7. High-throughput computational proteomics – challenges and solutions in the creation of the Genome Annotating Proteome Pipeline (GAPP).

Ian Shadforth

Engineering Doctorate Student, Cranfield University

Proteomics based on tandem mass spectrometry is a powerful tool for identifying novel biomarkers and drug targets. The computational challenge for high-throughput proteomics is to automatically generate high-confidence protein identifications, including post-translational modifications, in a searchable format covering multiple datasets. Methods developed to provide such a platform are presented here.

A major bottleneck in high-throughput proteomics is that the quantity of data generated is immense, and the computational techniques needed to reliably identify proteins from proteomic data currently lag behind the ability to collect that data. This is a problem that Cranfield University began to address in 2001, in a joint project with GlaxoSmithKline. The project was carried out by the author, and resulted in the development of a highly effective proteomic data analysis pipeline that is now installed within GSK. In 2004, Cranfield was awarded a BBSRC research grant to further improve this pipeline technology and develop it into the Genome Annotating Proteomic Pipeline (GAPP) – a system which builds and maintains a catalogue of all human proteins observed in publicly available proteomic datasets. The GAPP system will be launched at ECCB 2005.

GAPP scours remote databases containing proteomic mass spectra, identifying peptides observed in the data using a collection of robust peptide assignment algorithms, including a PTM detection loop. The resulting database of observed peptides can be searched via the GAPP website, and is also made available as Ensembl DAS tracks. Each peptide has an associated confidence score, along with a record of the chain of evidence that led to its identification. The pipeline has been designed from the outset to be fully compliant with HUPO PSI data standards, ensuring maximum interoperability with other proteomics resources, and is also prepared for quantitative proteomic data.

The primary challenges in developing this pipeline have been: a) Ensuring that the protein identifications are accurate; b) Improving the identification rates of post-translationally modified peptides; c) Enabling the experimentally confirmed peptides to be used to improve genomic annotation; and d) Linking datasets such that the results of many experiments can be queried simultaneously. Recently a number of systems have emerged that attempt to answer some of these challenges, such as PeptideAtlas and The Global Proteome Machine. The GAPP system provides solutions to all four of these issues. This presentation will focus on the innovations implemented in GAPP to achieve these aims.