Towards unbiased biomarker discovery

High-throughput molecular profiling technologies are routinely applied for biomarker discovery to make the drug discovery process more efficient and enable personalised medicine. The ability to find novel biomarkers rests on large-scale data generation and unbiased data analysis.

By Dr Jens Hoefkens

Over the past decade, biomarker discovery has become important in many aspects of drug discovery and development. Biomarkers predict toxicity and help to save hundreds of millions of dollars with the early termination of costly clinical trials. Many biomarkers play important roles in diagnostic applications, where they help to detect or predict diseases. Prominent examples of the last category are surely the PSA test for early detection of prostate cancer and the BRCA gene tests for the prediction of breast cancer susceptibility. Most recently, biomarkers have been at the foundation of a move towards personalised medicine.

Highly specific biomarkers hold great promise for drug development and the safe application of drugs. Not surprisingly, significant effort goes into the discovery of novel biomarkers. For this reason, researchers rely on a number of molecular profiling technologies ranging from whole genome and gene expression analyses to proteomics and metabolomic studies. As seen in several recent large biomarker studies (eg Innomed-Predtox.org), it is not a single marker but a combination of transcripts, proteins and metabolites that yields the high specificity required for predictive biomarkers.

Gene expression analysis has been one of the first high-throughput molecular profiling technologies with widespread adoption for biomarker discovery. Where reverse transcription polymerase chain reaction (RT-PCR)-based methods allowed for the targeted analysis of few, previously-known genes, microarrays enabled the simultaneous analysis of tens of thousands of genes. With the introduction of two-dimensional gel electrophoresis and mass spectrometry for high-throughput protein biomarker discovery, mass spectrometry for metabolite biomarker discovery and next-generation sequencing technologies for transcript biomarkers, pharmaceutical researchers now have a large toolbox of molecular profiling technologies for discovery and validation of biomarkers.

The rapid proliferation of high-throughput molecular profiling technologies is not without challenges. First, there are data volume and complexity. Modern profiling technologies can easily generate 25 gigabyte of data per sample, meaning that researchers conducting medium-sized biomarker discovery studies are easily faced with terabytes of raw data. Identifying relevant biomarkers is thus akin to finding a needle in a haystack. Where previously, molecular profiling data could easily be analysed in spreadsheet applications, modern experiments require specialised software systems to handle the immense data volume and complexity. Today's gold standard in data analysis is that data pre-processing and data mining should take less time than the original data acquisition – commercial software systems meeting this requirement are available for a wide range of applications and budgets.

Another challenge in large-scale biomarker discovery studies stems from the difficulty of integrating data streams originating from different technologies. As indicated earlier, often it is not a single marker that yields the required specificity but a combination of markers from different biological processes. Successfully finding such markers requires the ability to integrate and mine complex

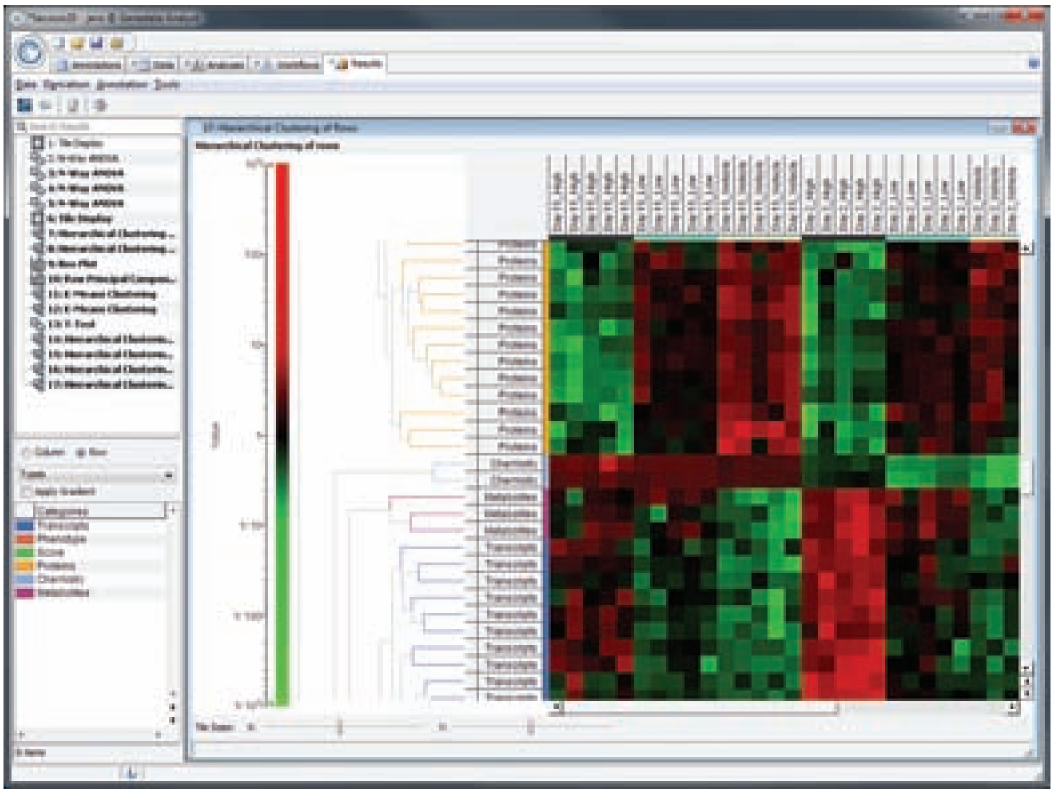


Figure 1
Predictive toxicology requires the integration of transcriptomics, proteomics and metabolomics data with serum clinical chemistry. The image shows a small set of biomarkers from a large predictive toxicology study involving thousands of samples and multiple technologies. The hierarchical clustering shows different biomarker signatures with distinct responses to time and dosage. Shown are proteins (orange), serum chemistry (light blue), metabolites (purple) and transcript (blue)

data sets from different sources. As with the problem of data volume, the issue of data integration has been addressed by commercial software vendors and data mining software packages with full support for cross-technology mapping are becoming available and are being accessible to biologists (**Figure 1**).

The focus here is on the question of how functional annotation is used in molecular profiling studies. This question originally came up in microarray-based gene expression experiments but also applies to high-throughput proteomics studies. The availability of annotation can – often inadvertently – bias statistical analysis towards marker sets that include known genes, proteins or metabolites of interest. At the same time, annotation can also bias analysis towards previously known biomarkers, hereby significantly reducing the chance for identifying novel markers. This problem will be explored in the context of transcript and protein biomarker discovery, but the conclusions transfer to metabolomics-based and other biomarker applications as well.

Transcriptomics

High-throughput biomarker discovery using gene expression microarrays started to see widespread use in the late 1990s. While the first microarrays were limited to relatively few genes and Expressed Sequence Tags (ESTs), more recent microarrays allow the simultaneous measurement of expression of all known genes. Key to this method is the vendor's probe selection algorithms because it determines the specificity and accuracy of the transcript measurements. The static nature, however, of the technology means that once probes are selected, the array content is fixed and updated information on genes and transcripts cannot be included in the measurement. Recently, vendors have tried to address this problem with tiling arrays where the probe selection is based simply on the reference genome and not on *a priori* knowledge of known genes and transcripts. Tiling arrays allow an interpretation of measurements based on changing annotation. Currently, however, most studies searching for transcript biomarkers continue to rely on gene expression data from fixed layouts and are therefore limited to a pre-defined set of known transcripts.

As functional annotation linking probes to genes and transcript are readily available, data analysis can be biased by *a priori* expectations. Statistical analyses will sometimes be tweaked to yield results that match or include specific markers, regardless of whether the actual data justify these conclusions using scientific rigor. This bias can easily be avoided by excluding functional annotation from the initial statistical data analysis. By analysing experimental data without functional annotation, but with statistical, one can ensure an unbiased selection of biomarker candidates. To facilitate the biological interpretation of the results, functional annotation can be added back into the data set. It should be noted that this is not a new problem. Blinded and double-blinded studies have been used in many fields of scientific research to alleviate the problem of human bias. Moreover, double-blinded studies are considered the gold standard in clinical trials. Blinded biomarker discovery does add overhead to the process, and so it is often avoided.

Another limitation for identifying novel biomarkers is the use of gene expression microarrays as most microarrays use a fixed probe selection. Doing so limits the search space for potential biomarkers to those transcripts that were known at the time of probe selection or have been included in the probe selection for different reasons. In other words, it is impossible to use these microarrays for the identification of entirely new biomarkers. While the use of tiling arrays tries to alleviate some of these problems, it is only with the introduction of next-generation sequencing technologies that the inherent bias can be overcome. Unlike microarrays with their fixed probe selection, RNA sequencing promises to uncover the complete space of potential transcript biomarkers. By mapping short sequence reads to a reference genome, the technology allows unbiased data-driven biomarker discovery unaffected by functional annotation. Not surprisingly, several studies have shown that unannotated regions of the genome contain many highly specific biomarkers (**Figure 2**).

Proteomics

While transcript biomarkers are relatively easy to find and validate using gene expression microarrays and PCR assays, protein biomarkers hold a much greater potential for biomarker discovery. First, the proteome is much larger than the corresponding transcriptome – especially when accounting for post-translational modifications. Thus, the space of potential markers in the proteome is much larger and there simply are more potential protein markers than transcript markers. In addition, the proteome is usually a better reflection of the underlying

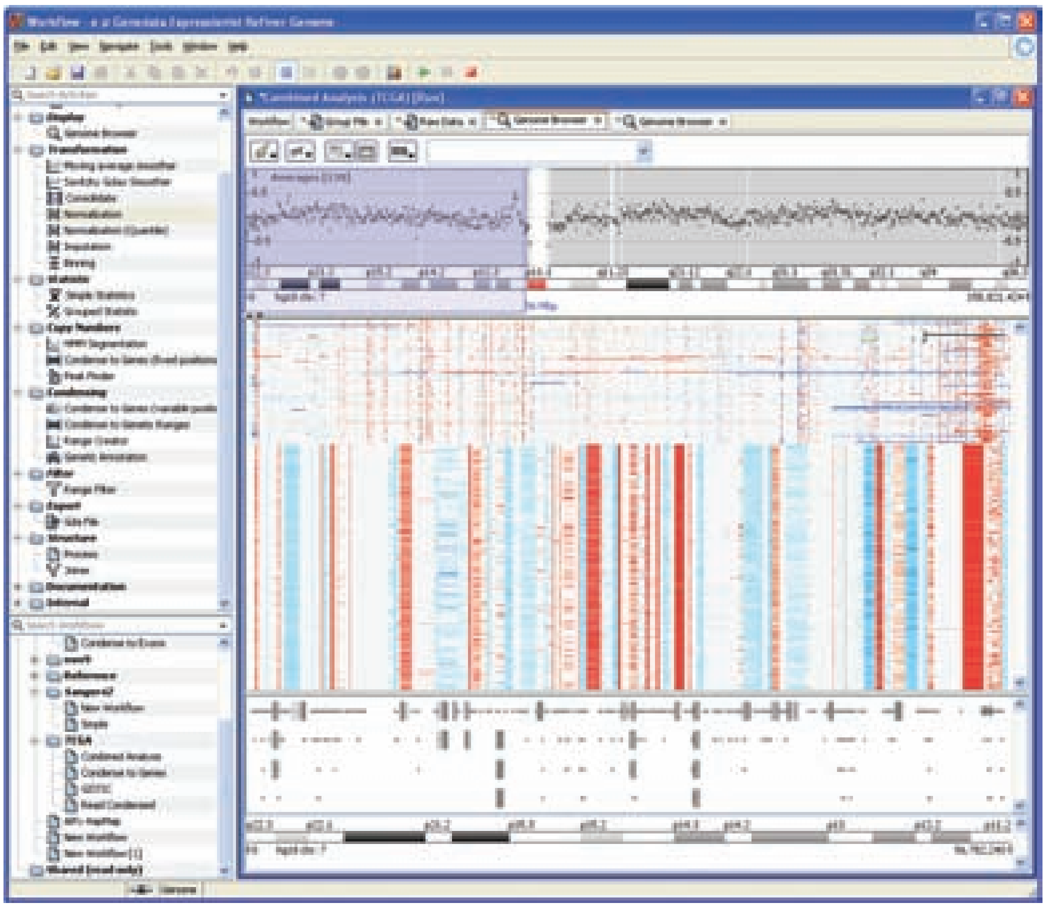


Figure 2
Next-generation sequencing and related technologies promise to revolutionise the identification of genomic biomarkers. Shown is genome browser view showing chromosome seven with data measuring copy number variation and gene expression for several hundred patient samples. For easier interpretation, known genes and transcripts are shown below the data

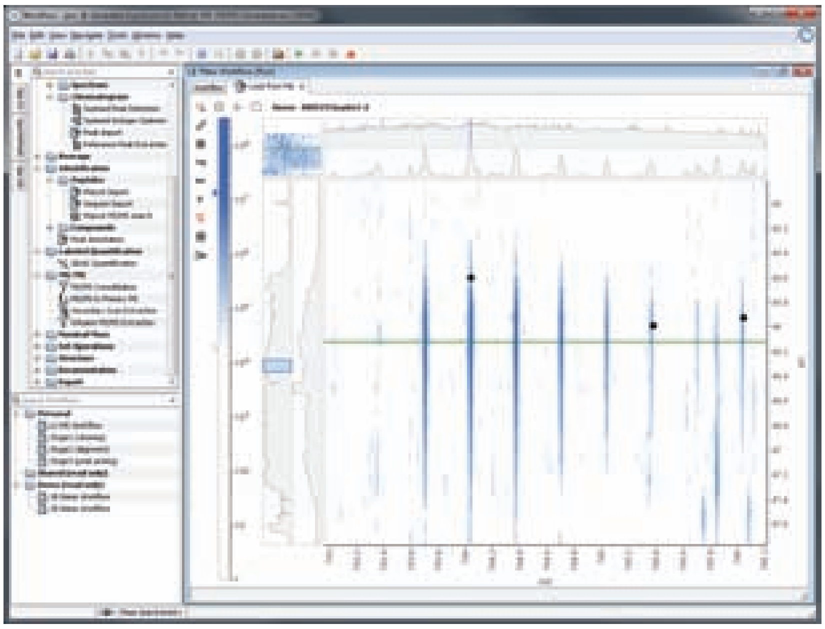


Figure 3A: A typical LC-MS chromatogram visualised in a two-dimensional view with colour-coded MS intensity values where darker regions indicate higher intensity. Selected region of the chromatogram covers about three minutes (vertically) and three Daltons (horizontally). Parallel dark blue lines represent single peptide peak and black dots denote retention times and masses of fragment precursors

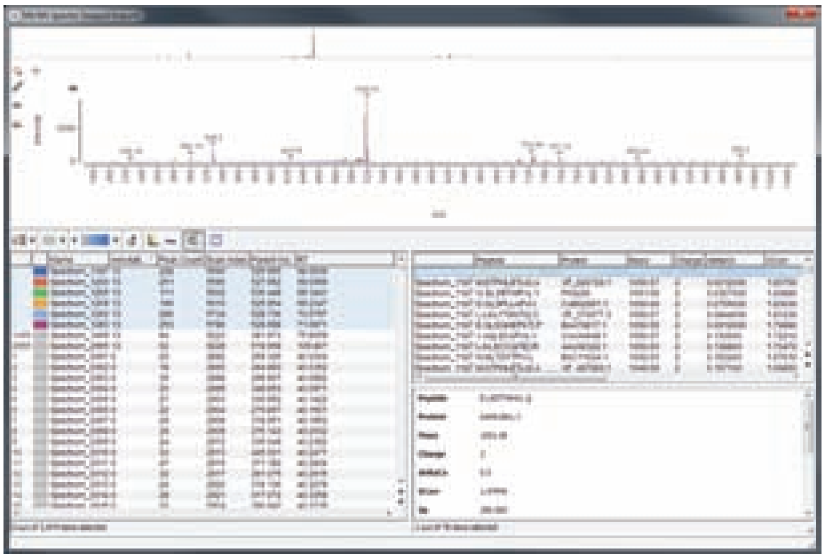


Figure 3B: Selected fragment spectra of peptide peaks and their respective peptide identifications as determined by SEQUEST search engine

biological processes and it is possible to see changes on the protein level (eg post-translational modifications) that would be missed by transcriptomics.

For discovery efforts, however, proteomic experiments are more difficult to conduct than the corresponding transcriptomics experiments. Protein arrays tend to be small and are rarely suitable for biomarker discovery because as with gene expression arrays, they require *a priori* selection of proteins. Traditionally, protein biomarker discovery has been conducted with two-dimensional gel electrophoresis. After separation and identification of significant protein spots using software tools for image analysis, spot matching and statistical analysis, in-gel digestion of selected spots and subsequent identification of the peptides and proteins by mass spectrometry and proteomics search engines such as Mascot or SEQUEST are used. Preparation of gels is time-consuming and labour-intensive with limited potential for automation. In addition, matching spots across gels is a difficult computational problem. Software tools often fail to scale well to hundreds of samples. Thus, in the last few years many drug discovery organisations have abandoned their protein biomarker discovery using two-dimensional gel electrophoresis in favour of liquid chromatographic mass spectrometry (LC-MS) experiments.

LC-MS experiments typically start with removal of highly abundant proteins like albumin. Samples are then subjected to a tryptic digest, splitting long proteins into shorter peptides. This is followed by liquid chromatography to provide another dimension for peptide separation in addition to the mass to charge ratio. For typical proteomics applications, the chromatographic gradient lasts between one and two hours and sample elucidating off the column is used as input for a continuously running mass spectrometry instrument. While the exact timing depends on the MS instrumentation used, individual MS scans are acquired every few seconds. As such, a typical LC-MS experiment consists of a few thousand mass spectra, which are typically treated as a single chromatogram. LC-MS experiments are now the most commonly used technologies for the identification of protein biomarkers.

The LC-MS experiments provide the researcher with intensity values associated with peptide peaks. For wide ranges of peptide concentrations, there is a linear relationship between peak heights and corresponding peptide concentrations. Combined with the fact LC-MS experiments are easily automated, liquid chromatography mass spectrometry is well-suited for protein biomarker discovery. However, as proteins are digested into many peptides, it is difficult to correlate peaks back to proteins. The most common approach uses secondary mass spectrometry with ions collected from chromatographic peaks and fragmentation to obtain MS spectra that can be submitted to proteomic search engines such as Mascot and SEQUEST. Those identified peptides are then used to determine differential expression, say between diseased and normal tissue. However, because this technique utilises only the identified peptides for analysis (typically only 5-10% of all peaks), the breadth of the comparison is by necessity limited. In addition, the peptide identification step can also be limiting, due to its inability to detect low-abundance proteins. The absence of these low-abundance proteins often overlooks some of the most important biomarkers that might be used to identify or characterise a given disease state.

To avoid this limitation and the bias for highly expressed peptides, recent informatics approaches currently gaining favour among researchers feature the employment of screening techniques prior to peptide identification, using parametric statistical analysis tools and methods. In university settings, for example, Schikowski et al from the University of Washington, Seattle, provided an excellent workflow for this approach in their seminal paper on signal maps using MS. On the commercial side,

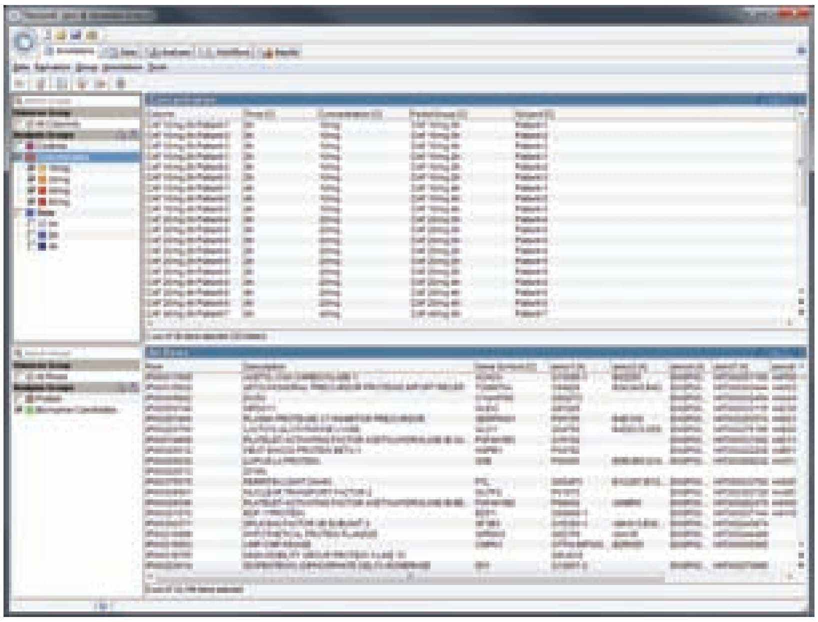
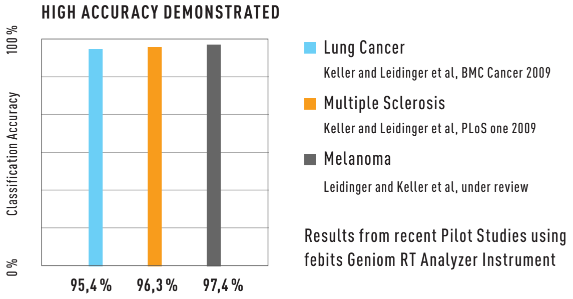


Figure 3C: Statistical analysis of protein data after peaks and peptides have been rolled up to the protein level. Tabular view shows protein identifiers and corresponding functional annotation (including gene symbol assignments based on protein identifications)



a number of vendors are using these new methodologies and tools in their respective solutions.

By its nature, mass spectrometry-based protein biomarker discovery is a two-step process: 1) Detection and quantification of chromatographic peaks; 2) Secondary experimentation is required to identify peaks and correlate them back to peptides and proteins. As such, it would seem that protein biomarker discovery avoids the two main sources of discovery bias discussed in the context of transcriptomics.

In the interest of convenience, many mass spectrometry instrument vendors are offering instruments that combine the primary and secondary mass spectrometry using a so-called 'data-dependent mode'. In this case, the instrument control software investigates the primary spectra and automatically determines appropriate pre-cursor masses and retention times for secondary mass spectrometry. Thus, using this experimental set-up the starting point for subsequent biomarker discovery efforts is a list of peak intensities and corresponding peptide and protein annotation. As this approach avoids the need for a secondary experiment, it is relatively popular with researchers conducting biomarker discovery studies.

From the previous discussion on transcriptomics biomarker discovery it should, however, be clear that the approach of including peptide annotation in proteomics studies is not without pitfalls. Not only does it open the door to biologically-biased data analysis, it can also limit the search for biomarkers to those that have been identified by database searches. In any given experiment, only about half the peaks are covered by automatically positioned fragment spectra, the in-line acquisition of fragment spectra also has an inherent bias towards high-intensity peaks – on the other hand, it is known that many interesting protein biomarkers are of relatively low abundance and will therefore not be covered by automatically acquired fragment spectra. Last but not least, it should be noted that the in-line acquisition of fragment spectra usually comes at the expense of primary scanning time and reduced quantification accuracy (especially for peptide peaks of low abundance). Coupled with the fact that 95% of all fragment spectra are in peaks that are not biomarker candidates, it would seem that in-line fragmentation provides very little benefit in protein biomarker identification studies.

For protein biomarker discovery it is advantageous to separate the primary LC-MS experiment from a secondary experiment focusing on protein identification. Doing so usually leads to improved quantification results, more statistically significant results, and avoids bias towards known proteins (**Figure 3**).

Conclusion

Biomarkers play an ever-increasing role in today's drug discovery process and high-throughput molecular profiling technologies have the potential to rapidly accelerate the discovery of highly specific biomarkers for drug safety and efficacy and diagnostic applications.

For biomarkers to fulfill their promise of safer and more effective drugs, it is important to find highly specific markers that are easily translated into laboratory tests. To find novel markers offering the required specificity, unbiased search methods and data analysis approaches are requisite. Next-generation sequencing techniques and liquid chromatography coupled with mass spectrometry can make tremendous contributions to this effort. There is, however, a need for unbiased biomarker discovery that is not restricted by previous results and *a priori* expectations. DDW

Dr Jens Hoefkens is Head of the Genedata Expressionist Business Unit that develops and markets Genedata Expressionist, the premier biomarker discovery platform in life science R&D. He joined Genedata after earning a dual major PhD in Physics and Mathematics from Michigan State University for work on self-validating computational methods.