Informatics and Statistics for Molecular Biologists (MOLB 7900)

Jay Hesselberth

2019-05-09

Contents

Course Overview

Informatics and Statistics for Molecular Biologists (MOLB 7900) is offered at the University of Colorado School of Medicine and teaches students how to design and analyze common molecular biology experiments.

Syllabus

0.1 Course Overview

Informatics and Statistics for Molecular Biologists (MOLB 7900) teaches students to design and analyze experiments commonly used in molecular biology. The course is organized around the Central Dogma (DNA > RNA > Protein) wherein each block presents 2-3 experimental approaches. Each week, a new experiment is introduced with a discussion of appropriate design and statistical considerations. The remaining weeks' classes are devoted to digging into the analysis of a sample data set.

The course begins with a 4 week "boot camp" designed to get students familiar with and bring them up to speed on using software for shell programming and data analysis with R and Python. We also establish accounts on Github for problem set submission.

0.1.1 Block 1: Bootcamp

```
A: Shell programming
B: R Studio
C: Integration with Git
D: Python
```

0.1.2 Block 2: DNA

```
A: Chromatin accessibility
B: Motif finding
C: Variant Calling
```

0.1.3 Block 3: RNA

```
A: Quantitative PCR
B: Bulk mRNA-seq
C: RNA-protein interaction (CLIP)
```

0.1.4 Block 4: PROTEIN

```
A: Mass spectrometry (counts)
B: Densitometry of gel images
C: Fluorescent protein localization (Chad)
```

0.2 Course Objectives

- Cater to students of many backgrounds (more computational or more biological)
- Be able to formulate questions that are testable with computational techniques
- Understand the limitations of sequencing-based techniques
- Be fluent in statistical considerations for different approaches and design well-controlled experiments that can be analyzed using statistical tests
- Be fluent in command-line programming, scripting (Python) and data analysis / viz (R / R Studio)
- Understand the value of internet-based analysis tools (NCBI BLAST etc)
- Use reproducible software development approaches (Github) and dynamic documents (Rmarkdown)
- Independently conceive of and implement a soup to nuts reanalysis of an existing data set, which are presented to the class.

0.3 Class Schedule

Monday, Wednesday, and Friday, 1-2:30pm from Sept 10 to Dec 10.

Chapter 1

Bootcamp

This 4 week "boot camp" is designed to get students familiar with and bring them up to speed on using software for shell programming and data analysis with R and Python. We also establish accounts on Github for problem set submission. In future blocks, students will build upon the foundations they acquire in the bootcamp and combine it with domain-specific knowledge of specific experiment types, data, and statists for each major step in the central dogma.

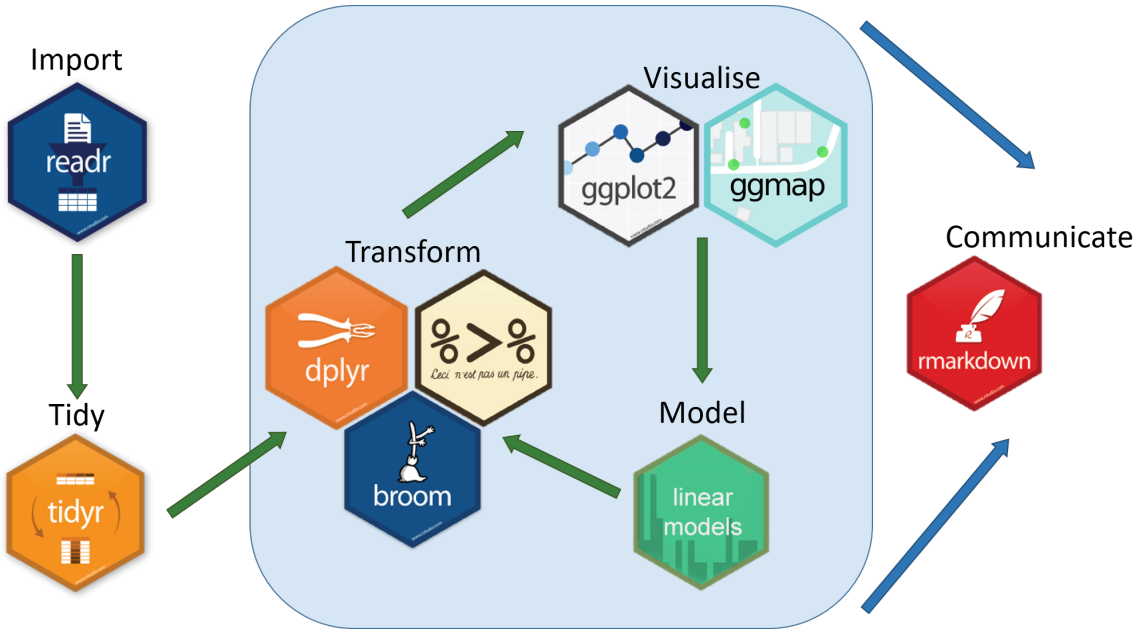
1.1 Shell programming

Students will become familiar with the operating in the Shell, which is closely related to command-line/terminal. The Shell is a program which runs other programs rather than doing calculations itself. Bash is the default shell on most modern implementations of Unix, Mac, and in most packages that provide Unix-like tools for Windows.

After this section, students will be able to navigate directories, create an organized directory structure for a project, and install and run software.

1.2 Rstudio

Students will learn to use Rstudio, which is an integrated development environment that makes it way easier to code in R. Within Rstudio, students will use Rmarkdown to produce high-quality reports, presentations, and documents in a highly reproducible manner. Students will become fluent in importing, processing, and transforming data using a collection of R packages designed for data science tidyverse.



After achieving exellence in basic data handling, students will be introduced to a broad range of commonly used statistical tests and an underlying conceptual framework for deciding the appropriate statistical test. We will emphasize concepts unique to genomic/big data. These statistical concepts and tests will be revisited during applied sections of the course in which specific technology and data are introduced and re-analyzed.

After this section, students will be able to import, process, transform, visualize data; use statistical tests to analyze basic tabular data; genarate html/pdf reports of their work.

1.3 Git

Modern scientists need to write code as part of their research. This code needs to be documented just as bench experiments need to be logged in a lab notebook. Version control systems like Git, and online hosting site, GitHub, are critical tools to address these important issues. These tools allow students to track iterative changes made to their code, revert to a specific previous version, and share their code with the broader scientific community. Altogether these tools are a cornerstone of reproducible research practices that is conveniently integrated within Rstudio and Rmarkdown.

After this section, students will be able to use git within Rstudio to easily version and share their code. They will be expected to turn in assignments using git in Rstudio.

1.4 Python

Python is an easy to learn language that combines the flexibility of bash along with the conveniences of higher level languages like R. This useful for solving problems for which software does not exist, which is important for students to learn in order to anticipate unmet needs and data from new technologies/assays. Many tools for the analysis of modern datasets are available as python packages, allowing the incorporation of pre-built analysis tools within custom, made-from-scratch, code. Furthermore, python's suitability as

a scripting language makes it a natural fit for APIs of other common analysis tools. For example, image analysis pipelines can be constructed in Fiji, allowing the batch processing of many files at once.

After this section, students will be able to write basic software and scripts that will enable them to derive meaning from the large datasets typical of modern biology.

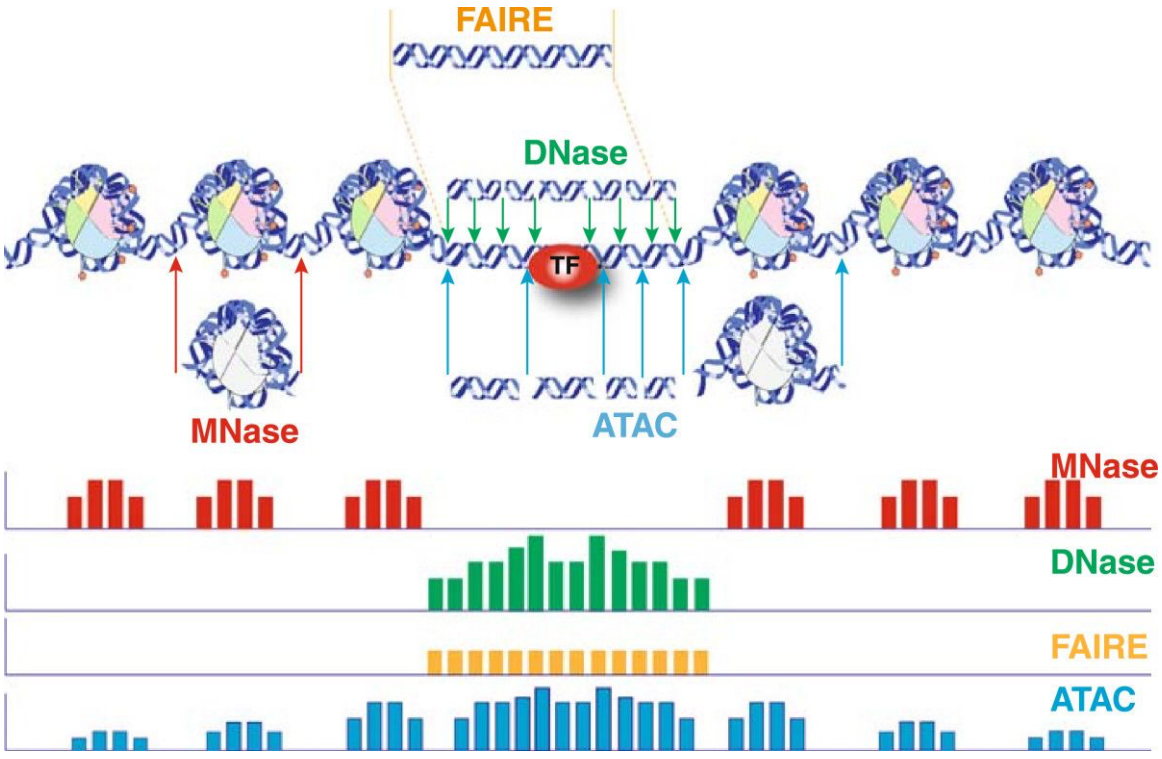
Chapter 2

Block 2: DNA

This 3 week section covers specific experiments performed on DNA sequences and the types of molecular insights they can provide.

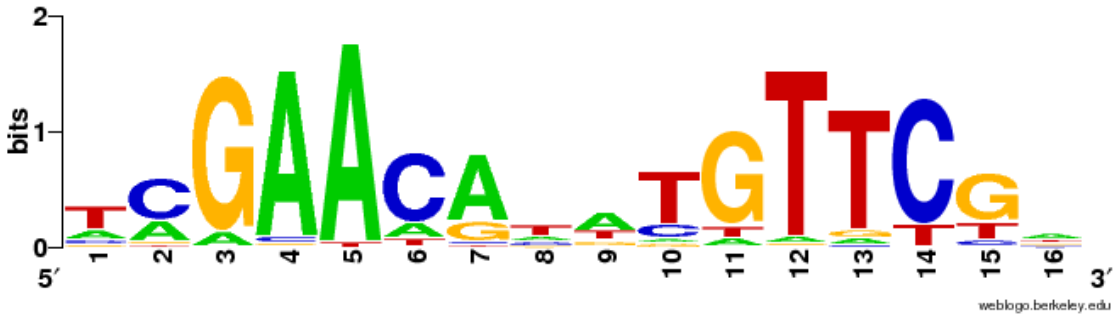
2.1 Week 1: Chromatin accessibility

We will cover different experimental techniques to study chromatin accesibility, how they relate to each other, their biological implications, and important data analysis considerations (figure from https://doi.org/10.1186/1756-8935-7-33).



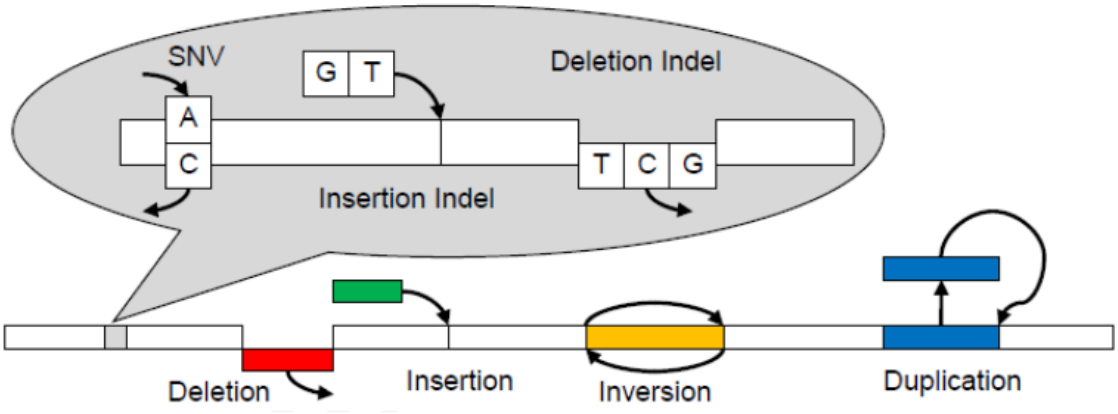
2.2 Week 2: Motif finding

Sequence elements determine the specificity of many DNA regulatory mechanisms. We will cover key principles and considerations for the discovery of these regulatory motifs from large sets DNA sequences (figure from Wikipedia).



2.3 Week 3: Variant Calling

Reliable identification and interprtation of DNA variation is important to advance our understanding of the genetic basis of disease. Here we will cover methods and considerations in calling different types of DNA variants from DNA sequence data (figure from https://doi.org/10.1093/gigascience/gix091).


SNV
G
T
Deletion Indel
A
C
T
C
G
Insertion Indel
Deletion
Insertion
Inversion
Duplication

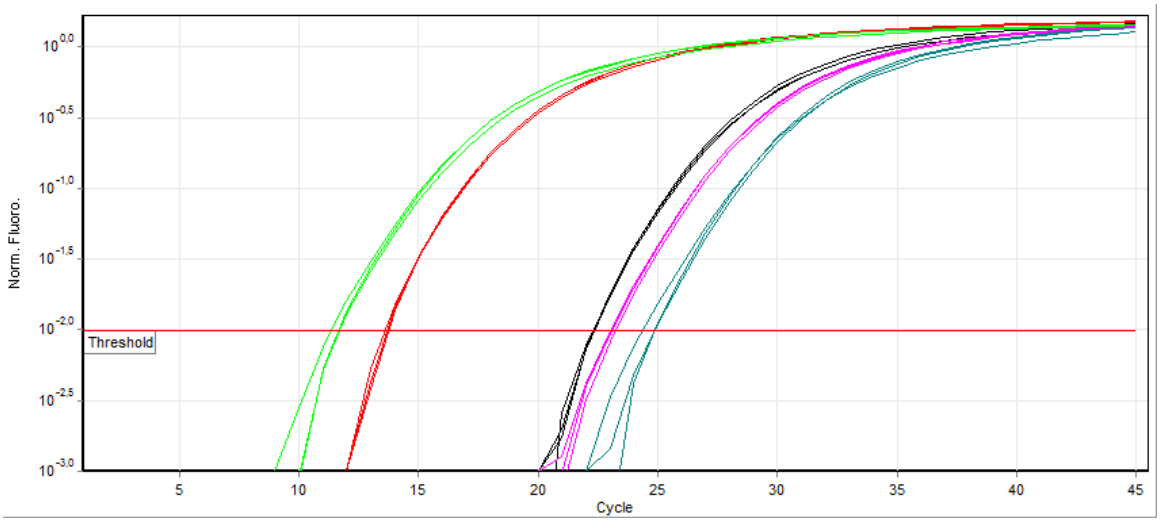
Chapter 3

Block 3: RNA

This 3 week section covers specific experiments performed on RNA sequences, protein-RNA interactionsand the types of molecular insights they can provide.

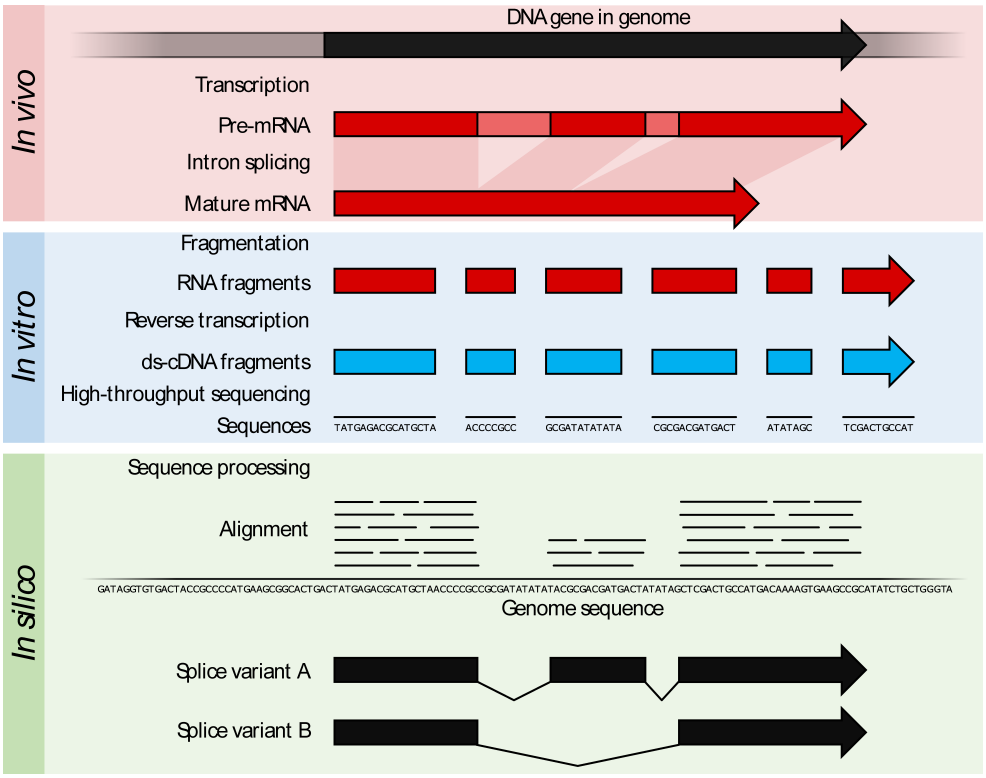
3.1 Week 4: Quantitative PCR

This is one of the most commonly used (and abused) techniques to measure the relative expression levels of specific RNAs. We will cover common pitfalls and important data analysis considerations (figure from Wikipedia).



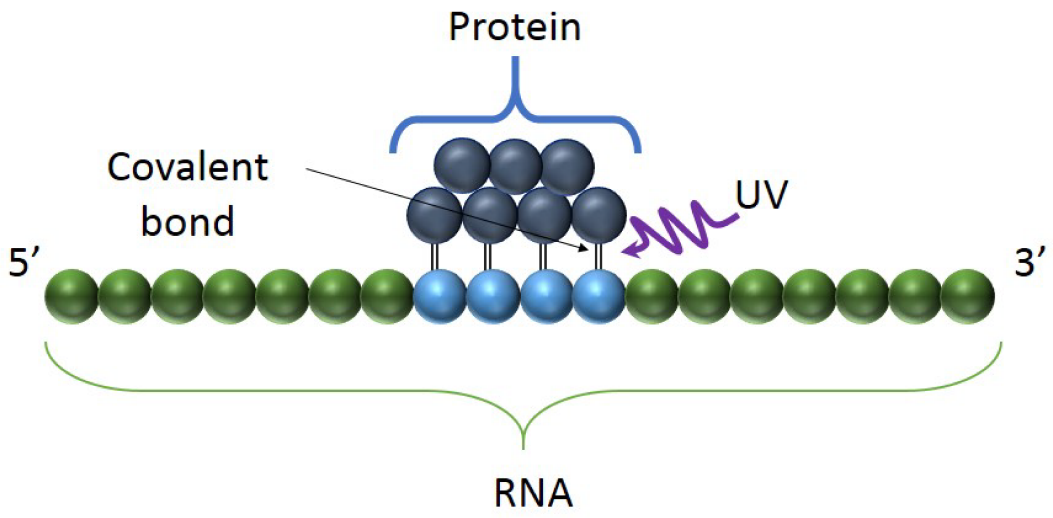
3.2 Week 5: RNA-seq

A major breakthrough in the last decade has been the introduction of RNA-sequencing, which has allowed us to identify and count the sequences of all RNAs in a biological sample. We will cover the basic principles and most common RNA-seq experimental designs and appropriate analysis methods and considerations (figure from Wikipedia).



3.3 Week 6: RNA-protein interaction (CLIP)

RNA-binding proteins (RBPs) regulate each step in the life of an mRNA. This includes key decisions like RNA splicing, translation, and degradation. We will cover methods that identify RBP-RNA interactions and the considerations in analyzing and interpreting these data (figure from Wikipedia).


Protein
Covalent
bond
UV
5'
3'
RNA

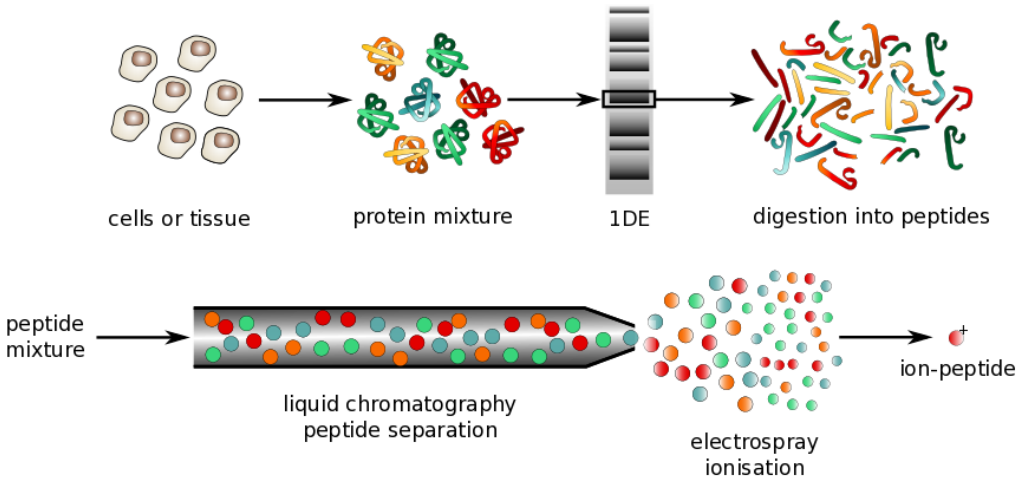
Chapter 4

Block 4: Protein

This 3 week section covers specific experiments performed on proteins and the types of molecular insights they can provide.

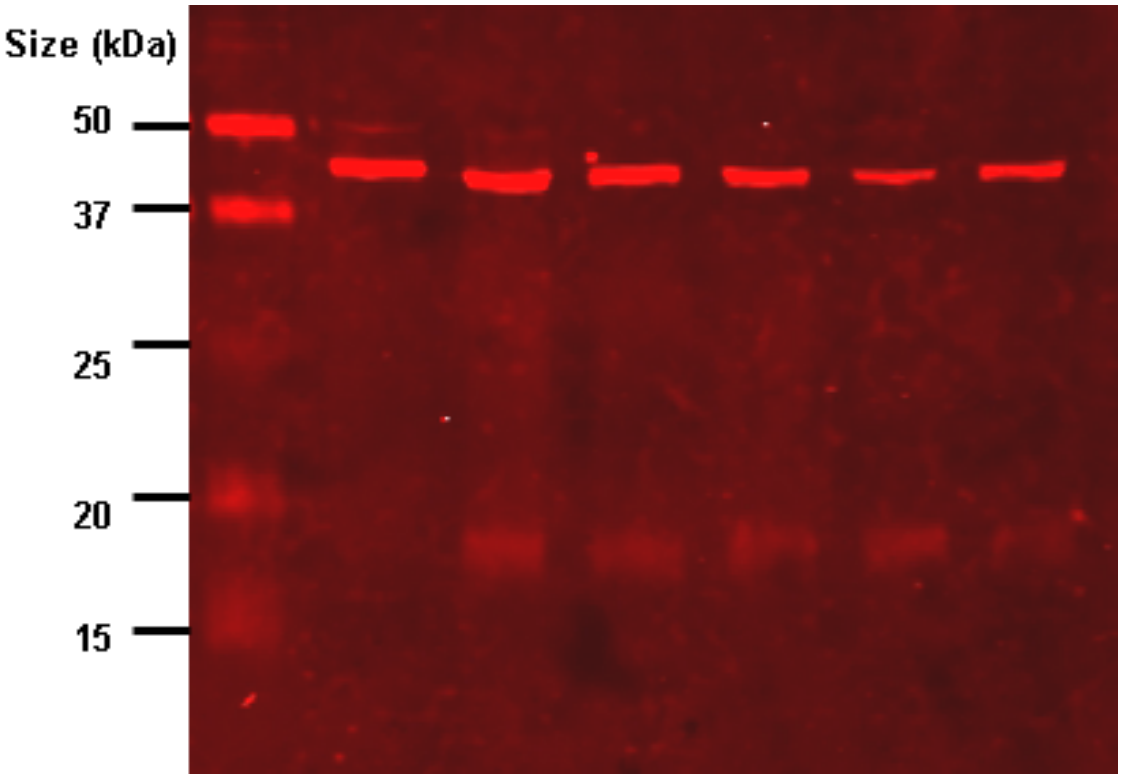
4.1 Week 7: Mass Spectrometry

Mass spectrometry (MS) is a powerful technique used to identify and quantify proteins by analyzing peptides generated from samples. We will cover the principles, variety of methods, and important data analysis considerations of MS experiments (figure from Wikipedia).



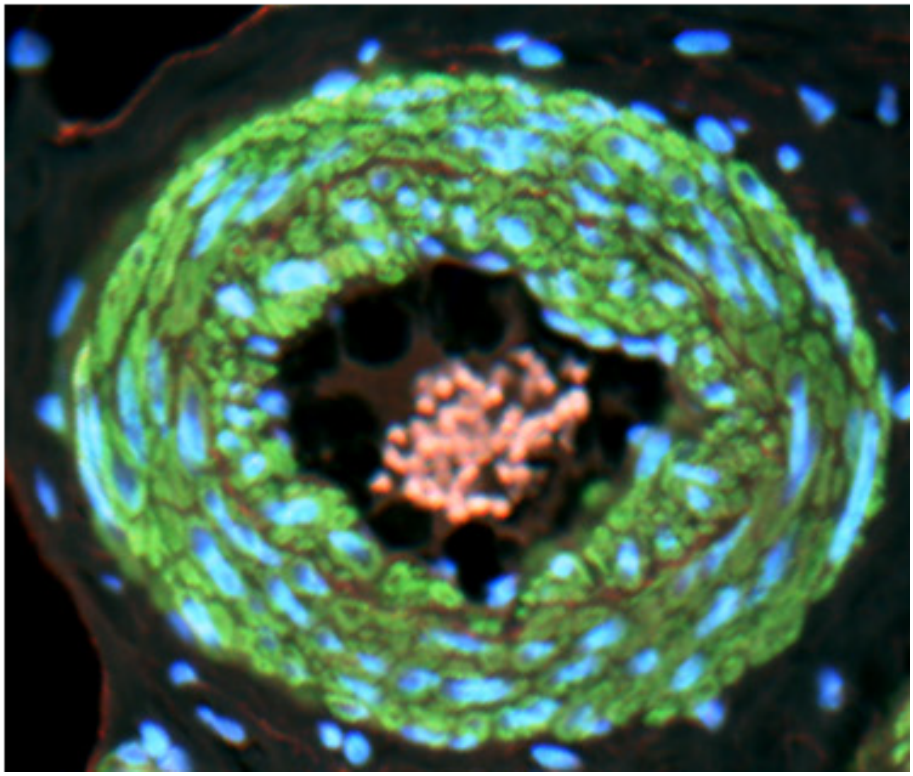
4.2 Week 8: Densitometry

A recurrent principle in molecular biology is to seprate molecules by molecular mass using gel electrophoresis, more specifically Western bloting for protein analsis. This is very common and we will cover methods to perform analysis and quantification of these experiments using densitometry methods (figure from Wikipedia).



4.3 Week 9: Immunofluoresence

Determining the precise location of a specific protein either within a cell or organ provides critical insights. We will cover important considerations and common pitfalls for the analysis of protein localization by immunofluoresence experiments (figure from Wikipedia).



Chapter 5

Exercises