Research Review

Characterisation of Biopharmaceuticals

Derek Bradley[1]; Mary Burke[1]; James J. Roche[2]

[1] *BioClin Research Laboratories, IDA Technology Park, Garrycastle, Athlone, Co. Westmeath, Ireland*
[2] *Biosciences Research Institute, Athlone Institute of Technology, Dublin Road, Athlone, Co Westmeath, Ireland*

Corresponding author: derek.bradley@bioclinlabs.com

Introduction

Biopharmaceuticals are drugs which are primarily (glyco)protein in nature, that are produced in living cells using recombinant DNA technology – the combination of genetic material from multiple sources, creating sequences that would not otherwise be found in the genome. They have become an essential component of modern pharmacotherapy, and are the only effective treatment option available for many severe, often life-threatening, diseases. Biopharmaceutical products have seen unprecedented sales growth in the last decade, particularly when compared with the conventional drug market. Most of the world's top-selling drugs in 2015 were biopharmaceuticals, accounting for over 70% of the sales revenue for the top ten drug products in that year – see Table 1. The global biopharmaceuticals market currently enjoys a compound annual growth rate of approximately 9%, and is expected to reach an estimated value of over US$250 billion by 2017.[1]

Table 1: Global sales revenue of the top 10 pharmaceutical products in 2015

Rank	Product	Active Ingredient	Company	Sales - US$ million
1	Humira	Adalimumab	AbbVie	14,012
2	Harvoni	Ledipasvir and Sofosbuvir	Gilead Sciences	13,864
3	Enbrel	Etanercept	Amgen / Pfizer	8,697
4	Remicade	Infliximab	Johnson & Johnson / Merck	8,355
5	MabThera/Rituxan	Rituximab	Roche	7,115
6	Lantus	Insulin Glargine	Sanofi	7,029
7	Avastin	Bevacizumab	Roche	6,751
8	Herceptin	Trastuzumab	Roche	6,603
9	Revlimid	Lenalidomide	Celgene Corporation	5,801
10	Sovaldi	Sofosbuvir	Gilead Sciences	5,276
		Total revenue for top 10 drugs of 2015		***83,503***

	US$ million	Percentage of total
Biopharmaceuticals	58,562	70.1
Small molecule drugs	24,941	29.9

Data source: http://www.pharmacompass.com/pharma-news/top-drugs-by-sales-revenue-in-2015-who-sold-the-biggest-blockbuster-drugs. Accessed 18/05/2016

Biosimilars represent a distinct class of biopharmaceuticals that are essentially 'copy-versions' of innovator biologics that emerge upon patent expiry of the innovator drug. By the end of 2015, important innovator biologics with a combined global annual revenue in excess of US$50 billion, including Herceptin®, Rituxan® and Remicade® had lost patent protection, and many more key patents are set to expire in the

period up to 2020 (frequently termed the 'patent cliff'). This creates an attractive opportunity for big pharmaceutical companies to develop biosimilars, enabling them to diversify their thinning pipelines.

Due to the highly complex nature of biologics, and their dependence on biological processes for production, biosimilars cannot be considered to be identical to innovator drugs. Therefore, legislation put forward for licensure of generics (such as the Drug Price Competition and Patent Term Restoration Act of 1984 in the United States of America (USA), often referred to as the Hatch-Waxman Act) is inadequate for biosimilars. Unlike biologics, generic versions of conventional drugs contain active substances whose safety and efficacy profiles are well-established. The FDA definition of a generic is that it should be comparable to the reference product in dosage, strength, route of administration, quality, performance characteristics, and intended use.[2] Generics' developers only need to prove average bioequivalence in order to obtain approval. However, considerably more data is required for biosimilars, as slight differences between biosimilar and innovator may have significant consequences, such as eliciting a potentially dangerous immune response, when administered to patients. Therefore, specific legislation for biosimilar approvals was required.

In the European Union (EU), the "Guideline on similar biological medicinal products"[3] was published in 2005, which introduced the concept of a biosimilar medicine, and describes the approach for demonstrating biosimilarity of a proposed biosimilar. The first biosimilar was licensed in EU in 2006 (Sandoz Inc.'s Omnitrope®; a recombinant somatropin), and more than 20 biosimilars have since gained authorisation. As such, considerable experience has been gained on biosimilars in the EU, not only from a conceptual perspective, but also from available data.

While biosimilar licensing legislation was pioneered in the EU, the US has significantly lagged behind in developing its own legislation. Indeed, some products that have been registered as biosimilars in the EU, such as Omnitrope®, have gained approval in the US via the full 'Biological Licensing Application' pathway before US biosimilar regulations were even created. However, in recent years, there has been a surge of regulatory activity in the US laying out the route for approval of biosimilars. A major step forward in this process was the issuing by the FDA of three draft guidance documents in 2012 which cover quality considerations, scientific considerations, and FAQ's regarding the implementation of the legislation relating to biosimilars.[4-6] In an exciting development in March 2015, the FDA announced a first biosimilar approval in the US – Sandoz's Zarxio®, a biosimilar to Amgen's Neupogen® (filgrastim), used to combat chemotherapy-induced neutropaenia in cancer patients. More recently in April 2016, the FDA approved Hospira's Inflectra® (infliximab-dyyb), a biosimilar to Janssen Biotech, Inc.'s Remicade® (infliximab), which is used to treat a range of autoimmune disorders. These recent approvals indicate that the licensure pathway for biosimilars in the US is well-established at this point.

Biosimilar development can take advantage of these abbreviated licensure pathways based upon characterisation programs that demonstrate sufficient similarity between the biosimilar and innovator product. This can offer enormous time and cost savings when compared to development of a new molecular entity (NME) as there is less dependence on lengthy and costly pre-clinical and clinical studies. According to the Tuft's Centre for the Study of Drug Development (an highly-reputable independent, non-profit organisation dedicated to researching drug development), the cost of taking an NME from concept to market can exceed US$2.6 billion.[7] By contrast, a biosimilar development programme typically costs in the region of US$100 million – 250 million.

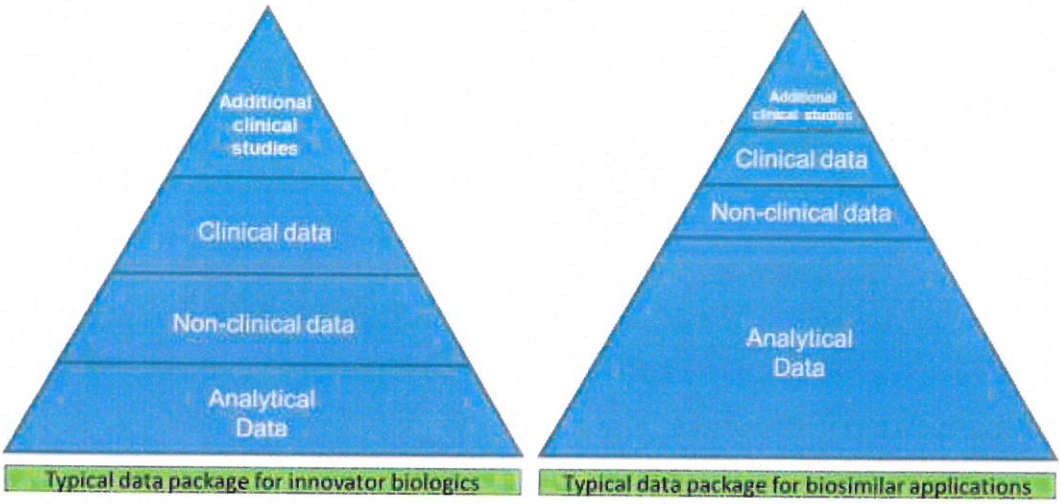


Figure 1: Data requirements for innovator biologics and biosimilar licensing applications

The importance of having analytical methods that can fully characterise biologics lies in the fact that even minor differences between a biosimilar and its reference product can have significant patient safety implications. Indeed, even slight changes in production processes may introduce subtle differences between batches of commercial biological product, and these differences may have potential for a clinical impact. Therefore, the analytical technologies developed for characterisation and release testing of biopharmaceuticals should be capable of detecting differences in product structure where they occur. Guidance from the International Council for Harmonisation on the characterisation of biopharmaceuticals states that new technologies (and improvements or modifications of existing technologies) are continually being developed and that these should be used if they can provide additional information or discriminating power when characterising biologics.

Our interests lie in the exploration and development highly discriminating analytical strategies for characterising biologics, in order to support biosimilar licensing applications. The research is being conducted through a joint industrial/academic setting, in a collaborative programme involving BioClin Research Laboratories and the BioSciences Research Institute at Athlone Institute of Technology (AIT), both of which are in close proximity to one another in Co. Westmeath, Ireland. The research is part-funded by a scholarship award from the Irish Research Council under the 'Employment-Based Postgraduate Programme". An impressive suite of *state-of-the-art* analytical technologies exist between BioClin and the Research Hub at AIT, and are at our disposal for performing this research. Undertaking this research at BioClin has enabled the company to diversify service offerings to include biopharmaceutical characterisation programs, which ensures they remain current and competitive as an analytical service provider in a rapidly-changing pharmaceutical market.

Analytical strategies for characterisation of biologics

Marketing authorisation for a new biopharmaceutical or a biosimilar, requires that all characteristics of the proposed drug that may have an impact on safety or efficacy are fully evaluated. Proteins may exhibit a high degree of heterogeneity due to the biosynthetic processes that living cells use to produce them. Owing to this heterogeneity, and the diverse and complex structure of biopharmaceuticals, there is no 'one-size-fits-all' analytical strategy for characterising them; therefore, each biologic requires a tailored approach. The goal for the analytical laboratory is to develop methods and technologies that can characterise all attributes of a biological drug. The International Council for Harmonization (ICH) Topic Q6B – "Specifications: Test Procedures and Acceptance Criteria for Biotechnological/ Biological Products",[8] which was adopted in

1999, provides guidance on tests and specifications that are appropriate for characterising biologics. The guidance indicates that applications for licensing of biologics should be supported by a comprehensive analytical package which characterises all the critical quality attributes (CQA's) of the drug. It provides direction on the setting of specifications and acceptance criteria that will ultimately serve as release test specifications, which will be a condition of approval for the drug. The document details that the reduction in dependence on clinical data permitted for a biosimilar application depends on the 'weight-of-evidence' from analytical studies that no functionally important differences exist between biosimilar and the reference drug. The following sections highlight the range of strategies typically applied in line with ICH Q6B guidelines, and comments on the structural elements that can be explored and the limitations of each strategy.

Confirmation of primary structure

Amino acid compositional analysis

The relative amount of each amino acid in a protein provides a characteristic profile for each biopharmaceutical, and can therefore confirm identification and support structural elucidation. Results from quantitative amino acid determination can be used for a precise determination of protein quantity in a sample (without the need for a reference standard), and this information can be further used to determine the extinction co-efficient – an important characteristic of a protein. Furthermore, amino acid analysis results can help to evaluate digestion strategies for peptide mapping and aid in identifying the presence of atypical amino acids that may have been incorporated into the protein. The test is often used to demonstrate comparability and consistency between batches for lot release of finished products.

The analysis involves the hydrolytic degradation of the protein into its constituent amino acids, followed by separation and quantitation of the free amino acids. Prior purification is essential, as buffer components can interfere with hydrolysis. Additionally, high-purity materials are required, as some reagents may be contaminated with low levels of amino acids that can distort the results – e.g. general reagent-grade hydrochloric acid (HCl) frequently has relatively high levels of amino acids such as glycine present.[9] Glassware and other consumables used for analysis must be free from contaminants – e.g. depyrogenised by baking at 500°C for 4 hours, or certified pyrogen-free disposable glassware should be procured.

Hydrolysis is typically performed by heating to 110°C for 24-72 hours in the presence of 6 M HCl (constant boiling), during which peptide bonds are hydrolysed, releasing the free amino acids. The wide range of treatment durations is due to the fact that some peptide bonds (such as those between isoleucine and valine) are more difficult to break than others. As such, hydrolysis duration is dependent upon amino acid content and sequence, and should be determined empirically for each unique protein. Hydrolysis can be performed in either the liquid phase (in which the protein is dissolved in HCl), or the vapour phase (in which only HCl vapours come into contact with the sample) – the former approach can reduce sample contamination from low-grade HCl. Accelerated methods involving higher temperatures for a shorter duration, or the use of microwave energy can significantly reduce hydrolysis times – in part, this research is evaluating rapid hydrolysis methods based on the latter.

One of the drawbacks of acid hydrolysis is that amino acids vary considerably in their stability to such treatment. Tryptophan is completely destroyed, and asparagine and glutamine are both deamidated to aspartic acid and glutamic acid, respectively. The complete loss of these three therefore limits analysis to 17 of the 20 common amino acids. Also, serine and threonine are partially destroyed; some amino acids are prone to oxidation, and others, such as tyrosine, can become halogenated. Certain treatments can address many of these concerns (e.g. addition of phenol can prevent halogenation of tyrosine, and removal of oxygen from the headspace of the reaction tube can reduce oxidation). However, the addition of agents to

protect one amino acid from degradation or alteration can often be at the expense of another amino acid. Therefore, where the sequence is available for a test protein, it should be consulted in order to evaluate the best approach for hydrolysis, and a time-course study initiated to determine appropriate hydrolysis duration. As part of our research, we characterise the rate of degradation of all naturally occurring amino acids subject to a range of hydrolysis treatments. This information will prove useful when interpreting experimental results from subsequent test samples.

Following hydrolysis of proteins, the free amino acids must be separated from each other and detected, in order to allow the relative quantities of each to be determined. Ion-exchange or reverse-phase HPLC are the methods of choice for separating the amino acids, followed by ultra-violet or fluorometric detection. In order to increase sensitivity and enhance detection, amino acids are typically derivatised, either pre-column (with ninhydrin or *o*-phthalaldehyde) or post-column (with phenyl isothiocyanate, 6-aminoquinolyl-*N*-hydroxysuccinimidyl carbonate, or other agents). Based on available technologies, we perform this analysis using pre-column methods employing both of the derivatisation agents specified above. Phenyl isothiocyanate (PITC) reacts with free amino acids to form phenylthiocarbamoyl derivatives, which are then separated on a reverse-phase octadecylsilane column with detection at 254 nm. Reaction with 6-aminoquinolyl-*N*-hydroxysuccinimidyl carbamate produces amino acid-urea derivatives that fluoresce strongly at 395 nm (excitation wavelength 250 nm) – a representative chromatogram from amino acid analysis is shown in Figure 2.

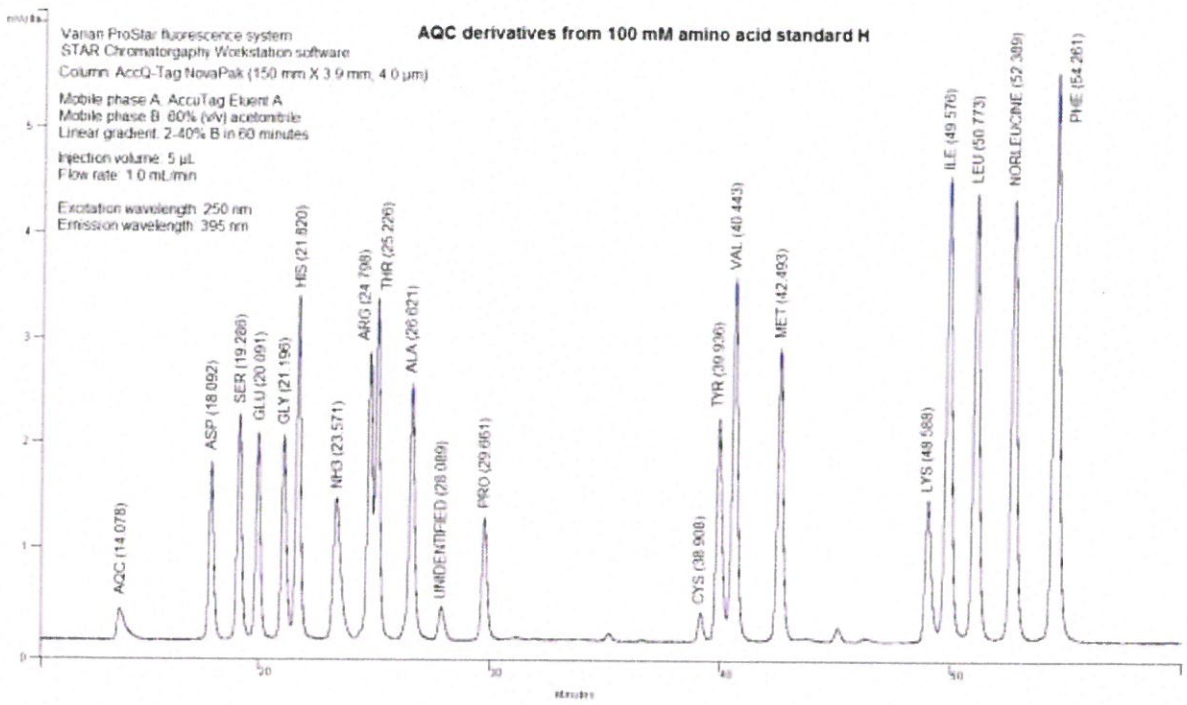


Figure 2: Representative chromatogram from amino acid analysis

Peptide mapping

Peptide mapping is an indispensable tool for characterising biologics. The technique is sensitive enough to detect even very minor changes to primary or secondary structure in a biopharmaceutical, and allows localisation of where this change has occurred. The technique can be performed using standard HPLC alone;

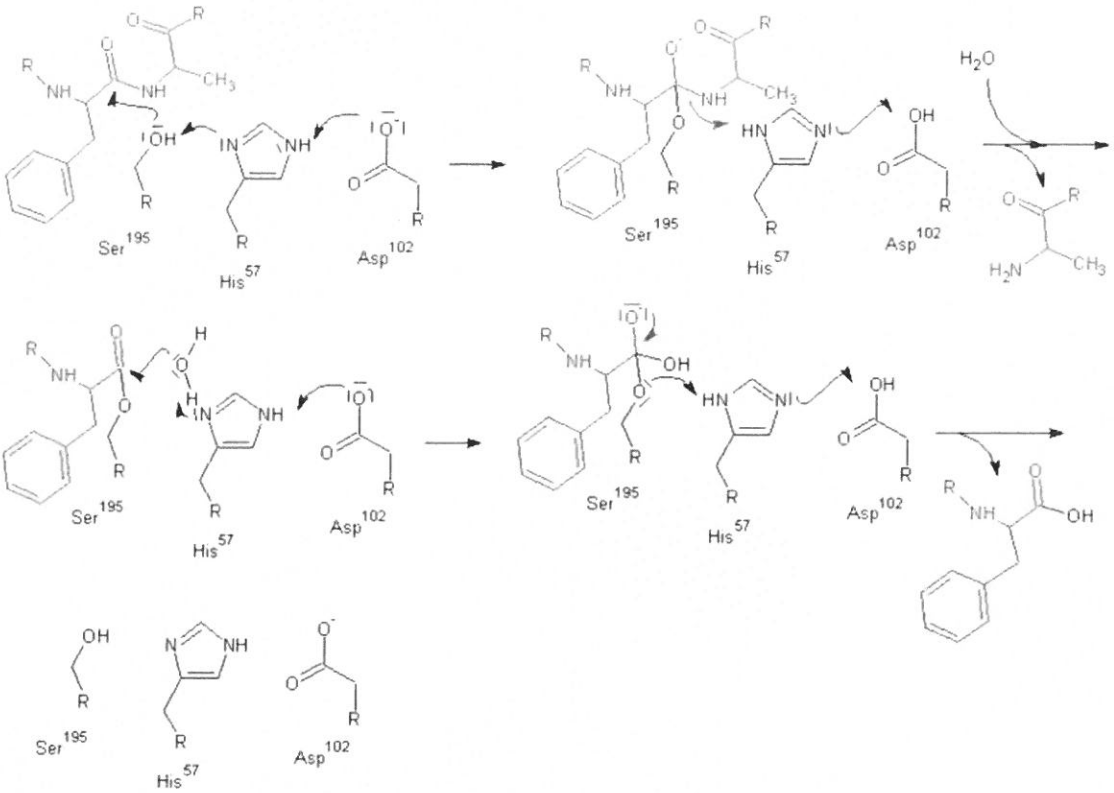
however, HPLC combined with mass spectrometry (LC-MS) releases enormous characterisation potential, as the mass information provided by mass spectrometry can reveal the nature of any change that is detected. Peptide mapping is one of the core techniques that is applied to release testing for biopharmaceuticals, and should be extensively used when comparing a biosimilar to the innovator drug. The technique is also used for assessing genetic stability of production cell lines, as mutations in the encoding gene often result in a change in the amino acid encoded – this change can be readily identified from peptide mapping data.

Peptide mapping makes use of commercially-available 'sequencing grade' proteolytic enzymes, such as trypsin (which cleaves at the *C*-terminal side of arginine and lysine residues) and endoproteinase Glu-C (which cleaves at the *C*-terminal side of hydrophobic amino acids). These enzymes (being protein in nature themselves) are often chemically treated in order to eliminate or reduce autolytic activity (where the enzyme cleaves neighbouring enzyme molecules). Chemical fragmentation strategies are also possible, such as treatment with cyanogen bromide, which cleaves at the *C*-terminal side of methionine residues. Table 2 details the commonly used enzymatic and chemical cleavage agents, and their respective specificities.

Table 2: Cleavage agents commonly used for peptide mapping

Type	Agent	Specificity
Enzymatic	Trypsin	*C*-terminal side of arginine and lysine residues
	Chymotrypsin	*C*-terminal side of leucine, methionine, alanine, tyrosine and tryptophan
	Pepsin	Non-specific digest
	Lysyl endopeptidase	*C*-terminal side of lysine
	Endoproteinase Glu-C	*C*-terminal side of glutamic acid and aspartic acid
	Endoproteinase Asp-N	*N*-terminal side of aspartic acid
	Endoproteinase Arg-C	*C*-terminal side of arginine
Chemical	Cyanogen bromide	*C*-terminal side of methionine
	2-Nitro-*thio*-cyanobenzoic acid	*N*-terminal side of cysteine
	O-Iodosobenzoic acid	*C*-terminal side of tyrosine and tryptophan
	Dilute acid (<0.1 M HCl)	Aspartic acid and proline
	3-Bromo-3-methyl-2-(2-nitrophenylthio)-3*H*-indole (BPNS-skatole)	Tryptophan

These agents cleave the biopharmaceutical at specific sites along the protein backbone with very high regularity. Treatment therefore results in the protein being broken down into a number of fragments dependent upon the sequence of amino acid residues of the parent protein. As all proteins have a unique amino acid residue sequence, treatment of different proteins with the same enzyme will give rise to distinctive peptide maps. An example of the mechanism of action of enzymatic cleavage of a peptide bond is illustrated in Figure 3, for the serine protease, chymotrypsin.



The active site of chymotrypsin consists of a triad of amino acid residues (shown in black) - histidine-57, aspartic acid-102 and serine-195. The enzyme cleaves peptide bonds by attacking the unreactive carbonyl carbon with serine-195, which is a powerful nucleophile. The enzyme associates non-covalently with the polypeptide substrate. Next, H+ is transferred from Ser to His, which forms a tetrahedral transition state with the enzyme. H+ is then transferred to the C-terminal fragment which is released by the cleavage of the C-N bond. The N-terminal peptide is bound through an acyl bond to serine. A water molecule binds to the enzyme in place of the departed polypeptide. This water molecule then transfers its proton to His-57 and its -OH group to the remaining substrate fragment. This forms a second tetrahedral transition state. The second peptide fragment is then released - the acyl bond is cleaved, the proton is transferred from His back to Ser, and the enzyme returns to its initial state.

Figure 3: Mechanism of action of chymotrypsin.
Image credit: Felix Plasser CC-BY-SA 3.0.

Prior to digestion with an enzyme, it is often necessary to break disulphide bonds within the protein in order to ensure the proteolytic enzymes have access to all cleavage sites. This is typically accomplished by treating the protein with an agent that reduces disulphide bonds, followed by alkylation of the free thiol groups to prevent the disulphide bonds from reforming. Dithiothreitol or β-mercaptoethanol are commonly used reducing agents and iodoacetamide is a frequently used alkylating agent for this purpose. It is important to remember that these treatments increase the mass of peptide fragments – e.g. alkylation with iodoacetamide increases the mass of each fragment by approximately 58 Da for each cysteine present.

When the amino acid residue sequence of a protein is known (which is usually the case for biopharmaceuticals) the fragments that should be generated from treatment with a given cleavage agent can be predicted using *in silico* digestion. Many software applications and online resources are available for performing this task, such as the freely-accessible 'PeptideMass' from the website http://www.expasy.org.[10] The experimentally obtained peptide map results can then be compared to those predicted from the *in silico* digestion.

When HPLC alone is used for analysing results from peptide mapping, differences (between a biosimilar and its reference biologic, for example) are revealed by shifts in retention time of proteolytic fragments. When the nature and precise location of differences is to be determined, LC-MS is used instead of HPLC. With this method of analysis, each fragment is detected with a characteristic retention time, but the mass of

each eluting fragment is also obtained. With this additional information, the analyst can determine which peak corresponds to each fragment from the *in silico* digestion. If any changes in mass are observed, there will be a shift in the mass of the fragment that corresponds with the nature of the change (likely also accompanied by a shift in retention time). For example, where a single amino acid residue in a peptide fragment is oxidised, the overall mass of the fragment will increase by approximately 18 *Da*, or where a lysine has been substituted for an alanine, there will be a mass decrease of approximately 85 *Da* due to the difference in mass between these two amino acid residues. This information allows the analyst to pinpoint with high accuracy the location of any observed modification. Subsequently, the exact nature of the modification can be further explored by using tandem mass spectrometry (LC-MS/MS). Figure 4 illustrates a typical BioClin workflow for peptide mapping using LC-MS, while Figure 5 shows experimental peptide mapping data for asparaginase following digestion using trypsin. Asparaginase is an enzyme (Enzyme Commission (EC) Number: 3.5.1.1.) which is used to treat acute lymphoblastic leukaemia (ALL) in children. Asparaginase exploits the observation that acute lymphoblastic leukaemia cells cannot synthesise asparagine, and therefore depend on it being supplied in the bloodstream. Asparaginase catalyses the conversion of circulating asparagine to aspartic acid and ammonia, which deprives leukemic cells, leading to cell death. Table 3 shows the predicted peptide fragments from *in silico* digestion of asparaginase.

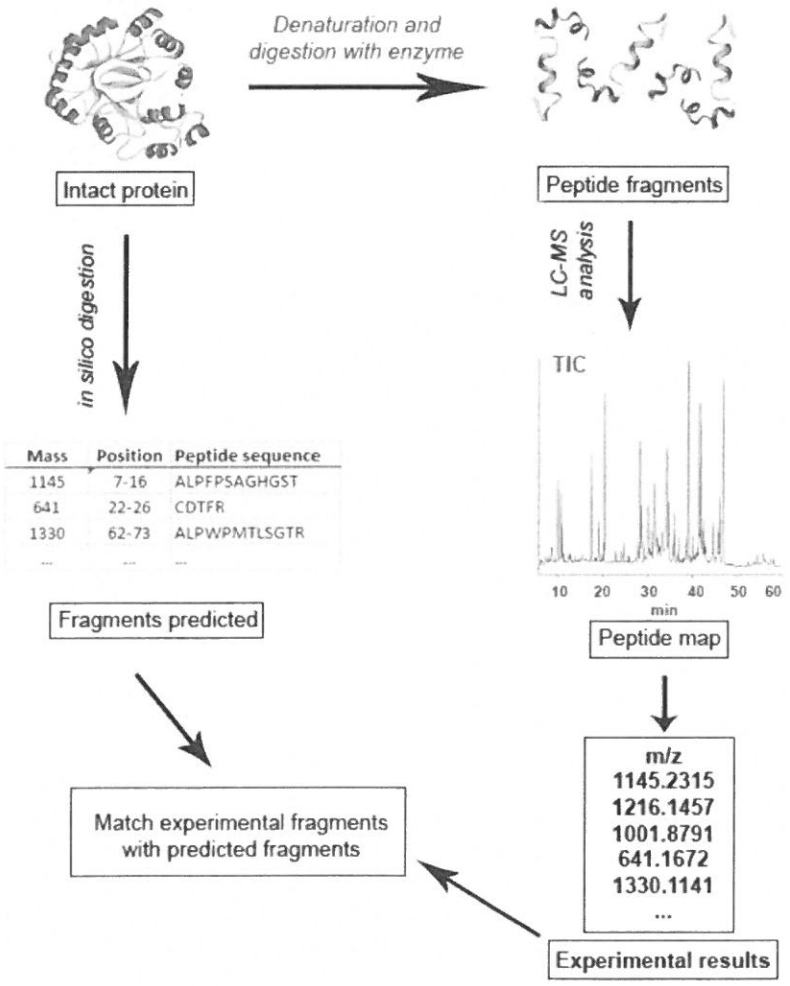
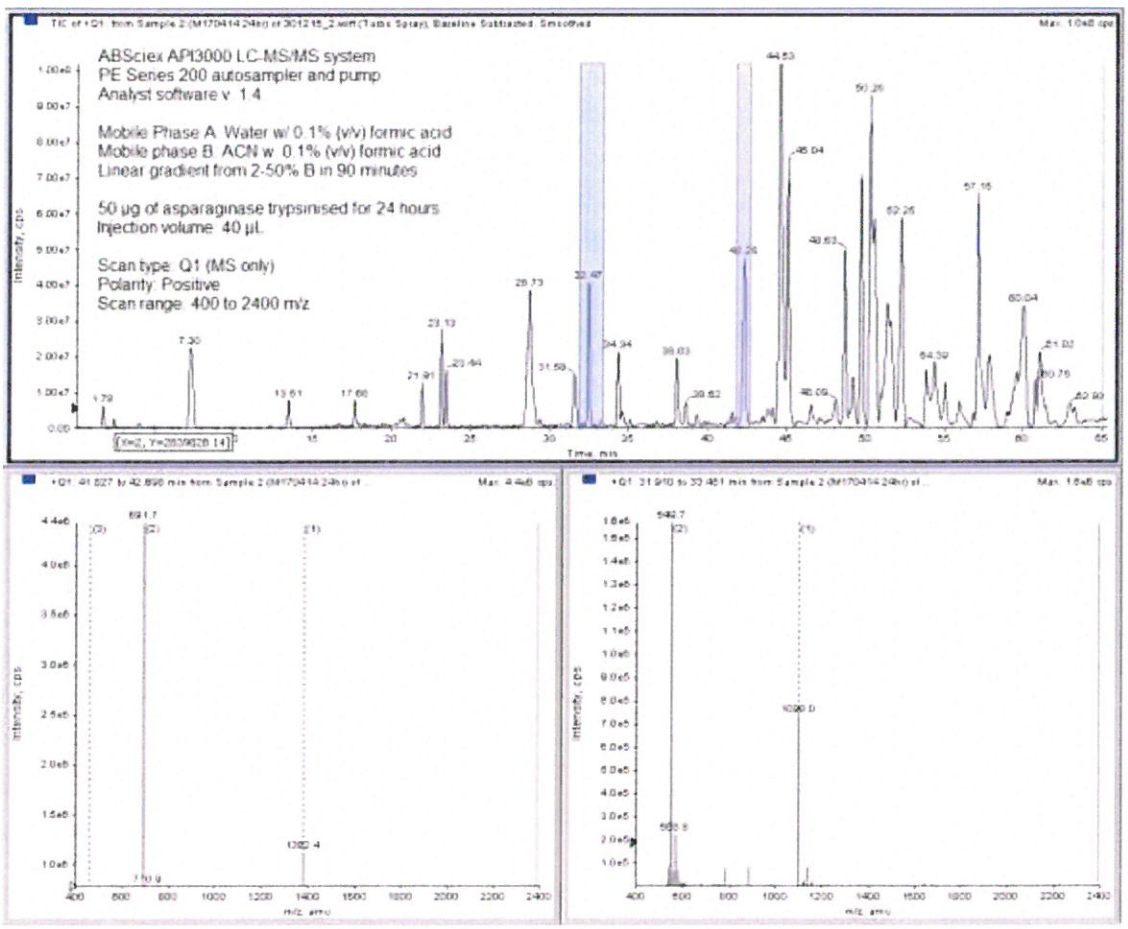


Figure 4: Peptide mapping using LC-MS

Table 3: *In silico* digestion of asparaginase using trypsin

Mass - [M+H]+	Position	Sequence
3642.7	127-161	CDKPVVMVGAMRPSTSMSADGPFNLYNAVVTAADK
3472.9	7-44	TALAALVMGFSGAALALPNITILATGGTIAGGGDSATK
2804.3	102-126	TDGFVITHGTDTMEETAYFLDLTVK
2431.2	72-93	GEQVVNIGSQDMNDNVWLTLAK
2153.1	252-273	ALVDAGYDGIVSAGVGNGNLYK
1694.9	236-251	VGIVYNYANASDLPAK
1617.8	295-310	VPTGATTQDAEVDDAK
1521.7	337-348	DPQQIQQIFNQY
1518.8	195-208	SVNYGPLGYIHNGK
1487.8	167-180	GVLVVMNDTVLDGR
1479.9	52-65	VGVENLVNAVPQLK
1381.7	***311-323***	***YGFVASGTLNPQK***
1233.6	219-229	HTSDTPFDVSK
1227.8	326-336	VLLQLALTQTK
1123.6	274-284	SVFDTLATAAK
1097.5	***185-194***	***TNTTDVATFK***

Data in bold italics correspond to peaks that are highlighted in Figure 5 below



Peptide mapping results for asparaginase from *E. coli*. The top pane shows the total ion current (TIC) from the mass spectrometer, while the bottom panes show the respective peptide masses for two well-resolved peaks in the TIC - 1382.4 and 1098.0, both of which are predicted from the *in silico* digestion (See Table 3). The slight differences in masses reported are related to the resolution and calibration of the mass detector used.

Figure 5: Peptide mapping results of asparaginase

Continues…..P50

LENNOX

The one source for all your chemical needs.

PH Buffers & Conductivity Standards

Lennox offers a comprehensive range of pH Buffers and Conductivity solutions for the calibration, monitoring and qualifying of pH and conductivity instruments. All of Lennox pH and Conductivity solutions are traceable against SRM of NIST.

Volumetric Solutions

Volumetric solutions from Lennox are ready-to-use solutions manufactured in large lots that will save you the time and expense of preparation and standardization. We offer a full range of Base and Acid solutions. Lennox ready-to-use volumetric solutions are manufactured to stringent specifications and utilise Quality Control procedures to reduce lot to lot variability, are labelled with expiration date and available in several packaging options.

Custom Manufacturing

Lennox offers a flexible custom manufacturing service to produce quality products. Our lab routinely manufactures solutions to meet research, pilot scale and full scale production requirements. We have extensive experience in this area and can manufacture from 100ml to 1000lt. Contact our sales team to discuss your chemical custom manufacturing needs now.

Ethanol

We can supply from stock a full range of Ethanol Absolute & Ethanol Denatured (IMS) in a large range of volumes and concentrations.

Contact us on 01455 2201 or email cs@lennox for more information on Lennox Chemicals.
www.lennox.ie Article

LENNOX

Amino acid sequencing

The sequence of a protein refers to the linear arrangement of amino acid residues in that protein. Due to the *in vivo* processing (and possible post-translational modification, explained later) associated with biopharmaceuticals, the *N*-terminal and *C*-terminal sequences of a protein are not always readily predictable from the gene sequence. As such, sequence analysis of a biopharmaceutical is typically limited to the *N*-terminal and *C*-terminal ends of the protein. *N*-terminal sequencing can reveal if truncation of the protein has occurred. Such truncations may result from the presence of trace levels of expression cell proteases that were not removed through downstream processing (involving a series of chromatographic steps which result in successively higher purity), and truncated protein may not exhibit the desired therapeutic effect. The sequence information obtained is used to confirm consistency between batches and for demonstrating similarity between biosimilars and reference drugs. There are two main approaches presently used by industry for the terminal amino acid sequence analysis of biopharmaceuticals – Edman degradation chemistry and mass spectrometry.

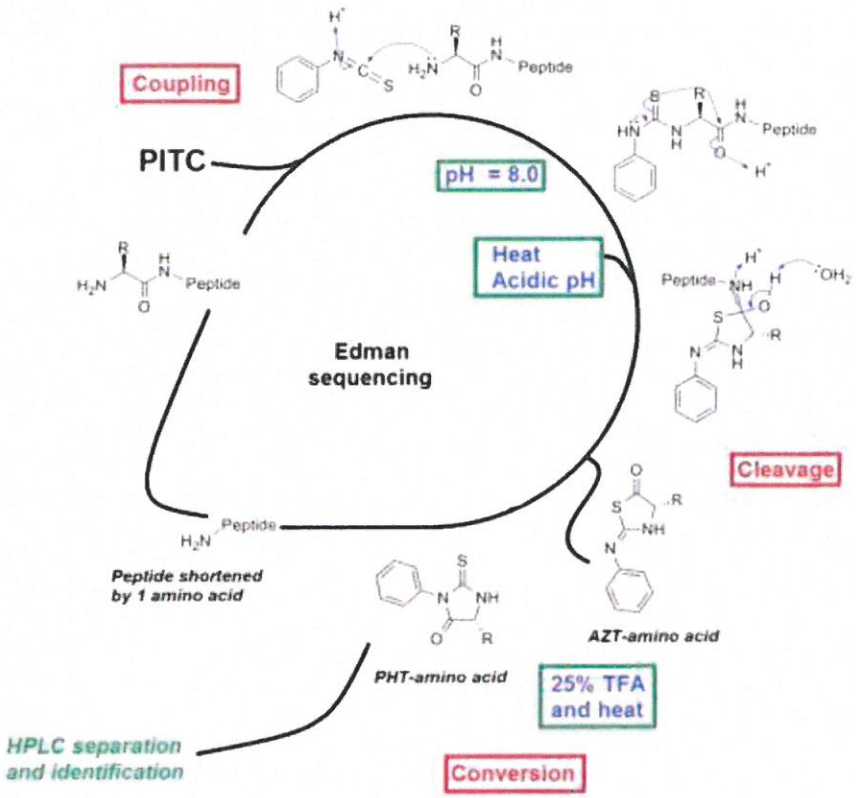


Figure 6: Edman degradation chemistry used for sequencing of proteins

Edman sequencing, illustrated in Figure 6 above, involves derivatisation and cleavage of one amino acid at a time from the *N*-terminus, followed by analysis and identification of the derivatised amino acid. The protein is first combined with a reagent that will selectively react with the *N*-terminal amino acid. PITC (which is also used for amino acid compositional analysis, described earlier) is combined with the protein sample under mildly alkaline conditions, where it selectively reacts with the uncharged terminal amino group to give a phenylthiocarbamoyl derivative. Subsequently, under acidic conditions, a thiazolinone derivative of the amino acid is liberated, leaving the remainder of the polypeptide chain intact, but shortened in length

from the *N*-terminus by one amino acid. The thiazolinone derivative is then extracted into an acidified organic solvent to give a more stable phenylthiohydantoin (PTH) amino acid derivative, which can subsequently be identified using HPLC. This procedure is continued until the entire length of the polypeptide chain has been sequenced.

Theoretically, it should be possible to sequence an entire protein using the Edman sequencing method, however, in practise this is not achievable. The efficiency of each cycle of the reaction is approximately 98%,[11] which means that only approximately 50 amino acid residues can be reliably sequenced using this method. In practise, no more than 30 residues should be sequenced by the Edman method in order to produce reliable and reproducible results. However, some automated sequencers claim 99% efficiency and manufacturers claim that they can reliably sequence 50-100 residues or more.[12]

This limitation of the Edman method can be overcome using a method developed by Frederick Sanger in 1955, for which he received the Nobel Prize in Chemistry in 1958.[13] Sanger's method involves selectively cleaving the original protein into smaller peptide fragments using trypsin, then separating these fragments using electrophoresis. Each fragment can then be sequenced using the Edman method. The order in which these fragment sequences are to be recombined is determined by cleaving the protein with a second proteolytic enzyme with a different specificity, such as chymotrypsin, followed by Edman sequencing of the fragments as previously described. Overlapping the partial sequences obtained from the two different digests allows the correct order of the fragments to be determined, thereby allowing reconstruction of the complete protein sequence.

Another important limitation of the Edman method is that it requires an unmodified amino group at the *N*-terminal of the molecule. Unlike prokaryotic cells, which are not thought to significantly post-translationally modify proteins, evidence suggests that up to 80% of intact proteins from eukaryotic organisms, such as Chinese hamster ovary (CHO) cells, have modified *N*-terminal amino groups.[14, 15] This has potential to result in significant complications for biologic characterisation, as the majority of biologics in present-day use are expressed in eukaryotic cells.[16] Several methods for unblocking these amino groups to facilitate Edman sequencing exist, but they require comparatively large amounts of protein, and don't produce consistent results, particularly when the nature of the blocking group is unknown. This limitation is less significant when enzymatic fragmentation of the protein is employed, as each of the internal polypeptide fragments would have unmodified N-terminal amino groups, and are therefore amenable to Edman sequencing.

Tandem mass spectrometry (MS/MS) is emerging as one of the most powerful methods of sequencing proteins.[17-22] It also offers the most reliable approach for C-terminal sequencing of proteins (analysing only the C-terminal peptide from the proteolytic digest of the protein). Using this approach, proteins are first cleaved into smaller fragments of approximately 20 amino acid residues or less (as for peptide mapping). Trypsin is the enzyme of choice for cleaving proteins prior to sequencing by MS/MS, as it typically gives rise to peptide fragments from 8-20 amino acid residues long, which is the ideal range for most mass spectrometers. The resulting series of peptide fragments can then be analysed via (MS/MS) to gain sequence information. In a typical sequencing workflow, proteolytic digests are infused directly to the mass spectrometer, and each predicted fragment (from the *in silico* digestion) is selected independently in the first quadrupole (see Figure 8). Selected peptides are then passed into the second quadrupole, where they are broken down through a process called collision induced dissociation (CID), and the resulting ions analysed in the third quadrupole. During CID, bond breakage most frequently occurs through the lowest energy pathways in the molecule, which for peptides are the amide bonds. Therefore, the major fragments generated differ from each other by a single amino acid. Roepstorff-Fohlmann-Biemann nomenclature is used to describe the ions that are produced in this process.[23, 24] Using this system, the fragments are called b-ions when the charge is retained by the amino terminal fragment, and y-ions if the charge is retained by the

carboxy terminal fragment, with ions being labelled consecutively from the original amino terminus – other fragments are also possible, albeit with a lower frequency - see Figure 7. As with the Edman method, reconstruction of the partial sequences to build up the overall sequence of the parent protein is accomplished by using a second enzyme, by overlapping both sets of results. Many software platforms and online resources are available for simplifying this task. Sequencing via mass spectrometry is illustrated in Figure 8.

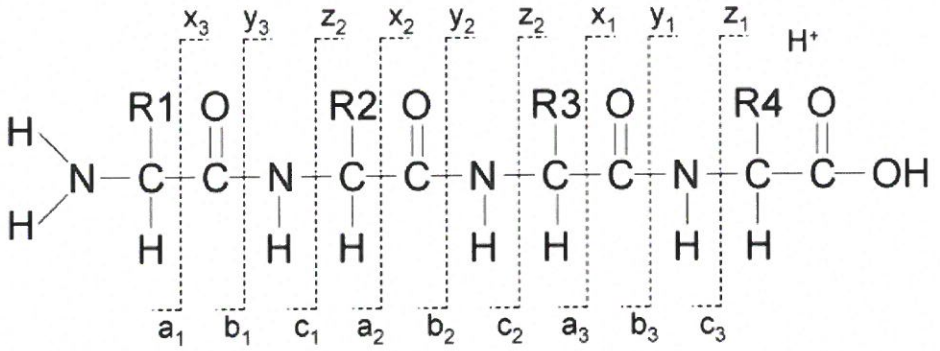


Figure 7: Roepstorff-Fohlmann-Biemann nomenclature.
Ionsource.com. Available at http://www.ionsource.com/tutorial/DeNovo/full_anno.htm.

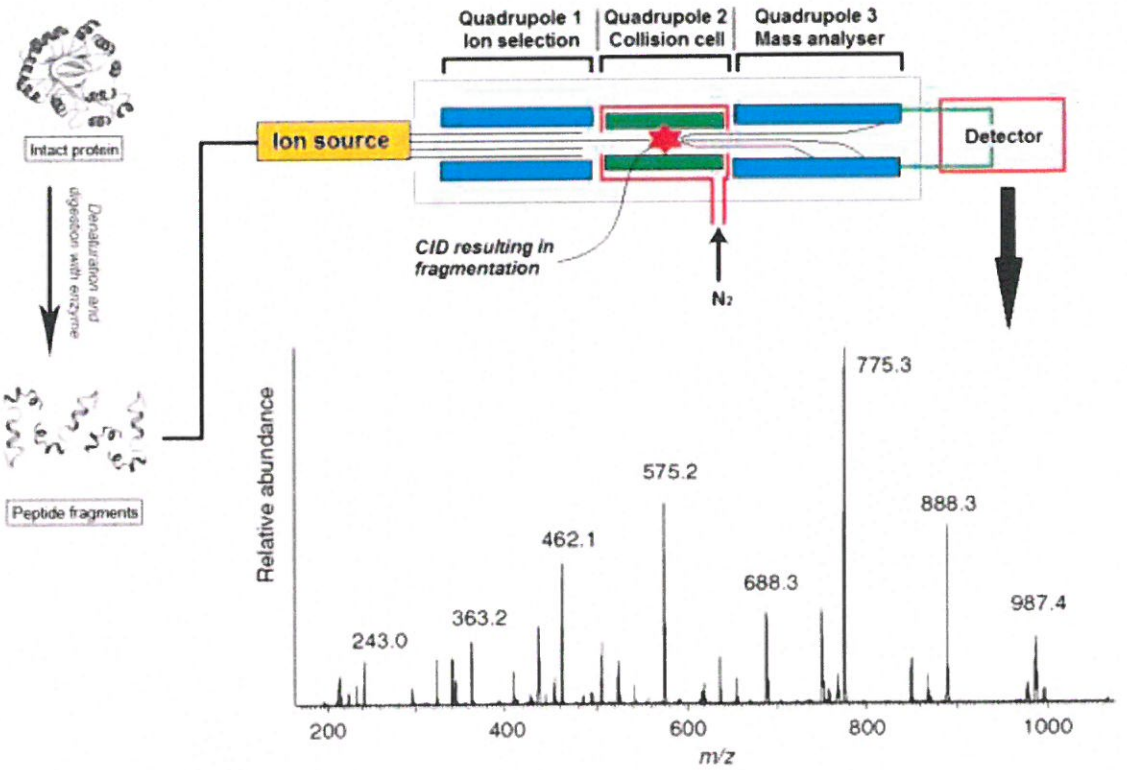


Figure 8: Workflow for protein sequencing via mass spectrometry

Sulphydryl groups and disulphide bonds

The side chains of cysteine residues contain thiol groups that can react with one another to form covalent disulphide bonds (denoted S-S) through a process called oxidative folding. An enzyme called 'protein disulphide reductase' (PDI) oxidises the thiol group of cysteine residues, thereby catalysing the formation of S-S bonds. Disulphide bonds can form between cysteines in the same polypeptide chain (intramolecular) or between cysteines from separate polypeptide chains (intermolecular). Intramolecular S-S bonds are responsible for stabilising the tertiary structure of a protein, while intermolecular S-S bonds are attributed to stabilising quaternary structure in complex proteins consisting of two or more subunits. While methionine also contains a sulphur atom, only cysteine residues can form S-S bonds. Reduced cysteines will react with each other if they are in close proximity, even if the protein isn't properly folded. Therefore, the greater the number of cysteines in a protein, the greater the potential for mismatched disulphide bonds. This is likely to result in a protein which does not act as intended when used as a biopharmaceutical. Therefore, in order to ensure drug safety and efficacy, it is important that the location of cysteine residues and disulphide bonds are determined, as this cannot be accurately predicted from the gene sequence.[25]

Conventional methods such as high-field NMR have been used to characterise disulphide bonds in proteins.[26] However, such methods typically require high concentrations of protein, which may not always be available at the early drug development stage. Mass spectrometry using soft ionisation techniques, such as electrospray ionisation, are gaining in popularity for disulphide bond analysis. Typical approaches involve performing peptide mapping via LC-MS under both reducing and non-reducing conditions. Where non-reduced protein is digested for peptide mapping, the disulphide linkages will keep fragments covalently combined, giving rise to larger molecular mass fragments. The reduced sample will result in more fragments, as fragments connected by disulphide bonds will be separated from one another when those bonds are broken. Data from these analyses can be combined and used to determine where disulphide bonds are present. This process is illustrated in Figure 9 below.

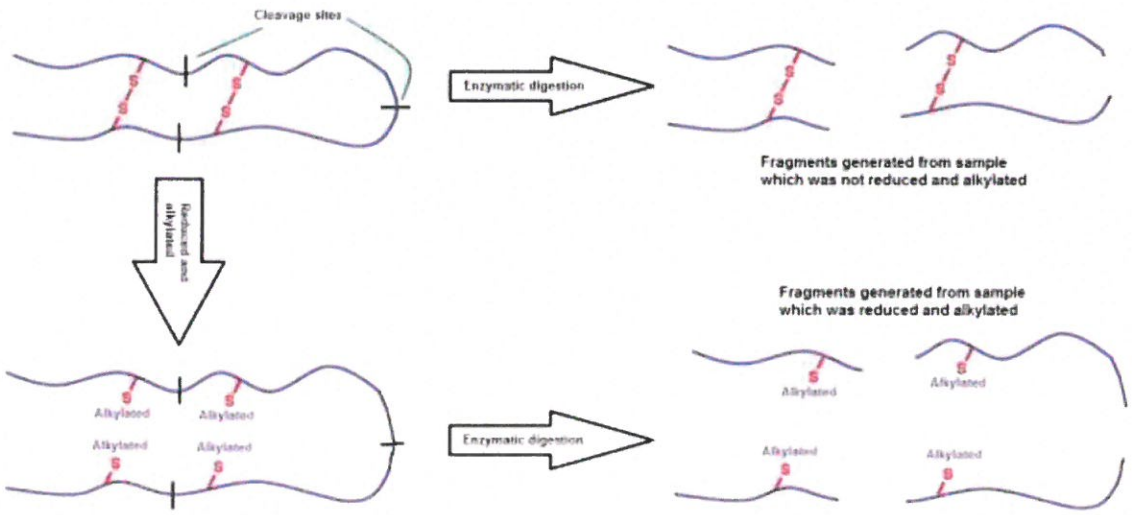


Figure 9: Disulphide bond analysis using reducing and non-reducing peptide mapping

Glycosylation analysis

Approximately two-thirds of the biopharmaceuticals in the current market are glycoprotein in nature. The carbohydrate content of proteins often plays a significant role in the function of the protein, having an impact on physiochemical properties and thermal stability, and helping to mediate effects such as circulating half-life and their reactivity towards target receptors (and therefore their pharmacological efficacy). Indeed, proteins with unanticipated glycan structure may promote a potentially-harmful immune response. Therefore, it is imperative that the constancy of carbohydrate moieties is maintained to ensure the safe and efficient use of glycosylated biopharmaceuticals. Glycosylation can occur at any number of sites on a protein molecule. *N*-glycosylation is the attachment of glycans to the carboxamido nitrogen on asparagine, and this is the most common type of glycosylation seen. *O*-glycosylation involves the attachment of glycans to serine or threonine residues. Other forms of glycosylation are less common. No universally applicable approach for glycan structural analysis of proteins is available, and a combination of methods is typically employed, many of which exploit various forms of mass spectrometry.[27] Glycosylation analysis presents a significant analytical challenge, as glycans typically form highly-branched structures (unlike the simpler linear arrangement of amino acids in peptides). As many of these variant arrangements of sugars will be isobaric (have the same molecular mass) it can be particularly challenging to tease out the actual glycan structure. However, three general approaches are typically employed, often in combination: (i) *characterization of glycans in intact glycoproteins*; analysis is by means of techniques including capillary isoelectric focussing, capillary zone electrophoresis, or mass spectrometry – results give partial characterisation of glycan profile and often serve as the starting point for glycan analysis; (ii) *characterization of glycopeptides derived from the protein*: proteins are digested and fragments analysed via capillary zone electrophoresis and tandem mass spectrometry – this provides information on glycosylation sites and some information on the nature of glycans present; (iii) *analysis of glycans that have been chemically or enzymatically removed from the protein*: acid hydrolysis of proteins or treatment with enzymes with specificity for various forms of glycosylation (such as PNGase F) releases the glycans which can be subsequently analysed by HPLC and other methods – this provides data on total glycan content.

Characterisation of physiochemical properties

Molecular weight or size

The molecular weight of a biopharmaceutical is often the first analysis applied during a characterisation programme. A molecular weight that differs from the predicted molecular weight immediately highlights a difference in the actual and expected structure. Many methods are used for this determination, ranging from size exclusion chromatography (SEC-HPLC), sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE), and mass spectrometry. These techniques differ significantly in their resolution, and the type of technique chosen is often dependent on the equipment that is available at the test laboratory. Resolution is sacrificed with SDS-PAGE, where proteins can typically be estimated to within approximately 500 *Da*; however, this technique has the benefit that it allows early visual inspection for other protein-based contaminants (of different molecular weights) in the sample. Also, separation of proteins *via* SDS-PAGE effectively removes impurities and formulation excipients, so proteins can be excised from the gel for downstream analysis, such as peptide mapping. For determination of molecular mass, the highest resolution is obtained with mass spectrometry, particularly matrix-assisted laser desorption ionisation coupled with time-of-flight mass spectrometry (MALDI-TOF-MS), which can determine protein mass down to the sub-Dalton range.

Electrophoretic patterns

SDS-PAGE, native PAGE (in which the protein is not denatured prior to electrophoresis, but run in its native form), isoelectric focussing (IEF), capillary electrophoresis (CE), capillary isoelectric focussing (cIEF) and Western blotting are some of electrophoretic techniques commonly used to provide data on identity, homogeneity, and purity of a biopharmaceutical. These techniques are widely employed as part of a biopharmaceutical characterisation programme in support of new biologic or biosimilar licensing applications. IEF and SDS-PAGE are also included in the battery of specialised tests for release testing of biotherapeutics, according to USP General Chapter <1045>.

SDS-PAGE provides molecular mass information, while also revealing the presence of protein contaminants that may be present in a sample, which would appear as additional bands on the gels. SDS-PAGE has applications in forced degradation studies, where samples are extracted at various time-points during the degradation study and subject to SDS-PAGE analysis.[28] Sample degradation would be revealed through shifts in the location or changes in the appearance of the bands on the gel. For example, where a protein undergoes significant fragmentation during forced degradation, the band corresponding to the protein would be seen to decrease in intensity, and the presence of additional bands or smearing of the main band corresponding to the protein of interest may be observed. These observations provide insight into degradation pathways and products, which can serve as a basis for more detailed analysis using other techniques, such as mass spectrometry.

Aggregation refers to the non-covalent association of protein molecules in solution. Aggregation of proteins can be a significant problem during storage of finished product, as it can be intimately tied to protein folding. Physical factors (such as light exposure and temperature excursions) and chemical factors (including the impact of formulation excipients and pH shifts) may promote the development of aggregates,[29] and these may have an adverse impact on the biological activity of the protein, particularly where immunogenic responses to the protein are considered.[30] The presence of aggregates may be determined using native PAGE, as such aggregates would not dissociate during the relatively mild sample preparation steps employed in this analysis. The appearance of bands corresponding to 'multiple times' the mass of the protein of interest on native-PAGE gels would be indicative of aggregate formation.

IEF enables the detection of charge heterogeneity of a biopharmaceutical preparation. Charge variants have the potential to influence stability and biological activity of biopharmaceuticals, particularly those of the monoclonal antibody (mAb) class.[31] IEF can reveal the presence and ratio of charge variants due to post-translational modifications, as these differences can shift the isoelectric point of a protein considerably. The technique involves application of the sample to an immobilised pH gradient, then applying current to the system. Due to the variability in number of charged side groups in proteins, they generally have a charge at a given pH. Therefore, when an electrical current is applied to the IEF gel, they will migrate through the pH gradient towards either the anode or cathode, until they arrive at a pH at which the native charge is suppressed (i.e. where the net charge on the protein is zero) – refer to Figure 15. If the protein then diffuses away from this point in the immobilised pH gradient, it will gain a charge again, causing it to migrate back in the opposite direction under the influence of the electric current – see Figure 10. This has the effect of 'focussing' the sample to a very sharp band within the gradient, allowing high resolution detection of charge variants, which would be present as additional bands in the gel. IEF can be performed in gel format, on IPG strips (solid matrix with immobilised pH gradients), or in capillary format. The latter has a greater resolution than the first two techniques, allowing charge variants that differ by less than 0.1 pH units to be readily resolved. However, where information of the charge heterogeneity is available (e.g. from prior analysis) resolution in the gel format can be greatly improved by using narrower gradients (e.g. using a pH gradient from 7 to 9, instead of from pH 3 to 10).

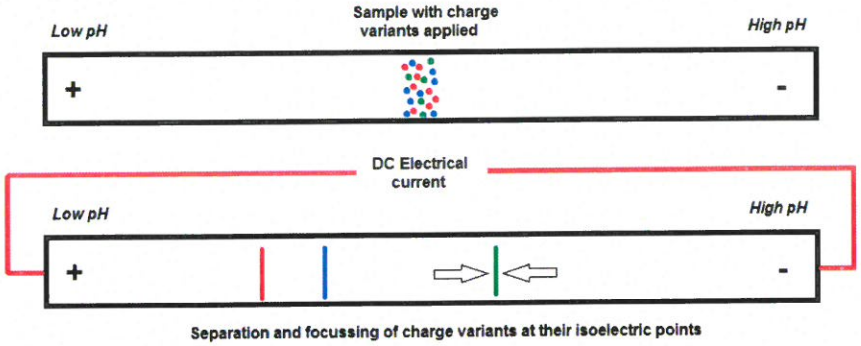


Figure 10: Isoelectric focussing for analysis of charge heterogeneity

Western blotting allows for unequivocal confirmation of protein identity. The technique is particularly useful for mixtures of complex proteins, as it allows detection of only the protein of interest in that mixture.[32] It can also reveal whether or not additional bands that are observed on a gel are related to the protein of interest (e.g. dimers or truncated protein) or if they are unrelated proteins that may be of host-cell origin, for example. Proteins are first separated using an electrophoretic method such as SDS-PAGE, with subsequent electrophoretic transfer to an inert membrane (such as PVDF or nitrocellulose), where they bind tightly to the membrane, becoming immobilised. The immobilised proteins can then be probed with a primary antibody ('anti-drug' IgG) that binds specifically to an epitope (the part the immobilised protein to which the antibody attaches itself) on the target protein. Following this, a secondary antibody (anti-IgG) (which has a 'reporter molecule' that generates a signal which can be detected) is added to the blot where it binds to bound primary antibody. A common example of a reporter molecule is the enzyme horseradish peroxidase – this catalyses the conversion of a chromogenic substrate to a derivative that can be detected via colorimetry, fluorimetry or through luminescence.

Chromatographic patterns

The identity and heterogeneity of formulated biotherapeutics should be thoroughly evaluated, with a number of important characteristics requiring consideration. Because of the high molecular complexity of biologics, they can be very sensitive to even minor changes in any detail of production. Examples include production process changes (either intentionally or inadvertently introduced), changes in batches of raw materials, or changes induced or promoted by formulation excipients. Other product changes may evolve slowly over time during storage, such as methionine oxidation; a primary degradation pathway for biologics.[33] Even minor changes (such as modification of a single susceptible amino acid), or more significant changes such as mismatched disulphide bonds, can result in significant peak shifts in chromatographic profiles. These retention time shifts allow rapid detection of variant forms that may be present in a biopharmaceutical preparation. For many biologics, particularly those based on monoclonal antibodies, a certain degree of heterogeneity is expected; however, this heterogeneity should be thoroughly characterised between production batches in order to ensure that it remains stable over time. Ion exchange (IEX) chromatography, size-exclusion chromatography (SEC) and hydrophobic interaction chromatography (HIC) are orthogonal approaches, which offer excellent selectivity and resolution for separating charge variants, size variants and hydrophobicity variants, respectively, in biopharmaceutical preparations.

IEX separates proteins based on charge heterogeneity. Among the various different modes of IEX described in the literature, cation exchange chromatography is the most appropriate mode for biologics. A typical approach involves elution of the sample using a linear salt concentration gradient, with charge variants eluting in order of increasing binding charge. A modified approach (termed 'chromatofocusing') was introduced by Sluyterman *et al.* between 1977 and 1981.[34] This approach involves the use of a pH gradient that can be generated internally in an IEX column. The column is packed with beads of highly-cross-linked poly(styrene–divinylbenzene) (PS/DVB) – selected for its stability across a broad pH range (pH 2 to 12). This alternative approach allows for high resolution separation of isoforms with very minor differences in their isoelectric points.

Unlike many forms of separation, SEC offers a significant advantage in that the comparatively milder aqueous mobile phases used allow biologics to be characterised with minimal impact on their native conformation. The technique is widely employed for the qualitative and quantitative determination of protein aggregates in biopharmaceuticals. Proteins are separated based on their hydrodynamic radius, using a column packed with spherical, porous beads with strictly-controlled pore size. Larger proteins and aggregates cannot diffuse into the pores, and so pass through the column unimpeded, eluting first in the chromatographic run. Smaller proteins diffuse into the pores of the beads, and so take longer to elute. SEC can also be used for approximating protein size by plotting a standard curve of molecular mass (of a range of proteins of different masses) versus retention time.

HIC exploits the hydrophobicity of proteins, which enables their separation on the basis of hydrophobic interactions between the non-polar regions of proteins and immobilised hydrophobic ligands present on the column packing. The adsorption of protein on the column is greater with higher salt concentrations. Therefore, proteins are separated by decreasing the salt concentration of the mobile phase over time. Proteins are not significantly altered using this separation technique – indeed, the technique is often used to purify proteins which maintain biological activity from formulated biopharmaceuticals; such bioactive proteins can subsequently be used for specific activity bioassays.

Reversed-phase HPLC is gaining in popularity for separating variant forms of biotherapeutics, owing to its compatibility with mass spectrometry (LC-MS) and its high resolving power. Hydrophobic proteins adsorb onto a hydrophobic solid support in an aqueous (polar) mobile phase. Increasing the organic solvent in the mobile phase decreases its polarity, and this reduces the hydrophobic interaction between the proteins and the stationary phase, resulting in desorption. The more hydrophobic the protein, the higher the concentration of organic solvent that is required to promote desorption.

Developments and advancements in column and separation technology (such as UHPLC) are greatly accelerating the use of chromatographic techniques for characterising biologics. These techniques offer improvements over more classical techniques (such as electrophoresis) from the perspective of analysis time, precision, selectivity, resolution and a range of other considerations.[35] This is due to the fact that most testing laboratories would have HPLC systems available, and only need to acquire new specialist columns to enable them to perform biopharmaceutical characterisations. UHPLC is particularly attractive during the early development stage (when available sample can often be as low as microgram quantities) as it has extremely small sample requirements when compared to more conventional HPLC.

Spectroscopic profiles

The three-dimensional structure of a protein not only determines size and shape, but also dictates physiological behaviour and biological activity: for example, solubility of a protein depends on a predominance of polar variable groups on the outside of the protein, where they interact with the aqueous

environment, and biological activity is intrinsically linked to the shape of the active site of a protein. Indeed, incorrectly folded proteins may elicit potentially harmful immune reactions, or cause loss of efficacy of the drug as a consequence of antibody response.[36, 37] As such, the higher-order structural elements of biopharmaceuticals need to be thoroughly evaluated in order to ensure drug efficacy and safety. Protein higher-order structure is typically evaluated using a range of spectroscopic analyses, including circular dichroism (CD) spectroscopy, nuclear magnetic resonance (NMR), and infrared (IR) spectroscopy. Stabilisation of protein higher order structure so as to provide practical shelf lives is an ongoing challenge for formulation scientists.[38] Spectroscopic techniques represent the most frequently encountered approach for determining stability of higher order structure as part of product development programmes. Many excipients exhibit their own characteristic responses in CD and FT-IR spectroscopy. Software applications to 'subtract' these exist, but the best approach is to extract the target analyte (i.e. the therapeutic protein) from formulations using solid phase extraction, and perform analysis in the absence of excipients.

CD analysis (in the 'far UV' spectral region; 190-250 nm) can be used for estimating the secondary structural elements (such as the α-helix or β-sheet) of proteins in solution. Each secondary structural element gives rise to spectra of a characteristic profile, as shown in Figure 11.

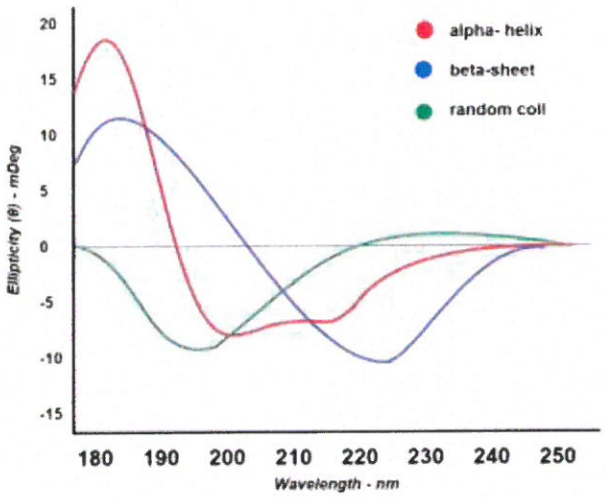
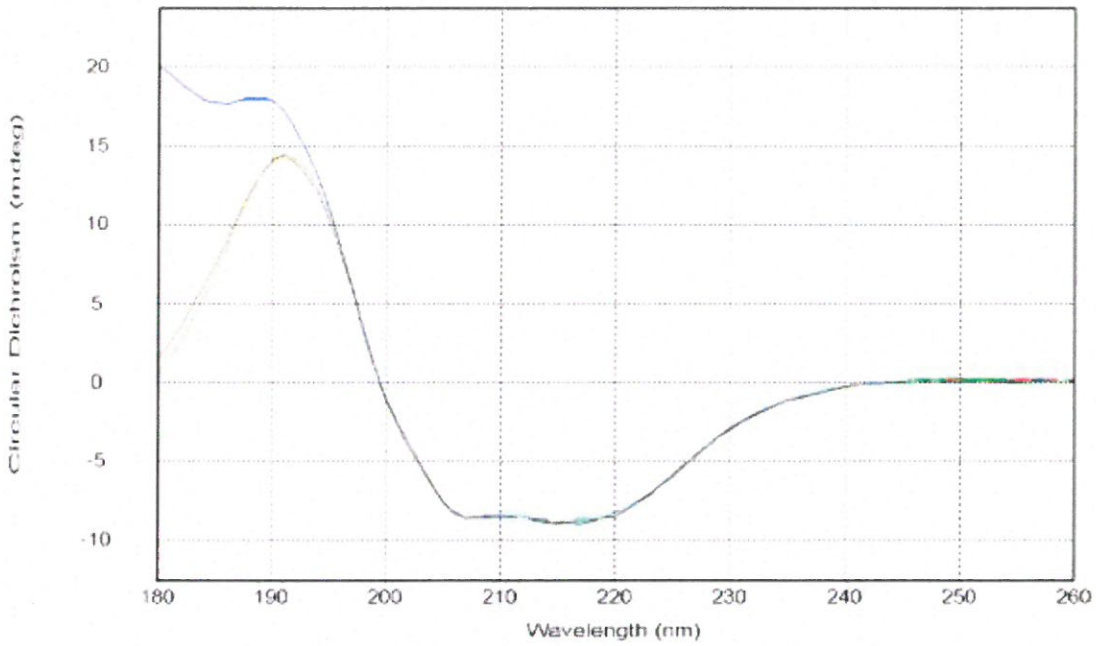


Figure 11: Typical spectra associated with secondary structural features in proteins

Since secondary structures in proteins are subject to denaturation upon exposure to physical or chemical stresses, CD analysis also offers a convenient method to determine thermal stability and formulation stability (including pH) of biopharmaceutical products. For instance, CD is frequently used to explore solvent conditions that increase the melting temperature (and reversibility of thermal unfolding) of proteins, which greatly facilitates development of formulations that prolong product shelf life. CD analysis has very minimal sample requirements (typically in the order of micrograms), and analysis can generally be completed in a matter of hours. The technique measures the difference in adsorption of left-handed and right-handed circularly polarised light by the asymmetric centres of chiral molecules. Ordered secondary structures within biomolecules result in a CD spectrum which can contain both positive and negative signals. Where no ordered structure is present, CD analysis results in a 'zero-intensity' signal.

A limitation of the CD technique is that it does not provide the residue-specific information that can be obtained with high-field NMR or X-ray crystallography analysis – the CD signal obtained for a protein sample represents the average signal for the entire population of molecular chiral centres. However, this

information is complimentary to information from other spectroscopic techniques, enabling detection of differences in a protein that may not be revealed through other spectroscopic analyses – see Figure 12, showing experimental data for three asparaginase preparations, highlighting a difference detected in the 180-200 nm region of the spectra. Therefore, CD analysis can provide information on the proportions of secondary structural elements (for instance, 50% α-helix), but it cannot reveal which specific residues are involved in those structures. CD in the far UV region may also be sensitive to certain elements of tertiary structure,[39, 40] providing information on whether denaturation occurs in a single step (i.e. with simultaneous loss of both tertiary and secondary structure), or if it occurs in a two-step process.



Analysis performed by Applied Photophysics, Leatherhead, UK on the Chirascan-plus qCD Spectrometer with 'Global 3 analysis' software

Figure 12: Normalised CD spectra of three assumed-identical preparations of asparaginase (Native, Recombinant, and Lyophilised).

Fourier transform infrared spectroscopy (FT-IR) is becoming increasingly recognised as a valuable tool for investigating protein structure. It can provide a wealth of information on folding, unfolding and misfolding of proteins. Information provided is complimentary to other methods of higher order structure analysis – e.g. some molecular features that produce a weak signal in CD analysis may produce a much stronger signal using FT-IR. Practically all biological molecules absorb infrared light, and the wavelength and magnitude of infrared light absorption by proteins produces characteristic spectra. The repeating units present in polypeptides and proteins give rise to nine highly-conserved IR absorption bands; amide bands A and B, and seven bands denoted by Roman numerals (I–VII). The amide I and II bands are the two most important vibrational bands in protein IR spectra.[41, 42] Other amide bands result from molecular force fields, the side chains present, and hydrogen bonding; these are often very complex and are of little practical use in the protein conformational studies.[43] FT-IR can be used to analyse proteins in either solution or solid-state, and is not significantly impacted by other sample components, such as buffer salts. Sample preparation, analysis and data interpretation can be completed relatively quickly (from our experience in as little as 15 minutes), and the presence of changes in protein secondary structure can be readily identified from IR spectra.

Pioneering work by Ernst and Wüthrich[44] led to NMR spectroscopy becoming one of only two techniques currently available that can provide structural data with resolution to the level of single atoms (X-ray crystallography also offers this resolution, but this technique is not dealt with in this article). Analysis involves sending radio signals across a range of frequencies through a sample contained in a powerful magnetic field, and measuring the absorption of the different frequencies by protons and isotopically-labelled atoms in the molecule. Absorption of these radio frequencies by atomic nuclei depends upon the local molecular environment, and on how atoms are covalently linked, arranged and move with respect to one another in three-dimensional space. Absorption signals may be perturbed by the presence of neighbouring nuclei, and these perturbations can provide an estimate of the distance between adjacent nuclei. These distances can be reconstructed to determine overall protein structure. Limitations of NMR include that proteins can only be analysed when in highly purified solution (and must not form aggregates when in solution); sample quantity requirements are much higher than for other spectroscopic techniques, and the upper practical limit of protein size is in the order of 50 kDa due to problems arising from overlapping spectra for larger proteins. Also, data collection times can often extend to days, and data processing and interpretation can be a very cumbersome process. More recent developments in NMR analysis of proteins detail solid-state analysis, where proteins are not required to be in solution for analysis,[45, 46] and adaptations of the technique to make it applicable to larger proteins[47] – however, these adaptations still require significant development, with multidimensional tools to the forefront in facilitating signal assignment.

A rationale for biosimilar characterisation

Conventional, small-molecule drugs (such as ibuprofen) depend entirely on easily-regulated chemical manufacturing methods, and production processes for these drugs readily delivers a consistent quality (identity, potency, purity and physical characteristics) of drug between production batches. Upon patent expiry of chemically-synthesised drugs, generics manufacturers can readily replicate production of the active ingredient, and can therefore produce a drug product that is identical to the innovator product. However, biopharmaceuticals are mostly protein in nature, and exhibit molecular complexity several orders of magnitude greater than that for small-molecule drugs. Production of biopharmaceuticals depends upon biological systems (e.g. animal cell cultures), where interplay between metabolomic processes and the environment of the producing cells can result in a high degree of variability in the finished product. This is particularly important when post-translational modifications (which some production cells may perform while others do not) are considered. Therefore, 'copy-versions' of biopharmaceuticals cannot be considered generic, as they are very unlikely to be identical to the innovator, and the more appropriate term 'biosimilars' has been widely adopted to describe this class of drugs. Since minor differences in structure may produce a biosimilar that is not comparable to the innovator in terms of its safety or efficacy profile, it is important that they are thoroughly structurally characterised and compared to the reference product prior to being granted authorisation. Licensing of biosimilars depends upon having technology available that allows for all the characteristics of a biologic drug to be closely examined. The technologies used should be sensitive enough to be able to detect even minute changes in structure between the biosimilar and the reference product. Therein lies a technical challenge in the biosimilar drug development sector.

Due to the enormous development and production costs for innovator biologics, they are often prohibitively priced, with some treatments costing tens of thousands of US dollars per patient per annum. This puts enormous financial burden on patients, national healthcare systems, and insurance providers. When a sufficient degree of similarity has been demonstrated using analytical characterisation strategies, biosimilars can take advantage of abbreviated licensure pathways, which have reduced dependence on costly and time-

consuming clinical studies. This results in biosimilars typically arriving on the market faster than innovators, and also with substantially reduced development costs. These cost savings can be passed on to the end users, and therefore reduce the cost burden on patients and payers. However, biosimilars still have comparatively large production costs when compared with generic small-molecule drugs, so the cost of a biosimilar still commands typically 80% that of the innovator. Nonetheless, this cost reduction still increases availability of these ground-breaking treatment to a wider patient group.

With patents for a large number of innovator biologics with combined global annual revenue of over US$50 billion set to expire in the period up to 2020, development and production of biosimilars is fast gaining ground. While licensing legislation for biosimilars has been slow to develop in the Unites States, two biosimilars have recently been approved there by the FDA, Sandoz's Zarxio® (filgrastim) and Hospira's Inflectra® (infliximab-dyyb), and these approvals are likely to be followed by many more. This will see the biosimilars market expand enormously over the coming decade, as the US is the primary market for biological drugs. These facts have resulted in a large increase in demand for specialist testing services (and experts) such as mass spectrometry and spectroscopic profiling. This has potential to release significant business opportunities for contract research organisations, such as BioClin Research Laboratories, as the biosimilars market becomes increasingly crowded.

Biotherapeutics compared to conventional drugs

Production methods for biologics compared to conventional drugs

The vast majority of biotherapeutics in common use today are proteins, which are extracted from living cells following growth of those cells in large-scale bioreactors. Many are produced through recombinant DNA technology, involving the insertion of a gene which encodes the protein of interest into the genome of the target cell. When these target cells are cultured under strictly controlled conditions in a bioreactor, they produce the protein encoded by the inserted gene, and this can then be extracted and purified using a range of downstream processing steps. Production of a biopharmaceutical is a highly dynamic process, with parameters such as temperature, pH, and dissolved oxygen continually changing under the influence of metabolising cells, and these parameters need to be monitored and controlled during production. For example, the application or removal of heat is used to control temperature, and the addition of acids or bases in order to maintain the pH within pre-defined limits. There is also the requirement that the specifications of the growth media components and other raw materials used in producing biopharmaceuticals are tightly controlled. A failure to adequately control can impact cell cultures even more than the fermentation process.[48] Since the production of recombinant proteins depends on biological substrates and biological processes, even the slightest alteration in production process parameters can lead to changes in the final product that can affect the identity, safety and/or efficacy profile of the drug. Biopharmaceuticals are almost exclusively intended for parenteral administration which means that they must be sterile. However, being protein in nature, they are often highly sensitive to extremes of temperature, pH and other harsh conditions commonly used to control or eliminate microbial contamination. Therefore, manufacture of biologics must be done under custom-designed 'clean-room' conditions, and using sterile equipment and processes throughout production.

With conventional drugs, production depends on purely chemical means, whereby drugs are synthesised in large reaction vessels, typically through a number of intermediates, using chemicals as raw materials. For example, the production of acetylsalicylic acid (the active ingredient in aspirin) involves esterification of

salicylic acid by acetic anhydride; illustrated in Figure 13, below. The identity of the drug produced through such processes is dependent upon of stringent control over the quality of starting materials, detailed understanding of the stoichiometry involved, and characterisation of any possible side reactions. Process parameters such as temperature and pH are usually easily regulated, and are not subject to unpredictable or erratic drift, such as may be observed in a bioreactor during biopharmaceutical production. Once a process for production of a conventional drug has been appropriately defined and regulated, production of a consistent quality of drug is readily achieved. Indeed, any manufacturer with details of the production process, and the specifications of the starting materials, can readily manufacture the same quality of drug at a site remote from the original manufacturers' site. Clean-room conditions are not usually required, as most chemically-derived drugs are not designed for parenteral administration, or can be sterilised post-filling using heat or other treatments if necessary, as the active ingredient and other constituents in the final formulation are typically stable during such treatments.

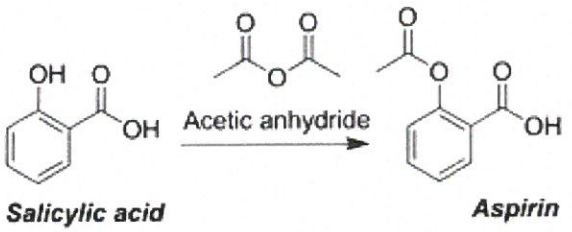


Figure 13: Synthesis of acetylsalicylic acid

Biopharmaceutical complexity

Biopharmaceuticals are much more complex than conventional drugs in many ways, the most obvious of which is the number of atoms of which they are composed and relative molecular mass, as illustrated between aspirin and a monoclonal antibody (a framework on which many biopharmaceuticals are based) in Figure 14, below. It is this complexity that makes biopharmaceuticals extremely difficult to characterise fully, when compared to the same task for small-molecule drugs. The following sections give a brief description of the structural levels of proteins.

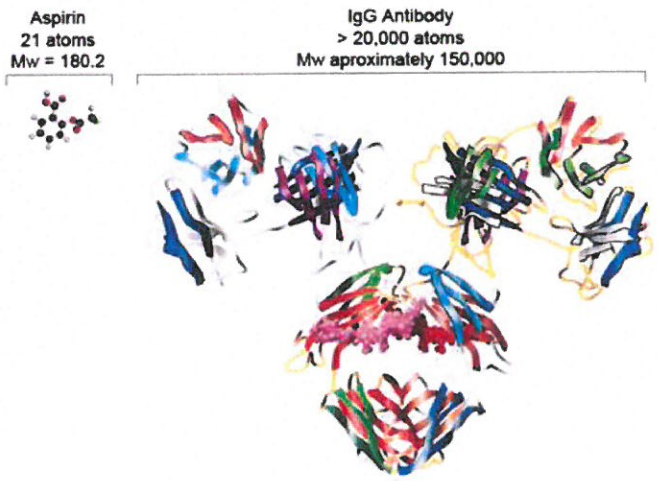


Figure 14: Comparison of size and complexity between conventional drugs and biologics. Source: http://www.amgenbiosimilars.com/the-basics/the-power-of-biologics/

Amino acids – the building blocks of all proteins

All amino acids (with the exception of proline) are primary amines, and all possess an asymmetric carbon (with the exception of glycine) and are therefore chiral. Of the naturally occurring amino acids, only L-amino acid enantiomers are naturally incorporated into proteins. All amino acids possess both a carboxylic acid group and an amine group, both of which are ionised at neutral pH in an aqueous environments (zwitterionic) with the carboxyl group having a net negative charge (COO^-), and the amino group having a net positive charge (NH_3^+) at physiological pH. This ionisation state varies as the pH of the solvent environment varies – in an acidic environment, ionisation of the carboxyl end is suppressed; in an alkaline environment, ionisation of the amino group is suppressed. Figure 15 illustrates the change in ionisation state of amino acids as pH changes in an aqueous environment.

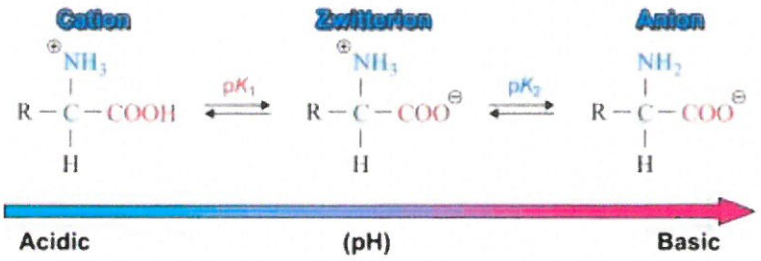
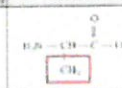
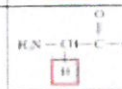
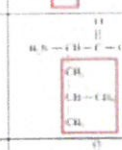
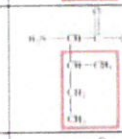
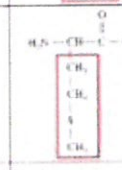
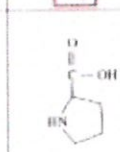
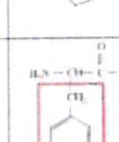
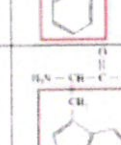
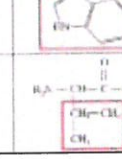


Figure 15: Ionisation state of amino acids in aqueous environments

Each amino acid is characterised by a variable side-group (or 'R'-group) which imparts a range of characteristics including size, charge, water solubility, and reactivity. Tables 4, 5 and 6 below group the amino acids into various categories, based on similar properties of their side groups, and also provides further information on important characteristics for each amino acid.

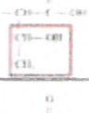
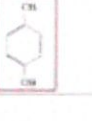
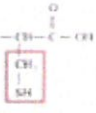
Table 4: Properties of amino acids containing non-polar side chains

Amino acids with non-polar side chains							
Amino Acid	Abbreviation (3-letter, 1-letter)	Structure	Molecular mass / Da	pKa1 (α-carboxyl)	pKa2 (α-amino)	pKa3 (side chain)	Properties
Alanine	Ala, A		71.1	2.35	9.87	N/A	(2S)-2-aminopropanoic acid: one of the simplest amino acids. Its side chain is very non-reactive, and so it is rarely directly involved in protein function. This non-reactive nature also means that the side chain of alanine can be found both on the inside and outside of a protein molecule.
Glycine	Gly, G		57.1	2.35	9.78	N/A	2-aminoacetic acid: the only achiral amino acid and also the smallest and most flexible. Within proteins, it often acts as a structure breaker because of the entropy required to restrain its flexibility. Glycine can play a distinct functional role, such as using its side chain-less backbone to bind to phosphates.
Leucine	Leu, L		113.2	2.33	9.74	N/A	(2S)-2-amino-4-methylpentanoic acid: the most common amino acid. It is hydrophobic due to the presence of the isobutyl side chain.
Isoleucine	Ile, I		113.2	2.32	9.76	N/A	(2S,3S)-2-amino-3-methylpentanoic acid: an isomer of leucine which differs from valine by just one methyl group.
Methionine	Met, M		131.2	2.13	9.28	N/A	(2S)-2-amino-4-methylsulfanylbutanoic acid: one of two sulfur containing amino acids, meaning it can interact with atoms such as metals. However, as the sulfur in methionine is bound to a methyl group, it is far less reactive than the sulfur in cysteine (which is bound to a hydrogen atom). Primarily found in the hydrophobic cores of globular proteins.
Proline	Pro, P		97.1	1.95	10.64	N/A	(2S)-pyrrolidine-2-carboxylic acid: the only amino acid where the side chain is connected to the protein backbone twice, forming a five-membered nitrogen-containing ring, meaning it is more correctly referred to as an imino acid. This structure means that it cannot adopt many of the main chain conformations. Therefore, it can often be found in very tight turns in protein structures such as where the polypeptide chain must change direction. Because of this, it is commonly found on the outer surfaces of proteins, despite being both aliphatic and hydrophobic.
Phenylalanine	Phe, F		147.2	2.20	9.31	N/A	(2S)-2-amino-3-phenylpropanoic acid: one of the three aromatic amino acids along with tyrosine and tryptophan. It absorbs UV light, all though not as well as the latter two. Phenylalanine is converted to a different amino acid, tyrosine by hydroxylation, and is a precursor for the biosynthesis of dopamine and norepinephrine. Its side chain is fairly unreactive and so it is rarely involved directly in protein function, however, it is frequently involved in protein binding or substrate recognition.
Tryptophan	Trp, W		186.2	2.46	9.41	N/A	(2S)-2-amino-3-(1H-indol-3-yl)propanoic acid: the largest and least abundant amino acid. It is a precursor of serotonin, through the precursor 5-hydroxy-tryptophan. Along with the other aromatic amino acids, it is frequently involved in 'stacking' interactions within the cores of protein structures.
Valine	Val, V		99.1	2.29	9.74	N/A	(2S)-2-amino-3-methylbutanoic acid: found in many proteins, primarily in the interior of globular proteins helping to determine three-dimensional structure.

Note: molecular masses in column 4 refers to the residue mass when the amino acid is incorporated into a protein (or peptide)

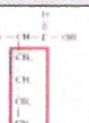
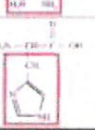
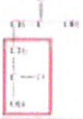
Amino acids are linked together by peptide bonds (peptide bond formation is described below). Two amino acid residues linked by a peptide bond is called a 'dipeptide', while three amino acid residues linked by peptide bonds is called a 'tripeptide'. When many amino acid residues are linked together by peptide bonds, the structure is called a polypeptide, and when a polypeptide is sufficiently long to exhibit higher levels of structure (detailed below), it is referred to as a protein. The possible diversity of proteins is enormous – for example, a simple octapaptide (peptide consisting of eight amino acid residues), there are over 2.5 billion (20^8) possible arrangements. Proteins are typically several hundred amino acid residues long with molecular weights of several tens of kilodaltons (k*Da*), and often, much greater. The correct functioning of a protein is dependent upon the polypeptide chain folding into the correct three-dimensional shape. This is dependent upon four levels of structure – denoted 'primary', 'secondary', 'tertiary', and 'quaternary'.

Table 5: Properties of amino acids containing uncharged polar side chains

Amino acids with uncharged polar side chains							
Amino Acid	Abbreviation (3-letter, 1-letter)	Structure	Molecular mass - Da	pKa1 (α-carboxyl)	pKa2 (α-amino)	pKa3 (side chain)	Properties
Serine	Ser S		87.1	2.19	9.21	N/A	(2S)-2-amino-3-hydroxypropanoic acid: can be synthesised from glycine or threonine, and is the only amino acid with a primary hydroxyl group. Acts as a nucleophile in the active site of many proteins including serine proteases such as trypsin and chymotrypsin. Can be a hydrogen bond donor or acceptor. Can undergo phosphorylation, and is the most commonly phosphorylated amino acid.
Threonine	Thr T		101.1	2.09	9.10	N/A	(2S,3R)-2-amino-3-hydroxybutanoic acid: the only amino acid with a secondary hydroxyl group. Can be phosphorylated, though less frequently than Serine. Threonine is common in protein functional centres. The hydroxyl group is fairly reactive, and can readily form hydrogen bonds with many polar substrates.
Asparagine	Asn N		114.1	2.14	8.72	N/A	(2S)-2,4-diamino-4-oxobutanoic acid: biosynthesised from aspartic acid and ammonia by the enzyme asparagine synthetase. Typically found interacting with the aqueous environment on the surface of proteins owing to its polar nature. Very frequently involved in binding sites as the polar side chain can interact with other polar groups on substrates. Can become hydrolysed to form aspartic acid. When identity is uncertain, it is often abbreviated to ASX to represent 'either aspartic acid or asparagine'
Glutamine	Gln Q		128.1	2.17	9.13	N/A	(2S)-2,5-diamino-5-oxopentanoic acid: differing by only a methyl group, it has similar characteristics and functions to those of asparagine. Can become hydrolysed to form glutamic acid. When identity is uncertain, it is often abbreviated to GLX to represent 'either glutamic acid or glutamine'
Tyrosine	Tyr Y		163.2	2.20	9.21	10.46	(2S)-2-amino-3-(4-hydroxyphenyl)propanoic acid: synthesized from phenylalanine and is the precursor of epinephrine, thyroid hormones, and melanin. Typically found in protein hydrophobic cores. The aromatic side chain means that it is often involved in stacking interactions with other aromatic side-chains. Contains a reactive hydroxyl group, thus making it likely to be involved in interactions with non-protein ligands.
Cysteine	Cys C		103.1	1.92	10.70	8.37	(2R)-2-amino-3-sulfanylpropanoic acid: the second of two common sulfur-containing amino acids. The role of cysteine in proteins depends on the cellular location of that protein. Within extracellular proteins, pairs of cysteines become oxidised to form disulfide bonds, which covalently link different parts of the polypeptide chain serving to stabilise protein structure. The reducing environment within cells makes the formation of disulfide bonds extremely unlikely, however, cysteines can still have an important structural function in this environment as the sulfhydryl side chain can effectively bind metals, which can be very important for enzyme functions. Cysteines are therefore prevalent in protein active and binding sites.

Note: molecular masses in column 4 refers to the residue mass when the amino acid is incorporated into a protein (or peptide)

Table 6: Properties of amino acids containing charged polar side groups

Amino acids with charged polar side chains							
Amino Acid	Abbreviation (3-letter, 1-letter)	Structure	Molecular mass - Da	pKa1 (α-carboxyl)	pKa2 (α-amino)	pKa3 (side chain)	Properties
Lysine	Lys K		128.2	2.16	9.06	10.54	(2S)-2,6-diaminohexanoic acid: most often found on the surface of proteins, lysines frequently play a part in determining protein structure - involved in formation of salt-bridges, where they pair with a negatively charged amino acid to create stabilising hydrogen bonds, that can be important for protein stability. Quite often found in protein binding sites.
Arginine	Arg R		156.2	1.82	8.99	12.48	(2S)-2-amino-5-[diaminomethylideneamino]pentanoic acid: the most basic amino acid, its positive charge is extensively delocalised meaning it can donate several H-bonds. It is frequently involved in determining protein structure, being frequently involved in forming salt bridges. The positive charge means it can interact with negatively charged non-protein atoms making it frequently found in protein active and binding sites.
Histidine	His H		137.1	1.80	9.33	6.04	(2S)-2-amino-3-(1H-imidazol-5-yl)propanoic acid: the only amino acid whose pKa is in the physiological range, meaning it is relatively easy to move protons on and off the side chain. This means that it can be found in both the core and surface of proteins and is the most common amino acid involved in protein active centres.
Aspartic Acid	Asp D		115.1	1.99	9.90	3.90	(2S)-2-aminobutanedioic acid: generally found on the surface of proteins exposed to an aqueous environment, although are often also found in hydrophobic cores involved in salt bridges where they pair with positively charged amino acids, stabilising protein structure. Often involved in active centres where they can bind positively charged non-protein atoms, such as zinc.
Glutamic Acid	Glu E		129.1	2.10	9.47	4.07	(2S)-2-aminopentanedioic acid: similar characteristics and functions to those of aspartic acid above – indeed, these two amino acids can often be found substituted for one another in proteins with no measurable impact on protein function or specificity.

Note: molecular masses in column 4 refers to the residue mass when the amino acid is incorporated into a protein (or peptide)

Continues….. p67

Protein primary structure

The primary structure of a protein refers to the linear sequence of amino acid residues in the polypeptide chain(s) (which is determined by the encoding gene sequence). The positions of cysteine residues and disulphide bonds, which are covalent bonds between cysteine residues in the polypeptide chain, are also considered as part of primary structure. Amino acids form unbranched polymers (polypeptides) through nucleophilic attack of the electrophilic carbonyl group at the carboxyl end of a polypeptide, by the amino group of the amino acid being added to that polypeptide. For this to take place, the carboxyl group must first be activated by adenosine triphosphate (ATP) to provide a better leaving group than the hydroxyl. The link formed in this reaction is called a peptide bond – illustrated in Figure 16. A single water molecule is liberated in the formation of each peptide bond, which means that amino acids once incorporated into a growing polypeptide, are more correctly referred to as amino acid residues. The primary structure of a protein is always written as the sequence of amino acid residues from the amino terminal end (*N*-terminus) to the carboxy terminal end (*C*-terminus). Post-translational modifications (PTM's) of a protein, such as phosphorylation and glycosylation, are also considered to be part of the protein primary structure, however, information on these cannot be derived from the encoding gene.

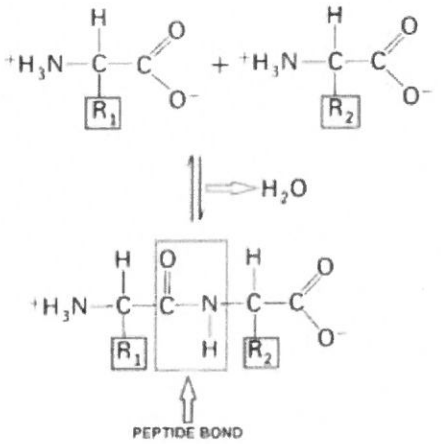


Figure 16: Peptide bond formation

Protein secondary structure

Secondary structure arises when the variable groups of amino acid residues interact locally with each other (and with their immediate environment) through hydrogen bonds and other non-covalent interactions. These interactions give rise to highly-stable structural motifs, the most prevalent of which include the alpha helix (α-helix) and the beta sheet (β-sheet). The α-helix forms between residues that are within close proximity to each other in the polypeptide chain, while β-sheets are formed between residues that are distant from each other. Figure 17 illustrates these structural elements.

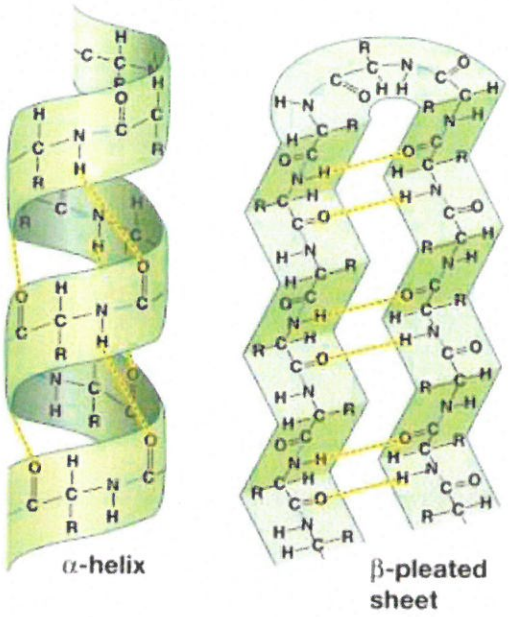


Figure 17: Common secondary structural elements of proteins

The α-helix, discovered in 1953,[49] is a 'right-handed' helix. The internal part of the helix is composed of the polypeptide backbone, with the R-groups of the amino acid residues projecting outwards on the helix. The structure is stabilised by hydrogen bonds between the carbonyl group of each amino acid residue with the amine group of the amino acid residue four resides away in the polypeptide chain. The length of α-helices in proteins can range from a few to several tens of residues, and the presence and number of α-helices in proteins can vary considerably. For example, globular proteins (which function as membrane receptors) tend to have greater α-helix content than other proteins.

The β-sheet structure which was elucidated in 1951,[50] is composed of two β-strands of polypeptide (either intermolecular or intramolecular). These strands are approximately five to ten residues long, and are associated with each other through hydrogen bonding between the carbonyl groups of one β-strand, and the amine groups of the adjacent β-strand. The alternate α-carbons between adjacent amino acid residues lie above and below the plane of the sheet, giving the structure a pleated configuration. The strands may both be aligned with their *N*-termini at the same end, in which case they are called parallel β-pleated sheets, or with the *N*-termini at opposite ends, in which case they are called antiparallel β-pleated sheets.

A number of well-ordered three-dimensional structural motifs are also commonly found in proteins, and these are collectively referred to as 'supersecondary structure'. These structural motifs facilitate correct folding of the protein. Examples of these include the 'helix-loop-helix' motif, which is composed of two α-helices joined by a loop; and the 'hairpin β-sheet motif', which contains two antiparallel β-sheets joined by a loop. Figure 18 illustrates some of the more commonly occurring supersecondary structural motifs.

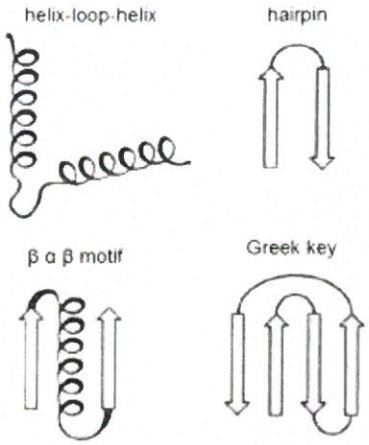


Figure 18: Protein supersecondary structural motifs

Torsion angles and the Ramachandran plot

The torsion angles in a polypeptide describe the rotation of the polypeptide backbone around two bonds – the bond between the α-carbon and nitrogen (called the phi, or φ), and the bond between the α-carbon and the carbonyl carbon (the psi, or ψ) – refer to Figure 19. These torsion angles are very important local structural parameters that control protein folding, because certain bond angles are restricted, as they would result in steric hindrance between atoms.

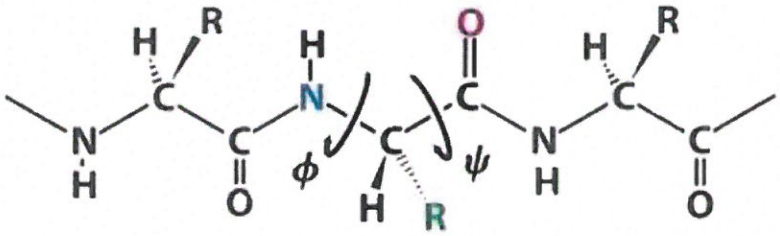


Figure 19: Torsion angles in polypeptide chains

The Ramachandran plot (illustrated in Figure 20), which was developed in 1963 by Ramachandran *et al.*,[51] is a way to visualise the distribution of all possible torsional angles in protein structure. It plots φ angles on the x-axis and the ψ angles on the y-axis which provides an overview of allowed and disallowed torsional angles. Due to steric hindrance, the allowed torsional angles are constrained within specific areas of the plot, particularly for secondary structures such as the α-helix or the β-sheet. In practical terms, the Ramachandran plot is a reliable method for predicting protein structure, but some proteins may include angles in the disfavoured regions – where this occurs, additional interactions will be present that help to stabilise the structure.[52] Ramachandran plots are often used to validate results obtained from structural analysis via X-ray crystallography or NMR analysis – problems with the experimentally-derived structures will be revealed where a large number of torsion angles are found in the disfavoured regions of the plot.

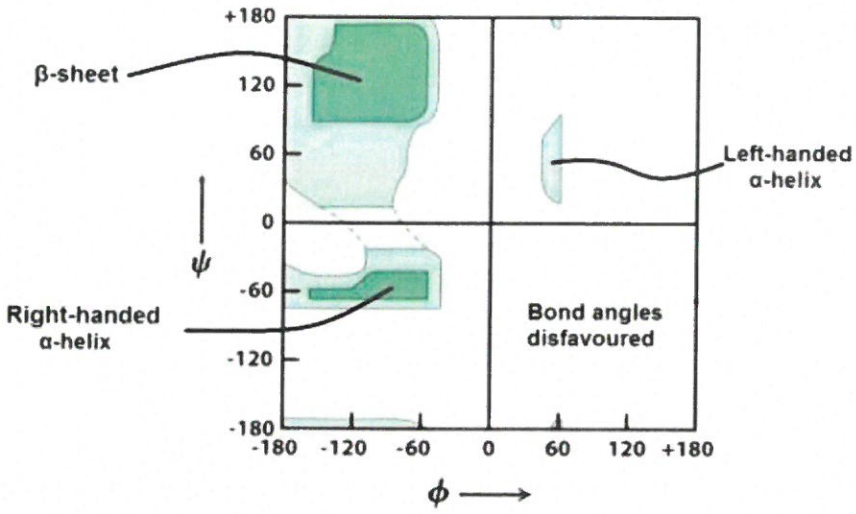


Figure 20: Ramachandran plot: dark green: low-energy regions where torsion angles are highly-favoured; light green: allowed regions; white: highly disfavoured regions.

Protein tertiary structure

Tertiary structure refers to the overall spatial arrangement of the polypeptide chain following the development of the secondary and supersecondary structural elements, to produce the compact globular shape of the protein. This conformation is determined by the combination of secondary structures to form protein 'domains'. It is generally accepted that the tertiary structure of a protein is the most thermodynamically stable arrangement. The tertiary structure of proteins is absolutely critical to their function, and for this reason is a critical characteristic of therapeutic protein that needs to be thoroughly characterised. Globular proteins generally have tertiary structures with hydrophobic residues at the core of the molecule, and a surface with hydrophilic residues exposed. This arrangement helps to stabilise the protein while also increasing its water solubility. Figure 21 shows a 'ribbon diagram', also known as a "Richardson diagram", which is a 3-D representation of protein tertiary structure in common use today.

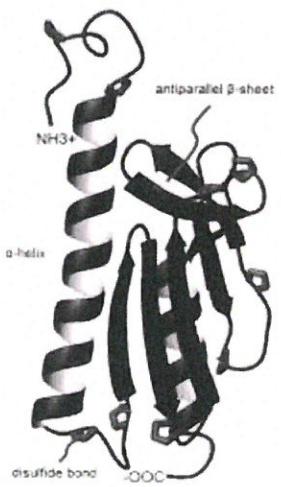


Figure 21: Richardson diagram illustrating protein tertiary structure

Protein quaternary structure

Quaternary structure refers to the covalent and/or noncovalent association of two or more protein subunits to form a functional protein. These subunits may have identical or different amino acid sequences and structures. Not all proteins exhibit quaternary structure (i.e. proteins composed of only one polypeptide chain). Monoclonal antibodies are examples of proteins that exhibit quaternary structure, being composed of four subunits; two identical 'heavy chains', and two identical 'light chains', which are linked together through a series of disulphide bonds and noncovalent interactions. In order to characterise proteins which exhibit quaternary structure, it is often necessary to first dissociate the subunits from one another to be characterised separately, then rebuilding the picture of the complete functional protein.

Post-translational modifications of proteins

Post-translational modifications (PTM's) of polypeptide chains can extend the functionality of proteins by covalently attaching chemical groups, or in some cases, cleaving chemical groups or signal peptides from the molecule. PTMs often play a critical role in the functionality of proteins,[53] and can also be important in regulating cellular functions – for example, many enzymes are activated through phosphorylation by kinases. PTMs are highly dependent on production cell lines. Prokaryotic organisms, such as *Escherichia coli*, (which has been used as an expression system for biopharmaceuticals for many years) do not significantly modify proteins following translation, however, eukaryotic cells often do modify proteins. For instance, recombinant asparaginase produced in *Escherichia coli* is non-glycosylated, whereas recombinant erythropoietin (used in treating anaemia) produced in Chinese hamster ovary (CHO) cells is heavily glycosylated, with carbohydrate accounting for up to 40% of the mass of the molecule.[54]

A large number of PTMs are commonly encountered including glycosylation, *S*-nitrosylation, methylation, *S*-palmitoylation, and many others. Glycosylation is acknowledged as one of the most significant PTMs, and can have an effect on protein secondary (and higher order) structure, function and stability.[53, 55, 56] Glycosylation involves the attachment of sugar moieties ranging from simple monosaccharides to highly complex branched polysaccharides. *S*-nitrosylation involves reaction of free cysteines with nitric oxide (NO) to form *S*-nitrothiols. This PTM has a major stabilising effect on proteins and also plays a part in regulating enzymes involved in gene expression.[57] Methylation involves the addition of methyl groups to nitrogen or oxygen (*N*- and *O*-methylation, respectively) or to amino acid R-groups, which increases the hydrophobicity of the protein, thereby enhancing cell membrane association.[58] *S*-palmitoylation attaches a C_{16} palmitoyl group to cysteine residues; this long hydrophobic group facilitates anchoring of the protein in the lipid membrane of cells.[59] Identifying, characterising and understanding the role that PTMs play in protein function is critical to the study of recombinant proteins as biopharmaceuticals.

References

1. Global Market For Biologics To Reach Nearly $252 Billion In 2017 [Internet] BCC Research. [Date accessed: 18/05/2016] Available from: http://www.bccresearch.com/pressroom/bio/global-market-biologics-reach-nearly-$252-billion-2017
2. Questions & Answers - Generic Drugs [Internet] US Food and Drug Administration 2015. [Date accessed: 08 Jun 2016] Available from: http://www.fda.gov/Drugs/ResourcesForYou/Consumers/QuestionsAnswers/ucm100100.htm
3. Committee for Medicinal Products for Human Use. "Guideline on similar biological medicinal products." London: European Medicines Agency (2005).
4. US Food and Drug Administration. "Guidance for industry: biosimilars: questions and answers regarding implementation of the Biologics Price Competition and Innovation Act of 2009." US Food and Drug Administration (2012).
5. US Food and Drug Administration. "Guidance for industry: quality considerations in demonstrating biosimilarity to a reference protein product." US Food and Drug Administration (2012).
6. US Food and Drug Administration. "Guidance for industry: scientific considerations in demonstrating biosimilarity to a reference product." US Food and Drug Administration (2012).
7. Tufts CSDD. Cost to develop and win marketing approval for a new drug is $2.6 Billion [Internet] Tufts Center for the Study of Drug Development 2014 [Date accessed: 15/11/2015] Available from: http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study.
8. International Conference for Harmonisation. Specifications: test procedures and acceptance criteria for biotechnological/biological products. 1999. CPMP/ICH/365/96.
9. Wolman Y, Miller S. Amino-acid contamination of aqueous hydrochloric acid. Nature. 1971;234(5331):548-549.
10. Wilkins M, Lindskog I, Gasteiger E, Bairoch A, Sanchez J, Hochstrasser D, *et al.* Detailed peptide characterization using PEPTIDEMASS - a World-Wide-Web-accessible tool. Electrophoresis. 1997;18(3-4):403-408
11. Berg J, Tymoczko J, Stryer L. Biochemistry. 5th Edition ed: W H Freeman; 2002.
12. Niall H. Automated Edman degradation: the protein sequenator. Methods in Enzymology. 1973;27:942-1010.
13. Stretton A. The first sequence. Fred Sanger and insulin. Genetics. 2002;162(2):527-532.
14. Brown J, Roberts W. Evidence that approximately eighty per cent of the soluble proteins from Ehrlich ascites cells are N^{α}-acetylated. Journal of Biological Chemistry. 1976;251(4):1009-1014.
15. Polevoda B, Sherman F. *N*-terminal acetyltransferases and sequence requirements for *N*-terminal acetylation of eukaryotic proteins. Journal of Molecular Biology. 2003;325(4):595-622.
16. Ecker D, Ransohoff T. Mammalian cell culture capacity for biopharmaceutical manufacturing. Advances in Biochemical Engineering / Biotechnology. 2014;139:185-225.
17. Bonthron D. L-asparaginase II of *Escherichia coli* K-12: cloning, mapping and sequencing of the *ansB* gene. Gene. 1990;91(1):101-105.
18. Hughes C, Ma B, Lajoie G. *De novo* sequencing methods in proteomics. Methods in Molecular Biology. 2010;604:105-121.
19. Coon J. Collisions or Electrons? Protein Sequence Analysis in the 21st Century. Analytical Chemistry. 2009;81(9):3208-3215.
20. Kaltashov I, Bobst C, Abzalimov R, Wang G, Baykal B, Wang S. Advances and challenges in analytical characterization of biotechnology products: mass spectrometry-based approaches to study properties and behavior of protein therapeutics. Biotechnology Advances. 2012;30(1):210-222.