibiotics resistance? If yes, against which antibiotic?

2.5e) What is the molecular function of the protein and how does it cause the resistance against the antibiotic?

Hint: In the protein "record" at the NCBI site there is a lot of information on the protein's function. Also many links to articles describing research on the protein are provided.

2.5f) Does a homologous protein exist in humans? If yes, in which cell organelle does the homologue reside?

Hint: Perform a Blast search while selecting in the "Search Set" the Organism "Homo sapiens". Then focus in the output list on a hit that represents a protein from the SWISSPROT database (their names all end with "_HUMAN") and read the accompanying description of the protein.

2.5g) Are the human protein and your query sequence (that you used for your BlastP search) homologous over the entire length of the protein?

Hint: Compare start and end points of the alignment with the lengths of the proteins involved.

Question 2.6

The article indicated below describes how *Staphylococcus aureus* developed resistance in a host against the antibiotic Vancomycin. This paper is a nice example of the use of multiple techniques and Bioinformatics concepts within a biomedical context. Answer the following questions, all dealing with this article.

Some definitions that will be of use:
Nonsynonymous mutations = mutations that change a protein sequence.
Synonymous mutations = mutations that do not change the encoded protein but do change the mRNA sequence, e.g. because they concern the third ("degenerate") position in a codon triplet.
Acquired genetic elements = Genes that ended up in a genome via so-called Horizontal Gene Transfer
Isogenic = having the same genetic background

Course Material

Literature:

- Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED, Tomasz A (2007) Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. Proceedings National Academy of Sciences, 104, 9451-9456.

2.6a) Horizontal gene transfer (HGT), in which genes of one organism are taken up by another, is playing a major role in the spread of drug resistance from one strain to another. Is HGT at stake in the S. aureus drug resistance development case in this article? Why do you think this is / is not remarkable with respect to drug resistance development within a single host strain?

2.6b) Why is it necessary to sequence multiple genomes from a single host in order to monitor the development of drug resistance? Why not simply compare the genome of a non-resistant strain with that of a resistant one?

2.6c) How many of the mutations detected by the authors are "synonymous" and how many are "non-synonymous"? Assuming that about 1/3 of the mutations would be synonymous, is the fraction of "non-synonymous" mutations that was observed more or less than you might expect? What could be the explanation for this apparent discrepancy?

2.6d) The authors found much more mutations in the RpoB gene of S.aureus than has been reported for other rifampin resistant S.aureus strains. How do they explain their findings.

2.6e) The authors found mutations in the gene SA1702. This gene is a so-called hypothetical gene; nothing is known about its function ("unknown function"). Despite this lack of data they come up with the suggestion that it is implicated in drug resistance. What argument in favour of this theory do they use?

2.6f) De cause of resistance against Daptomycin is not known. Which gene mutations that are reported in this article appear responsible for the resistance against Daptomycin in S.aureus?

2.6g) In paragraph "between JH5 and JH6" two types of mutations are mentioned that lead to a dramatic change in the protein. Which are they?

2.6h) In addition to mutations in protein coding regions the authors also found mutations in non-coding regions, in between genes. How could such mutations influence the phenotype?

2.6i) In table 2 in the article a number of proteins are stated to be involved in “transcription regulation”. For which of these genes is it possible, based on their domain composition, to corroborate a role in gene transcription regulation?

2.6j) In table 2 it is stated that protein SA1129 contains an RNA binding motif. Which of the domains in SA1129 is supposed to bind RNA (or DNA)? Is there any other protein domain within SA1129 that relates to DNA in some way?

Hint: analyse the domain composition of SA1129 and read and interpret the information on these domains using a protein domain database.

3 – Genomics data, pathway databases and the prediction of protein interactions

Introduction

In part 3 databases will be discussed in which the pathways per sequenced genome can be found. Furthermore, other types of genomics data than just sequences will be dealt with: protein-protein interaction data, gene expression data etc. Such data types can be used in combination with genome information to predict functional relationships between proteins and thus, indirectly, the functions of proteins.
In this part we will also deal with clustering of genomic data. Cluster analysis is one of the most important methods that is used to obtain insight into the meaning of genomics data. In this part it will be demonstrated how one type of cluster analysis, the so-called agglomerative cluster analysis, actually works. We will also explain how genomics data is being used in the diagnosis of breast cancer. Finally, some hands-on exercises will introduce you to the web server STRING that can be used to detect functional relationship between proteins.

Course Material

Some background reading (optional) to refresh your memory:
Molecular Biology of the Cell (4th edition) Alberts et al.

Chapter 3: Proteins
- Section: The shape and structure of proteins
  - The Protein Domain Is a Fundamental Unit of Organization
  - Proteins Can Be Classified into Many Families
  - Sequence Homology Searches Can Identify Close Relatives
  - Computational Methods Allow Amino Acid Sequences to Be Threaded into Known Protein Folds
  - Some Protein Domains, Called Modules, Form Parts of Many Different Proteins
  - The Human Genome Encodes a Complex Set of Proteins, Revealing Much That Remains Unknown
- Section: Protein Function
  - Proteins often form large complexes that function as protein machines
  - A Complex Network of Protein Interactions Underlies Cell Function

Chapter 8: Manipulating proteins, DNA and RNA
- Section: Analyzing Protein Structure and Function
  - Affinity Chromatography and Immunoprecipitation Allow Identification of Associated Proteins
  - Protein-Protein Interactions Can Be Identified by Use of the Two-Hybrid System
- Section: Studying Gene Expression and Function
  - Microarrays Monitor the Expression of Thousands of Genes at Once
  - Large Collections of Tagged Knockouts Provide a Tool for Examining the Function of Every Gene in an Organism

Chapter 15: Cell Communication
- Section: General Principles of Cell Communication
  - Interactions Between Intracellular Signaling Proteins Are Mediated by Modular Binding Domains

Predicting protein-protein interactions

Due to the progress in molecular techniques, experimentally unravelling the functions of genes is tremendously lagging behind the speed in which sequences are acquired and analyzed nowadays. This has triggered the development of methods that can be used to determine gene functions on a genome-wide scale. In addition to the new techniques also new methods have been developed to extract more information about the function of proteins from genomic sequence data. In general one could say that the genomic techniques and computational methods that are applied to the data not so much address the function of the protein itself, as is the case with homology searches, but more the interactions between proteins. This means that on the basis of these methods on cannot directly predict "protein A does this" or "protein A catalyses reaction B" but more something like "protein A interacts with protein B and because we know protein B participates in pathway C we predict that protein A will also exert its effect in pathway C". Therefore, this method is sometimes referred to as "guilt by association" method: because when protein A can be associated with protein B there is indication that A will participate in the same processes as B does. Although this may all sound rather indirect these methods have turned out to be quite successful, especially if they are combined with function predictions based on homologies. If one can predict in which process a particular protein will participate (via guilt by association methods) and what type of activity it will display (by means of homology detection) this allows the rather detailed design of a 'wet lab' experiment that can put the hypothesis to the test.

But even if there are no homology data, then still the guilt by association methods are quite useful. There are many proteins encoded by the human genome for which we have absolutely no clue what they might be doing. Without hypotheses on the probable function of such proteins it becomes quite hard to design experiments to probe for any function. In addition such methods can also be directly exploited to find genes that are associated with (heterogeneous) genetic diseases. It has become clear, for example, that proteins that have been associated via genomics data often appear to be involved in the same disease. Since they play a role in the same biological process, detrimental changes in any of the involved genes will affect this process in the body and hence all cause a similar disease phenotype.

As compared to DNA sequence determination, which generates extremely reliable data (>99,99%), the "functional genomics" methods yields data that are much less reliable; many of the interactions that are inferred or predicted appear to be not biologically relevant ("false positives"), and some of the biologically relevant interactions are not detected in these bioinformatics screens ("false negatives"). That is why one should consider these type of bioinformatics analyses as "hypothesis-generating": the results from such large-scale studies need to be put to the test in specific, small-scale, accurate experiments.

Guilt by association methods based on genome comparisons

There are several different "guilt by association" methods. A first set of methods makes use of the genome sequences itself, the so-called "genomic context" methods. Genomic context methods exploit the associations of genes within the genomes.

Gene fusion:
The gene fusion method assumes that if two genes appear fused in a genome this can be taken as evidence that the encoded proteins act together / interact, also in genomes where these genes are not fused. Thus, if we study a gene (A) with a hypothetical function and we detect a homolog of that gene to be fused with another gene (B) of which the pathway it is involved in is known, it is likely that gene A will also function in that pathway.

Gene order conservation:
This concept exploits the genome organization of prokaryotes. In prokaryotes the genes that are involved in the same biological process are often found next to each other, in clusters on the genome that are called operons. These genes are transcribed as a group and transcripts are often multi-cistronic messenger RNAs: translation results in multiple different proteins of which the encoding open reading frames (ORFs) were tandemly oriented on a single mRNA. Comparative genome analyses reveal that the more frequent genes / ORFs are residing next to each other in genomes (thus, the gene organization appears evolutionary conserved) the more likely it is that these genes / ORFs are involved in the same process.

Gene co-occurrence:
Proteins do not necessarily have to be encoded on genomic regions that are adjacent to be involved in the same process. But of course they should be present within the same species, thus both be encoded in the same genome. Gene co-occurrence methods are based on this thought. The idea is that if one would like to predict the function of protein A, one should look for another gene (B) that nicely occurs in genomes that have the gene for protein A, and that is absent in genomes that do not have gene A. This can be taken as evidence that A and B act together.

Guilt by association methods based on experimental data

Gene co-expression:
The first technique from the 'post-genomic era' is the measurement of the expression levels of many genes in one go through the use of so-called (micro-)arrays. By means of hybridization (of an mRNA-derived cDNA population as probe onto micro-array chips that contain gridded sequences representing a large number of different genes) one can perform a large-scale, parallel assessment of the relative abundance of each mRNA species in a 'total RNA population' for a given tissue, developmental stage or treatment condition. If genes display a similar expression profile – f.e. they all display high levels under certain conditions and all switch to low expression levels when conditions are changed – this can be viewed as an indication that these genes are involved in similar processes.

Protein interactions:
A quite direct form of guilt-by-association is of course an analysis based on protein-protein interactions. Various different tests have been developed that can measure on a genomic scale the physical interaction between proteins. For example, yeast two-hybrid screens and (tandem) affinity purification screens. For the latter, the protein of interest is fused to an affinity tag (via recombinant DNA techniques) so that it can be purified using appropriate "column material" together with its associating proteins. The individual proteins that co-purify in the complex with the tagged protein are then identified by means of Mass Spectrometry technology.

Comparative analysis of genomics data

The large-scale experimental methods to determine protein interactions or to predict interactions on the basis of comparative genomics analyses are still not very accurate. In general, interactions that are observed or predicted by multiple different approaches will be much more reliable than those that result from just one method. If f.e. two proteins interact in the yeast two-hybrid screen and in addition are co-expressed in many tissues and developmental stages, the chance that this concerns a meaningful interaction is much bigger than if it was based on just one of these findings.

A Guilt-by-association search engine: STRING.
Just as BLAST can be used to search for homologs in sequence databases, search engines have been devised that can search for proteins that are, one way or another, associated with your protein of interest. One of those databases / search engines is STRING (string.embl.de). You can simply enter a protein sequence and then it will look whether your protein or any ortholog of your protein is interacting with other proteins in one of the types of genomic data that have been mentioned previously. STRING also searches in abstracts of all published literature whether these is possibly some relevant mentioning of an interaction of a protein with your favourite one.

A pathway database: KEGG
Starting with a genome sequence one can predict which proteins do appear in the corresponding organism, and (based on available knowledge on the physiological role of many homologous protein) how those proteins are functionally related, for example in metabolic routes. There are a few databases where for all sequenced genomes one can retrieve the metabolic routes that have been identified. One of the best known databases of this type is the Kyoto Encyclopedia of Genomes and of Genes (KEGG; www.genome.jp/kegg). Here one can search which component of known pathways (e.g. glycolysis) are present in human or in species like *Escherichia coli*, or *Staphylococcus aureus*.

Hands-on Exercises

Question 3.1

You will now perform, perhaps just once in your life, a clustering by hand and not using the computer. This will provide you with some insight in the techniques that are used. The clustering of data is not a prescribed procedure; it very much depends on the questions you would like to answer how exactly you should cluster data!

Below you will find a simple hypothetical example of the expression levels of four genes as measured under four different circumstances / conditions. These conditions for example could be four different tissues, four different developmental stages, four different time points in the cell cycle or two healthy control and two carcinoma tissues. For the current assignment you have to cluster the tissues and the genes. In doing so you will stepwise go through the typical phases that also are used in the computer-based clustering procedures (but then these remain hidden for the investigator). The expression levels of the genes are relative expression levels: 8 means that the gene is expressed 8 times higher than the reference gene.

	Tissue I	Tissue II	Tissue III	Tissue IV
Gene A	4	8	¼	¼
Gene B	4	1/8	1/8	4
Gene C	4	4	¼	¼
Gene D	8	¼	8	8

3.1a) Transform the data onto a logarithmic scale (log base 2, thus ¼ becomes -2, 8 turns into 3, ½ equals -1, 1 transforms in 0, etc.). This type of transformations is usually done to prevent that extreme values, of f.e. one particular gene under one particular condition, will become highly influential in the whole clustering procedure.

	Tissue I	Tissue II	Tissue III	Tissue IV
Gene A				
Gene B				
Gene C				
Gene D				

3.1b) Now make two distance matrices with these data; one for the distance between the genes, one for the distance between the tissues. There are multiple ways to create such a distance matrix. The simplest is to sum up all the absolute differences per condition for a set of two genes. Another, often used, way is to calculate the correlation between two genes over all experiments. The distance between gene A and gene B is the same as that between gene B and A; the matrices thus are symmetrical and you only have to fill one 'triangle'.

	Gene A	Gene B	Gene C	Gene D
Gene A	0			
Gene B		0		
Gene C			0	
Gene D				0

	Tissue I	Tissue II	Tissue III	Tissue IV
Tissue I	0			
Tissue II		0		
Tissue III			0	
Tissue IV				0

3.1c) Which genes resemble each other the most in their behaviour? And which ones the least? Similarly, which tissues appear most distant and which look most related?

3.1d) Now use the distance matrices to construct two dendrograms; one for the genes and one for the tissues. You may use the UPGMA method for that; the distance between two clusters is the mean of the distances between all data points in the two clusters. Thus; determine the average value of the four distances for the distance between two clusters of two genes.

First for the genes (the diagonal is 0, the distance from a gen to itself is 0)

	Gene	Gene	Gene
Gene	0		
Gene		0	
Gene			0

	Gene	Gene
Gene	0	
Gene		0

And then for the tissues:

	Tissue	Tissue	Tissue
Tissue	0		
Tissue		0	
Tissue			0

	Tissue	Tissue
Tissue	0	
Tissue		0

Question 3.2

There is a protein in *Yersinia pestis* (which causes bubonic plague a.k.a. Black Death) that in fact represents a fusion protein of two separate proteins present in *Rickettsia prowazekii* (the cause of typhus). Using various databases you will be able to connect protein domain function and disease.

3.2a) Determine which protein in *Y. pestis* looks like a fusion of two *Rickettsia prowazekii* proteins, the sequence of one of which is given below. Compile a drawing that shows where the gene from *Y. pestis* is homologous to the two genes in *R. prowazekii*.

```
mtnktitlnl gpqhpathgv lrlilemdge vvnnadphig llhrgtekli ehktylqaip
yfdrldyvsp mcqehafala vesllecsvp rraqfirvlf seltrilnht lnigsqaldi
gattpllwlf eerekimefy ervsgsrmhs nyfrpggvae dlpenlledi nkfieqfpsk
lndienllne nrlwkqrlvd igvvsqkdam dwgfsgpmlr gsgiawdlrk snpydvydem
dfevpigkng dcydrylvri lemyesikii kqcivkmpkg qvktddpklt pptrgkmkes
meamihhfkl ytegydvpig etykaveapk gefgvylysq ggnkpyrcri kapgfahlqg
lnfmskghli advitiiatl divfgeidr
```

Hint: You will have to “Blast” this *R. prowazekii* sequence against the proteins of *Y. pestis*. Subsequently, to find the second fusion gene, you will have to use the *Y. pestis* protein sequence in a Blast against *R. prowazekii* proteins. This then should return two hits.

3.2b) What are the functions of the two genes? Are these functions related?

Hint: Remember the respiratory chain you dealt with during biochemistry classes.

3.2c) Make a drawing of the domain composition of the *Y. pestis* protein using SMART. Are the protein domain regions in the *Y. pestis* protein more or less comparable with the parts that are homologous to the proteins in *R. prowazekii*?

3.2d) Are there any homologous proteins in human? If yes, what is the subcellular localisation? And what is the function? Do these human proteins interact with each other?

Question 3.3

Urease is a well known virulence factor that is found in many pathogens like Helicobacter pylori that is associated with ulcers. Below you will find the sequence of a urease protein:

```
>Urease
mmsnisrqay admfgpttgd kirladtelw ieveddltty geevkfgggk virdgmgqgq
mlsagcadlv ltnaliidyw givkadigvk dgrifaigka gnpdiqpnvt ipigvsteii
aaegrivtag gvdthihwic pqqaeealts gittmigggt gptagsnatt ctpgpwyiyq
mlqaadslpv nigllgkgnc snpdalreqv aagviglkih edwgatpavi ncaltvadem
dvqvalhsdt lnesgfvedt ltaiggrtih tfhtegaggg hapdiitaca hpnilpsstn
ptlpytvnti dehldmlmvc hhldpdiaed vafaesrirq etiaaedvlh dlgafsltss
dsqamgrvge vvlrtwqvah rmkvqrgplp eesgdndnvr vkryiakyti npalthgiah
evgsievgkl adlvlwspaf fgvkpativk ggmiamapmg dingsiptpq pvhyrpmfaa
lgsarhrcrv tflsqaaaan gvaeqlnlhs ttavvkgcrt vqkadmrhns llpditvdsq
tyevringel itsepadilp maqryflf
```

3.3a) From which species is this protein sequence originating? Is it also present in Helicobacter pylori?

3.3b) As you will see in the BLAST output resulting from your search for the species from which the urease sequence originates, there is not much consistency in the naming of this protein. Provide at least two different names for this protein (e.g. subunit X or Urease Z).

Hint: just scroll through the Blast output.

3.3c) Does the urease protein function on its own or is it part of a protein complex?

Hint: To be able to answer this question you can use feed the protein into STRING (string.embl.de). STRING is a tool to detect functional relationships between proteins based on f.e. gene cluster conservation in sequenced genomes, protein-protein interactions in genomics data, or interactions that have been published in PubMed. Enter the sequence in the appropriate window and click on Go. STRING will now search first to which homology group your sequence belongs and then will ask you to confirm this. The subsequent output list will contain proteins that on the basis of several different types of data are expected to be functionally related to Urease.

3.3d) List the proteins that Urease is functionally related to. What sorts of "proof" form the basis for the suggested functional relationship? Provide the names of the five proteins that yield the highest score with Urease and mention one type of genomics-type proof for the relationships per protein.

3.3e) Are all these Urease "top five hits" part of the same protein complex? Or is there a different reason why these genes are often found in the same operon? If yes, which reason?

Hint: To answer this question you cannot escape from reading some literature. Often the introductory parts of papers on urease already contain the required information for these aspects. For example, have a look at the introduction of the paper by Chen & Burne (J. Bacteriology 2003) or that by Palinska, Jahns, Rippka and Tandeau de Marsac (Microbiology 2000). In general, effective analysis and interpretation of genomics data is requiring that one couples biological knowledge to the output results. Since you will not always have this required biological background information, you will need to acquire it by reading literature "on the fly".

Question 3.4
In the article we used for Question 6 in the previous Chapter (Mwangi et al, 2007) the authors reasoned that the *Staphylococcus aureus* gene SA1702 would be involved in Vancomycin resistance base on the fact that the gene is located in the operon that contains gene VraR. The VraR gene has already been shown to be causative in vancomycin resistance.

3.4a) Will it be possible to exploit comparative genomics to generate additional proof that SA1702 and VraR indeed are involved in the same pathway (e.g. co-expression, protein-protein interaction, gene order conservation)

Hint: Run STRING (string.embl.de).

3.4b) Is there another protein that, on the basis of the comparative genomics data, could be predicted to be involved in the same pathway as that VraR and SA1702, and thus also in drug resistance?

Hint: Exploit STRING data

Additional practice (optional)

Question 3.5
The following protein in *Y. pestis* is homologous to a part in a *Mycobacterium tuberculosis* protein. In fact, the *M. tuberculosis* protein represents a fusion of two *Y. pestis* proteins.

```
>YPO4125
mnlnatilgq aiafvlfvif cmkyvwppim aaiekrqqei adglssaera kkdldlaqan
atdqlkkaka eaqviieqas krkaqildea kaeaeqernk ivaqaqaeid aerkrareel
rkqvamlaia gaekiiersv deaansdivd klvael
```

3.5a) Determine the two *Y.pestis* proteins and the *M. tuberculosis* protein and make a drawing symbolizing the homology between these three protein.

3.5b) What are the functions of the two *Y.pestis* proteins, and are these functions related (think about the way how mitochondria generate ATP)?

3.5c) Make a drawing of the domain composition of the *M.tuberculosis* protein. Does it bear resemblance to the homology relationships between the *M.tuberculosis* protein and the two *Y.pestis* proteins?

4 – Analyzing genomes: Genome Browsers

Introduction

In this chapter we will explore the UCSC human genome browser, a tool to visualize and analyze the ever increasing amount of genomic data which is being produced. Within the browser one can look at regions of interest in their genomic context. This includes a.o. information about the genomic DNA sequence, intron-exon structure, alternative transcripts, protein sequences and structures, ESTs, SNPs and regulatory sites (e.g. TFBS) but also links are provided to external experimental data, like expression data. Furthermore, comparisons of genomic regions in different species are facilitated.

Course Materials

Some background reading (optional) to refresh your memory:
Molecular Biology of the Cell (4th edition) Alberts et al.

Chapter 7: Control of gene expression
Section: How genomes evolve
Genome Alterations are Caused by Failures of the Normal Mechanisms for Copying and Maintaining DNA
The Genome Sequences of Two Species Differ in Proportion to the Length of Time That They Have Separately Evolved
The Chromosomes of Humans and Chimpanzees Are Very Similar
A Comparison of Human and Mouse Chromosomes Shows How The Large-scale Structures of Genomes Diverge
It Is Difficult to Reconstruct the Structure of Ancient Genomes
Gene Duplication and Divergence Provide a Critical Source of Genetic Novelty During Evolution
Duplicated Genes Diverge
The Evolution of the Globin Gene Family Shows How DNA Duplications Contribute to the Evolution of Organisms
Genes Encoding New Proteins Can Be Created by the Recombination of Exons
Genome Sequences Have Left Scientists with Many Mysteries to Be Solved
Genetic Variation within a Species Provides a Fine-Scale View of Genome Evolution

Chapter 14: Energy conversion: Mitochondria and Chloroplasts
Section: The Genetic Systems of Mitochondria and Plastids
Mitochondria and Chloroplasts Contain Complete Genetic Systems
The Genomes of Mitochondria and Chloroplasts Are Diverse
Mitochondria and Chloroplasts Probably Both Evolved from Endosymbiotic Bacteria
Mitochondrial Genomes Have Several Surprising Features
Section: The Evolution of Electron-Transport Chains
The Photosynthetic Electron-transport Chains of Cyanobacteria Produced Atmospheric Oxygen and Permitted New Life-Forms

Human genome browsing

Although information about genes can also be found in databases like Genbank, there are several advantages in using a genome browser. In the first place they provide one point of entrance to several kinds of data about a gene. In one glance you can analyze and visualize the mRNAs that are transcribed from the gene, the proteins that are encoded by the gene, the functions that have been ascribed to the gene and its protein(s), which nucleotides and amino acids vary between individuals in a population (SNPs), etc. Furthermore, you can identify orthologs in other species, for example to find the gene in Macaque that corresponds to the human alpha-globin gene.

Another advantage of genome browsers is the possibility to visualize the genes in their genomic context. You can look at promoter areas to see if there are known transcription factor binding sites (TFBS). You can look at the introns within a gene; in some cases these may harbour important sequences that regulate alternative splicing or gene expression. Regulatory elements can also be found outside transcribed gene regions. However, still little is known about such elements and consequently these have not been annotated extensively. One could perhaps consider comparing genomic regions between species to find conserved regions outside genes that might be worthwhile investigating experimentally.

The browser also displays the neighbouring genes. In contrast to prokaryotes, eukaryotes do not have polycistronic mRNAs (with a few exceptions) but genomic context can still be very important. Genes that are located close together can sometimes be jointly regulated, like the HOX genes and the beta-globin gene family. In the human genome genes are not always neatly positioned one after another, there are also overlapping genes present as well as gene pairs that are transcribed from the same promoter but in different directions. Sometimes a collection of genes is close together while other genes lie relatively isolated on the genome. To know the positions of all the genes in a specific area of the genome is especially important when studying genes involved in a genetic disease.

Genome browsers are updated regularly so that newest data gets incorporated. Recently is has become clear that there is much transcriptional activity in mammalian DNA areas that do not code for proteins. We don’t know the function of these intergenic transcripts yet but at least the genome browsers can show us where these areas are. Another finding is that certain pieces of chromosome can have copies on several places in the genome, and that different individuals can have a different number of copies of such pieces of DNA – the so-called copy number variations (CNVs). The CNVs can contain genes, which then leads to a different number of copies in the population. CNVs do not occur randomly but tend to have a preference for certain genomic areas which seem to correlate with genomic instability, a feature that also can be analyzed in the genome browser. Some genetic diseases, like Prader-Willi and Angelman syndrome, are caused by deletion of pieces of genome. Deletions and duplications also often occur in different forms of cancer. This area of research is relatively new, a.o. because it is not clear which mutations cause the cancer and which mutations are an effect of the cancer.

The UCSC human genome browser

There are several genome browsers available:

- Ensembl (http://www.ensembl.org/index.html)
- Mapviewer (http://www.ncbi.nlm.nih.gov/mapview/)
- The UCSC genome browser (http://genome.ucsc.edu/)

In this course module we are going to work with the UCSC human genome browser. The following text can also be found online, at https://cgwb.nci.nih.gov/goldenPath/help/hgTracksHelp.html, together with the UCSC user guide (see **course materials**).

As vertebrate genome sequences near completion and research re-focuses on their analysis, the issue of effective sequence display becomes critical: it is not helpful to have 3 billion letters of genomic DNA shown as plain text! As an alternative, the UCSC Genome Browser provides a rapid and reliable display of any requested portion of genomes at any scale, together with dozens of aligned annotation tracks (known genes, predicted genes, ESTs, mRNAs, CpG islands, assembly gaps and coverage, chromosomal bands, mouse homologies, and more). Half of the annotation tracks are computed at UCSC from publicly available sequence data. The remaining tracks are provided by collaborators worldwide. Users can also add their own custom tracks to the browser for educational or research purposes.

The Genome Browser stacks annotation tracks beneath genome coordinate positions, allowing rapid visual correlation of different types of information. The user can look at a whole chromosome to get a feel for gene density, open a specific cytogenetic band to see a positionally mapped disease gene candidate, or zoom in to a particular gene to view its spliced ESTs and possible alternative splicing. The Genome Browser itself does not draw conclusions; rather, it collates all relevant information in one location, leaving the exploration and interpretation to the user.

The Genome Browser supports text and sequence based searches that provide quick, precise access to any region of specific interest. Secondary links from individual entries within annotation tracks lead to sequence details and supplementary off-site databases. To control information overload, tracks need not be displayed in full. Tracks can be hidden, collapsed into a condensed or single-line display, or filtered according to the user's criteria. Zooming and scrolling controls help to narrow or broaden the displayed chromosomal range to focus on the exact region of interest. Clicking on an individual item within a track opens a details page containing a summary of properties and links to off-site repositories such as PubMed, GenBank, Entrez, and OMIM. The page provides item-specific information on position, cytoband, strand, data source, and encoded protein, mRNA, genomic sequence and alignment, as appropriate to the nature of the track.

In addition to the Genome Browser, the UCSC Genome Bioinformatics group provides several other tools for viewing and interpreting genome data:

- BLAT - a fast sequence-alignment tool similar to BLAST.
- Table Browser - convenient text-based access to the database underlying the Browser.
- Genome Graphs - a tool that allows you to upload and display genome-wide data sets such as the results of genome-wide SNP association and linkage studies and homozygosity mapping.
- Gene Sorter - expression, homology, and other information on groups of genes that can be related in many ways.
- Proteome Browser (accessible from Known Genes details pages) - protein property data and links to a wealth of related information.

Course Materials

For these tools and further detailed information we refer to the UCSC webpages and documentation:
- http://genome.ucsc.edu/
- https://cgwb.nci.nih.gov/goldenPath/help/hgTracksHelp.html

Hands-on Exercises

IMPORTANT: Be sure not to forget that the UCSC browser will keep your selections for displaying information during one browser session (by using cookies). This might come in handy in some cases, but not always. If you want to start with the default settings of the browser you simply have to select the button "default tracks" in the lower half of the screen.

Question 4.1

The following exercises will help you to get a short introduction to the UCSC genome browser.

4.1a) Look at the Open Helix tutorial "UCSC Genome Browser: introduction" (see link below).

4.1b) Perform the exercises provided in the Open Helix tutorial "UCSC Genome Browser: an Introduction". This tutorial contains a step-by-step handout to lead you through.

Course Materials

The material can be found at http://www.openhelix.com/cgi/tutorialInfo.cgi?id=27:
- Tutorial: Choose *Launch online tutorial*. In case you do not have access to audio you can study the accompanying presentation on the same webpage (choose *Download powerpoint slides*). Hint: The *Exercises* part of the tutorial is the most informative, you may want to skip other parts in sake of time (the total tutorial is ca. 1 hr).
- Exercises: Choose *Download hands-on exercises*.

Please note that Open Helix http://www.openhelix.com/ provides lots of interesting (often freely available) tutorials about all kinds of bioinformatics tools.

Question 4.2 – Analyzing a gene: genomic context and alternative transcripts

4.2a) Find with the UGSC genome browser the human TP53 (tumor protein p53) gene. On which chromosome does it reside?

Hint: Notice that there are multiple hits, indicating that there are entries with "TP53" somewhere in their name. Sometimes there are transcripts of different lengths, corresponding to different isoforms of the protein. You always need to analyze the hits yourself and select the one you need for your current research question.

4.2b) View this gene in its genomic context. In which direction is it transcribed?

4.2c) Which gene lies at its 5′ (upstream) side, and from which strand is it transcribed?

4.2d) From which strand is the downstream gene (the one residing at TP53's 3′ end) read?

4.2e) What is the closest gene that is transcribed in the same direction / from the same strand as p53?

One single gene can lead to multiple transcripts. Two ways in which this can take place is the use of alternative promoters and alternative splicing. In case of alternative promoter use, transcription starts at different positions on the DNA. In case of alternative splicing, the transcripts of one gene contain different combinations of exons. Sometimes there is even alternative use of poly-A addition signals, leading to different 3′ ends of mRNAs.

4.2f) Find the alternative transcripts that have been found for the TP53 gene. Compare and analyze the splice variants in more details. Are there splicing differences between the transcripts? Are there transcripts that start at an alternative promoter?

Hint: Under "Genes and Gene Prediction Tracks" select UCSC Genes, Alt Events, Refseq Genes. Study the different splicing events given in purple *("Alternative splicing, Alternative promoter and similar events in UCSC genes").* Under "mRNA and EST Tracks" select Human mRNAs, Spliced ESTs and see which transcribed variants have been found.

Question 4.3 – Genomics made easier

Read the paper "Genomics made easier" by Peter Schattner (see course materials). Read at least until page 8 (in the author reprint version of the paper). Now try to answer the research questions given on pages 4 and 5 (in the author reprint version of the paper), by following the instructions in the paper with the help of the UCSC browser.

Hints: When needed, follow one of the Open Helix tutorials. Document your findings (also for yourself for later use) by making a small report containing the search strategies you applied and including some screen dumps (Alt-PrtScn) of the UCSC output.

Course Materials

- Schattner, P (2009) Genomics made easier: an introductory tutorial to genome datamining. Genomics 93 (3): 187-195.

Additional practice (optional)

Question 4.4 - Genomic variation: Single Nucleotide Polymorphisms

Find the Alzheimer gene presenilin-1 on the human genome. Analyze the SNPs in this gene.

4.4a) Study the display modes of the SNP track.
What do you see when choosing respectively 'dense', 'squish', 'pack' and 'full'.
What is the definition of the colouring scheme?

Hint: In the 'Variations and Repeats' part click on one of the blue hyperlinks named SNP(xxx), where xxx is the build number. You will arrive at the SNP track settings where you can check the "coloring options".

4.4b) Analyze the SNPs in presenilin-1.
Give two SNPs that are synonomous (mutations that do not cause an amino acid change) and two SNPs that are non-synonomous (mutations that do cause an amino acid change)

Hint: Choose a specific color for both kinds of mutations in the "Coloring Options" part.

Question 4.5 - Unusual genes

4.5a) Very long genes: The length of a gene on the genome can vary strongly from shorter than a kilobase to longer than a megabase. Examine the human DMD (dystrophin) gene in its genomic context. Now examine this gene in the UCSC browser. Can you view the entire gene? What is the size of the gene?

Hint: The length of the gene can be seen in the gene record in the UCSC browser.

4.5b) Overlapping genes: Some genes overlap each other on the genome. Examine the human RTDR1 (rhabdoid tumor deletion region gene 1) gene in its genomic context. Can you find the GNAZ (guanine nucleotide binding protein (G protein), alpha z polypeptide) gene in this area? In which direction is it transcribed?

4.5c) Fusion proteins: Some genes are even weirder. Examine the human NME1 (non-metastatic cells 1, protein (NM23A) expressed in) gene, and zoom out to view the surrounding part of the genome. Where is the NME2 gene? Is it part of the NME1 gene? You can examine the gene records to get the full "story".

Question 4.6 - Gene clusters

Sometimes clusters of homologous genes are found in the same genomic area. Examine the human gene PCDHB13 (proto-cadherin beta 13) in its genomic context. Zoom out until you can see ca. 1 Mb.

4.6a) Study the genes in this area. Are they related?

HINT: If needed, you can look at their gene records but their names already reveal some information. Many of these genes overlap and share exons. You can wonder if these are all separate genes of alternative transcripts from the same genes.

Question 4.7 - Conserved genomic elements

The UCSC browser allows you to see the degree of evolutionary conservation of a genomic sequence based on multiple sequence alignment of vertebrate genomes. This enables you to find conserved genomic elements even if they are not annotated.

4.7a) Examine the human LHX9 (LIM homeobox 9) gene in the UCSC browser. Zoom out a little (about 1.5x) so that you can also see the neighboring genomic regions. Make sure the "Conservation" track is switched on (it is by default). What is the level of conservation just downstream of this gene? Note: The region in question lies in the last intron of the longest altenative LHX9 transcript.

4.7b) Make sure that the "Non-Human Refseq Genes" track and the "Other mRNAs" track (within the subsection "mRNA and EST Tracks") are both switched the "pack" level. Can you find RefSeq transcripts or mRNAs from other species that include the region of 4.7a? What do you think the evolutionary conservation may indicate? *[Hint: You can switch tracks off or on at different levels of detail using the little drop-down menus underneath the main genome browser display. Do not forget to click on the "refresh" button afterwards to activate your changes]*

5 – Analyzing 3D protein structures

Introduction

There are many forms of genetic defects resulting in a hereditary disease. Among the possible defects are so-called point mutations, places in our genome where one nucleotide has been substituted by another. The majority of these mutations do not lead to phenotypic differences or lead to "unimportant" differences like length, eye colour etcetera. These are called single nucleotide polymorphisms or SNPs. From comparisons of the various genomic sequence reads from individuals it is currently estimated that there will be over 20 million SNPs in the human population, a considerable amount given our genome size of three billion base pairs.

About 1% of the SNPs reside in coding parts of the genome and about two-third of those are so-called non-synonymous; they give rise to nonsense (introduce a stop codon), missense (cause an amino acid substitution) or frame-shift (alter the reading frame) mutations which may have severe consequences for the functioning of the protein and, consequently, our health. One way to investigate possible SNP effects at the protein level is to try to decipher the position and role of the affected amino acid in the protein 3D structure of a healthy person and try to predict what the changes are upon mutating this amino acid. For example, if an amino acid is located in the DNA binding domain of the protein, or in the active site, it can very well be imagined that there is a severe effect when this amino acid is mutated, and a disease can be the result. Thus, to study the effect of mutations in more detail we have to start looking at the proteins involved at the atomic level.

Course Materials

Some background reading (optional) to refresh your memory:
Molecular Biology of the Cell (4^{th} edition) Alberts et al.

Chapter 3: Proteins
    Section: The shape and structure of proteins
        The Shape of a Protein Is Specified by Its Amino Acid Sequence
        Proteins Fold into a Conformation of Lowest Energy
        The α Helix and the β Sheet Are Common Folding Patterns
        The Protein Domain Is a Fundamental Unit of Organization

Chapter 10: Membrane Structure
    Section: Membrane proteins
        Membrane Proteins Can Be Associated with the Lipid Bilayer in Various Ways
        In Most Transmembrane Proteins the Polypeptide Chain Crosses the Lipid Bilayer in an α-Helical Conformation
        Some β Barrels Form Large Transmembrane Channels

Chapter 11: Intracellular compartments and protein sorting
    Section: The Compartmentalization of Cells
        Signal Sequences and Signal Patches Direct Proteins to the Correct Cellular Address

Protein functions rely on protein structures

When you study a protein (and its – disease-related – mutants), you are usually interested in its function. And since there is a tight relationship between the structure and the function of a protein, you will turn to analyzing the 3D structure as soon as you find an interesting feature at the level of the protein sequence — perhaps a motif, an evolutionarily conserved segment or a disease-causing mutation.

Typical questions you want to ask are:

- Are these amino acids located at the surface of my protein molecules?
- Are these amino acids directly involved in the function of the protein (e.g. are they in the active site)?
- Are these amino acids involved in the binding of another molecule (e.g. DNA or specific metal ions)?
- Can I explain the change in behaviour of the mutant protein now that I know or can predict its role in the wild-type protein?

In this chapter, we show you how to look at your protein from a 3-D perspective. We will not discuss predicting structural features of a protein based on its amino acid sequence, such as secondary structure, domain structure, transmembrane helices, signal peptides, localization signals, solvent accessibility. This has been discussed already in chapter 2 of this manual but for those interested we have included some nice excersises addressing these aspects in the Additional practice section.

The Brookhaven Protein Databank (PDB)

The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules (proteins, nucleic acids) and their complexes, including ligands (e.g. enzyme-substrate, dna-protein, protein-drug etc.). All structures can be accessed freely through the PDB website (www.pdb.org). The data contained in the archive include atomic coordinates, crystallographic structure factors and NMR experimental data. Aside from coordinates, each deposition also includes the names of molecules, primary and secondary structure information, sequence database references, where appropriate, and ligand and biological assembly information, details about data collection and structure solution, and bibliographic citations.

Structures deposited in the PDB are assigned a unique four letter code which is often called PDB accession code or PDB code. Because of the PDB's importance as the central repository for biological macromolecular structures, the PDB code is often used in the scientific literature to refer to a particular structure which has been used in a study. By convention, the PDB code consists of a single numeric digit followed by three alphanumeric characters. The PDB code is not case sensitive, i.e. 1abc and 1ABC refer to the same structure.

Please read the article given below and also do the PDB tutorial. After this you are ready for the exercises where you will download, visualize, and analyze the 3D characteristics of your protein of interest.

Course Materials

- Goodsell, D. S. (2010) The Protein Data Bank: Exploring Biomolecular Structure. Nature Education 3(9):39

What if the structure of our protein of interest is not known?

Although the number of proteins for which the 3D structure is known is rising, the structure of your protein of interest may not have been elucidated yet. However, if your protein has significant homology with a protein for which the structure is known, this structure can be used to make a 3D model of your protein. This process is called **homology modelling**. If you are interested you can read more about this in the Additional practice section.

Preparing for 3D structure visualization

A) Preparing your browser for 3D structure visualization within a webpage

Before you will be able to look at structures within a webpage (e.g. from the PDB website www.pdb.org) take the following steps:

- Almost all features of the PDB web site require a modern web browser with JavaScript and cookies enabled. See http://www.pdb.org/pdb/static.do?p=home/faq.html#goodBrowsers for information about browsers that have been tested. Check if your browser is OK on this webpage: http://www.pdb.org/pdb/browser/browsercheck.do.

- Molecular viewers require Java 1.7 or above. You can check the Java version installed on http://www.java.com/en/download/testjava.jsp and/or download Java if necessary on http://www.java.com/en/.

When all preparations have been made you should be able to view the structures on the pages mentioned in exercises 5.1 and 5.2 below. The structures in this module will be displayed using Jmol (http://wiki.jmol.org/index.php/Main_Page). Jmol is an open source molecule viewer written in Java. It runs as a standalone application and as a web browser applet. The basic mouse commands in Jmol are:

Action	**Windows**	**Macintosh**
Rotate X, Y*	Left	Unmodified
Translate X, Y*	Ctrl-Right	Command
Rotate Z*	Shift-Right	Shift-Command
Zoom	Shift-Left	Shift
Slab Plane**	Ctrl-Left	Ctrl
Menu***	Right	Hold Down

* The X and Y axes are left-right and up-down on your screen, respectively. The Z axis is perpendicular to your computer screen.
** The Slab Mode option must be activated in the Jmol script before this option can be used. Think of the slab as a plane parallel to your computer screen. Anything on the other side of the slab plane from you can be seen. The Slab Plane option moves the slab plane closer or further away from you when you move the mouse up or down.
*** Opening the Jmol menu only requires a right mouse click (Windows) or held mouse click (Macintosh). The menu gives a multitude of options to manipulate the display which are best learned by practice. Trying the options in the Color and Display submenus will provide the most striking changes to the display.

However, if you really are planning to manipulate a 3D structure you might consider using Yasara (see below).

B) Downloading and installing YASARA to work with 3D structures

- Go to www.yasara.org and click on 'Products'
- Choose 'freely download now' behind YASARA View
- Fill out the form. At 'department' you can fill the name of your university. Your email address is only used to send the download-link
- In your email you will receive a download-link
- Install YASARA according the instructions.
- If all went well, you can now click on 'Help > Play help movie > Working with YASARA' and spend five minutes to see how the program works.
- A short introduction of YASARA can also be found on this website http://www.cmbi.ru.nl/~hvensela/yasara/

YASARA shortcuts:

Moving objects around	
LeftButton+MouseMovement	Rotate object or scene or move label
Ctrl+LeftButton+MouseMovement	Rotate object or scene, centered on the marked atom
LeftButton+RightButton+MouseMovement	Move object or scene along the X/Y-axes
MiddleButton+MouseMovement	Move object or scene along the X/Y-axes
Cursor keys	Move object or scene along the X/Y-axes
RightButton+MouseMovement	Move object, scene or label along the Z-axis
LeftClick on a label	Mark label and move it around
Changing the scene style	
F1	Set scene style to balls
F2	Set scene style to balls&sticks
F3	Set scene style to sticks
F4	Set scene style to Calpha trace
F5	Set scene style to tube
F6	Set scene style to ribbon
F7	Set scene style to cartoon
F8	Change side-chain style

Course Materials

- If you experience trouble with downloading YASARA (e.g. because of a slow Internet connection, or because of lack of administration privileges for installation), please contact C.vanGelder@cmbi.ru.nl. You will then receive the manual of 2012/2013, where all exercises that follow below will be with Jmol instead of with YASARA.

Exercise 5.1 - Introduction to protein structure viewing

5.1 Go to the webpages http://proteopedia.org/wiki/index.php/Main_Page and http://proteopedia.org/wiki/index.php/Proteopedia:Table_of_Contents. Look around a bit, choose a molecule you like and play around with it. When you click on the green links in the text the structure in the screen will change. But you can also use your mouse to manipulate the structure, using the commands given in the table above.

Exercise 5.2 - Finding a protein structure in PDB

We are going to look for the 3D structure of the transcription factor p63, a p53 family member that plays a role in epithelial cell development, cell cycle arrest, apoptosis, and tumorigenesis. Point mutations, primarily in the DNA binding domain (p63DBD), lead to malformation syndromes as you have learned in question 2.2.

5.2a) Go to www.pdb.org and type "p63" in the search window. Please choose the tab MacroMolecule (with this tab your search molecules by name) and click on the icon with the magnifying glass on the right. You will see the search results (14 in June 2013). Scroll down to the hit list. The first hit is 1RG6. This structure only contains the C-terminal domain of P63 and is not relevant for our study. Look further in the list for 3QYN, click on this PDB code to go to the *Structure Summary* page for this protein.

5.2b) On that *Structure Summary* page, look at the title. Note that this structure is the DNA-binding domain of p63 complexed with a piece of DNA.

5.2c) Find the primary citation to learn where and when this structure was first published. Note that it was published in 2011 by Chen et al. in PNAS.

5.2d) Scroll down to the Molecular Description section to find out how many molecules are in this structure and what they are. Note that there are 4 protein chains (A through D) and 2 DNA chains (E and F) in the 3QYN structure file.

5.2e) At the top of the Molecular Description section you will find Classification information. Here you can find out what the protein component of this structure does, in broad functional terms. Note that this protein is involved in activation of DNA transcription (it states Transcription Activator/DNA). Under *Ligand chemical component* you can see that a Zn (zinc) ion is also present.

5.2f) Read the Source section to find the species in which this protein is normally found in and the expression system host species. Note that it is a human protein and that *E. coli* was used as an expression system.

5.2g) Click on "View in 3D", underneath the image on the top right hand of the Structure Summary page. Rotate the interactive image and examine it to see how clearly you can differentiate between the DNA and protein components.

Exercise 5.3 - Downloading a protein structure from the PDB

When you want to study your protein of interest in more detail you may want to display it independently from the PDB website. In that case you will have to download the structure to your computer so it can be used as input for other applications.

5.3a) On the 3QYN page (exercise 5.2), the upper right corner shows the option "Download Files".

5.3b) To download the structure information you will have to choose "PDB file (Text)" or "PDB file (gz)". In the first case you will get a file that you can directly import in other programs, in the second case you get a zipped file that you have to unzip before further use.

5.3c) Take the first option and download the file 3QYN.pdb to your computer. IMPORTANT: when downloading and using PDB structures be aware to always give the file the extension **.pdb**. All software programmes that can read and display molecular structures need this file extension.

The pdb file can be used as input for YASARA, Jmol and for other structure viewing programs (see Exercises 5.4 and 5.6).

Course Materials

Further reading about the functionality of the PDB is possible through:

- Video tutorials on PDB website, e.g. :
  http://www.pdb.org/pdb/static.do?p=general_information/screencasts.jsp
- Use google to find PDB tutorials. For example on:
  http://amrita.vlab.co.in/?sub=3&brch=273&sim=1440&cnt=1

Exercise 5.4 - Exploring the p63 structure and its mutations
There is wealth of structural information on p63. Here we will focus on four amino acids that, when mutated, each cause EEC syndrome:

- Histidine 208 has been mutated to Alanine (His208Ala or H208A)
- Cysteine 269 has been mutated to Alanine (Cys269Ala or C269A)
- Serine 272 has been mutated to Alanine (Ser272Ala or S272A)
- Arginine 304 has been mutated to Alanine (Arg304Ala or R304A)

EEC syndrome (acronym of Ectrodactylyl, Ectodermal dysplasia and Cleft lip syndrome) is an inherited disorder that is characterised by a series of anomalies. Some symptoms are lacrimal duct abnormalities, urogenital problems, conductive hearing loss, facial sysmorphism, chronic/recurrent respiratory infections and developmental delay. The most prominent features of the syndrome are a cleft lip, skin, hair, nail, and tooth malformations and severe hand and foot malformations.

5.4a) What can you say about these mutations without any knowledge of the protein 3D structure? Thus, what are the properties of the wild-type and mutant amino acids? If you don't know the amino acids by heart, try to find more information about them (e.g. the CMBI-wiki on http://wiki.cmbi.ru.nl, search for amino acid name). Write down how size, charge, hydrophobicity, and other striking features of wildtype residues are changed by these mutations.

1. Mutation His 208 (Ala)
   Properties Histidine:
   Properties Alanine:
2. Mutation Cys 269 (Ala)
   Properties Cysteine:
   Properties Alanine:
3. Mutation Ser 272 (Ala)
   Properties Serine:
   Properties Alanine
4. Mutation Arg 304 (Ala)
   Properties Arginine:
   Properties Alanine:

5.4b) Analysis of the P63 mutations

Now it is your turn to study this protein and its interactions with DNA. Start YASARA and load the 3QYN.pdb file (File > Load > PDB file). Play around with scaling the complex and the different display styles. When looking at the complete complex, F5, F6 or F7 are good displays. When zooming in on particular parts of the structure you should also display the individual atoms (e.g. with F3, sticks display)

Move the pointer to the bottom of the window. The sequence selector appears. Click on the top left pin-icon to fix it. You can select amino acids in the sequence selector and they will be indicated in the screen and vice versa.

Push <Insert> to show the head-up display (HUD) where you can see all info about the atoms (left) and the molecules (right) that are currently loaded. Pressing <Insert> again will hide the HUD.

5.4c) There are multiple molecules in the file. Try to figure out what and where all molecules are in the display. It is good to use ball-and-stick display for analysing a structure in detail.

Look up the mutated residues and try to analyse their role in the protein structure.

Hints of YASARA options, you might want to:

- Colour the DNA red: View > Color > Molecule > NucAcid > OK > Red > Apply unique color
- Hide the water molecules from the display (Note: you only see the red dots showing the oxygen atoms of all the water molecules): View > Hide atoms> Molecule > Water > OK
- Hide protein chains B, C, D: View> Hide > Molecule > B, C, D (use ctrl to select multiple molecules in the menu) > OK
- Color a mutated residues: View > Color > Residue > His208 > OK > Yellow > Apply unique color
- Show Zn as a ball in stead of a point in space: View > Style atoms > Ball > Residu > Zn > OK
- Label e.g. the DNA chains: Effects > Label > Molecule > NucAcid > OK
- Show hydrogen bonds: First add hydrogens with Edit > Add > hydrogen to all. After this display the hydrogen bonds with View> show interactions > hydrogen bonds of > All.
- Save nice displays while you are playing along, so you can get them back easily: File > Save as > Yasara scene. This will result in a .sce file that can be reopened with File > Load > Yasara Scene.

5.4d) Having viewed and analyzed the structures under 5.4c), can you now explain why the four mutations cause EEC syndrome? And if you are ready go to the site for "Project HOPE" (http://www.cmbi.ru.nl/hope) and see if your answers correspond to HOPE's answers.

Additional practice (optional)

5.5 - Predicting structural features from the amino acid of a protein

In addition to the exercises described above there are several other things that life science researchers can do when they have obtained a protein sequence, like predicting secondary structure elements, domain structures, transmembrane helices, signal peptides or other localization signals, solvent accessibility etcetera. A plethora of online tools is available for such purposes. Check out the following websites if you want some more information on this:

- http://www.predictprotein.org/
- http://swift.cmbi.ru.nl/teach/DNA/

Choose *Servers* in the upper left frame (which points you to the one-day course about protein webservers), and do the exercises.

5.6 - Other protein 3D visualization and modelling programs

Please note that there are several free modelling programs available for viewing and analyzing 3D structures of proteins, nucleic acids (DNA, RNA) and their complexes. Thus far you have used YASARA (developed at the CMBI (UMC St Radboud, Nijmegen)) . There are also other freely available packages, you might want to take a look at:

- SwissPDBViewer (http://spdbv.vital-it.ch/)
- Cn3D (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml)

5.7 - Homology modeling

Since until 2011 the structure of the p63 protein in complex with DNA was not known, models for interaction between p63 and DNA were predicted on the basis of the structural data for the DNA-binding domain of the homologous protein p53 (see the paper by Celli et al). This represents an example of so-called 'homology modeling', a powerful approach to come to predictions for proteins who's structure is yet to be defined.
Some interesting material you may like to study is suggested below.

Course Materials

- Celli J, Duijf P, Hamel BCJ, Bamshad M, Kramer B, Smits APT, Newbury-Ecob R, Hennekam RCM, van Buggenhout G, van Haeringen A, Woods CG, von Essen AJ, de Waal R, Vriend G, Haber DA, Yang A, McKeon F, Brunner HG & van Bokhoven H. Heterozygous Germline mutations in the p53 homolog p63 are the cause of EEC syndrome. Cell 99 (1999) 1-20.
- Introduction to homology modelling on http://www.cmbi.ru.nl/~hvensela/, choose 'Homology Modelling' in the left 'Information' frame.
- Homology Modeling (author version) by Venselaar, Krieger and Vriend. This is chapter 25 from the 2009 book Structural Bioinformatics (2nd edition; Bourne PE & Gu J eds).

Some more additional reading.

For the ones that look forward to read about current and future bioinformatics applications in the field of structure-function relationships, please have a go at the following review by Marco Wiltgen. Parts 1 to 10 and from part 14 onward contain the most useful items – unless you like quantum theory and the lot, of course.

Course Materials

- Wiltgen M. (2009) Structural Bioinformatics: From the Sequence to Structure and Function. Current Bioinformatics 4: 54-87.

6 – The many 'omics' data

Introduction

Already at the turn of the century it was obvious that the Molecular Life Sciences field was preparing for a so-called post-genomic era; in addition to the complete blueprint of the genetic make-up of a cell, tissue or organism, new techniques allowed the probing of large sets of RNA molecules (transcriptome), proteins (proteome) or metabolites (metabolome) as well. Various functional genomics, proteomics and computational techniques have been developed, both for qualitative and quantitative assessments, and all have in common that they result in the production of extremely large datasets as output material. Needless to say that Bioinformatics approaches are essential here to make some sense out of this data avalanche.

These 'omics' approaches triggered new issues in the statistical analysis of data. It also became clear that interoperable data repositories (thus based on standard annotations, infrastructures and services) are needed to support the pooling and meta-analysis of data and their comparison to earlier experiments. Furthermore, quality standards had to be formulated to safeguard the scientific community from improper data floods that would obstruct effective Bioinformatic analyses.

Before we deal with these 'omics' strategies, you may prefer to have a look at the following textbook pages on which several of the sequencing, hybridisation, purification and mass spectrometric techniques that underly 'omics' approaches are briefly explained.

Course Materials

Some background reading (optional) to refresh your memory:
Molecular Biology of the Cell (4th edition) Alberts et al.

Chapter 1: Cells and Genomes
- Section: Genetic Information in Eucaryotes
  - The expression levels of all the genes of an organism can be monitored simultaneously

Chapter 7: Control of gene expression
- Section: An Overview of Gene Control
  - Different Cell Types Synthesize Different Sets of Proteins (esp. the accompanying figures)
- Section: DNA-Binding Motifs in Gene Regulatory Proteins
  - The DNA Sequence Recognized by a Gene Regulatory Protein Can Be Determined
  - A Chromatin Immunoprecipitation Technique Identifies DNA Sites Occupied by Gene Regulatory Proteins in Living Cells

Chapter 8: Manipulating proteins, DNA and RNA
- Section: Fractionation of cells
  - More Than 1000 Proteins Can Be Resolved on a Single Gel by Two-dimensional Polyacrylamide-Gel Electrophoresis
  - Selective Cleavage of a Protein Generates a Distinctive Set of Peptide Fragments
  - Mass Spectrometry Can Be Used to Sequence Peptide Fragments and Identify Proteins
- Section: Analyzing Protein Structure and Function
  - Affinity Chromatography and Immunoprecipitation Allow Identification of Associated Proteins
  - Protein-Protein Interactions Can Be Identified by Use of the Two-Hybrid System
- Section: Studying Gene Expression and Function
  - Microarrays Monitor the Expression of Thousands of Genes at Once
  - Large Collections of Tagged Knockouts Provide a Tool for Examining the Function of Every Gene in an Organism

21st century approaches

Genomics

More than two decades ago the significance of determining the entire human genome sequence was recognized. It was apparent that this goal was achievable with existing sequencing, cloning and mapping techniques. Tiling arrays of large-insert clones - supplemented with cosmid and phage lambda clones - were subjected to Sanger's dideoxy chain termination technique by the Human Genome Project at NIH. The Institute for Genomics Research (TIGR) and Celera Genomics, led by J. Craig Venter, rather employed a random shotgun sequencing method that heavily relied on computer analysis to assemble the overlapping sequences.

In the course of these endeavours new DNA sequencing technologies (e.g. pyro-sequencing and single-molecule sequence-by-synthesis) were repeatedly developed that made sequence analysis much cheaper and faster. The thousand dollar human genome is now within reach and the price for deciphering a prokaryotic genome is down to a few dollars. On NCBI and EBI websites you will encounter many completely sequenced genomes of a wide variety of species, all amenable to Bioinformatics analyses. The so-called next generation sequencing (NGS) methods (that allow simultaneous sequencing of millions of fragments) and their many implications will be discussed in the next chapter.

Transcriptomics

The human genome was estimated to code for over 100,000 transcripts. However, Bioinformatics approaches led to a considerable revision of the transcriptome size; estimates range from 20,000 to 25,000, which only just exceeds that of *C.elegans*. It is more useful, of course, to investigate which of these genes are indeed expressed in a particular cell/tissue type. For that reason oligo-nucleotides corresponding to every open reading frame (ORF) were arrayed onto high-density slides, or chips, and hybridized with labelled cDNA populations generated from different RNA isolates. Carefully conducted microarray-based transcriptome studies (on the GEO and ATLAS sites) now serve as rich hunting grounds for researchers that use Bioinformatics to gain insight into gene expression changes as a function of disease and/or developmental stage.

RNA interference (RNAi), proposed in 1998, added a complexing layer to the regulation of the transcriptome. The impact of non-coding RNAs or microRNAs (miRNAs) in the hierarchy of gene regulation is underscored by the findings that about half of the transcripts in a cell are under miRNA control. Hence, profiling of miRNA expression patterns should complement the assessment of the ORF transcriptome. Although array-based platforms do exist for global profiling of miRNAs (microR-Nomics) their significance is still difficult to extract; hundreds of target sequences in the human genome can be potentially regulated by a specific miRNA and one target site can be under the surveillance of multiple different miRNAs. Hopes are on Bioinformatic methodologies that may allow high-throughput identification of target-miRNA combinations and faithful predictions of outcomes on transcriptome profiles.

Proteomics

Ultimately, the challenge is to determine structure - function relationships for the collection of proteins that make cells tick. Confoundingly, most of our genes produce multiple transcripts, e.g. due to alternative splicing, and thus encode different protein products which may have related, opposite, or even entirely different functions. Therefore, initial proteomics studies focussed on prokaryotes or unicellular eukaryotes (yeast), applying two-dimensional (2-D) gel electrophoresis and subsequent identification of the protein spots by using protein fragmentation and mass spectrometry. Techniques such as difference gel electrophoresis (DIGE) enabled the use of a single gel to separate a mixture of several differentially labelled protein samples: relative abundance of individual proteins was reflected by the ratio of the fluorescent tags in the merged images.

Thanks to the application of different tagging strategies (e.g. iTRAQ, isobaric tag for relative and absolute quantitation; SILAC, stable isotope labeling by/with amino acids in cell culture) quantitative proteomics has come of age. Furthermore, capillary electrophoresis or chromatography in tandem with highly improved mass spectrometric strategies is now enabling the analysis of much more complex protein mixtures. However, the dynamic range of protein concentrations in mammalian cells still forces investigators to select out a specific part of the proteome they want to invest. For example, 'protein machines' representing multiprotein complexes are usually first affinity purified using

strategies like TAP-tagging (application of two distinct tags on a 'bait' protein to allow Tandem Affinity Purification). By systematically TAP-tagging each protein one could decipher the interactome within a cell.
A complexing issue arises when thinking about the way how proteins are identified in mass spectrometric analyses. MS identification relies on a series of perfect matches of atomic weights of peptide fragments with predicted ones based on protein databases. Thus, new proteins - or proteins from organisms for which the 'ORFeome' is far from completed - might be missed. Furthermore, simple analyses will not incorporate the (N-terminal) processing, phosphorylation, methylation, acetylation, glycosylation, and ubiquitination status of individual proteins. As we speak MS-based methods are being developed or improved to investigate these post-translational modifications or effectively incorporate their possibility in the search algorithms used for protein identification. Of course, selecting out the modified proteins or protein complexes onto (antibody) beads for high-throughput sequencing by mass spectrometry is an appealing route.

Interactomics
Simply documenting the genome, transcriptome and proteome within a cell will not suffice to understand their functioning. The next step would be to document the combinatorial complexity of the biomolecules. Thus we should map out the interactome. Traditionally, co-immunoprecipitation experiments have been used for this and the advent of the yeast two-hybrid assay enabled a more large-scale investigation of protein interaction networks. The power of TAP tagging in combination with MS identification has been mentioned already. But clearly there is a need of more high-throughput versions to tackle the complexity of the mammalian cell's interactome. Whole-proteome antibody microarrays, to see what individual proteins ended up in complexes created with reversible cross-linking agents, come to mind but are as yet farfetched. Microscopic methods that rely on fluorescence resonance energy transfer (FRET) would perhaps allow interaction studies among a large set of proteins in the living cell, but without automation such approaches are too laborious. The analysis of membrane protein complexes poses an additional problem because agents that disrupt the membranes may also change the protein complexes themselves. Perhaps chemical cross-linking may help out here.

Metabolomics
Of course, the cell is composed out more than just DNA, RNA and proteins and in all areas of cell biology the decisive influence of metabolites on cellular processes is well recognized. Therefore, the study of the cell's small-molecule metabolite profile is gaining momentum. Techniques have evolved from biochemical use of dyes, spectroscopy and NMR to dedicated bio(reporter) assays and most notably mass spectrometry. In 2007 the Human Metabolome Project completed a first draft of the human metabolome, consisting of a database of approximately 2500 metabolites, 1200 drugs and 3500 food components, and similar projects are underway for several other organisms. As of 2011, the metabolomics web database METLIN contains over 40,000 metabolites as well as the largest repository of tandem-MS data.

Course Materials

- Kandpal R, Saviola B, Felton J. The era of 'omics unlimited. Biotechniques 46 (2009) 351-355.
- Burgun A, Bodenreider O. Accessing and Integrating Data and Knowledge for Biomedical Research. Yearb Med Inform (2008) 91–101.
- van Noort V, Snel B, Huynen MA. Exploration of the omics evidence landscape: adding qualitative labels to predicted protein-protein interactions. Genome Biol 8 (2007) R197.

Hands-on exercises

Question 6.1

APL is a type of leukemia that is caused by the translocation of two chromosomes which results in the expression of the oncofusion protein PML-RAR. PML-RAR is a transcriptional regulator that is thought to recruit epigenetic repressor proteins to *promoters* of genes thereby making them transcriptional silent. This hypothesis is based mainly on the analysis of the promoter of the RARbeta gene. Here you will determine whether this hypothesis holds true for all PML-RAR binding sites on chromosome 20.

Location of data: mb04.azn.nl/data/mb03/FG1 (or download from Blackboard)

Download:
- PML wiggle track chr20 (PMLRARchr20.wig)
- PML peak track (PMLRARpeaks.txt)
- Genomic Location files (GL-1.txt, GL-2.txt, GL-3.txt, GL-4.txt, GL-5.txt)

NOTE: To prevent overload of servers, pick randomly one of the following sites:
http://genome.brc.mcw.edu/
http://genome-mirror.duhs.duke.edu/
http://genome-mirror.bscb.cornell.edu/
http://genome.ucsc.edu/

6.1a) Determine the PML/RAR-binding sites on chromosome 20 in the available data set. Load the *ChIP-seq* data file **PMLRARchr20.wig** into the UCSC Genome Browser (Human genome, **build hg18 !!!**, March, 2006): Go to genome browser and select genomes. Select option human genome march 2006. Select option 'add custom tracks' and upload the data file (PMLRARchr20.wig) from http://mb03.extern.umcn.nl/FG1/. Select option 'Go to genome browser', hide all tracks except for 'base position' and 'UCSC genes', and view chromosome 20. The name of your track once uploaded in the Genome Browser should be PMLRARchr20.

Search for two good examples that appear to be binding sites to you. Describe why these are binding sites and explain the characteristics of binding regions versus nonbinding regions.

6.1b) In order to prevent that one has to go through the entire track manually to determine all possible binding sites in the genome, we use peak recognition programs to find peaks. This has been done for this track already, and the results are in de file **PMLRARpeaks.txt**. This file contains all PML-RAR peaks detected *genome-wide*.

Load the peak file PMLRARpeaks.txt into the Human Genome Browser and use the display option 'dense'. You now have all the 1850 binding sites for PML-RAR in the genome. Visualize your peaks on chr20 and check whether the manually chosen peaks from 6.1a) are included.

Why is one of the peaks located in region (chr20:25,822,283-26,393,096) not included?

6.1c) Where are your peaks located in relationship to annotated genes? Get a manual overview where in the genes your two peaks are located (e.g. only in the first intron, etc.). Define categories that your peak locations can fall into (e.g. 5' upstream regions or introns or…).

Upload the custom tracks (the GL-1, -2, -3, -4 and -5 **txt files**) using the 'add custom tracks' option and examine them in the genome browser. The 5 tracks each represent a category of genomic locations in relationship to annotated genes. What does each separate custom track represent?

6.1d) Select the option table browser and select 'custom tracks' in the section 'group'. At the option track you can find all your uploaded custom tracks. Select GL-1 and determine the number of base pairs covered by this category, by selecting the summary/statistics option. Select the PML-RAR peaks track (PMLRARpeaks) and intersect the PML-RAR peaks with the GL-1 track to determine how many peaks fall into this categories (the summary/statistics option) using the intersection tool.

Note that the intersection does not have to be restricted to chromosome 20 but can be performed *genome-wide*. Now, intersect the GL-1 track with the PML-RAR peaks. What is different between the two analyses?

6.1e) Perform a similar intersection for GL2-5: intersect the PML-RAR peaks with the GL tracks. How many peaks fall into each category? Does PML-RAR only bind promoters (as suggested by the hypothesis above)?

Question 6.2

A wealth of information is nowadays coupled to 'sequence database entries' and as an example you will take a look at the UNIGENE database (at the bottom of the 'search' drop-down list on the NCBI start page). At UNIGENE all kinds of expression databases are linked to your gene of interest.
Over the years, researchers have constructed copy-DNA (cDNA) libraries of entire mRNA populations from tissues or cell types, by reverse-transcriptase reaction and subsequent cloning into (plasmid/phage) vectors. By end-in sequencing all those cDNA library inserts, a good impression of the relative occurrence for each mRNA type was obtained (the short stretches of cDNA sequences can be unambiguously identified using database searches; such sequence reads are referred to as EST's, expressed sequence tags). Such random EST / cDNA libraries serve to extract an expression profile for the gene of interest by simply clicking "EST profile". The computer then counts in EST libraries the relative frequency of occurrence of the requested sequence. In this way an 'electronic Northern blot' is obtained on which expression patterns are visualised as ovals with different grey values (like 'band intensities' resulting from RNA blot hybridisations in former times). Nowadays, 'next generation sequencing' techniques allow direct sequencing of such cDNA pools in a single step ('deep sequencing') without the need to first clone them.

6.2a) Describe the pattern of expression (highly or lowly expressed, is it everywhere – ubiquitous – or rather tissue or developmental stage specific?) that emerges from the EST profile of p53.

Hint: start on the Unigene page for human TP53 (tumor protein p53; UniGene ID (UGID) 2723799; UniGene entry number Hs.654481).

6.2b) Describe the pattern of expression that emerges from the EST profile of myod1.

Hint: you will need the Unigene page for human MYOD1 (Myogenic differentiation 1; UGID 159276; entry Hs.181768).

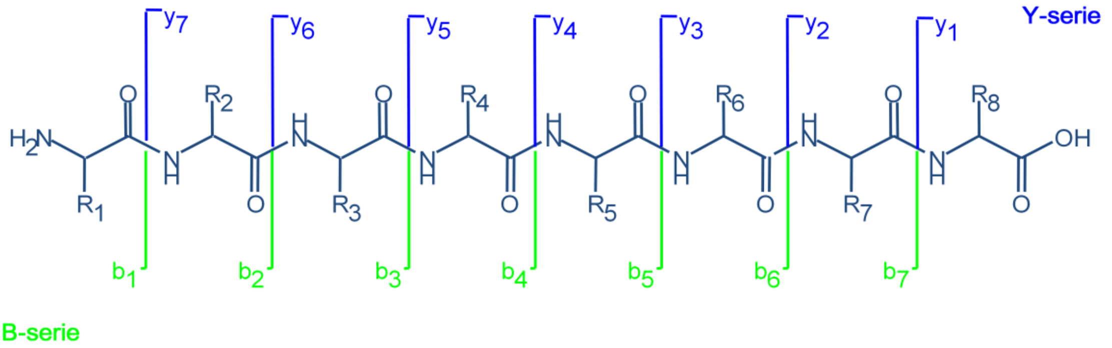
Question 6.3

Suppose you ran a two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) with a sample that contains proteins that were affinity-purified using your favourite protein as bait. Consequently you now have a pretty good idea about the apparent molecular weight (Mw) and the isoelectric point (pI) of the interacting proteins that stained on your gel. A Bioinformatics step towards the identification of these proteins would be to use tools available on the Expasy proteomics server that also hosts the UniProt protein database (consisting of the manually annotated SwissProt and the automatically annotated TrEMBL databases).

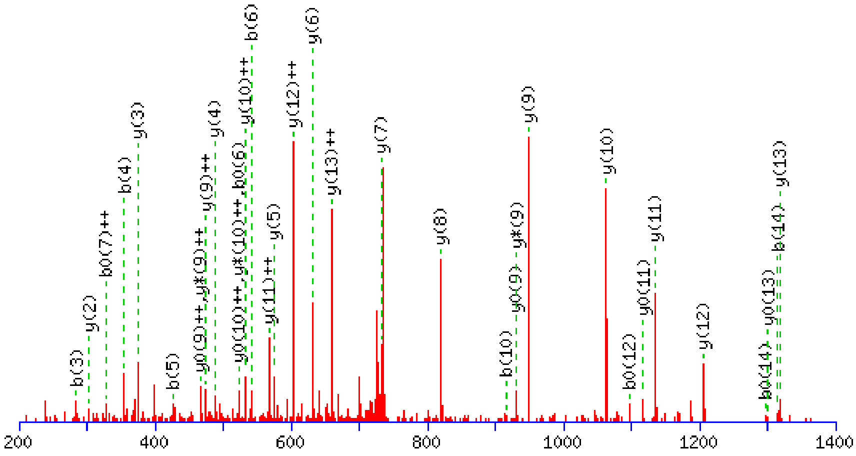
6.3a) Use the TagIdent tool (under 'Resources') to generate a list of candidates for the human (species 9606) protein that displayed a pI value of 4.9-5.1 and an Mw value of 200 kDa (within a 5% range) on the 2D gel.

Question 6.4

In "bottom–up" proteomics experiments, proteins are typically digested with trypsin prior to mass spectrometric (MS) analysis and subsequently the masses of the tryptic peptides and their fragmentation products are measured. Tryptic peptides usually fragment in many shorter peptides with positive charges. Peptides fragment into so-called Y- and B-ions, which are determined by the location of the positive charge of the fragment ion, either at the side chain of R/K (Arg or Lys) residues at the C-terminus, or at the N-terminus respectively. The mixture of Y and B ion fragments provide sequencing information since the mass difference between two subsequent Y/B fragment ions equals the mass of one amino acid building block.



In a human sample we acquire the following spectrum of a peptide with m/z=744.906438, and charge =2+.



mass	Y-ion
1205.612	12
1134.575	11
1063.538	10
948.5109	9
819.4683	8
732.4363	7
631.3886	6
574.3671	5
487.3351	4
374.251	3
303.2139	2

6.4a) Determine the sequence of the peptide from the Y-ion fragment ions using the *mono isotopic weights* of the amino acids from http://www.ionsource.com/Card/aatable/aatable.htm.

6.4b) Then, use a search alignment tool at http://expasy.org/tools/#proteome to identify the protein.

6.4c) Alternatively, the protein can be identified directly with the MS search program PEPFRAG at http://prowl.rockefeller.edu/prowl/pepfrag.html.

HINT: Limit your search to 'Homo sapiens' proteins and provide input on the 'Enzyme' used, the 'Mass of parent peptide' and its charge state (in the dropdown box on the same line), and the 'Fragment ion masses' (list them as separate lines in the entry field) from the table.

7 – Outlook

Introduction

DNA sequencing platforms that are bought today will be out of date within a few years and already commercial centres are offering complete genome sequencing services to individuals. Mass spectrometry set-ups have undergone a tremendous face lift and their application within the life sciences is still expanding. Many other automated and high-throughput analysis systems are constantly being developed and improved, e.g. multidimensional live cell video microscopy applications. All these techniques produce massive amounts of data that all need to be stored and analysed. The limiting factor is no longer the acquisition of data; bottlenecks in data storage, exchange, analysis and visualization that is now determining the speed of things. Needles to say that this is urging for up-to-date bioinformatics solutions.

Importantly, these 'new generation' of giant data sets require the development of new algorithms to address f.e. statistical issues in Whole genome sequencing (WGS) approaches to link complex disease phenotypes to predisposition genes. And how to visualise such giga-amount of data so that one can make sense out of it? High-throughput cell transfection experiments that involve qualitative and quantitative measurements with high-content microscopic set-ups ask for high-tech image analyses software to process the data to useful databases filled with parameters that can be subjected to cluster analyses and coupled to GO databases etc. The many 'whole cell' descriptive techniques that produce all those data allowed a new field to emerge: systems biology. Systems biology focuses on the modelling of a biological system of complex interactions (e.g. a cell) based on the available parameters and then use the model to extract properties of the system (i.e. run a 'simulation' in silico) that subsequently can be put to the test in experiments. As such it is a hypothesis-generating part of the molecular life sciences, that will be instrumental in translating all the 'omics' data into biological processes and principles.

Course Materials

- Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J Genet Genomics 38 (2011) 95–109.
- Verveer PJ, Bastiaens PIH. Quantitative microscopy and systems biology: seeing the whole picture. Histochem Cell Biol 130 (2008) 833–843.
- Laubenbacher R, Hower V, Jarrah A, Torti SV, Shulaev V, Mendes P, Torti FM, Akman S. A Systems Biology View of Cancer. Biochim Biophys Acta 1796 (2009) 129–139.

Some future developments

Next generation sequencing
The Sanger sequencing method is based on the application of dideoxyribonucleotides to induce premature stops in newly synthesized DNA strands, and it proved to be quite suitable for automation. Nowadays, the next generation sequencing (NGS) platforms – although still using the same biochemical reaction - produce orders of magnitude more data through massive parallel sequencing. This has become feasible due to the following changes. Preparation of DNA samples is now different: the first generation apparatuses needed clonal templates for each individual reaction; nowadays a pool of templates can be sequenced 'in one go' by means of (semi-)solid single molecule synthesis on millions of locations on one carrier (chip). Broadly speaking, the template DNA is sheared to create (genomic or cDNA) fragment libraries that are subjected to massive parallel clonal amplification of individual DNA molecules. These are then sequenced to generate short reads (30 to 500 bp, depending on the platform), from which the starting template sequence is reconstructed. Although the read lengths from Sanger sequencing may even exceed 1 kbp, it is the sheer magnitude of reads in NGS that provides the benefit. The current computing powers and the availability of reference sequences make that the many short reads can be efficiently aligned into 'genome-wide contigs'. As a result, NGS can generate whole-genome data sets, such as miRNA and ChIP-Seq, and detects somatic mutations much more sensitively, at ≤1% frequency.

As we speak, third-generation platforms are being developed (zero-mode waveguides, semiconductor and nanopore sequencing) that promise even larger and faster data generation. It is therefore important to realize the potential applications of 'collecting more and more sequence data'. The possibility to generate complementary data sets - from genome DNA sequencing, miRNA, both transcriptome sequencing and transcriptome quantification to epigenetic changes of DNA methylation, histone post-translational modifications and even protein translation (ribosome profiling) - of a given sample in a 'simple overnight experiment' opens up many avenues that are already being pondered. Data from NGS of whole human exomes are being successfully applied to identify mutations in genes underlying rare Mendelian disorders and to increase resolution of genetic linkage in complex trait genetics studies. Determination of a patient's full genetic make-up may allow the prescription of a medicine for a disease that fits the personal profile: efficacious, the right dose, with the least risk of adverse effects: personalised medicine or targeted therapy. Application areas also include molecular diagnostics research (e.g. to determine cell fate in clinical gene therapy studies) and re-analysis of preclinical trials where drugs have failed despite relatively high rates of efficacy. NGS proves a powerful tool in tumour diagnostics: identification and cataloguing of mutations can implicate new pathways in tumourigenesis, or identify tumour-specific mutations that may allow targeted therapies. It will aid clinical decision making and guide the choice of drug treatment. NGS can also be used as a quality-control tool to detect adventitious viruses in vaccines, using a metagenomics approach.

As mentioned previously, huge challenges in storage, transfer and analysis of NGS data lie ahead. At an institutional level, the investments in NGS instruments should be equalled by investments in upgrading the infrastructure and hiring staff for adequate data storage (servers, backup), transfer (network bandwidth) and analysis. Many short read analysis tools are available, each with its own limitations. Massive parallelism or cloud computing may solve hardware limitations. It is desirable to standardise NGS protocols and data formats, and indeed a sequencing quality control (SEQC) project was recently initiated by the FDA. The robustness and reproducibility of NGS facilitates large knowledge-gaining experiments, including metagenomics to compare different disease states or patient variability, genome sequencing of model organisms. Key to successful application, however, will be the ability to extract biological meaning from the increasingly detailed data sets, requiring development of sophisticated algorithms.

Quantitation in cell biology

Apart from sequences, also other experimental set-ups within the life sciences are now at a level that they can generate enormous amounts of data that require proper bioinformatics applications. Especially, automated image acquisition in microscopy, together with the introduction of the third dimension and time-lapse video imaging possibilities, and combined with multi-parameter readouts, have resulted in an explosion of data requiring adequate processing. Most automated image analysis systems usually are tailored for specific microscopes and a few standard applications. Pattern recognition software, originally developed for remote sensing, is now becoming a promising alternative. Instead of fixed algorithms and tuning parameters to process images, the computer is rather trained to recognize patterns in images. This more general approach may enable data mining in image repositories. One of the most popular image analysis packages is actually freeware and benefits from input coming from the worldwide scientific community: ImageJ (at an NIH sever). A slicker version is mounted on a server at the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden, Germany: Fiji (Fiji Is Just ImageJ).

The development of software-based (semi)automated analyses of biological data created the field of high-content screening: a discovery method based on the collection and analyses of multiple images / parameters of living cells. Fluorescent dyes, antibodies or tagged proteins in cells enable the measuring of changes in cellular properties that are triggered by f.e. chemical inhibitors or RNA interference. Use of microplates and robotic handling allows so-called high-throughput screening: many different compounds (shRNA libraries etc) can be simultaneously tested for effects on cells, in an automated fashion. The acquisition of multiple parameters of course is not limited to microscopy; also in flow cytometry (FACS) and quantitative (immuno)blotting, ELISAs etc multi-parameter analyses can be devised. Conglomerates of labs are nowadays collaborating to compile dozens of data types for hundreds of different conditions / time points to collectively build a descriptive database for a particular cellular or organismal behaviour. A nice example is The Alliance for Cellular Signaling (AfCS) that performs comprehensive experimental analyses of selected signalling systems and archives these data for (free) use by the research community (see f.e. at their Data Center the detailed ligand screen performed with a macrophage cell line). Such data serve as input for *systems biology* approaches.

Systems Biology

Only a decade ago the term 'systems biology' started penetrating the literature, so one may state that the field is still in its infancy. The omics-type of experimental data is expanding exponentially and cannot be analyzed or interpreted at the same pace with current bioinformatic methodology. Partly as a response to that, researchers are now developing frameworks by which they can model biological systems from that plethora of systematic measurements. But systems biology is more than just dealing with the 'omics' data workload. It is the field where almost philosophical thoughts lead to the development of model systems, consisting of a collection of equations and parameters, with the purpose to mimic real-life; to be able to simulate the biological processes and faithfully predict the outcome of experimental data. Once such a model can stand the test of 'reality' it turns into a hypothesis-generating machine: in silico alterations of parameters may predict unusual or unprecedented characteristics of the biological process that subsequently can be put to the test experimentally.

An important principle emerging from systems biology research is that in addition to mapping out the physical components and interactions of a system it is equally important to investigate how the system behaves in response to perturbations. Very stimulating examples come from the application of modular response analysis (MRA) to study cellular regulatory networks. The complexity of, for example, signal transduction circuits in cells must be enormous given the multitude of protein isoforms and post-translational modifications involved. MRA simplifies the analysis by assuming a modular network organization. Steady-state responses are then expressed as inter-modular interactions and processes operating within the modules can be ignored. Growth factor-induced mitogenic signalling through the MAP kinase cascade, for example, has been studied in this way and the systems biology approach yielded paradigm examples of regulatory feedback circuits.

Please don't confuse systems biology for synthetic biology. Systems biology aims at modelling natural biological systems; synthetic biology is about constructing new genetic and biochemical systems.

End of the Module - How to Proceed

As stated before, this Introductory Bioinformatics Module was created to help you to prepare for the upcoming MMD Masters Program. By reading this course manual and the additional literature and through hands-on experience with essential bioinformatics tools on the web you will now have a basic understanding of the potential of Bioinformatics for the molecular life science field.

You have documented your struggle to solve the bioinformatics questions that were posed in this manual. Please **upload** these files onto **Blackboard**; a dedicated "Assignment" for each chapter has been created for this purpose (http://blackboard.ru.nl/webapps/portal/frameset.jsp). Simply follow the Blackboard Assignment instructions on the screen. We will get back to you then, by sending our pre-assembled answers for comparison.

We also kindly ask you to fill out the **Questionnaire** that is available on **Blackboard**, under "Assignment". Since this is version 1.1 of the MMD Bioinformatics pre-Module we highly welcome any **feedback** from your side to help us improve the course. What parts are obsolete, what is lacking, how about the required time to perform the various parts, is it doable with slow internet connections, etc. – any comments on these and other issues will be deeply appreciated. This will allow us to improve the Module in order to serve future students even better.

On behalf of the MMD Bioinformatics pre-Module team we would like to thank you for your participation and feedback, and hope that you will experience a pleasant and successful start of your Masters study in Nijmegen.