Proteins: Biochemistry and Biotechnology, Second edition is a definitive source of information for all those interested in protein science, and particularly the commercial production and isolation of specific proteins, and their subsequent utilization for applied purposes in industry and medicine.

Fully updated throughout with new or fundamentally revised sections on proteomics, bioinformatics, protein glycosylation and engineering, as well as sections detailing advances in upstream processing and newer protein applications such as enzyme-based biofuel production, this new edition has an increased focus on biochemistry to ensure the balance between biochemistry and biotechnology, enhanced with numerous case studies.

This second edition is an invaluable text for undergraduates of biochemistry and biotechnology, but will also be relevant to students of microbiology, molecular biology, bioinformatics and any branch of the biomedical sciences who require a broad overview of the various medical, diagnostic and industrial uses of proteins.

• Provides a comprehensive overview of all aspects of protein biochemistry and protein biotechnology
• Includes numerous case studies
• Increased focus on protein biochemistry to ensure balance between biochemistry and biotechnology
• Includes new section focusing on proteomics as well as sections detailing protein function and enzyme-based biofuel production

www.wiley.com/go/walsh/proteinsbiochemistry

Cover design: Cylinder
Proteins
Proteins
Biochemistry and Biotechnology
Second Edition

Gary Walsh
Industrial Biochemistry Programme,
CES Department,
University of Limerick, Ireland

WILEY Blackwell
This book is dedicated to a most precious collection of proteins, my children Eithne, Shane and Alice.
Contents

Preface xi
About the Companion Website xiii

Chapter 1 Proteins and proteomics 1
1.1 Proteins, an introduction 1
1.2 Genes, genomics and proteomics 2
1.3 Bioinformatics 12
1.4 Proteomics: goals and applications 14
Further reading 22

Chapter 2 Protein structure and engineering 25
2.1 Primary structure 25
2.2 Higher-level structure 36
2.3 Protein classification on the basis of structure 41
2.4 Protein structural stability 45
2.5 Higher-order structure prediction 47
2.6 Protein folding 48
2.7 Intrinsically disordered proteins 50
2.8 Protein engineering 51
2.9 Protein post-translational modification 54
Further reading 62

Chapter 3 Protein sources 65
3.1 Recombinant versus non-recombinant production 65
3.2 Approaches to recombinant protein production 67
3.3 Heterologous protein production in E. coli 72
3.4 Heterologous production in bacteria other than E. coli 77
3.5 Heterologous protein production in yeast 77
3.6 Heterologous protein production in fungi 78
3.7 Proteins from plants 80
3.8 Animal tissue as a protein source 84
3.9 Heterologous protein production in transgenic animals 85
3.10 Heterologous protein production using animal cell culture 86
3.11 Insect cell culture systems 87
Further reading 88
Chapter 4  **Protein purification and characterization**

4.1 Protein detection and quantification  
4.2 Initial recovery of protein  
4.3 Removal of whole cells and cell debris  
4.4 Concentration  
4.5 Chromatographic purification  
4.6 Protein inactivation and stabilization  
4.7 Protein characterization  
Further reading

Chapter 5  **Large-scale protein production**

5.1 Upstream processing  
5.2 Downstream processing  
5.3 Therapeutic protein production: some special issues  
5.4 Range and medical significance of impurities potentially present in protein-based therapeutic products  
Further reading

Chapter 6  **Therapeutic proteins: blood products, vaccines and enzymes**

6.1 Blood products  
6.2 Anticoagulants  
6.3 Thrombolytic agents  
6.4 Additional blood-related products  
6.5 Vaccine technology  
6.6 Therapeutic enzymes  
Further reading

Chapter 7  **Therapeutic antibodies**

7.1 Antibodies  
7.2 IgG structure and activity  
7.3 Antibody therapeutics: polyclonal antibody preparations  
7.4 Antibody therapeutics: monoclonal antibodies  
7.5 Therapeutic applications of monoclonal antibodies  
7.6 Antibody conjugates  
7.7 Bispecific antibodies  
7.8 Antibody fragments  
7.9 Engineering the antibody glycomponent  
7.10 Fc fusion proteins  
Further reading

Chapter 8  **Hormones and growth factors used therapeutically**

8.1 Insulin  
8.2 Glucagon  
8.3 Gonadotrophins  
8.4 Growth hormone
8.5 Erythropoietin 246
8.6 Other hormones 247
8.7 Growth factors 249
Further reading 253

Chapter 9 Interferons, interleukins and tumour necrosis factors 257
  9.1 Regulatory factors: cytokines versus hormones 257
  9.2 Interferons 258
  9.3 Interleukins 264
  9.4 Tumour necrosis factors 271
Further reading 274

Chapter 10 Proteins used for analytical purposes 277
  10.1 The IVD sector 279
  10.2 The basis of analyte detection and quantification 280
  10.3 Enzymes as diagnostic/analytical reagents 281
  10.4 Biosensors 289
  10.5 Antibodies as analytical reagents 295
Further reading 309

Chapter 11 Industrial enzymes: an introduction 311
  11.1 Sales value and manufacturers 313
  11.2 Sources and engineering 314
  11.3 Environmental benefits 315
  11.4 Enzyme detection and quantification 315
  11.5 Immobilized enzymes 316
  11.6 Extremophiles 319
  11.7 Enzymes in organic solvents 324
  11.8 Industrial enzymes: the future 325
Further reading 325

Chapter 12 Industrial enzymes: proteases and carbohydrases 327
  12.1 Proteolytic enzymes 327
  12.2 Carbohydrases 340
Further reading 367

Chapter 13 Additional industrial enzymes 371
  13.1 Lipases 371
  13.2 Penicillin acylase 375
  13.3 Amino acylase and amino acid production 378
  13.4 Cyclodextrins and cyclodextrin glycosyltransferase 380
  13.5 Enzymes and animal nutrition 382
  13.6 Enzymes in molecular biology 387
Further reading 390
## Chapter 14  **Non-catalytic industrial proteins**

14.1 Functional properties of proteins 393
14.2 Milk and milk proteins 397
14.3 Animal-derived proteins 408
14.4 Plant-derived proteins 411
14.5 Sweet and taste-modifying proteins 412

Further reading 414

---

**Index** 417
Preface

This textbook aims to provide a comprehensive and up-to-date overview of proteins, both in terms of their biochemistry and applications. The first edition was published over a decade ago and in the intervening period this field has continued to rapidly evolve. The new edition retains the overall structure of the original one. Chapters 1–4 are largely concerned with basic biochemical principles. In these chapters issues relating to proteomics, protein sources, structure, engineering, purification and characterization are addressed. The remaining 10 chapters largely focus on the production of proteins and their applications in medicine, analysis and industry.

Despite the similarity in overall structure, the new edition has been extensively revised and updated to reflect recent progress in the area. Relative to the earlier edition there is greater emphasis on protein biochemistry, engineering and proteomics. The production of proteins via fermentation and animal cell culture are considered in new sections, which better balance the subsequent consideration of protein purification. The protein application chapters have been updated to reflect recent trends and developments. Thus, for example, recent bioprocess developments such as the use of disposable bioreactors are considered, there is greater relative emphasis on recombinant production systems and engineered products, therapeutic antibodies now are considered in a full dedicated chapter, and newer industrial applications such as the use of enzymes in biofuel generation are also included. The chapters considering protein applications have also been strengthened via the incorporation of numerous specific commercial product case studies.

The text caters mainly for advanced undergraduate and graduate students undertaking courses in applied biochemistry/biotechnology, but it should also be of value to students pursuing degrees in biochemistry, microbiology, or any branch of the biomedical sciences. Its scope also renders it of interest to those currently working in the biotechnology sector.

A sincere note of thanks is due to a number of people who have contributed to the successful completion of this project. Thank you to J.J. Tobin, Tewfik Soulimane and Jayne Murphy for useful scientific discussions and to Angela Boyce, Madlen Witt, Martin Wilkinson, Brigit Hogan and Jimmy Kelly for helping provide many of the photographs included. I am grateful too to John Wiley & Sons for their professionalism, efficiency and never-ending patience as I spectacularly over-ran my manuscript submission date.

Gary Walsh
Limerick, June 2013
About the companion website

This book is accompanied by a companion website:

   www.wiley.com/go/walsh/proteinsbiochemistry

The website includes:

- Powerpoints of all figures from the book for downloading
- PDFs of all tables from the book for downloading
Proteins and proteomics

Throughout this book, I will consider various aspects of protein structure, function, engineering and application. Traditionally, protein science focused on isolating and studying one protein at a time. However, since the 1990s, advances in molecular biology, analytical technologies and computing has facilitated the study of many proteins simultaneously, which has led to an information explosion in this area. In this chapter such proteomic and related approaches are reviewed.

1.1 Proteins, an introduction

While we consider protein structure in detail in Chapter 2, for the purposes of this chapter it is necessary to provide a brief overview of the topic. Proteins are macromolecules consisting of one or more polypeptide chains (Table 1.1). Each polypeptide consists of a chain of amino acids linked together by peptide (amide) bonds. The exact amino acid sequence is determined by the gene coding for that specific polypeptide. When synthesized, a polypeptide chain folds up, assuming a specific three-dimensional shape (i.e. a specific conformation) that is unique to the protein. The conformation adopted depends on the polypeptide's amino acid sequence, and this conformation is largely stabilized by multiple, weak interactions. Overall, a protein's structure can described at up to four different levels.

- **Primary structure**: the specific amino acid sequence of its polypeptide chain(s), along with the exact positioning of any disulfide bonds present.
- **Secondary structure**: regular recurring arrangements of adjacent amino acid residues, often over relatively short contiguous sequences within the protein backbone. The common secondary structures are the α-helix and β-strands.
- **Tertiary structure**: the three-dimensional arrangement of all the atoms which contribute to the polypeptide. In other words, the overall three-dimensional structure (conformation) of a polypeptide chain, which usually contains several stretches of secondary structure interrupted by less ordered regions such as bends/loops.
- **Quaternary structure**: the overall spatial arrangement of polypeptide subunits within a protein composed of two or more polypeptides.
Proteins derived from eukaryotes undergo covalent modification either during, or more commonly after, their ribosomal synthesis. This gives rise to the concept of co-translational and post-translational modifications, although both modifications are often referred to simply as post-translational modifications (PTMs), and such modifications can influence protein structure and/or function. Proteins are also sometimes classified as 'simple' or 'conjugated'. Simple proteins consist exclusively of polypeptide chain(s) with no additional chemical components being present or being required for biological activity. Conjugated proteins, in addition to their polypeptide components, contain one or more non-polypeptide constituents known as prosthetic groups. The most common prosthetic groups found in association with proteins include carbohydrates (glycoproteins), phosphate groups (phosphoproteins), vitamin derivatives (e.g. flavoproteins) and metal ions (metalloproteins).

Table 1.1  Selected examples of proteins. The number of polypeptide chains and amino acid residues constituting the protein are listed, along with its molecular mass and biological function.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Polypeptide chains</th>
<th>Total no. of amino acids</th>
<th>Molecular mass (Da)</th>
<th>Biological function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin (human)</td>
<td>2</td>
<td>51</td>
<td>5800</td>
<td>Complex, but includes regulation of blood glucose levels</td>
</tr>
<tr>
<td>Lysozyme (egg)</td>
<td>1</td>
<td>129</td>
<td>13,900</td>
<td>Enzyme capable of degrading peptidoglycan in bacterial cell walls</td>
</tr>
<tr>
<td>Interleukin-2 (human)</td>
<td>1</td>
<td>133</td>
<td>15,400</td>
<td>T-lymphocyte-derived polypeptide that regulates many aspects of immunity</td>
</tr>
<tr>
<td>Erythropoietin (human)</td>
<td>1</td>
<td>165</td>
<td>36,000</td>
<td>Hormone which stimulates red blood cell production</td>
</tr>
<tr>
<td>Chymotrypsin (bovine)</td>
<td>3</td>
<td>241</td>
<td>21,600</td>
<td>Digestive proteolytic enzyme</td>
</tr>
<tr>
<td>Subtilisin (Bacillus amyloliquefaciens)</td>
<td>1</td>
<td>274</td>
<td>27,500</td>
<td>Bacterial proteolytic enzyme</td>
</tr>
<tr>
<td>Tumour necrosis factor (human TNF-α)</td>
<td>3</td>
<td>471</td>
<td>52,000</td>
<td>Mediator of inflammation and immunity</td>
</tr>
<tr>
<td>Haemoglobin (human)</td>
<td>4</td>
<td>574</td>
<td>64,500</td>
<td>Gas transport</td>
</tr>
<tr>
<td>Hexokinase (yeast)</td>
<td>2</td>
<td>800</td>
<td>102,000</td>
<td>Enzyme capable of phosphorylating selected monosaccharides</td>
</tr>
<tr>
<td>Glutamate dehydrogenase (bovine)</td>
<td>~40</td>
<td>~8300</td>
<td>~1,000,000</td>
<td>Enzyme that interconverts glutamate and α-ketoglutarate and NH₄⁺</td>
</tr>
</tbody>
</table>

The term ‘genome’ refers to the entire complement of hereditary information present in an organism or virus. In the overwhelming majority of cases it is encoded in DNA, although some viruses use RNA as their genetic material. The term ‘genomics’ refers to the systematic study of the entire genome of an organism. Its core aims are to:

- sequence the entire DNA complement of the cell; and
- to physically map the genome arrangement (assign exact positions in the genome to the various genes and non-coding regions).

Prior to the 1990s, the sequencing and study of a single gene represented a significant task. However, improvements in sequencing technologies and the development
of more highly automated hardware systems now renders DNA sequencing considerably faster, cheaper and more accurate. Cutting-edge sequencing systems now in development are claimed capable of sequencing small genomes in minutes, and a full human genome sequence in a matter of hours and for a cost of approximately $1000. By early 2014, the genomes online database (GOLD; www.genomesonline.org), which monitors genome studies worldwide, documented some 36,000 ongoing/complete genome projects, and the rate of completion of such studies is growing exponentially. From the perspective of protein science, the most significant consequence of genome data is that it provides full sequence information pertinent to every protein the organism can produce.

The term 'proteome' refers to the entire complement of proteins expressed by a specific cell/organism. It is more complex than the corresponding genome in that:

- at any given time a proportion of genes are not being expressed;
- of those genes that are expressed, some are expressed at higher levels than others;
- the proteome is dynamic rather than static because the exact subset of proteins expressed (and the level at which they are expressed) in any cell changes with time in response to a myriad of environmental and genetic influences;
- for eukaryotes, a single gene can effectively encode more than one polypeptide if its mRNA undergoes differential splicing (Figure 1.1);
- many eukaryotic proteins undergo PTM.

The last two points in particular generally signify that the number of proteins comprising a eukaryotic organism’s proteome can far exceed the number of genes present in its genome. For example, the human genome comprises approximately 22,000 genes whereas the number of distinct protein structures present may exceed 1 million, with any one cell containing an estimated average of approximately 10,000 proteins.

Traditionally, proteins were identified and studied one at a time (Figure 1.2) (see Chapters 2, 3 and 4). This generally entailed purifying a single protein directly from a naturally producing cellular source,
or from a recombinant source in which the gene/cDNA coding for the protein was being expressed. While this approach is still routinely used, a proteomic approach can potentially yield far more 'global' protein information far more quickly.

Proteomics refers to the large-scale systematic study of the proteome or, depending on the research question being asked, a defined subset of the proteome, such as all proteome proteins that are phosphorlated or all the proteome proteins that increase in concentration when a cell becomes cancerous. It is characterized by the integrated study of hundreds, more usually thousands or even tens of thousands of proteins. This in turn relies on high-throughput techniques/processes that facilitate the production, purification or characterization of multiple proteins rapidly and near simultaneously, usually by using automated/semi-automated and miniaturized processes/procedures. Standard techniques of molecular biology, for example, allow convenient global genome protein production (Figure 1.3) as well as facilitating the attachment of affinity tags to the proteins (as discussed later in this chapter and in Chapter 4), thereby enabling high-throughput purification efforts. Proteomics relies most of all on techniques that allow high-throughput analysis of the protein complement under investigation. Among the more central techniques in this regard are two-dimentional electrophoresis, high-pressure liquid chromatography (HPLC) and mass spectrometry (MS).

Before we consider the goals and applications of proteomics in more detail, it is worth reviewing these analytical techniques. In the context of proteomics, they are often applied in combination to characterize a target proteome, with electrophoretic and/or HPLC-based methods initially used to separate individual constituent proteome proteins from each other, followed by MS-based analysis. These techniques can also be used for the detailed analysis of individual proteins characteristic of classical protein science studies or, for example, as part of a quality control process for commercial protein preparations such as biopharmaceuticals. Such applications will be considered further in later chapters.

### 1.2.1 Electrophoresis

Electrophoresis is an analytical technique that separates analytes from each other on the basis of charge. The technique involves initial application of the analyte mixture to be fractionated onto a supporting medium (e.g. filter paper or a gel) with subsequent activation of an electrical field. Each charged substance then moves towards the cathode or the anode at a rate of migration that depends on the ratio of charge to mass (i.e. the charge density) of the analyte as well as on any interactions with the support medium. As described in Chapter 2, proteins are charged species, with their exact charge density being dependent on their amino acid sequence.

The most common electrophoretic method applied to proteins is one-dimensional polyacrylamide gel electrophoresis (PAGE) run in the presence of the negatively charged detergent sodium dodecyl sulfate (SDS-PAGE), and is most often used to analyse protein purity (see Chapter 4). In the case of PAGE, migration occurs through a polyacrylamide gel, the average pore size of which is largely dependent on the concentration of polyacrylamide present. A sieving effect therefore also occurs during PAGE so that the rate of protein migration is influenced by its size/shape as well as charge density.
Incubation of the protein with SDS has two notable effects: (i) it denatures most proteins, giving them all approximately the same shape, and (ii) it binds directly to the protein at the constant rate of approximately one SDS molecule per two amino acid residues. In practice this confers essentially the same (negative) charge density to all proteins. Separation of proteins by SDS-PAGE therefore occurs by a sieving effect, with the smaller proteins moving fastest towards the anode (Figure 1.4).

1.2.1.1 Isoelectric focusing

Isoelectric focusing is an additional form of electrophoresis. A modified gel is used which contains polyacrylamide to which a gradient of acidic and basic buffering groups are covalently attached. As a result an immobilized pH gradient is formed along the length of the gel. The gel is normally supported on a plastic strip. The protein solution to be applied is normally first incubated with a combination of urea and a non-ionic detergent such as Triton or CHAPS and a reducing agent to break any disulfide linkages present. This ensures that all sample proteins are completely disaggregated and fully solubilized. On application of the protein sample, the proteins present migrate in the gel until they reach a point at which the pH equals their isoelectric point (pI) (Figure 1.5).

Neither SDS-PAGE nor isoelectric focusing, by themselves, can fully separate (resolve) very complex mixtures of proteins, such as would characterize an entire cell’s proteome. Each separation mode can individually resolve about 100 protein
Figure 1.4  Separation of proteins by SDS-PAGE. Protein samples are incubated with SDS (as well as reducing agents, which disrupt disulfide linkages). The electric field is applied across the gel after the protein samples to be analysed are loaded into the gel wells. The rate of protein migration towards the anode depends on protein size. After electrophoresis is complete individual protein bands may be visualized by staining with a protein-binding dye.

Figure 1.5  Proteins are amphoteric molecules, displaying a positive, negative or zero overall net charge depending on the pH of the solution in which they are dissolved. Contributing to the overall charge of a protein are all the positive and negative charges of its amino acid side chains as well as the free amino and carboxyl groups present at its amino and carboxyl termini, respectively. The state of ionization of these groups is pH dependent. The pH at which the net number of positive charges equal the net number of negative charges (i.e. the protein has an overall net electric charge of zero, and hence will not move under the influence of an electric field) is known as its isoelectric point (pI).
bands, but when combined about 1000–2000 bands can be resolved. As such, combining them into so-called two-dimensional electrophoresis (Figure 1.6) can achieve far better resolution of a complex protein mixture, and hence this approach is often used to achieve initial separation of a protein set prior to additional proteomic analysis and individual protein identification/sequencing (usually via MS). In this context, two-dimensional electrophoresis has a number of strengths, including:

- high-resolution separation;
- straightforward technique;
- relatively inexpensive.

However, it also has a number of potential drawbacks, in particular:

- exact reproducibility of gel banding patterns often challenging to consistently achieve;
- not amenable to genuine high-throughput experiments.

1.2.1.2 Capillary electrophoresis

Capillary electrophoresis (CE) is yet another electrophoretic format, and separates molecules on the basis of charge density. In this case, however, electrophoretic separation occurs not in a polyacrylamide gel but along a narrow-bore capillary tube usually containing a conductive buffer (Figure 1.7). Typically, the capillary will have an internal diameter of 50–75 µm and be up to, or greater than, 1 m in length. The dimensions of this system yield greatly increased surface area to volume ratios (when compared with polyacrylamide gels), hence greatly increasing the efficiency of heat dissipation from the system. This in turn allows operation at a higher current density, thus speeding up the rate of migration through the capillary. Sample analysis is usually completed within 15 minutes. In some ways CE is more similar to liquid chromatography (see section 1.2.2) than conventional electrophoresis. It exhibits very high resolving power, and its short analysis time and simple instrumentation is amenable to high-throughput analysis. CE is most typically used in proteomics to achieve separation of a peptide or a protein mix, with the separated species being fed into a mass spectrometer for analysis (CE-MS).

1.2.2 High-pressure liquid chromatography

Chromatography refers to the separation of individual constituents of a mixture via their differential partitioning between two phases: a solid stationary phase and a liquid mobile phase. In the context of protein chromatography, the stationary
Phase is usually chromatographic beads, packed into a cylindrical column, and the mobile phase is usually a buffer and chromatographic separation takes advantage of differences in protein characteristics such as size and shape, charge or hydrophobicity.

Chromatography can be used at a preparative or analytical level, and both applications are considered in detail in Chapters 4 and 5. Preparative chromatography in particular is usually performed under relatively low pressures, where flow rates through the column are generated by low-pressure pumps (low-pressure liquid chromatography or LPLC). Fractionation of a single sample on such chromatographic columns typically requires several hours to complete. Low flow rates are required because as the protein sample flows through the column, the proteins are brought into contact with the surface of the chromatographic beads by direct (convective) flow. The protein molecules then rely entirely on molecular diffusion to enter the porous gel beads. This is a slow process, especially when compared with the direct transfer of proteins past the outside surface of the gel beads by liquid flow. If a flow rate significantly higher than the diffusional rate is used, protein band spreading (and hence loss of resolution) will result. This occurs because any protein molecules which have not entered the bead will flow downward through the column at a faster rate than the (identical) molecules which have entered into the bead particles. Such high flow rates will also result in a lowering of adsorption capacity as many molecules will not have the opportunity to diffuse into the beads as they pass through the column.

One approach that allows increased chromatographic flow rates without loss of resolution entails the use of microparticulate stationary-phase media of very narrow diameter. This effectively reduces the time required for molecules to diffuse in and out of the porous particles. Any reduction in particle
diameter dramatically increases the pressure required to maintain a given flow rate. Such high flow rates may be achieved by utilizing HPLC systems (also often known as high-performance liquid chromatographic systems). By employing such methods sample fractionation times may be reduced from hours to minutes, and when experimental conditions are optimized chromatographic peak width is generally reduced compared with low-pressure systems and hence resolution power is higher (Figure 1.8).

The successful application of HPLC was made possible largely by (i) the development of pump systems which can provide constant flow rates at high pressure and (ii) the identification of suitable pressure-resistant chromatographic media. Traditional soft gel media utilized in low-pressure applications are totally unsuited to high-pressure systems due to their compressibility. Traditionally, HPLC bead diameter was typically in the 3–5 µm range (although beads with diameters up to 50 µm can be used in some applications). More recent advances in bead chemistry have allowed the development of mechanically stronger, even smaller beads (diameter <2 µm). Coupled with refined high-pressure pump design, this has still further improved flow rate (speed) and resolution, and is sometimes termed ultra performance liquid chromatography (UPLC). The high resolving power of HPLC, together with fast running times, makes it a suitable proteomic technique for achieving protein separation from complex mixtures, with individual protein peaks usually being fed directly to mass spectrometers (LC-MS) for further analysis. If the protein sample being analysed is very complex, the use of so-called multidimensional LC prior to MS analysis may be required. This generally entails contiguous separation by two HPLC modalities (e.g. ion-exchange-based HPLC, followed by reverse-phase HPLC separation of various fractions eluting from the initial ion-exchange column).

### 1.2.3 Mass spectrometry

MS is the analytical technique most intimately associated with proteomics. MS separates a mixture of (vaporized and ionized) analytes on the basis of their mass to charge ratio. It can very accurately determine the molecular mass of analytes and its basic principle of operation is outlined in Figure 1.9.

MS now finds routine application in protein science, both in the context of high-throughput proteomic analysis and in the analysis of single proteins. Although applied in areas such as characterization and quality control of biopharmaceuticals,

---

**Figure 1.8** HPLC-based chromatographic separation generally gives rise to better-resolved protein peaks (a) than do low pressure-based systems (b).
Proteins (see Chapter 5), the focus in this chapter is on its use in proteomics. However, overall MS is used to:

- determine protein mass;
- generate partial or full amino acid sequence data for a protein;
- quantify the amount of protein present in a sample;
- detect and identify protein PTMs;
- detect protein modification such as oxidation, deamidation or proteolysis;
- provide some information on protein structural detail.

Ultimately, these applications rely on the fact that all the amino acids, or other constituent biomolecules of the protein (e.g. specific sugars in the case of glycoproteins), have known molecular masses, and that potential modifications to a protein's structure (e.g. a PTM or the oxidation of an amino acid) will have predictable effects on the protein's molecular mass.

### 1.2.3.1 Ionization methods

Various methods can be used to ionize analytes for the purposes of MS, including the following commonly used approaches.

- **Electron ionization** (EI), which involves bombarding the analytes with electrons.
- **Chemical ionization** (CI), in which analytes are collided with a reactive gas.
- **Fast atom bombardment** (FAB), in which the analytes are bombarded with argon gas.
- **Electrospray ionization** (ESI), in which the analytes are sprayed into an electric field.

- **Matrix-assisted laser desorption ionization** (MALDI), in which the analytes are co-crystallized with a matrix substance (a UV-absorbing substance such as sinapinic acid), followed by exposure to an electric field and a pulsed laser beam. The matrix molecules absorb the laser photons, become excited and are transferred into the gas phase along with the neighbouring analyte molecules. A proportion of both matrix and analyte molecules become ionized by this process and the applied electric field accelerates the ions towards the analyser.

The exact ionization (and subsequent analyser mode; see Figure 1.9) chosen will depend on the research question posed. Ionization methods can be classified as ‘soft’ or ‘hard’. Soft ionization methods such as ESI and MALDI can achieve ionization while leaving the protein intact (and thus are usually used if a protein's molecular mass is to be determined; this is known as ‘top-down’ MS). Hard ionization methods such as EI and FAB result in protein fragmentation as well as ionization, yielding a fragment fingerprint analysed by mass (‘bottom-up’ MS).

### 1.2.3.2 Protein molecular mass determination

‘MALDI-TOF’ MS is a popular approach for determining the molecular mass of an intact protein. As described above, the MALDI approach achieves ionization of the intact protein, which is then fed into a time of flight (TOF) analyser. As they enter the analyser tube all the protein ions have essentially the same kinetic energy and charge. Because of this, the time required for each protein ion to reach the
detector reflects its molecular mass, with smaller proteins travelling fastest. A sample size of as little as a few femtomoles \((10^{-15}\text{ mol})\) of protein is all that is required for analysis.

Alternatively, ESI-MS can be used to determine the mass of an intact protein. It is also a soft ionization method, and even non-covalent protein complexes can remain intact (giving rise to the potential for some protein interaction analysis). It is often used with a quadrupole analyser (which contains four rod metal electrodes, which effectively serve as a mass filter). As ESI processes analytes in solution, the sample can be pumped into the mass analyser continuously and thus it can be connected directly to LC or CE instruments and used for high-throughput analysis. Because the sample must be co-crystallized (dry powder) for MALDI operation, MALDI cannot be used in continuous format with pre-separation LC/CE methods.

1.2.3.3 **MS-based protein identification**

While accurate determination of a protein’s molecular mass is one application of MS, the approach finds more routine use in the identification of proteins and the determination of a partial/full amino acid sequence. Protein identification obviously forms a central element of proteomics, but these techniques can also be used to better characterize a single protein isolated via a classical protein science approach, or can be used as quality control checks on purified biopharmaceutical products in order to verify identity/sequence. The more common approaches for achieving these objectives are outlined below. As these approaches involve initially fragmenting the intact protein, followed by mass analysis of the peptide fragments, they are termed bottom-up MS analyses.

Peptide mass fingerprinting is an approach commonly used to identify proteins (Figure 1.10).
The intact protein sample is initially treated with either a proteolytic enzyme (e.g. trypsin) or a chemical (e.g. CNBr) which selectively cleaves specific peptide bonds along the protein's backbone, thereby generating a peptide mix. As each protein has its own unique amino acid sequence, each generates its own unique peptide map or fingerprint. The peptides generated are then further analysed by MS using soft ionization techniques (MALDI or ESI) that do not further fragment them. This generates a peptide mass spectrum. Identification of the protein is then undertaken by using specialist computer software that compares the experimentally determined peptide masses with theoretical digestion data for all the proteins whose amino acid sequence is known and has been deposited in sequence databases (see section 1.3).

A variant approach that is more ‘information rich’ and which can often generate complete/near complete amino acid sequence information of the protein under investigation is that of tandem MS (MS/MS) analysis. The basic approach, as the name suggests, involves interrogation of the protein using two mass analysers in sequence, in other words in tandem, separated only by a collision cell. In the case of MS/MS, the protein to be sequenced is first chemically or enzymatically fragmented. The fragments are separated along the first analyser tube. One peptide ion fragment is selected at a time and fed (alone) into the collision tube, where it collides with inert gas molecules (He or Ar). This promotes further fragmentation into a range of complementary peptides that are separated on the basis of mass in the second tube. Computerized analysis of the mass of each fragment generated in the second tube can yield nearly complete/complete sequence data.

### 1.3 Bioinformatics

A central characteristic of genomics and proteomics is the vast amount of biological data, such as gene and protein sequences, that it generates. This provides two challenges: (i) how to store all this information and (ii) how to analyse, interrogate and use this data in order to understand its actual biological significance, apply it to research questions and generate new knowledge. Bioinformatics represents the scientific discipline that addresses these challenges. It is a multidisciplinary field that concerns itself with storing, retrieving and analysing biological data and draws expertise mainly from biology, mathematics and computer science. Bioinformatics is thus underpinned by two main activities: (i) the establishment of computer databases in which raw biological information (e.g. genome and protein sequences) are deposited and stored, and (ii) the development and operation of computer programs that allow users to interrogate, analyse and derive new understanding/information.

While there are many specialist databases available worldwide (some of which we will encounter in subsequent chapters, e.g. enzyme-based databases outlined in Table 11.6), there are three main global, publically accessible databases that serve as repositories for DNA sequence data. Each deposited sequence is given a unique, internationally recognized accession number and these repositories share information deposited on a daily basis, so all contain virtually the same data. The three databanks are GenBank, the European Molecular Biology Laboratory (EMBL) database and the DNA database of Japan. These databases are hosted by the National Center for Biotechnology Information (NCBI) in the USA, the European Bioinformatics Institute (EBI) and the (Japanese) National Institute of Genetics (Table 1.2). Nucleotide sequence information can be used to generate protein sequence information, as does direct protein sequencing efforts. Protein sequence databanks are therefore also maintained by these host bioinformatic institutes.

In addition to maintaining sequence databases themselves, the host organizations generally maintain (and often develop) bioinformatic computer software programs/tools which facilitate data analysis and/or cooperate with additional organizations that maintain databases and/or develop bioinformatic analytical tools used to derive biological knowledge from primary sequence information. As a result numerous bioinformatic resources are available for public use (usually via dedicated websites), hosted by various organizations and capable of providing/generating often overlapping sets of bioinformatic information. Generally, such protein-focused bioinformatic
web-based resources can be grouped in terms of their use as follows.

- **Sequence databases**: house primary DNA/protein sequence information (see, for example, Table 1.2).
- **Protein family databases/resources**: classify proteins into families based on sequence similarities. This can, for example, help elucidate potential functional and structural characteristics of a specific protein, as well as establishing likely evolutionary relationships.
- **Protein structural databases/resources**: organize and store experimentally determined protein three-dimensional structures or generate putative structural models of a protein based on sequence similarities to proteins whose structure has been determined experimentally.
- **Protein function databases/resources**: maintain information about protein function, most commonly relating to metabolic pathways and protein interactions.
- **Proteomics databases**: store proteomic MS and two-dimensional electrophoretic data.

Some bioinformatic resources can be applicable to more than one of the categories above and the number of databases established, as well as which databases will most conveniently answer a particular research question posed, can be somewhat confusing. However, some of the main international bioinformatic organizations maintain ‘gateway resource portals’ on their homepages, which serve as single entry points into multiple specific databases/resources and/or allow a simultaneous search of such multiple databases/resources with a specific search term (such as a protein’s name). For example, the Swiss Institute of Bioinformatics maintains a bioinformatics resource portal called ExPASy (Box 1.1), while the NCBI maintain a portal called Entrez.

### Table 1.2

<table>
<thead>
<tr>
<th>Database name</th>
<th>Host</th>
<th>Web address</th>
</tr>
</thead>
<tbody>
<tr>
<td>The EMBL database</td>
<td>The European Bioinformatics Institute (EBI)</td>
<td><a href="http://www.ebi.ac.uk/embl">www.ebi.ac.uk/embl</a></td>
</tr>
<tr>
<td>The DNA Database of Japan</td>
<td>The (Japanese) National Institute of Genetics (NIG)</td>
<td><a href="http://www.ddbj.nig.ac.jp">www.ddbj.nig.ac.jp</a></td>
</tr>
</tbody>
</table>

### Box 1.1 ExPASy

ExPASy (www.expasy.org) is the Swiss Institute of Bioinformatics resource portal that serves as a single search system/entry point for a whole range of bioinformatic databases and software tools. The databases and tools are categorized under a number of headings, including proteomics, genomics, structural bioinformatics, systems biology and population genetics. Specifically under the proteomics category, over 30 databases and some 250 tools are listed. Examples of both databases and tools, as well as the type of information provided/generated by these, are listed below and these resources generally focus on:

- protein sequences, similarity and identification;
- protein characterization and function;
- protein families;
- protein structure;
- protein–protein interaction;
- post-translational modifications;
- mass spectrometry and two-dimensional electrophoretic data.

The collection of databases and resources are collectively searchable using a key word or an accession number. Thus, for example, a key word search of the site (limited to the proteomics category) using the term ‘cellulase’ reveals almost 16,000 hits, some 14,000 of which are derived from the UniProtKB resource (see below). Each entry in UniProtKB provides information on a specific cellulase, including its source, size...
We will encounter some of the better-known protein-focused bioinformatic databases/tools in some subsequent chapters.

**Examples of proteomic-focused databases and tools which are accessible/searchable via ExPASy**

**Databases**
- UniProtKB: functional information on proteins
- STRING: protein–protein interactions
- Swiss Model repository: protein structure homology models
- PROSITE: protein domains and families
- Enzyme: enzyme nomenclature
- GlycoSuiteDB: glycan database

**Tools**
- APSSP: advanced protein secondary structure prediction
- BLAST: sequence similarity searches
- ClustalW: multiple sequence alignment
- FindMod: protein PTM prediction
- InterProScan: family domain database search
- Mascot: protein identification for MS data
- Peptide cutter: protein cleavage site prediction
- PredictProtein: prediction of protein physico-chemical properties
- RasMol: molecular graphics visualization
- T-Coffee: sequence and structure multiple alignment
- TargetP: subcellular localization prediction
- Swiss model workspace: structure homology modelling

It is important to note that there is overlap between these areas, for example changes in protein expression levels in response to a specific stimulus can provide valuable information about a protein’s likely function, while structural information can also provide insight into protein function.

These areas of proteomic analysis are operationalized by the application of a wide range of analytical (‘wet chemistry’) techniques. Some such techniques, including electrophoretic, chromatographic and MS-based analyses, have already been introduced while others, such as yeast two-hybrid systems and protein microarrays, are described in sections 1.4.2.1 and 1.4.2.2. It is also important to emphasize that such direct analytical approaches can be complemented by bioinformatic-based approaches. Thus, for example, computer programs exist which facilitate the assignment of a putative function to a protein based on amino acid sequence comparisons to proteins of known function. Similarly, bioinformatic tools exist which facilitate prediction of a protein’s likely three-dimensional structure based on amino acid sequence comparisons to those found in proteins of known (experimentally determined) three-dimensional structure. Some such bioinformatics programs will be considered in the next chapter.

### 1.4 Proteomics: goals and applications

While a central goal of proteomics is to separate and identify/record individual proteins constituting a cell or organism’s proteome (or a subset of the proteome), proteomics also incorporates additional goals of protein analysis.

- **Expression proteomics**: allows analysis of the expression of individual proteins in the proteome, and how these change in response to stimuli such as genetic or environmental factors.
- **Functional proteomics**: aims (ultimately) to assign a biological function to each protein in the proteome.
- **Structural proteomics**: aims to gain as much information as possible relating to the three-dimensional structure of proteome proteins.

1.4.1 **Expression proteomics**

Various classical techniques (e.g. immunoassays, see Chapter 10) may be used to detect and quantify the concentration of a specific protein in a biological sample. Quantitative or expression proteomics focuses on the simultaneous detection and quantification of many different proteins in a proteomic sample or, more usually, the simultaneous detection and quantification