Chapter 1
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: Biochemistry and Biotechnology, Second Edition. Gary Walsh.

Companion Website: www.wiley.com/go/walsh/proteinsbiochemistry

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.

Protein	Polypeptide chains	Total no. of amino acids	Molecular mass (Da)	Biological function
Insulin (human)	2	51	5800	Complex, but includes regulation of blood glucose levels
Lysozyme (egg)	1	129	13,900	Enzyme capable of degrading peptidoglycan in bacterial cell walls
Interleukin-2 (human)	1	133	15,400	T-lymphocyte-derived polypeptide that regulates many aspects of immunity
Erythropoietin (human)	1	165	36,000	Hormone which stimulates red blood cell production
Chymotrypsin (bovine)	3	241	21,600	Digestive proteolytic enzyme
Subtilisin (*Bacillus amyloliquefaciens*)	1	274	27,500	Bacterial proteolytic enzyme
Tumour necrosis factor (human TNF-α)	3	471	52,000	Mediator of inflammation and immunity
Haemoglobin (human)	4	574	64,500	Gas transport
Hexokinase (yeast)	2	800	102,000	Enzyme capable of phosphorylating selected monosaccharides
Glutamate dehydrogenase (bovine)	~40	~8300	~1,000,000	Enzyme that interconverts glutamate and α-ketoglutarate and NH_4^+

The majority of 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).

1.2 Genes, genomics and proteomics

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

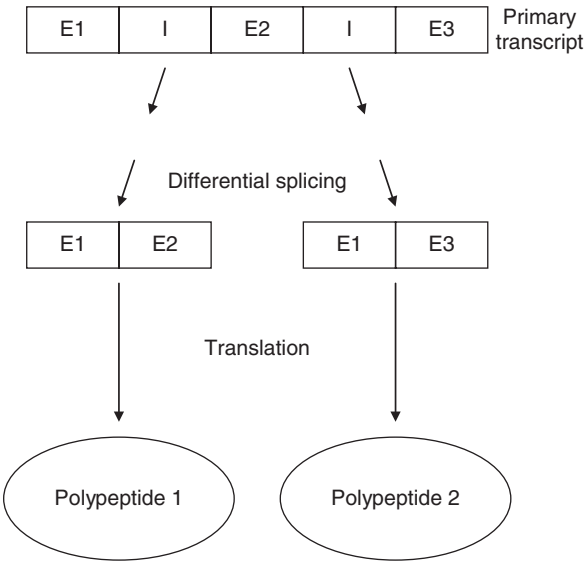

Figure 1.1 Differential splicing of mRNA can yield different polypeptide products. Transcription of a gene sequence yields a 'primary transcript' RNA. This contains coding regions (exons) and non-coding regions (introns). A major feature of the subsequent processing of the primary transcript is 'splicing', the process by which introns are removed, leaving the exons in a contiguous sequence. Although most eukaryotic primary transcripts produce only one mature mRNA (and hence code for a single polypeptide), some can be differentially spliced, yielding two or more mature mRNAs. The latter can therefore code for two or more polypeptides. E, exon; I, intron.

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 sigify 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,

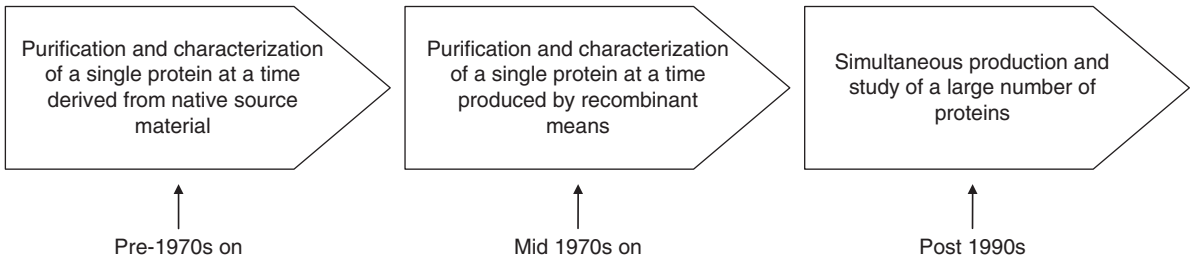

Figure 1.2 Evolution of the various approaches used to study proteins. Refer to text for details.

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 phosphorylated 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.

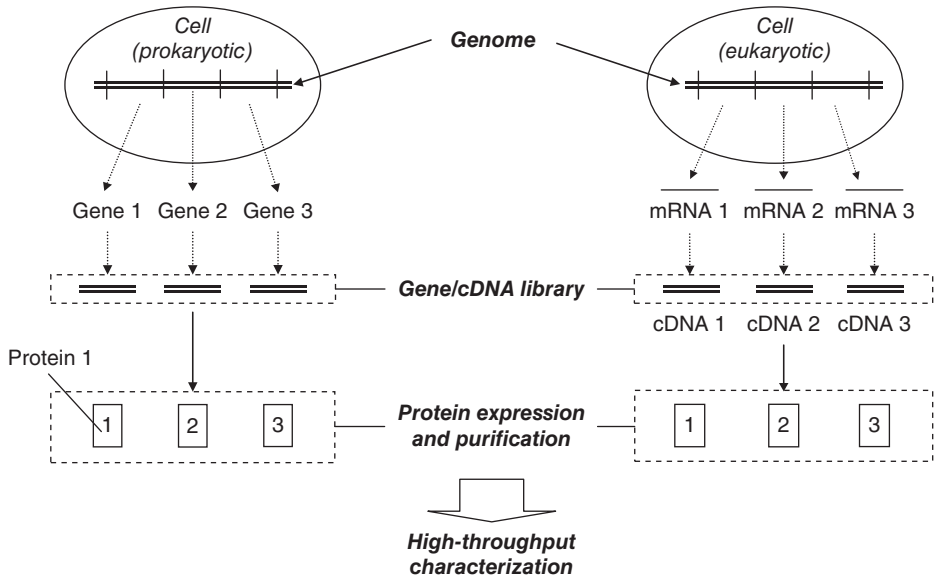

Figure 1.3 Global proteomics approach. While target proteins may be obtained from native (i.e. naturally producing) source material, they are most commonly obtained by recombinant means via the construction of gene/cDNA libraries. In the case of a prokaryotic cell source, a collection of individual genes can be isolated and cloned by standard molecular biology techniques, forming a genomic library (consisting of just three genes in the simplified example portrayed here). Eukaryotic genes generally consist of coding sequences (exons) interrupted by non-coding sequences (introns), while processed mRNA transcripts derived from those genes reflect the coding sequence for the final polypeptide product only. Isolation of total cellular mRNA followed by incubation with a reverse transcriptase enzyme yields complementary double-stranded DNA (cDNA) sequences, directly encoding the polypeptide sequences of the complement of expressed genes, thereby generating a cDNA library. Again by using standard molecular biology techniques the gene/cDNA library products can be expressed, yielding the recombinant protein products. The proteins, in turn, can be purified and characterized via techniques considered in subsequent sections of this chapter, as well as in Chapters 4 and 5.

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

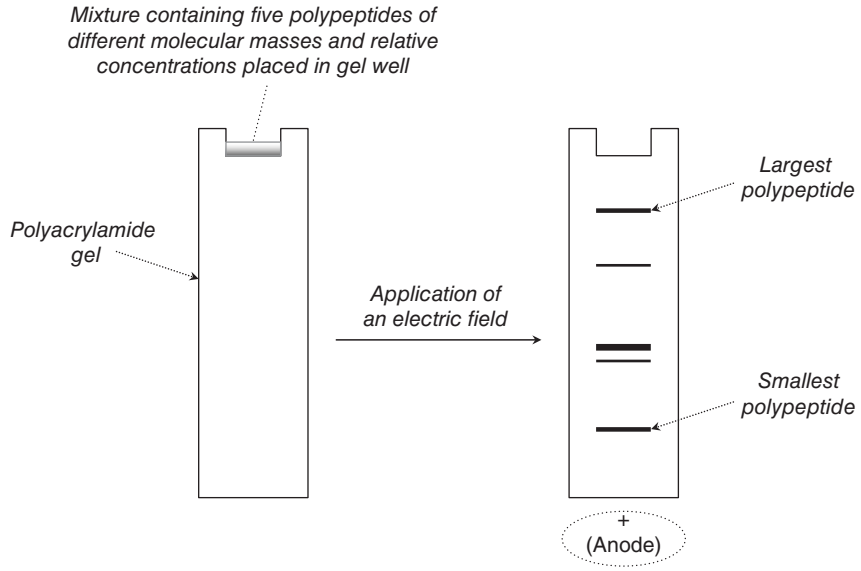

Figure 1.4 Separation of proteins by SDS-PAGE. Protein samples are incubated with SDS (as well as reducing agents, which disrupt disulfide linkages). The electric field is applied across the gel after the protein samples to be analysed are loaded into the gel wells. The rate of protein migration towards the anode depends on protein size. After electrophoresis is complete individual protein bands may be visualized by staining with a protein-binding dye.

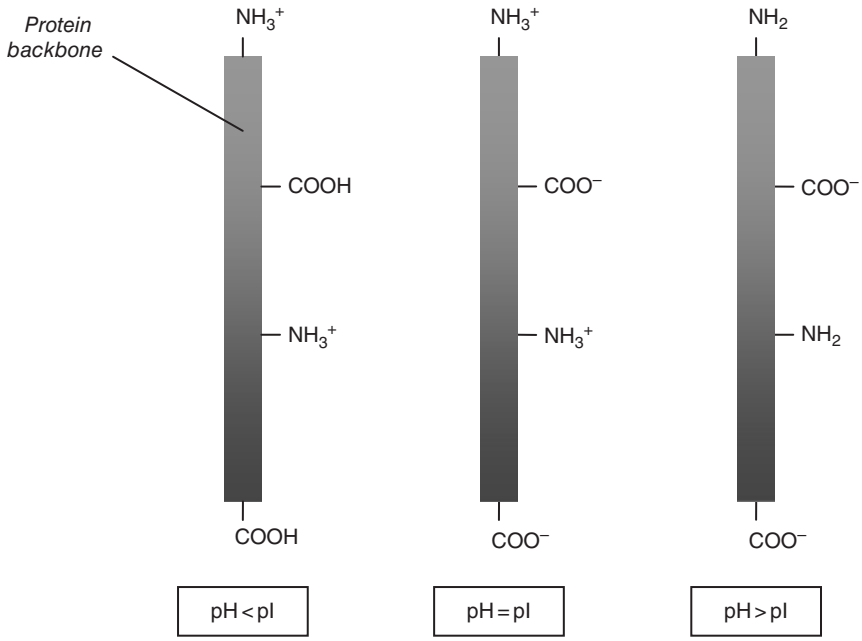

Figure 1.5 Proteins are amphoteric molecules, displaying a positive, negative or zero overall net charge depending on the pH of the solution in which are they dissolved. Contributing to the overall charge of a protein are all the positive and negative charges of its amino acid side chains as well as 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).

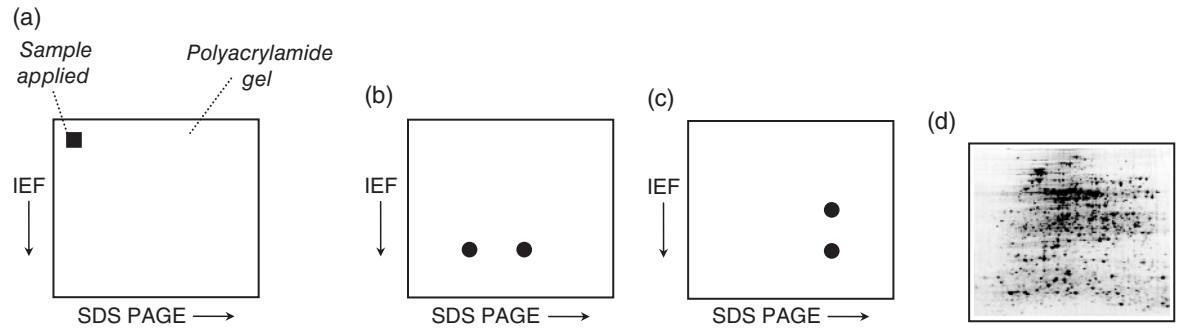

Figure 1.6 Principle of two-dimensional gel electrophoresis. The protein sample is applied to the polyacrylamide gel and first subjected to isoelectric focusing (IEF). After this is complete the protein bands are subjected to SDS-PAGE in the perpendicular direction (a). This combination has greater resolving power than either technique alone. (b) Resolution of two proteins with equal pI values but different molecular masses. (c) Resolution of two proteins of equal molecular mass but differing pI values. (d) Example of a two-dimensional gel in which a microbial proteome has been resolved.

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

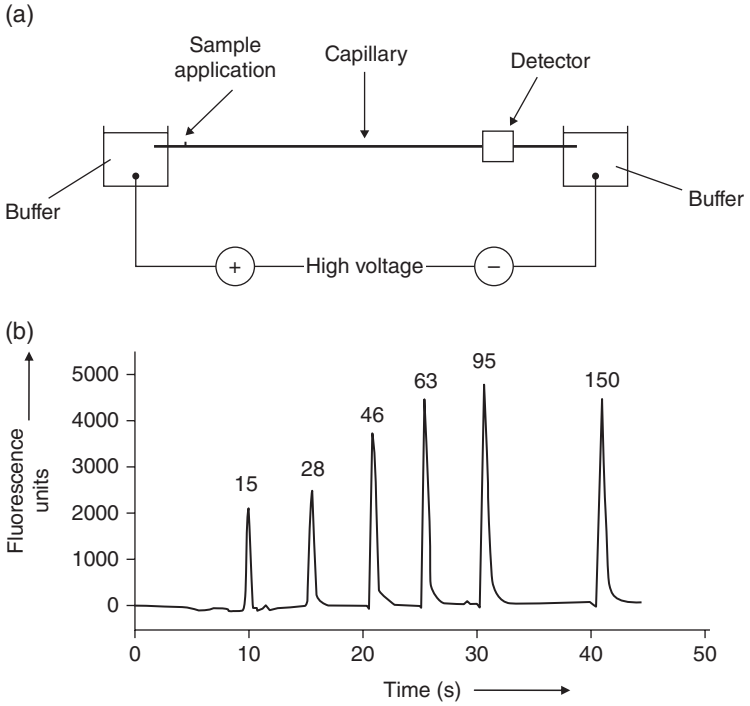

Figure 1.7 (a) Schematic representation of capillary electrophoresis. After sample application, a high voltage is applied and the proteins migrate under the influence of the resultant electric field. Visualization of proteins eluting is achieved using an in-line UV/visible, fluorescence or other appropriate detector. (b) Separation of individual constituents of a protein mixture, with the molecular mass of individual proteins (kDa) indicated above each peak.

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

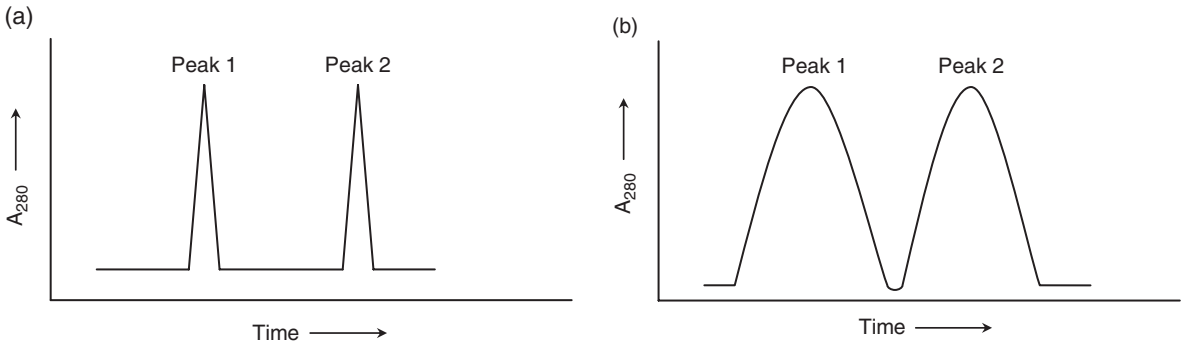

Figure 1.8 HPLC-based chromatographic separation generally gives rise to better-resolved protein peaks (a) than do low pressure-based systems (b).

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 has for many years been a central technique for determining the molecular mass of small molecules. Its routine application to protein work has only been made possible relatively recently, principally by the development of suitable ionization techniques that allow generation of gas-phase ionized proteins. It can determine the mass of proteins up to 500 kDa, with an accuracy of better than 0.01%.

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

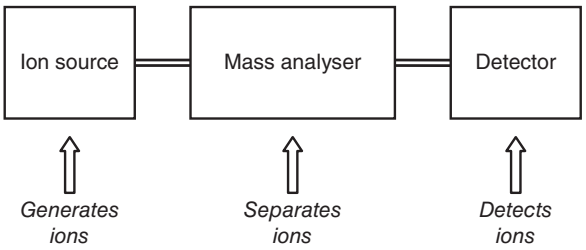

Figure 1.9 Basic principle of mass spectrometry. The system is composed of three essential components: an ion source which generates gas phase-ionized analytes; a mass analyser, which sorts the ions by mass via the application of, for example, electric or magnetic fields; and a detector, which detects and quantifies the ions. Data from the detector thus provides the mass and abundance of each ion present. Refer to text for further detail.

(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

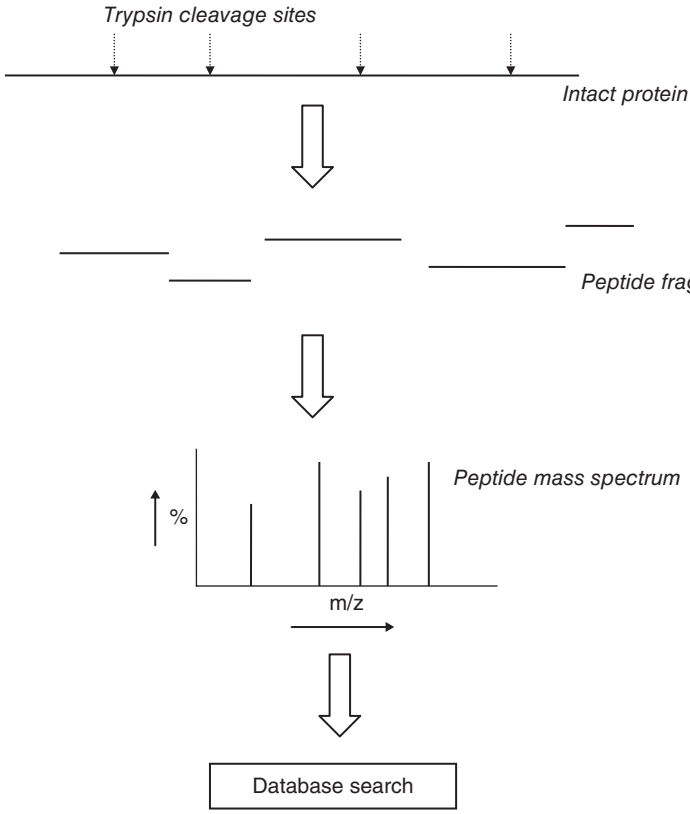

Figure 1.10 Schematic representation of a common approach to protein identification via MS-based peptide mass fingerprinting. Refer to text for detail.

detector reflects its molecular mass, with smaller proteins travelling fastest. A sample size of as little as a few femtomoles (10^{-15} 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

Table 1.2 The three main global sequence databases, their host organizations and web addresses. Refer to text for details.

Database name	Host	Web address
GenBank	The (USA) National Center for Biotechnology Information (NCBI)	www.ncbi.nlm.nih.gov/genbank
The EMBL database	The European Bioinformatics Institute (EBI)	www.ebi.ac.uk/embl
The DNA Database of Japan	The (Japanese) National Institute of Genetics (NIG)	www.ddbj.nig.ac.jp

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 Entrex.

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

and sequence as well as a list of literature references.

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

We will encounter some of the better-known protein-focused bioinformatic databases/tools in some subsequent chapters.

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.

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.

1.4.1 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

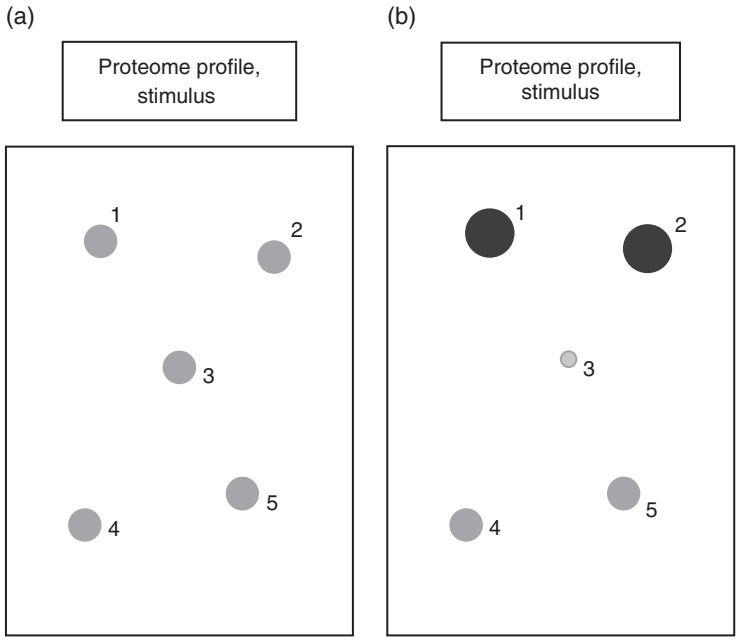

Figure 1.11 Diagrammatic representation of the quantitative proteomic approach as illustrated by two-dimensional gel electrophoretic-based analysis. In this simplified illustrative example, the 'proteome' consists of just five proteins derived from a biological source material under investigation (e.g. a specific cell type exposed to two different stimuli). It is clear that, relative to stimulus (a), stimulus (b) results in an increase in the concentration of proteins 1 and 2, a decrease in the concentration of protein 3, while making no difference to the concentration of proteins 4 and 5. In reality, proteomic samples analysed would generally contain hundreds or thousands of different proteins.

of differences in concentrations of many different proteins in two or more different proteomic samples that have been exposed to different stimuli. Electrophoretic, chromatographic and MS-based techniques may all be applied to such analyses (Figure 1.11).

Detecting and identifying changes in the expression levels of specific proteins/groups of proteins in response to a specific stimulus can of course provide clues as to protein function. At a basic research level therefore this approach can for example be used to identify groups of proteins likely involved in specific cellular processes. From an applied perspective, studying the changes in proteome expression profiles of clinical samples can provide potentially useful medical information, and such 'clinical proteomics' now forms a valuable part of medical-based research and development. For example, the approach can be used to potentially identify biomarkers for specific diseases/conditions and identify potential new drugs.

A biomarker is a specific measurable characteristic of a biological system whose quantity correlates in some way with a biological process. In the context of clinical science, most biomarkers are biomolecules whose levels in biological samples (e.g. blood, urine or tissue samples) are correlated with some disease or condition. Biomarker detection and measurement over time can therefore reflect the occurrence of a disease/condition, how it is progressing with time and perhaps how it is responding to therapy. Many established biomarkers are proteins (e.g. the gonadotrophic hormone hCG serves as a biomarker for pregnancy; Chapter 10). The high throughput and rapid nature of proteomics provides a very powerful tool for the identification of potential new disease biomarkers. Once identified and validated, standard classical diagnostic assays (e.g. immunoassays) for the biomarker can be developed and used in clinical chemistry laboratories (see Chapter 10). Moreover, comparative proteomic analysis of, for example, a cancer cell versus an untransformed cell of the same type could lead to the identification of cellular proteins fuelling the cancer phenotype. Such proteins could therefore represent targets for future anticancer drugs.

1.4.2 Functional proteomics

Genome sequencing studies have generated enormous amounts of protein sequence information. However, the function of the majority of such proteins remains to be elucidated, and assigning such functionality represents a major challenge. For example, the function of the majority of protein sequences identified by the Human Genome Project remains unknown and, even in the case of very well-studied organisms (e.g. *Escherichia coli*) function remains unassigned for a significant minority of proteins. Various genomic/bioinformatic/proteomic approaches may be pursued in an effort to assign function.

At a purely bioinformatic level, and as already mentioned, computer programs exist which help assign a putative function to a protein based on amino acid sequence comparisons to proteins of known function (see Chapter 2). At a genomic level, for example, knockout studies can be employed. Such studies entail the disruption of a specific gene with subsequent analysis of the effect on the organism.

At a proteomic level, analysing changes in the expression levels of specific proteins/groups of proteins in response to a specific stimulus can, as mentioned previously, provide clues as to protein function. However, the core laboratory-based approaches adopted in functional proteomics attempt to identify protein–protein interactions (the 'interactome'), usually by using a protein of interest as a 'bait molecule' to fish out proteins capable of interacting with it from a proteome of interest ('prey molecules'). The resultant 'prey' proteins recovered are likely to be functionally related to the bait protein. Careful experimental design and execution is required to ensure that any proteins recovered are interacting with the bait protein in a biospecific manner. If non-specific binding occurs, the assumption that the proteins are functionally related will of course be inaccurate.

Various experimental approaches may be pursued in order to identify protein interactions. One approach involves incubating the bait protein with the proteome of interest to allow the formation of interactions with prey protein partners. Antibodies raised against the bait protein are then added, which precipitate the bait–prey complex out of solution; the precipitate can then be fractionated by SDS-PAGE, with subsequent analysis of the protein components present via MS. An alternative approach entails immobilizing the bait protein on a chromatographic bead, followed by incubation with the proteome of interest (Figure 1.12).

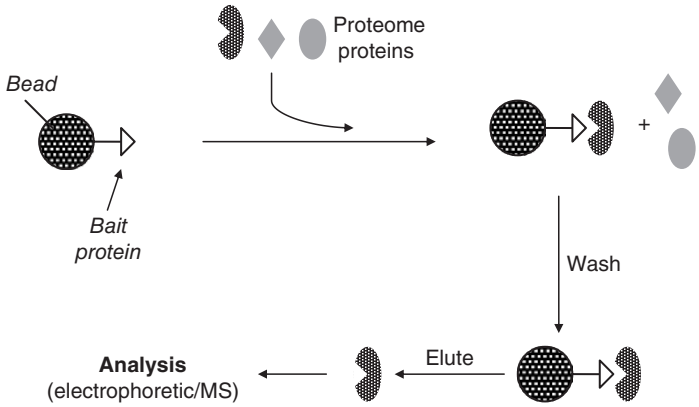

Figure 1.12 Approach to interactome studies using a bait protein immobilized on a chromatographic bead. Beads can be incubated with the target proteome (which in this simplified example contains only three proteins). Only proteins interacting with the bait molecule in a biospecific manner will be retained on the column. After washing away the additional (non-binding) proteins, the captured (prey) protein(s) can be eluted from the bead and analysed in order to establish prey protein identity.

Among the most prominent interactome techniques are the yeast two-hybrid (Y2H) system and protein microarrays. Before we consider these approaches it is important to recognize that the goals of functional proteomics are broader than simply assigning function to an individual protein. These goals also incorporate identification of the subcellular location in which the protein functions, determination of the composition and function of macromolecular complexes, and promotion of a broader understanding at a molecular level of cellular mechanisms/processes in which proteins participate and how some processes are interlinked.

1.4.2.1 Yeast two-hybrid system

The Y2H system is a molecular biology technique developed to investigate protein–protein interaction. The technique is based on the fact that gene expression requires the presence of a transcription activator (a protein that binds DNA, thereby stimulating transcription of a nearby gene, usually by facilitating/enhancing RNA polymerase binding). Transcription activators typically consist of two domains: a DNA-binding domain (DBD), which docks the protein at a specific DNA sequence, and an activator domain (AD), which actually facilitates transcription of the target gene(s) downstream of the DBD domain.

Using this system, as overviewed in Figure 1.13, the bait protein of interest is expressed as a fusion protein which incorporates the transcription factor's DBD domain. A possible interacting protein (prey protein) is expressed as a fusion product incorporating the transcription activator's AD domain. If bait–prey interaction does indeed occur, the transcription factor's DBD and AD domains are effectively reunited in the resultant protein complex (Figure 1.13). This in turn triggers expression of the downstream reporter gene.

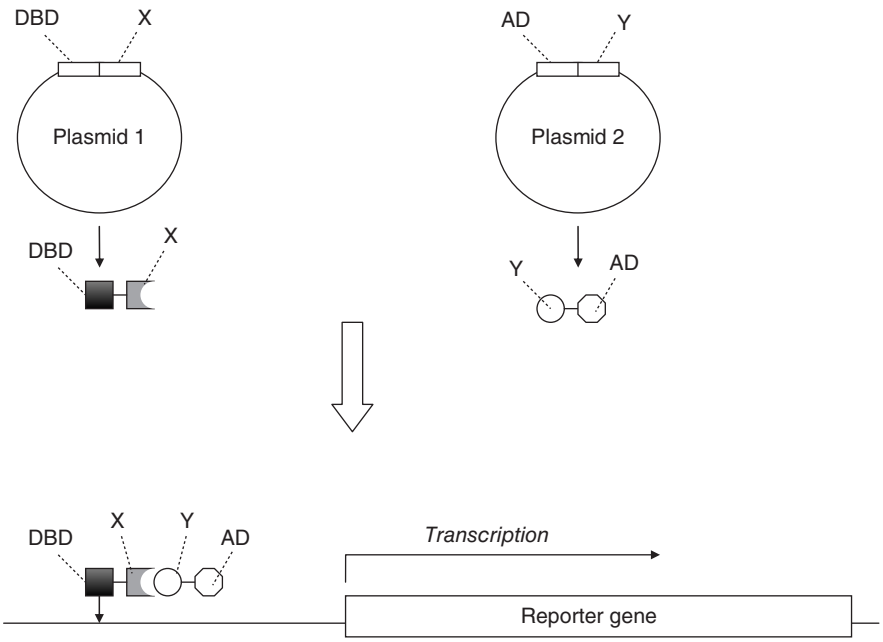

Figure 1.13 The basis on which the yeast two-hybrid (Y2H) system detects protein–protein interactions. Plasmid 1 (in yeast 1) contains a fusion construct housing a nucleotide sequence coding for a transcription activator DNA-binding domain (DBD) fused to a nucleotide sequence coding for the bait protein (X). Plasmid 2 (in yeast 2) contains a fusion construct housing a nucleotide sequence coding for a transcription activator domain (AD) fused to a nucleotide sequence coding for a possible prey protein (Y). The yeast are allowed to mate (or are transformed), bringing both plasmids into the one cell. If the bait and prey proteins (X–Y) actually do interact, they bring the transcription factor DBD and AD domains together in the one complex, which in turn specifically activates the downstream reporter gene. Refer to text for exact detail.

Reporter gene expression leads to some observable change in cellular phenotype, facilitating straightforward detection. Among the most common reporter genes is the *lacZ* gene, coding for β-galactosidase, which turns expressing yeast colonies blue by degrading the chromomeric substrate X-gal. Additional reporter genes include the *HIS3* gene (encodes a dehydratase enzyme essential in the biosynthesis of histidine and which therefore allows expressing cells to grow on a media devoid of histidine) and the *luc* gene (encodes a luciferase enzyme which can oxidize luciferin to produce green light). Sometimes a combination of reporter genes is used.

The Y2H system is amenable to high-throughput screening, making it particularly useful from a proteomic perspective. For example, a large 'prey' cDNA library (encoding the sequence of all proteins in the proteome of interest) can be generated and subsequently screened against any specific bait protein of interest.

1.4.2.2 Protein microarrays

Protein–protein interaction can also be investigated in high-throughput mode (i.e. simultaneous analysis of many different proteins derived from a proteome/proteome subset of interest) using a protein microarray (protein chip) approach. This approach entails:

- initial immobilization of the collection of proteins with which you wish to probe samples of interest for interacting proteins, thereby generating the actual protein array;
- exposure of the protein array to the sample you wish to analyse;
- subsequent analysis of the array to detect and identify any binding partners/interactions.

The collection of proteins immobilized will be dictated by the research question posed, but one common broad approach would be to source these proteins from a library of an organism's genome via recombinant production (Figure 1.3). By using this approach it is also possible to incorporate an affinity tag at one or other end of all the proteins produced, which can subsequently facilitate both affinity-based protein purification and affinity-based immobilization of the purified proteins (Figure 1.14a). Affinity tags will be discussed in Chapter 4 but, briefly, one such common tag is a short sequence of histidine residues (usually six, i.e. His-6) attached at the end of the protein. The His tag binds to divalent metals such as nickel (Ni^{2+}), which can therefore act as a capture ligand. In the context of protein purification, a chromatographic column containing Ni^{2+} capture ligand can selectively purify the tagged protein, while Ni^{2+} immobilized on an appropriate surface can act as an affinity anchor for individual proteins of the protein array.

Once the gene/cDNA library coding for the tagged proteins that will constitute the array is constructed (each protein-encoding gene/cDNA being present in a single engineered recombinant cell), each recombinant protein can be expressed, purified and immobilized onto a solid surface, often made from glass or nitrocellulose (Figure 1.14a), thus producing the protein array. The different protein samples are typically applied using robotic microspotting equipment (arrayers). Individual spots will contain hundreds to thousands of individual (identical) copies of one particular protein. The pitch (i.e. distance between any two spots) can be a little as 300 μm, facilitating the printing of up to 20,000 individual protein spots on a single glass microscope slide (Figure 1.14b).

The use of affinity tags provides a convenient means of protein immobilization on the array support surface. Moreover, the tag itself acts as a spacer arm, keeping the protein at a (short) distance from the support surface and ensuring that all the protein molecules are oriented in an identical direction. This usually maximizes the ability of interacting proteins to, in turn, bind to the array proteins during interaction analysis. However, an alternative immobilization approach involves the direct covalent linkage of the proteins to the solid support. This can be conveniently undertaken by using supports containing chemically reactive groups (e.g. aldehydes or activated esters) which are capable of forming direct covalent linkages with functional groups commonly found on proteins (e.g. amino, carboxyl or thiol groups). The covalent nature of such links prevents protein leakage

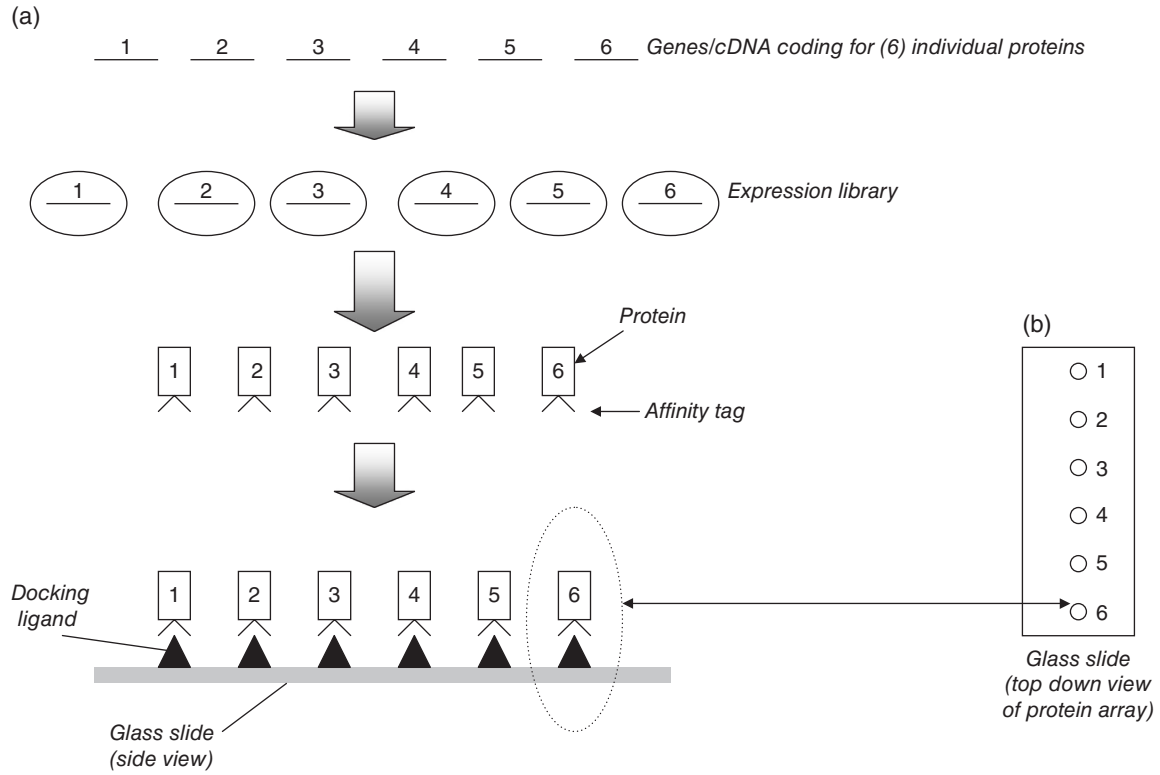

Figure 1.14 Generation of a protein array. (a) The genes/cDNAs coding for the proteome of interest (only six proteins in this simplified example) are expressed in a recombinant microbial library (i.e. individual genes/cDNAs are inserted into individual microbial cells which, when grown individually, will produce the recombinant protein product). This molecular biology approach also allows the attachment of affinity tags at the end of each protein, facilitating affinity-based protein purification subsequent to protein expression. The tags also allow the docking (attachment) of the proteins to a solid support (e.g. a glass slide) if a docking ligand for the affinity tag is first immobilized on that support (b). Refer to text for further details.

(desorption) from the array and the approach can be undertaken to immobilize proteins devoid of affinity tags (e.g. non-recombinant proteins). However, proteins are immobilized in direct contact with the support surface and at random orientations, which can potentially negatively affect protein–protein interaction when the array is in use.

Detection of the interaction between proteins from the sample being analysed with array proteins may be achieved in different ways (Figure 1.15). In some instances an array may be designed to detect a specific protein type such as an enzyme or an antibody. Under such circumstances, interaction detection may rely on some inherent characteristic of the molecule captured by the array. If the array were designed to capture a specific enzyme (Figure 1.15a), the enzyme captured from samples analysed could be detected using a chromogenic substrate (a molecule the enzyme is able to catalytically transform into a coloured product). Likewise, if the array were designed to detect specific antibody molecules in for example human blood, a second antibody which specifically binds to human antibodies and to which a fluorescent tag has been attached could be used (Figure 1.15 b).

However, a more widespread approach is to first pretreat the samples to be analysed such that a tag (usually a fluorescent molecule) is attached to all analyte molecules in the sample. After such samples are incubated with the array (and the array is subsequently rinsed in order to remove any unbound tag present), bound molecules can be detected via a

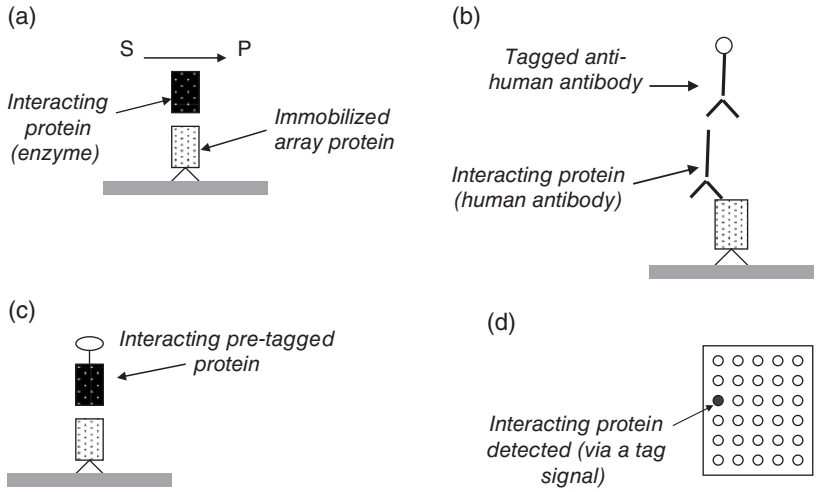

Figure 1.15 Some approaches that facilitate the detection of protein interactions. (a) The use of a substrate molecule which generates a coloured/fluorescent product if the array is designed to interact with a specific enzyme. (b) The use of a labelled (tagged) antibody capable of binding to human antibodies if the array is designed to interact with human antibodies. (c) Interaction detection via the use of a sample in which all analytes are pre-tagged. (d) A (simplified) array image in which molecules from a sample analysed have bound to one specific array protein. Refer to text for further detail.

fluorescent signal (Figure 1.15c). Signal can be visualized using a microarray laser scanner. This generates an image of the microarray spots in which those participating in interactions generate a fluorescent or other signal (Figure 1.15d).

Once the occurrence of protein interactions has been established, the next step in protein array experiments is normally aimed at identifying the interacting proteins. This is most often achieved by subjecting the proteins interacting with the array to MS analysis in order to establish identity/sequence. Protein microarrays may also be used to identify protein–non-protein interactions (e.g. protein–DNA or protein–carbohydrate interactions) by pursuing the same approach as in the case of protein–protein analysis.

Array technology may be used for applied as well as academic purposes. For example, antibody-based arrays have been developed to simultaneously detect various cytokines or other molecules of diagnostic/prognostic value present in clinical samples.

While the high-throughput miniaturized nature of protein array technology renders it an attractive analytical technique, the approach is not without its limitations. For example, the occurrence of non-specific binding reactions lead to false-positive results. Moreover, many if not most proteins are relatively labile molecules and array construction/storage prior to use can trigger protein modification and/or denaturation. This can prevent normal interactions (generating false-negative results) or can lead to artefactual interactions, leading to false-positive results.

Another limitation of array technology is the difficulty in obtaining sufficiently pure protein to construct large arrays. While the generation of libraries expressing perhaps thousands of different proteins (Figure 1.14) can be relatively straightforward, subsequent purification of each recombinant protein, even when using tag-based affinity purification systems, is usually more labour-intensive and complex (see Chapter 4). For example, affinity-based purification columns must often be followed by a second chromatographic step in order to fully purify the target protein. One approach which could potentially overcome this limitation is the development of so-called self-assembling protein microarrays. In this approach individual protein-encoding genes/cDNA (which are also engineered to contain an affinity tag at one end) are first immobilized on the

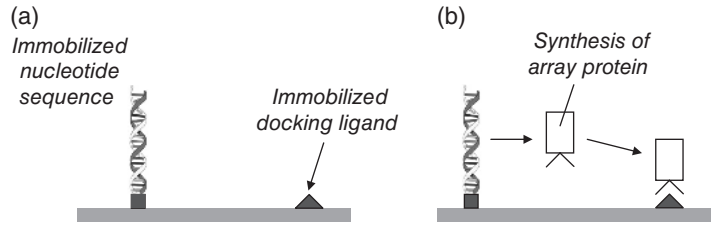

Figure 1.16 Self-assembling protein arrays. The nucleotide sequences (e.g. cDNA) coding for individual tagged proteins are immobilized on the array solid support surface, as are tag docking ligands. Only a single illustrative sample is shown here (a). A commercial cell-free expression system is then incubated on the array surface. Cell-free expression systems contain a cocktail of the components necessary to transcribe and translate a coding sequence (RNA polymerase, ribosomes, tRNA and ribonucleotides), thus allowing protein synthesis to occur *in vitro*. The result therefore is synthesis of the tagged array proteins, which then spontaneously immobilize on the solid support surface via binding to the docking ligands (b).

microarray support surface (Figure 1.16). These are then expressed *in situ* (i.e. directly on the support surface) using a cell-free expression system. The tagged proteins, once synthesized, then bind to a tag docking ligand, which has also been pre-immobilized onto the microarray support surface. This approach would bypass the need to purify individual array proteins. Furthermore, by undertaking the cell-free protein expression step immediately prior to array application, storage stability-related concerns may no longer be an issue. However, a potential drawback is that some proteins may not fold properly in this *in vitro* environment nor will this approach support protein PTM.

1.4.3 Structural proteomics

After their synthesis, proteins fold into a specific three-dimensional shape (specific conformation) and protein function normally depends on it retaining that conformation. The ultimate goal of structural proteomics is to provide a complete three-dimensional description of each protein constituting a proteome.

A detailed description of protein architecture and associated methods of determining three-dimensional structure is given in Chapter 2. However, briefly, X-ray crystallography is the technique most commonly used to resolve the three-dimensional structure of proteins. Nuclear magnetic resonance (NMR) can also be used to determine the three-dimensional structure of some, mainly smaller, proteins. Traditionally, attempts to study three-dimensional structure was undertaken on a protein-by-protein basis. The protein under study was first purified, either from a naturally producing biological source material, or from a recombinant system producing the protein. Structural analysis then ensued. The structural proteomic approach essentially pursues the same approach, but attempts to study a number of target proteome proteins at the same time. Thus, a structural proteomic starting point is often characterized by generation of a recombinant expression library expressing the target group of proteins. The recombinant proteins invariably include an affinity tag, which facilitates subsequent protein purification (see also Chapter 3). Once purified (which usually incorporates tag removal using a proteolytic enzyme), the proteins are subject to structural analysis.

The molecular biology element described above potentially allows the simultaneous/near simultaneous production and follow-on affinity purification of many proteins (i.e. has potential high-throughput characteristics). However, complications can arise including:

- the occurrence of low-level recombinant protein expression (making it difficult to source sufficient sample protein to conveniently work with);
- incomplete/no protein folding (i.e. the recombinant protein accumulates in a non-functional unfolded form, useless to structural studies);

- the extent of purity achieved by a single-step tag affinity-based purification system (highly purified protein is required).

In such instances considerable variation in experimental protocols may be required in order to optimize protein production and purification. For some proteins, an appropriate level of optimization may simply not be achieved.

Follow-on structural elucidation experiments often prove even less amenable to high-throughput automated analysis. For both X-ray crystallography and NMR spectroscopy, considerable protein-specific optimization of sample preparation is required. In the case of X ray crystallography, proteins must first be successfully crystallized, a process that again requires considerable protein-specific optimization and which ultimately may not prove successful. Moreover, the actual process of gathering and interpreting structural detail is often quite time-consuming. Overall, therefore, structural proteomics has some way to go before it becomes a genuinely automated high-throughput process.

Further reading

Altelaar, A.F.M. and Heck, A.J.R. (2012) Trends in ultrasensitive proteomics. *Current Opinion in Chemical Biology* **16**, 206–213.

Altelaar, A.F.M., Munoz, J. and Heck, A.J.R. (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics* **14**, 35–48.

Banci, L., Bertini, I., Luchinat, C. and Mori, M. (2010) NMR in structural proteomics and beyond. *Progress in Nuclear Magnetic Resonance Spectroscopy* **56**, 247–266.

Bencharit, S. and Border, M.B. (2012) Where are we in the world of proteomics and bioinformatics? *Expert Review of Proteomics* **9**, 489–491.

Berkelman, T. (2008) Quantitation of protein in samples prepared for 2-D electrophoresis. *Methods in Molecular Biology* **424**, 43–49.

Berrade, L., Garcia, A.E. and Camarero, J.A. (2011) Protein microarrays: novel developments and applications. *Pharmaceutical Research* **28**, 1480–1499.

Burgess, R.R. (2009) Use of bioinformatics in planning a protein purification. In: Burgess, R.R. and Deutscher, M.P. (eds), *Guide to Protein Purification*, 2nd edn, pp. 21–28. Academic Press, San Diego, CA.

Caufield, J.H., Sakhawalkar, N. and Uetz, P. (2012) A comparison and optimization of yeast two-hybrid systems. *Methods (San Diego, CA)* **58**, 317–324.

Chen, C., Huang, H. and Wu, C.H. (2011) Protein bioinformatics databases and resources. *Methods in Molecular Biology* **694**, 3–24.

Chung, L. and Baxter, R.C. (2012) Breast cancer biomarkers: proteomic discovery and translation to clinically relevant assays. *Expert Review of Proteomics* **9**, 599–614.

Cordero, P. and Ashley, E.A. (2012) Whole-genome sequencing in personalized therapeutics. *Clinical Pharmacology and Therapeutics* **91**, 1001–1009.

Espindola, F.S., Calabria, L.K., Alves de Rezende, A.A., Pereira, B.B., Santana, F.A., Rodrigues Amaral, I.M., Lobato, J., Franca, J.L., Mario, J.L., Figueiredo, L.B., dos Santos-Lopes, L.P., de Gouveia, N.M., Nascimento, R., Teixeira, R.R., dos Reis, T.A. and de Araujo, T.G. (2010) Bioinformatic resources applied on the omic sciences as genomic, transcriptomic, proteomic, interatomic and metabolomic. *Bioscience Journal* **26**, 463–477.

Friedman, D.B., Hoving, S. and Westermeier, R. (2009) Isoelectric focusing and two-dimensional gel electrophoresis. In: Burgess, R.R. and Deutscher, M.P. (eds), *Guide to Protein Purification*, 2nd edn, pp. 515–540. Academic Press, San Diego, CA.

Geiger, M., Hogerton, A.L. and Bowser, M.T. (2012) Capillary electrophoresis. *Analytical Chemistry* **84**, 577–596.

Gonzaga-Jauregui, C., Lupski, J.R. and Gibbs, R.A. (2012) Human genome sequencing in health and disease. *Annual Review of Medicine* **63**, 35–61.

Gonzalez-Gonzalez, M., Jara-Acevedo, R., Matarraz, S., Jara-Acevedo, M., Paradinas, S., Sayaguees, J.M., Orfao, A. and Fuentes, M. (2012) Nanotechniques in proteomics: protein microarrays and novel detection platforms. *European Journal of Pharmaceutical Sciences* **45**, 499–506.

Hu, S., Xie, Z., Qian, J., Blackshaw, S. and Zhu, H. (2011) Functional protein microarray technology. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* **3**, 255–268.

Lin, J.C.-H. (2010) Protein microarrays for cancer diagnostics and therapy. *Medical Principles and Practice* **19**, 247–254.

Loman, N.J., Constantinidou, C., Chan, J.Z.M., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R. and Pallen, M.J. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology* **10**, 599–606.

Manjasetty, B.A., Turnbull, A.P. and Panjikar, S. (2010) The impact of structural proteomics on biotechnology. *Biotechnology and Genetic Engineering Reviews* **26**, 353–370.

Maupin-Furlow, J.A., Humbard, M.A. and Kirkland, P.A. (2012) Extreme challenges and advances in archaeal proteomics. *Current Opinion in Microbiology* **15**, 351–356.

Pavlopoulou, A. and Michalopoulos, I. (2011) State-of-the-art bioinformatics protein structure prediction tools. *International Journal of Molecular Medicine* **28**, 295–310.

Popov, I., Nenov, A., Petrov, P. and Vassilev, D. (2009) Bioinformatics in proteomics: a review on methods and algorithms. *Biotechnology and Biotechnological Equipment* **23**, 1115–1120.

Rajagopala, S.V., Sikorski, P., Caufield, J.H., Tovchigrechko, A. and Uetz, P. (2012) Studying protein complexes by the yeast two-hybrid system. *Methods (San Diego, CA)* **58**, 392–399.

Righetti, P.G., Sebastiano, R. and Citterio, A. (2013) Capillary electrophoresis and isoelectric focusing in peptide and protein analysis. *Proteomics* **13**, 325–340.

Roepstorff, P. (2012) Mass spectrometry based proteomics: background, status and future needs. *Protein and Cell* **3**, 641–647.

Roncada, P., Piras, C., Soggiu, A., Turk, R., Urbani, A. and Bonizzi, L. (2012) Farm animal milk proteomics. *Journal of Proteomics* **75**, 4259–4274.

Sa-Correia, I. and Teixeira, M.C. (2010) 2D electrophoresis-based expression proteomics: a microbiologist's perspective. *Expert Review of Proteomics* **7**, 943–953.

Savino, R., Paduano, S., Preiano, M. and Terracciano, R. (2012) The proteomics big challenge for biomarkers and new drug-targets discovery. *International Journal of Molecular Sciences* **13**, 13926–13948.

Serpa, J.J., Parker, C.E., Petrotchenko, E.V., Han, J., Pan, J. and Borchers, C.H. (2012) Mass spectrometry-based structural proteomics. *European Journal of Mass Spectrometry* **18**, 251–267.

Shin, J., Lee, W. and Lee, W. (2008) Structural proteomics by NMR spectroscopy. *Expert Review of Proteomics* **5**, 589–601.

Stoevesandt, O., Taussig, M.J. and He, M. (2009) Protein microarrays: high-throughput tools for proteomics. *Expert Review of Proteomics* **6**, 145–157.

Urban, J., Vanek, J. and Stys, D. (2012) Current state of HPLC-MS data processing and analysis in proteomics and metabolomics. *Current Proteomics* **9**, 80–93.

van de Meent, M.H.M. and de Jong, G.J. (2011) Novel liquid-chromatography columns for proteomics research. *Trends in Analytical Chemistry* **30**, 1809–1818.

Xie, F., Smith, R.D. and Shen, Y. (2012) Advanced proteomic liquid chromatography. *Journal of Chromatography A* **1261**, 78–90.

Young, C.L., Britton, Z.T. and Robinson, A.S. (2012) Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. *Biotechnology Journal* **7**, 620–634.