1

Introduction

Numerous regulatory processes in an organism occur by changing the amount of a specific protein or a group of proteins, or by modifying proteins, and thereby making several variants of the same protein. Understanding these diverse processes is essential for progress in a long list of research areas, including biotechnology, biomedicine, and toxicology.

Given that proteins are the executive molecules in the organism, it follows that knowing which proteins are synthesized in which cells and under which conditions is vital. In other words, the identity of the proteins does not tell the whole story, of equal importance is the amount of proteins synthesized and occurring in the cells at a given time. Hence to understand the function and regulatory mechanisms of an organism, performing *protein quantification* is essential.

Before going into the details of protein quantification it is helpful to have a simple model of an organism, and knowledge about the most commonly used concepts.

1.1 The composition of an organism

An organism is a living object that can react to stimuli, grow, reproduce, and is capable of maintaining stability (homeostasis). Organism is used both for denoting individuals and a collection of individuals. Examples are viruses, bacteria, plants, animals, and individuals from these.

1.1.1 A simple model of an organism

A simple model of an organism includes the concepts *cell*, *tissue*, and *organ*, as shown in Figure 1.1.

Computational and Statistical Methods for Protein Quantification by Mass Spectrometry,
First Edition. Ingvar Eidhammer, Harald Barsnes, Geir Egil Eide and Lennart Martens.

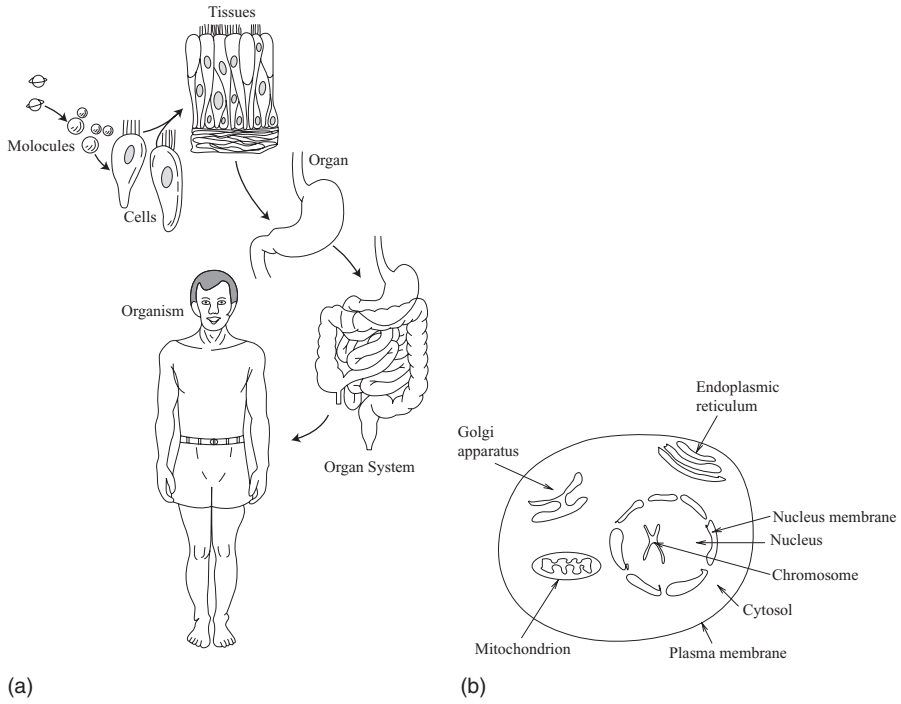

Figure 1.1 Schematic overview of an organism (a) and a eukaryotic cell (b). Figure (a) is reproduced from Norris and Siegfried (2011). This material is reproduced with permission of John Wiley & Sons, Inc.

Cell The cell is the basic structural and functional unit of all organisms. From the *Gene Ontology (GO)*[1] definition it includes the plasma membrane and any external encapsulating structures such as the cell wall and the cell envelope, as found in plants and bacteria. There are hundreds of different types of mature cells, and in addition there are all the intermediate states that cells can take during development. Cells can have different shapes, sizes, and functions.

Tissue A tissue is an interconnected collection of cells that together perform a specific function within an organism. All the cells of a tissue can be of the same type (a simple tissue), but many consist of 2–5 different cell types (a mixed tissue). Tissue types can be organized in a hierarchy, for example, the animal tissues are split into four main tissue types: epithelial tissues, connective tissues, muscle tissues, and nervous tissues. Blood is an example of a connective tissue. One assumes that there are less than 100 tissue (sub)types.

Organ An organ is a group of at least two different tissue types such that they perform a specific function (or group of functions) in an organism. The

[1] www.geneontology.org.

boundaries for what constitutes an organ is debatable, but some claim that an organ is something that one can capture and extract from the body. Organs are also sometimes collected in organ systems.

Note that the above definitions are not without issues. For example it is not always clear whether 'something' is an organ or a mixed tissue. However, they should be sufficient for our use.

1.1.2 Composition of cells

The traditional way of looking at the organization of a (eukaryotic) cell includes two main concepts, *compartment* and *organelle*:

Compartment A cell consists of a number of compartments (the term *cellular compartmentalization* is well established).

Organelle One definition of organelle, Alberts et al. (2002) is: *membrane-enclosed compartment in a eukaryotic cell that has a distinct structure, macromolecular composition, and function. Examples are the nucleus, mitochondria, chloroplast, and the Golgi apparatus*. Thus an organelle is always a compartment.

In the most general understanding of these terms one can say that there is nothing in a cell that does not belong to a compartment, but there can be something that is not an organelle. An organelle consists of one or several compartments, for example, a mitochondrion has four.

The Gene Ontology project provides a more formal definition of the composition of a cell. The top level is the *cellular component*:

The cellular component ontology describes locations, at the levels of subcellular structures and macromolecular complexes. Examples of cellular components include nuclear inner membrane, with the synonym inner envelope, and the ubiquitin ligase complex, with several subtypes of these complexes represented.

Further it is noted:

Generally, a gene product is located in or is a subcomponent of a particular cellular component. The cellular component ontology includes multi-subunit enzymes and other protein complexes, but not individual proteins or nucleic acids. The cellular component also does not include multicellular anatomical terms.

Note that the cellular component not only includes the 'static' component of the cell, but also *multi-subunit enzymes and other protein complexes*.

The Gene Ontology now contains more than 2000 cellular components organized hierarchically as a directed acyclic graph.

It should be noted that compartment is not a central term in the Gene Ontology, but it appears in a couple of places as part of a component, for example, *ER-Golgi intermediate compartment* and *replication compartment*.

1.2 Homeostasis, physiology, and pathology

Homeostasis, *physiology*, and *pathology* are frequently used terms when describing protein quantification experiments. The following definitions are from Wikipedia:

Homeostasis *is the property of either an open system or a closed system especially a living organism, that regulates its internal environment so as to maintain a stable, constant condition. Human homeostasis refers to the body's ability to regulate its internal physiology to maintain stability in response to fluctuations in the outside environment.*

Human physiology *is the science of the mechanical, physical, and biochemical functions of humans in good health, their organs, and the cells of which they are composed. The principal level of focus of physiology is at the level of organs and systems. Most aspects of human physiology are closely homologous to corresponding aspects of animal physiology, and animal experimentation has provided much of the foundation of physiological knowledge. Anatomy and physiology are closely related fields of study: Anatomy, the study of form, and physiology, the study of function, are intrinsically tied and are studied in tandem as part of a medical curriculum.*

Pathology *is the study and diagnosis of disease through examination of organs, tissues, bodily fluids, and whole bodies (autopsy). The term also encompasses the related scientific study of disease processes, called general pathology. Medical pathology is divided in two main branches, anatomical pathology and clinical pathology.*

1.3 Protein synthesis

The central dogma of molecular biology

$$DNA \rightarrow mRNA \rightarrow protein$$

gives a brief description of protein synthesis. DNA is located in the nucleus, and the mRNA is transferred to the cytoplasm, where protein synthesis takes place. The proteins can then again be transferred to other components, for example, to the nucleus or out of the cell.

1.4 Site, sample, state, and environment

Various molecular biology and medical experiments involve taking *samples* from a specific *site* from one or more research subjects. The subjects are most often of the same species, but cross-species experiments also occur.

A site is a specific 'part' of the subjects (one or several organs, tissues, or cell components), and is in one of several possible *states*. The state is specified by *features*

(or *attributes*), and the value of each feature. How to describe the possible states, that is, which features to include, depends on the context. This means that which features to include depends on the goals of the analysis.

Example Suppose that we want to analyze how some properties of a site depend on time, for example, how the amount of proteins varies during the day. Then the (only) feature used is time. Another example is investigating the effects of using a given medicine over a certain time period. In this case the features are the amount of medicine and the time elapsed.
△

When exploring how sites are affected by external stimuli, the features are often called *external features*.

The research subjects can be in a given *environment*, described by values of external features, for example, extremely low temperatures, resulting in the subject being in a frozen state.

1.5 Abundance and expression – protein and proteome profiles

Protein quantification concerns the determination of the amount of protein, relative, or absolute in gram or mole, in a sample of interest. In chemistry and biology however, *abundance* is used more or less synonymously with amount, and is in general the more popular term. We will therefore use abundance in favor of amount.

A related term is *concentration*, yet concentration is most often used to describe the ratio of the amount, mass, or volume of a component to the mass or volume of the mixture containing the component. For proteins, concentrations tend to be given in amount of protein per unit volume of mixture, for example, femtomole per microliter, fm/μl, given that most proteins are found in solution in living organisms. Note that for very low abundant proteins, the rather arbitrary unit of copy number per cell is sometimes used, for example, 100 copies per cell.

Another common term is *protein expression*. A strict understanding of this term in our context is *protein synthesis* (or *protein production*). It is however, also used for abundance, and the term *differentially expressed proteins* is often used for proteins with different abundances. We will here mainly use the term *differentially abundant*, to underline the difference between expression and abundance.

Protein abundances are commonly specified in *profiles*. We have two types of profiles, *protein profiles* and *proteome profiles*.

- A *protein profile* contains the abundances of a protein across a time series or across a set of sites/states, see Figure 1.2.
- A *proteome profile* consists of the abundances of a set of proteins from a site in a specific state.

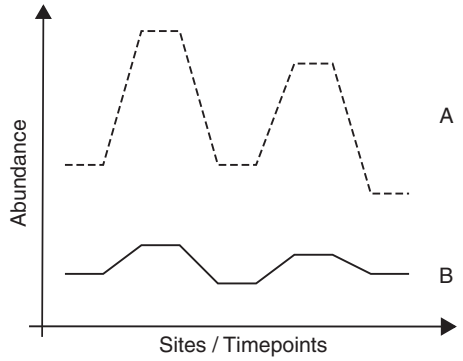

Figure 1.2 Illustration of the protein profiles for two proteins A and B.

The abundance of a protein can be determined for one research subject or for several subjects together (of the same species). If several subjects are included, the abundances are typically the average (or median) of the subjects, together with a measurement of the observed variation.

A protein profile thus contains the abundances of a single protein, while a proteome profile contains the abundances of a set of proteins.

Note that in the literature protein profile is sometimes (mis-)used for what we here define as a proteome profile. The understanding of the term should however be clear from the context in which it is used.

1.5.1 The protein dynamic range

The number of individual proteins occurring in a cell or a sample can vary enormously. It has been shown that in yeast the number of individual proteins per cell varies from fewer than 50 for some proteins to more than 10^6 for others. In serum, the difference between low abundance proteins and high abundance proteins can be from 10 to 12 *orders of magnitude*. One order of magnitude is ten to the first. The abundances of a specific protein can also vary enormously across sites/states.

To describe this variation the term *dynamic range* is often used. *Dynamic range* is a general term used to describe the ratio between the largest and smallest possible values of a changeable quantity. Dealing with large dynamic ranges is a challenge in protein quantification, and sets high requirements for the instruments used in order to be able to detect low-abundance proteins in complex samples.

1.6 The importance of exact specification of sites and states

Due to the large variations observed for protein abundances it is extremely important to clearly specify the sites where the samples were taken, and in which states the

sites were, such that the resulting profiles reflect what one wants to investigate and compare. If not, one can end up comparing profiles where unconsidered features have influenced the profiles. Any detected differences may then come from these unconsidered features, and not from the features under consideration.

Many experiments try to analyze how the profiles depend on changing the value of one or more features of the sites. The challenge is then to keep all other features (that may influence the profiles) constant. However, simply finding these features in the first place can be difficult.

An additional challenge when comparing profiles from different sites, or from the same site in different states, is to try to extract the proteins under exactly the same experimental conditions, for example, using the same amount of chemicals and the same time slots. Neglecting this point may result in differences in the profiles (partly) due to different experimental conditions.

We here divide the features into five types, and give examples of each.

1.6.1 Biological features

These are features for which the values are more or less constant, or change in a regular manner, and can most often be easily measured. Examples are age, sex, and weight.

1.6.2 Physiological and pathological features

These are features that describe the physiological state of an individual organism, such as hungry, stressed, tired, or drained. Physiological features relate to healthy individuals, while pathological features relate to diseased individuals.

1.6.3 Input features

These are features characterizing the chemical and physical elements the individual has been exposed to, over longer or shorter time periods. Examples are food, drink, medicine, and tobacco. Make-up, shampoo, electromagnetic radiation, and pollution are also examples of input features.

1.6.4 External features

These are features describing the environment in which the individual lives, such as temperature, humidity, and the time of day.

1.6.5 Activity features

These are features describing the activities of the individual over longer or shorter time periods, such as different degrees of exercising, sleeping, and working.

1.6.6 The cell cycle

During its lifetime an individual cell goes through several phases. Four distinct main phases are specified for eukaryotic cells (G_1-, $S-$, G_2-, $M-$phase). The length of a cell cycle varies from species to species and between different cell types, from a couple of minutes to several years. Around 24 hours is typical for fast-dividing mammalian cells.

For some genes the expression level depends on which phase of the cell cycle the cell is in. This results in two types of proteins:

Dynamic proteins are encoded by genes for which the expression level varies during the cell cycle.

Static proteins are encoded by genes for which the expression level is constant during the cell cycle.

Due to the dynamic proteins, the proteome profile of a cell depends on the cell phase. It is often difficult, but not impossible, to detect this information at the individual cell level. During the last couple of years methods have been developed to make this possible. However, generally a large number of cells are extracted in a sample, and an average state of all these cells is commonly (implicitly) used, often assuming homeostasis.

1.7 Relative and absolute quantification

Generally there are two types of protein quantification, *relative quantification* and *absolute quantification*. Both relative and absolute quantification are used in proteomics, depending on the goals of the experiments.

Relative quantification means to compare the abundance of a protein occurring in each of two or several samples, and determine the ratio of the occurrences between the samples.

Absolute quantification means to determine the absolute abundance of a specific protein in a mixture. This can thus be used even when analyzing just a single sample.

Due to the properties of the instruments used it is generally easier to perform relative than absolute quantification.

1.7.1 Relative quantification

Relative quantification is primarily used for comparing samples to discover proteins with different abundances (differentially abundant proteins). Usually the set of proteins occurring in the samples are very similar, and the (large) majority of them also have similar abundances in all samples.

The basic approach is to compare two samples, and the aim is to find the relative abundance of each protein occurring in the samples, and mainly those proteins with different abundances.

Relative protein abundance can be specified in different ways. Let a_1 and a_2 be the calculated abundances of a protein in sample 1 and 2 respectively. Then the relative abundance can be presented as:

- **Ratio.** Defined as $\frac{a_1}{a_2}$.
- **Fold change.** Basically the same as ratio. However, another definition is also used: $\frac{a_1}{a_2}$ if $a_1 > a_2$ otherwise $-\frac{a_2}{a_1}$. This is symmetric, but results in a discontinuity at 1 and -1, which can be a problem for the data analysis.
- **Log-ratio.** $\log \frac{a_1}{a_2}$, also called *logarithmic fold change*.

Example Let the normalized abundances of three proteins in a protein sample S_1 be (10, 20, 18) (of an undefined unit), and the same proteins in sample S_2 be (40, 20, 17). Then there is a four-fold increase in S_2 for the first protein, and little or no change in the other two proteins.
△

1.7.2 Absolute quantification

Absolute quantification means to determine the number of individual molecules of specific proteins in a mixture. A common unit used is *mole*/volume (where one mole is $6.022 \cdot 10^{23}$, Avogadro's number). Also mass/volume is used. When the mass of a protein is known it is straightforward to convert between the two.

For proteins, one needs an instrument able to measure the number of copies of the specified protein in a mixture, which (for existing instruments) is more complicated than measuring relative abundances.

Note that absolute quantification could be used to calculate relative quantification. However, since absolute quantification is not straightforward and the results often have large uncertainties, this is not very common.

An example where the absolute abundance of a given peptide is of interest is when looking for *biomarkers*, see Chapter 18, where the absolute abundance of a peptide biomarker will provide useful information about the suitability of different assays to detect this peptide in a subsequent diagnostic procedure. Several (more or less similar) definitions of biomarkers exists. A biomarker in medicine is anything (substance or method) that can be used to indicate that a specific disease is present, or that an organism has a (high) risk of getting the specific disease. It may also be used to determine a specific treatment for a specific individual having a specific illness. A biomarker can include one or several proteins.

1.8 *In vivo* and *in vitro* experiments

In vivo means 'within a living organism,' and *in vitro* means 'within the glass.' Exactly what is meant by these terms in relation to experiments can vary. Studies

using intact organisms, whether this be a mouse or a fruit fly, are always called *in vivo*. Studies where isolated proteins or subcellular constituents are analyzed in a tube, are always called *in vitro*. On the other hand, cell cultures grown in plastic dishes (a very common experimental technique) may be categorized as *in vivo* by some, while others, probably the majority, would define them as *in vitro*.

1.9 Goals for quantitative protein experiments

The overall goals for proteomics experiments are impossible to achieve with just one type of experiment. Several types of experiments (both small or large) can therefore be found. Examples of protein quantification goals are investigating how proteome or protein profiles:

- depend on changing states, for example, responding to different drugs, or different amounts of a drug;
- from similar states but varying sites differ;
- from corresponding sites from different categories of individuals differ, for example, between healthy and diseased persons;
- from corresponding sites from different species in similar states differ;
- from a site varies over time (under a constant environment);
- depend on the cell cycle.

A higher level goal is to explore if and how proteome or protein profiles can be used as biomarkers.

1.10 Exercises

1. Explain the difference between protein abundance and protein expression.
2. (a) Explain the difference between protein profile and proteome profile.

   (b) We have the following abundances (given in an arbitrary unit) for five proteins in four different cells:

	Cell 1	Cell 2	Cell 3	Cell 4
P_1	87	87	14	11
P_2	125	124	124	126
P_3	140	142	22	23
P_4	29	29	28	29
P_5	65	64	67	64

   Give examples of a protein profile and a proteome profile.

3. We have the following values for proteins in a cell: 21, 126, 2300, 560, 4700, 96 800, 19. What is the dynamic range? How many orders of magnitude are there?
4. Can you give examples of other features for each of the five types in Section 1.6?
5. A protein of mass 70 kDa has an absolute abundance of $5.1 \cdot 10^{10}$ pg/ml. What is the value in mole/ml? (Remember that 1 Da $= 1.66 \cdot 10^{-24}$ g, and that 1 mole $= 6.022 \cdot 10^{23}$.)