Dealing with the challenges of sequence variants

Dr RL Easton

BioPharmaSpec Ltd, Suite 3.1, Lido Medical Centre, St. Saviour, Jersey JE2 7LA, UK
and BioPharmaSpec Inc, 363 Phoenixville Pike, Malvern, PA 19355, USA

Introduction

An area that poses a significant challenge within the structural characterization of biopharmaceutical products is the assessment of sequence variants. But what exactly are sequence variants? In this article we are not talking about post-translational modifications (PTMs) or variation in glycosylation, but sequence variants that arise as a result of incorporation of an erroneous amino acid residue into the protein backbone in place of the amino acid that should be present at that position. Before we consider sequence variants I think it would be worthwhile to look at why amino acid sequencing is important and how the amino acid sequence of a protein can be determined.

Amino acid sequencing of proteins

It is not enough to know the sequence of the DNA used to produce a protein. The amino acid sequence must be confirmed to demonstrate the recombinant protein has been produced as expected and is a true product of the DNA, rather than a variant (e.g. modified, truncated, extended) version of the product. This is enshrined in the analytical guidelines of ICH Q6B, as well as the more recent regulatory authority biosimilar guidelines. As part of a "biosimilarity" assessment, it is necessary to show that the sequence of biosimilar and innovator are identical. For products that are not marketed under the umbrella of biosimilars, the sequence must be defined as part of the requirements for structural characterization. Based on the guidelines and our experience, it is wise to perform amino acid sequencing at an early stage of product development, to ensure that the clone selected has produced the expected product (in terms of primary sequence) and to thus avoid unwanted and costly surprises at a later stage.

Amino acid sequencing of a protein is not a trivial undertaking and if full sequence is to be identified, more than one method must be used. Using the two main methods available greatly helps with the task, since one technique can provide information that may be ambiguous or lacking from the other procedure. The use of the two methods also serves a secondary role by providing "orthogonal" data, which is something the regulatory authorities are very keen to see these days, due to the increased strength orthogonality gives to the package of structural investigative work.

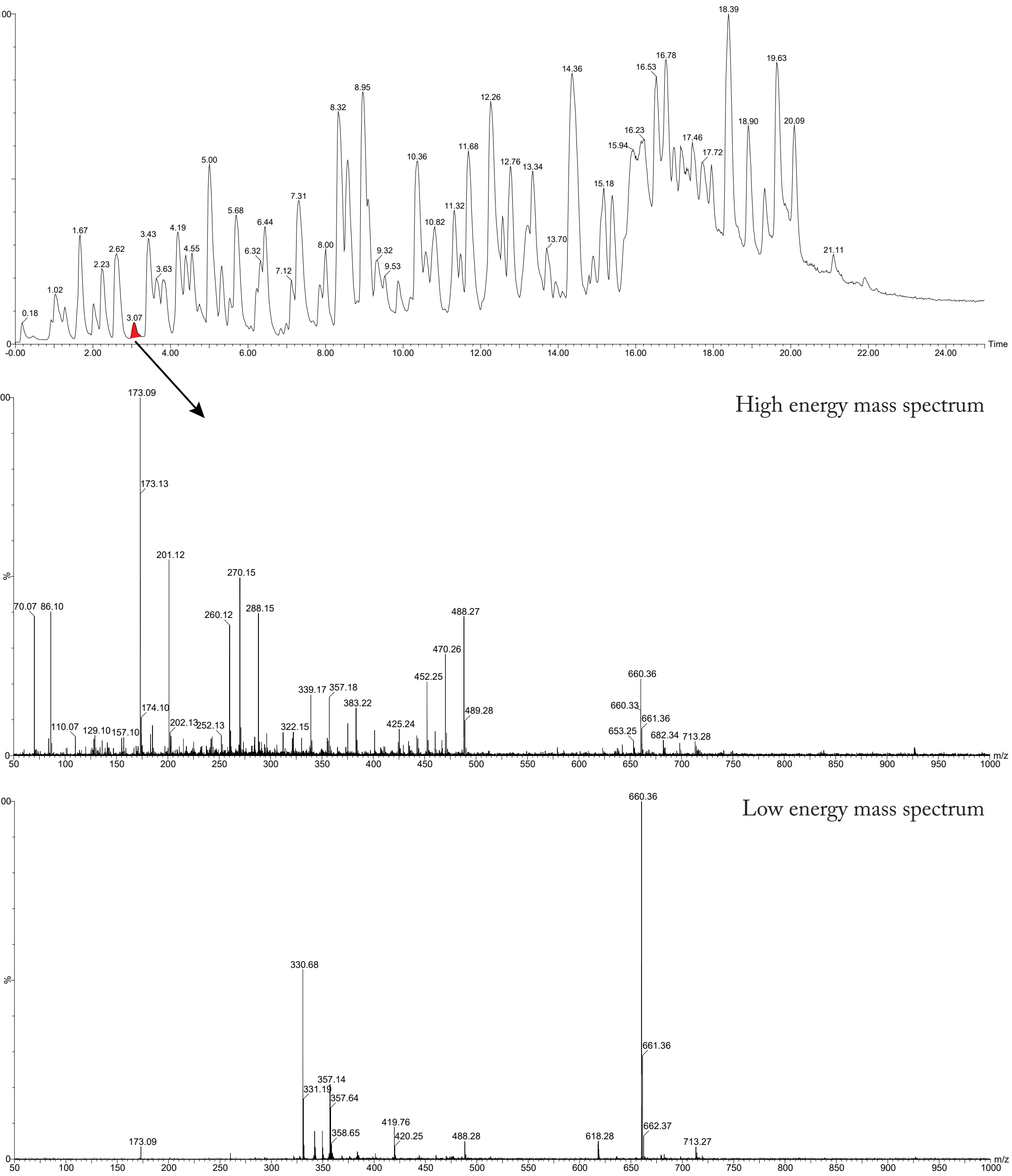
The main technique for amino acid sequencing is mass spectrometry and this is able to provide the vast majority of information regarding the primary amino acid sequence. The power of this technique lies in two key areas, both of which are necessary for success in any type of sequencing work. Firstly, mass spectrometers give accurate masses of peptides derived from a sequencing workflow. This is important for precise identification of the peptides in the first instance. Secondly, certain types of mass spectrometers such as the Q-TOF type instruments are capable of fragmenting peptides in real time and providing accurate mass information for the fragments (Figure 1). These types of instruments are most suitable for amino acid sequencing. Since the most readily produced fragment ions are generated as a result of cleavage across the amide backbone bonds in the peptide with both N- and C-terminal fragment ions produced, the masses of the fragment ions can be used to "read" the sequence of amino acids.

The production of overlapping peptides as a result of the use of different proteolytic enzymes in the methodology to produce the peptides results in overlapping sequences, which can then be "stitched" together to provide the full sequence (see Figure 2 for a description of this workflow). The use of mass spectrometry to determine the intact mass of the molecule (ideally in the absence of glycosylation for clarity of the protein mass component) will allow the endpoint to be determined, since the mass of the full sequence must total the mass of the intact protein.

It is very important to understand that peptide mapping alone is not the same as confirming peptide sequence. In other words, simply knowing the mass of a peptide is not the same as knowing the amino acid sequence. There may be mutations in a peptide that have not resulted in an overall mass change (e.g. a Tryptophan residue has the same mass as the dipeptides ValSer and GluGly) or indeed the identified peptide may be a different peptide altogether. So, peptide mapping must not be confused with amino acid sequencing by mass spectrometry.

Mass spectrometry is not suitable for identification of all amino acids and cannot be used as the sole means of sequence determination for this reason. The problem arises where amino acids have the same mass. Notably, this is the issue with Leucine and Isoleucine, which are isomers. Some types of mass spectrometer can use high energy fragmentation to produce fragment ions derived from the side chains of amino acids, which include unique mass fragment ions derived from the side chains of Leucine and Isoleucine. However, the ability to produce side chain fragment ions cannot be guaranteed (since other factors such as peptide sequence and size will affect the data produced) and thus assignment of Leucine or Isoleucine on this basis is often ambiguous. N-terminal sequencing procedures, using Edman chemistry, can be performed on collected peptides containing these amino acids. This will allow for categoric identification as a result of unique chromatographic run positions of the released derivatives. The combination of mass spectrometry and Edman chemistry has proven to be a very effective combination for amino acid sequencing and is capable of providing the full sequence of even large proteins such as monoclonal antibodies. Use of Edman chemistry alone (i.e. without mass spectrometry

Figure 1: A peptide map derived from an IgG. The upper image shows the total ion chromatogram and the lower two mass spectra show the identified C-terminal peptide (low energy mass spectrum) and the associated high energy mass spectrum shows the fragment ions derived from this peptide. It must be noted that whilst this is, in a relative sense, a high energy mass spectrum, the energy is not sufficiently high to create side chain fragmentation.



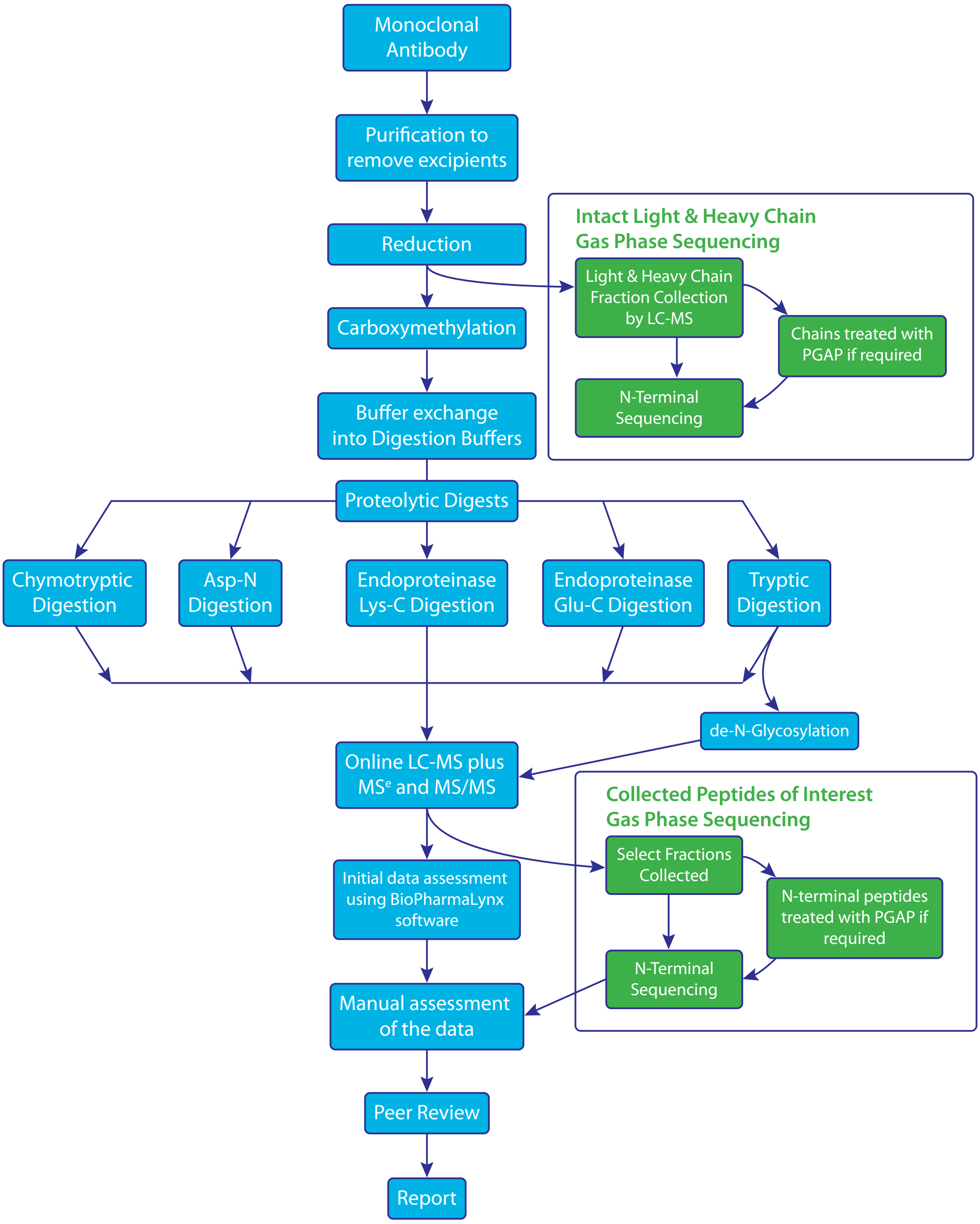
based sequencing) is not recommended since this would be a very time consuming and laborious process. Rather Edman chemistry can support a mass spectrometric sequencing strategy, where the mass spectrometer is unable to provide full sequence of a peptide or region or isomer ambiguity arises.

Sequence variant analysis

During the process of translation at the endoplasmic reticulum membrane, read errors can occur as a result of a mismatch between the mRNA and tRNA, resulting in the misincorporation of amino acids or amino acid derivatives (e.g. the mismatch of Guanine and Uracil on the mRNA codon and tRNA anticodon respectively will lead to a translation error and misincorporation at that residue position). This may be the result of the metabolic state or age of the cells within the culture medium or the characteristics of culture media (e.g. depletion of certain amino acids/amino acid metabolites) or may even be a result of the cellular metabolic activity itself.

As with amino acid sequencing, it is recommended that a screen for sequence variants is included in early phase assessments of product development and characterization. Early capture of sequence variants can then lead to clone or cell line reselection, if appropriate. Identifying sequence variants at an early

Figure 2: Generalized workflow showing the procedure for sequencing an IgG. There may be instances where other or alternate proteolytic procedures need to be used such as the use of different proteases but this will be dependent on the amino acid sequence itself.


Monoclonal Antibody
Purification to remove excipients
Reduction
Carboxymethylation
Buffer exchange into Digestion Buffers
Proteolytic Digests
Intact Light & Heavy Chain Gas Phase Sequencing
Light & Heavy Chain Fraction Collection by LC-MS
Chains treated with PGAP if required
N-Terminal Sequencing
Chymotryptic Digestion
Asp-N Digestion
Endoproteinase Lys-C Digestion
Endoproteinase Glu-C Digestion
Tryptic Digestion
de-N-Glycosylation
Online LC-MS plus MS^e and MS/MS
Collected Peptides of Interest Gas Phase Sequencing
Select Fractions Collected
N-terminal peptides treated with PGAP if required
N-Terminal Sequencing
Initial data assessment using BioPharmaLynx software
Manual assessment of the data
Peer Review
Report

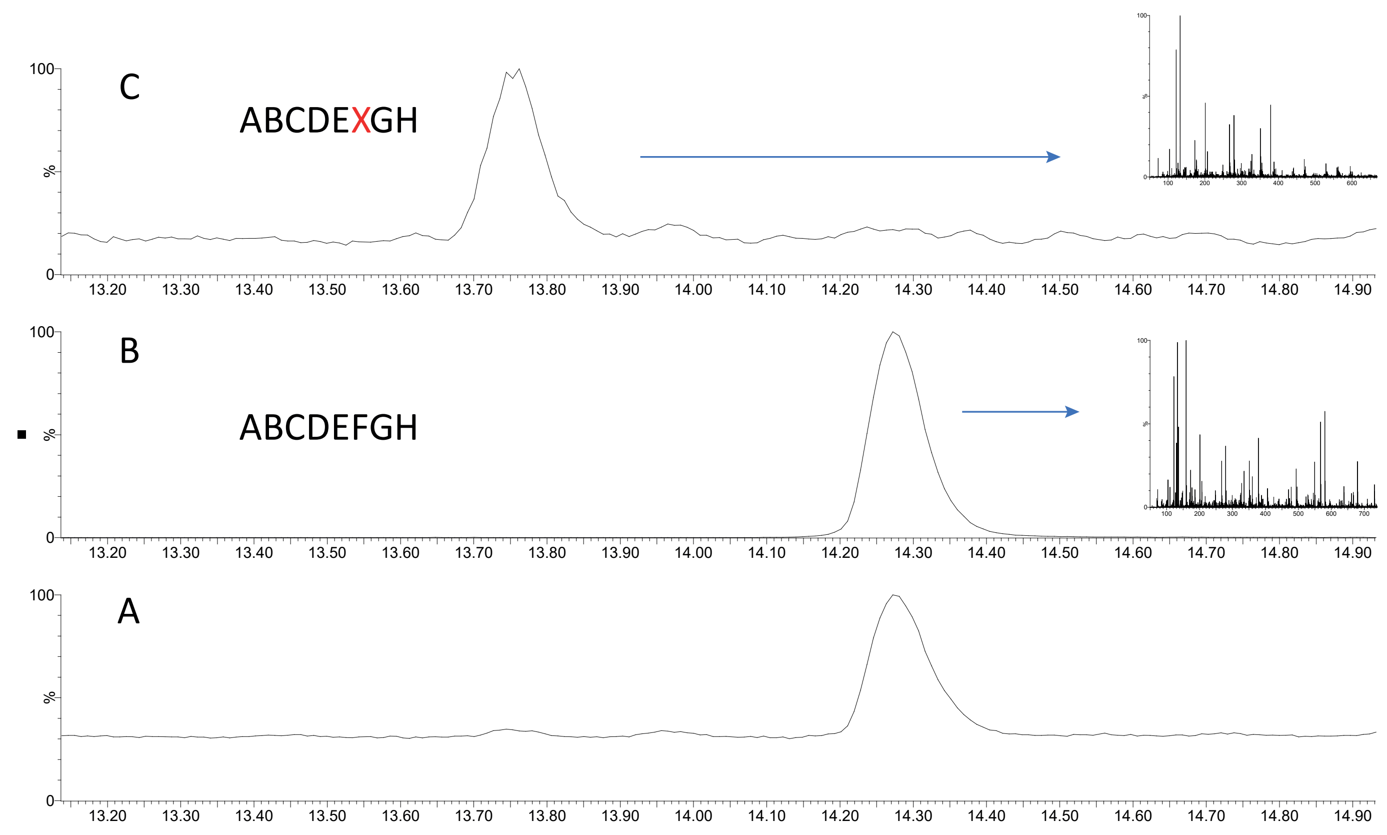


Figure 3: Schematic of sequence variant investigation from mass spectrometric data. The lower chromatogram (A) shows a small section of the total ion chromatogram obtained from the peptide mapping data. The central image (B) represents the extracted ion chromatogram produced from a search of the low energy chromatogram for one of the expected peptides from the digest with the inset associated high energy fragmentation data for this peak also shown. The upper chromatogram (C) shows the extracted ion chromatogram produced from a search of the low energy data for a sequence variant of the peptide (variant position marked in red) with the inset associated high energy data obtained from this variant peptide.

stage limits the financial and timeline impacts on product development. It may be that sequence variants (or a certain subset of them within predefined limits) are accepted as part of the make-up of the product and are shown to have no clinically meaningful effects.

Sequence variant analysis begins with mass spectrometric peptide mapping. The data is then screened for the presence of expected or predicted variants based on the predicted mass of the variant peptide. Sequence variants may be known through reports in the literature or expectations based on the production process itself, for example. Screening of data for variants can be performed in a targeted manner through specific mass searches if they are known, or are at least considered plausible in the protein sequence. Where there is no prior knowledge of sequence variants, the data can be assessed in an untargeted manner through the use of error tolerant data searches (described in more detail below) but this can produce a large volume of complex data requiring careful interpretation to avoid false positives.

Using mass spectrometers capable of generating real time fragment data may result in the presence of supportive fragment ions being detected for sequence variant peptides. This, of course, is not guaranteed since many variants that are present may be below the level where fragment ions can be detected. The predicted mass of the sequence variant peptide can be used to screen the peptide map data for the presence of the variant peptide. A screen should also be performed for the expected peptide and the elution position of the variant relative to the expected can be used to determine the validity of the signal (Figure 3).

Determination of the relative peak area of the variant to the expected peptide can be used as a measure of the relative abundance of that particular peptide. There are some points to bear in mind when performing analysis in this way. Firstly, the ionization efficiencies within the mass spectrometer may be different for the variant and native peptide, as a result of the presence of a different amino acid at a point within the sequence. Thus, the measured response can only be taken as a relative percentage of the native peptide response, rather than an absolute measure of amount. This amino acid variation could also lead to a difference in charge state distribution between variant and native peptide, depending on the charge propensity of the variant. Secondly, depending on the variant, the protein digest itself may have to be carefully considered if the variant is one which could affect the proteolytic cleavage site of the protein.

Analysis of MS/MS peptide mapping data using error tolerant searches can be performed, which is able to significantly assist in the investigation of sequence variants. During an error tolerant search of MS/MS data, relaxed data interrogation criteria and assessments of the data with sequential theoretical amino acid substitutions are performed *in silico* in an attempt to assign peptide signals that do not fit with the predicted amino acid sequence of the protein. Fragment ion patterns from the generated MS/MS data are compared to those which would be theoretically produced from substituted peptides to search for possible matches. In order for this to work,the data need to be sufficiently intense for MS/MS fragment ion signals to be detectable and clearly identifiable, which, as mentioned above, is not always the case with peptides containing sequence variants. Furthermore, some post translational modifications have the same mass difference as a sequence variant, which can lead to false positive assignments if care is not taken. For example, the mass of an oxidized Methionine residue is the same as a Phenylalanine residue and the 16Da increase in mass that oxidation provides is the same as the mass difference between several amino acids such as Alanine and Serine, Proline and Leucine/Isoleucine, Valine and Aspartic acid, Leucine/Isoleucine and Glutamic acid and also Phenylalanine and Tyrosine. Since the list of post translational modifications that can occur on proteins is extensive, careful screening of peptides needs to be performed and manual screening should be used as a means of confirming assignments if the data or peptide sequence give any cause for concern or ambiguity in the automatic assignment.

If there is a need to perform a more in-depth investigation into a particular variant or suspected variant, then the peptide peak of interest can be collected in sufficient levels for further mass spectrometric peptide sequencing using the procedure described above.

In any case, the assessment of relative mass spectrometric response outlined above can be used to compare product in batch to batch or biosimilar to innovator analysis program in a semi-quantitative sense.

Summary

Consideration should be given to performing sequence variant analysis during early stages of product development. This gives a greater understanding of the composition of the product and any product related impurities that may be present. It also allows time for any modification to the production process to eliminate or minimize variants of concern or that are above acceptable levels in the product. Mass spectrometric analysis provides a convenient and sensitive way of performing this type of investigation as part of a sequencing or peptide mapping investigation.

Dr. Richard Easton is Technical Director at BioPharmaSpec Ltd

Richard obtained his PhD in glycoprotein structural characterization using mass spectrometry from Imperial College of Science, Technology and Medicine. He subsequently spent several years there as a postdoctoral research scientist working in the field of glycoprotein structural characterization with emphasis on glycan elucidation. He moved to GlaxoSmithKline for a short time where he was head of mass spectrometry for the toxicoproteomics and safety assessment group. Richard joined M-Scan Limited (now part of SGS Life Sciences) as a biochemist and became the Team Leader for Carbohydrate Analysis before being appointed Principal Scientist. Richard joined BioPharmaSpec in 2016 as Technical Director for Structural Analysis and is responsible for management of all aspects of carbohydrate and glycoprotein characterization at the primary structure level.