Proteomics@Liverpool FAQ: Frequently asked questions

Can you perform sequencing de novo?

Yes.
High quality sequencing *de novo* * is best achieved using an instrument that generates high mass accuracy precursor (peptide) and product (peptide fragment) ions. THus, we would run such samples on the QExactive. Then, the datafile is best processed using Peaks 7 9BSI) which does an excellent job of reconstructing peptide sequences from the ion series.

However, even with a complete set of tryptic peptides, fully sequenced, it is still impossible to define the order of the peptides. This can be overcome in one of two ways. First, if there is a protein from a related species of known sequence, it may be possible to use homology matching to assemble the tryptic peptide sequences. Alternatively, if the same protein is digested with a different endopeptidase (such as GluC) we can generate a second set of peptide sequences that overlap with the tryptic series, and which can be used to reconstruct the protein sequence. One final point - some peptides will be too small to sequence, and it can also be helpful to digest a protein with endopeptidases LysC or ArgC.

Using these approaches, we have completely sequenced proteins up to about 200 amino acids. There is no reason why this cannot be extended to larger proteins.

And, one final caveat. Leu and Ile are isobaric, and cannot be discriminated by this approach.

* I believe that it is formally correct to say 'sequencing *de novo*' rather than '*de novo* sequencing'. This is akin to the common error

 © 2008-2014 Rob Beynon

of '*in vitro* studies' which should be 'studies *in vitro*'. Vlassical scholars might wish to enlighten me..

How can I have proteomics samples run?

We are always willing to talk to colleagues about the potential for running new analyses. These can vary from simple 'quick look' analyses to complete and complex, fully biologically replicated analyses. In all circumstances, we adopt a model of 'defend the mass spectrometer form the sample'! In fact, mass spectrometers are remarkably robust; it is the delivery of peptides through a nanoflow high pressure chromatography system that causes the problems.

Biological samples can be delivered in exotic and complex matrixes, either reflecting the biological context of the sample or the sample work up chemistry imposed by the user m(detergent, PEG and glycerol might seem like a dream extraction buffer to you, but we are never going to run that sample for you!). We are very reluctant to receive samples that contain insoluble material, high concentrations of detergents, polymers such a polyethylene glycol (PEG), for example. A nanoflow high resolution column (75um diameter and 150mm long) costs, with trap, nearly £1,000 and is time consuming to exchange and optimise. You can see why we're reluctant to take anonymous samples!

It is far, far better if you come to talk to us before you attempt to prepare proteomics samples.

Can you measure the mass of my protein?

In short, probably yes!

As of October 2011, we have set up a semiautomated system for the mass measurement of intact proteins. ThHis system will measure the mass of a protein to about 1Da in 10,000Da, and requires microgram quantities of protein. The mass is measured by electrospray ionisation mass spectrometry and thus, the protein molecular acquires a large and variable number of charges (protons). Each protein thus creates a multiply-charged envelope of ions that need to be deconvoluted by proprietary software using maximum entropy algorithms. The result is a true mass spectrum that can reveal the mass of the analyse protein and also, the mass of associated contaminants and possibly, fragments or modifications.

How much will it cost?

Impossible to answer until we have talked. However, we prefer to take responsibility for all aspects of the sample preparation (including the reduction, alkylation and digestion) as this greatly enhances the chances of useful data coming back. This takes a person's time, and we must recover that cost, as well as the cost of running the instrument, covering service charges, analysing the data and generating human-readable output. We're not cheap, but we're certainly not in microarray territory! The only person who can authorise these samples, and the attendant charges, is Rob Beynon. When we have a clear charging structure, this will be a lot simpler, but at this stage, we're only covering our costs.

What is the lowest limit of detection you can attain.

We would hope to be able to reach 100 attomol for the lowest abundance in a discovery

 © 2008-2014 Rob Beynon

experiment. To know whether that is enough, let's do some quick calculations. A detection limit of 100 attomol is equivalent to injecting 60 million molecules into the mass spectrometer - that sounds like a lot. If the sample was derived from yeast cells, we would have loaded of the order of 200,000 yeast cell equivalents onto the same column. THus, we can measure 60 million molecules, derived from 200,000 cells. From this, you will see that the limit of detection in an unprocessed sample is 60,000,000/200,000 = 300 copies per cell.

That sounds pretty good, right? But, if we have been using HeLa cells, the numbers are very different. The limit to what can be loaded on the hplc column is dictated by the total protein load - 200,000 cells gives us about 1000 nanograms of protein. However, each HeLa cell would contain 50 times as much protein as a yeast cell, approximately. Therefore, for a 1000 ng column capacity, we can only load about 4,000 HeLa cells equivalent onto the column. For the same detection limit of 100 attomol, we obtain a limit of detection of 60,000,000/4,000 copies per cell, or 15,000 copies per cell. Rather different!

To overcome such difficulties, it is necessary to resort to sample prefractionation and concentration steps, which not introduce 'lossy' steps but also increase the number of subsequent LC-MS/MS analyses that need to be conducted.

How many proteins can you detect in a typical proteomics analysis?

This is like the answer to 'how long is a piece of string?'. The four factors that dictate the length of the identification list in a typical proteomics experiment are:

- Complexity: the number of proteins in the sample
- Dynamic range: the range of concentrations of those proteins
- Available material
- LC-MS Instrument being used

Is label-free quantification the way to go?

This is a tricky question. Mass spectrometry is not an intrinsically quantitative method, and it is difficult to predict the relationship between the analyte and the intensity of the signal in the instrument. This is manifestly so when introducing a group of peptide ions into an instrument. For example, on MALDI-ToF, lysine terminated peptides are known to give much weaker signals than arginine terminated peptides (there are ways around this). When a complex mixture of peptides is electrosprayed into the source of an instrument, some peptides give very strong signals, but others can be pretty feeble.

It seems as though the major strength of label-free methods is in comparative (relative)

Accessing the MASCOT server (Liverpool Uni only)

 © 2008-2014 Rob Beynon

We no longer operate a MASCOT server for University of Liverpool users. Access is now limited to PFG and affiliates. THis regrettable, but there are no external input streams to support this package.
From the manufacturers 'web site:

Mascot is a powerful search engine which uses mass spectrometry data to identify proteins from primary sequence databases.
While a number of similar programs available, Mascot is unique in that it integrates all of the proven methods of searching. These different search methods can be categorised as follows:

- **Peptide Mass Fingerprint** in which the only experimental data are peptide mass values,
- **Sequence Query** in which peptide mass data are combined with amino acid sequence and composition information. A super-set of a sequence tag query,
- **MS/MS Ion Search** using uninterpreted MS/MS data from one or more peptides,

The general approach for all types of search is to take a small sample of the protein of interest and digest it with a proteolytic enzyme, such as trypsin. The resulting digest mixture is analysed by mass spectrometry.

Different types of mass spectrometer have different capabilities. A simple instrument will measure a set of molecular weights for the intact mixture of peptides. An instrument with MS/MS capability can additionally provide structural information by recording the fragment ion spectrum of a peptide. Usually, the digest mixture will be separated by chromatography prior to MS/MS analysis, so that MS/MS spectra from individual peptides can be measured.

The experimental mass values are then compared with calculated peptide mass or fragment ion mass values, obtained by applying cleavage rules to the entries in a comprehensive primary sequence database. By using an appropriate scoring algorithm, the closest match or matches can be identified. If the "unknown" protein is present in the sequence database, then the aim is to pull out that precise entry. If the sequence database does not contain the unknown protein, then the aim is to pull out those entries which exhibit the closest homology, often equivalent proteins from related species.

What is QconCAT?

QconCATs are standards for protein quantification, based on the two principles of surrogacy and stable isotope labelled internal standardisation. They were invented by Rob Beynon Simon Gaskell and Julie Pratt.

The QconCAT itself is an artificial protein, created by de novo gene synthesis and prepared by heterologus expression, usually in bacteria.

Can I have a copy of XXXXX software?

The answer is, in most, instances, no. The proprietary software that is provided with instruments often includes a free results viewer, but will lack most analytical capabilities. Further, the raw data that derive from the instrument are usually subject to high level

 © 2008-2014 Rob Beynon

processing for discovery or comparative analyses. We have paid a great deal of money for these packages, and they are usually copy-protected and restricted to single computers. These computers are usually very highly specified, running 64-bit OS, with every enhancement for handling large data files, and also for speed.

In general, external users will expect us to perform most of the analyses as well, and thus, we will access these packages on your behalf. If you wish to become adept at running the software yourself, that is no problem, but access to the computers is chargeable as well, and we will also have to make a charge for training. This is the only way we can operate without PFG subsidising other groups research.

 © 2008-2014 Rob Beynon