Transforming data into better healthcare

The healthcare industry is drowning in data for which it lacks established data analysis processes. This situation exists despite the fact that appropriate and efficient data analysis processes are keys to better decision-making and enhanced healthcare therapies, which can potentially result in multi-billion dollar savings. This article examines software solutions that enable researchers, who do not have a PhD in statistics, to better understand the ever-increasing data collected on patients and their diseases.

By Dr Jens Hoefkens

Translational medicine holds much promise for the healthcare industry, particularly for the continued ability of pharmaceutical companies to create safe and effective drugs. As an extension of evidence-based medicine, translation medicine relies on a data-driven approach. This approach requires the integration of many different data sources, the analysis of which is alternatively described as finding the needle in the haystack or drinking from the fire hose.

Clinical trials usually involve hundreds of patients from different cohorts. Previously, such clinical trials focused on a relatively small number of phenotypic parameters (ie patient measurements) such as body weight, blood pressure, cholesterol levels, etc. With new and affordable molecular profiling technologies such as next-generation sequencing and mass spectrometry, clinical trials now generate literally mountains of molecular profiling data for each patient.

Going forward, research and development organisations will be forced to turn the available data into actionable intelligence and find ways to make real-time decisions based on available patient data. Here we discuss how software solutions enable researchers to perform advanced statistical analysis on an ever-growing sea of data and show how statistical analysis can be combined with data visualisation for improved decision making, without requiring a PhD in statistics.

The process of clinical trial data analysis has long been established. Carefully planned experiments combined with well-defined clinical outcome and endpoint measurements have been part of the drug development process and have led to thousands of drugs. While clinical trials have always involved large numbers of patients, the number of data points per patient has historically been relatively small, combining a few dozen patient data with clinical outcome measurements.

The arrival of personalised medicine, however, has a dramatic impact on this scenario. Advances in molecular profiling technologies enable clinical researchers to monitor gene and protein expression levels for every trial subject. They can make baseline comparisons of metabolite profiles, and study the correlation between genomic variations and phenotype data. Given the human genome includes about 30,000 genes, the human proteome has between 20,000 and 25,000 non-redundant proteins, and there are about 10 million annotated single nucleotide polymorphisms (ie isolated gene variations or SNPs), it is easy to see how the healthcare industry is facing an unprecedented challenge in managing, analysing and interpreting this rising tide of data.

Fortunately, this does not necessarily require the invention of a new statistical methodology. Instead, existing statistical tools can be put into the hands of scientists. Most statistical methods to

analyse large (clinical) data sets are well established. Additionally, recent advances on the algorithmic side, combined with ever-increasing computational processing power, have made the application of statistics straightforward. However, there is a major change in data analysis software requirements: Statistical methods must be accessible to scientists and researchers.

The old paradigm in which clinical data was collected and sent to a dedicated team of statisticians who would then combine clinical data with preclinical data and other information, has proven to be inflexible and slow. Therefore, statistical tool development is shifting from tool provisioning to usability and ease of use.

As clinical data analysis is regulated and requires submission of data and methods to regulatory agencies (eg FDA, EMEA), there is little competitive advantage for doing in-house method development. Most lifescience companies rely on well-tested and established commercial software packages to explore and analyse clinical and translational data sets. That said, for any number of historical reasons, different organisations use different data management solutions and semantics to describe data. Given the fragmented and disconnected processes rooted in legacy systems and solutions, the ideal statistical analysis software package provides tools and Application Programming Interfaces (APIs) that enable the rapid connection and adaption to in-house systems.

Last but not least, the ability to deal with immense data sets requires tools that can scale to support the necessary analysis. With clinical trials easily generating a million data points per patient, simple spreadsheet applications can no longer analyse data. In fact, any tool not designed from the ground up that can scale to billions of data points will quickly drown in a torrent of data.

The following examines implications of the different challenges in more detail, and discusses how software tools can be designed to address individual challenges. We will focus on statistical methods, usability and scalability and how these affect the ability manage an ever-expanding sea of data.

Statistical methods

Multivariate statistical methods are the workhorse for analysing molecular profiling data. The fact that these methods have a long and successful track record, combined with their ability to study multiple observables across many different patients makes them the premier choice for clinical trials analysis. Additionally, established experimental designs and power analyses enable users to plan clinical trials so that hypotheses and endpoints can be validated with a pre-defined level of confidence. This important category of statistical methods includes Pearson's Correlation, Student's t-Test, ANOVA and MANOVA, Linear Models, Discriminant Analyses, Clustering, Self-Organising Maps, and the Principal Components Analysis (PCA).

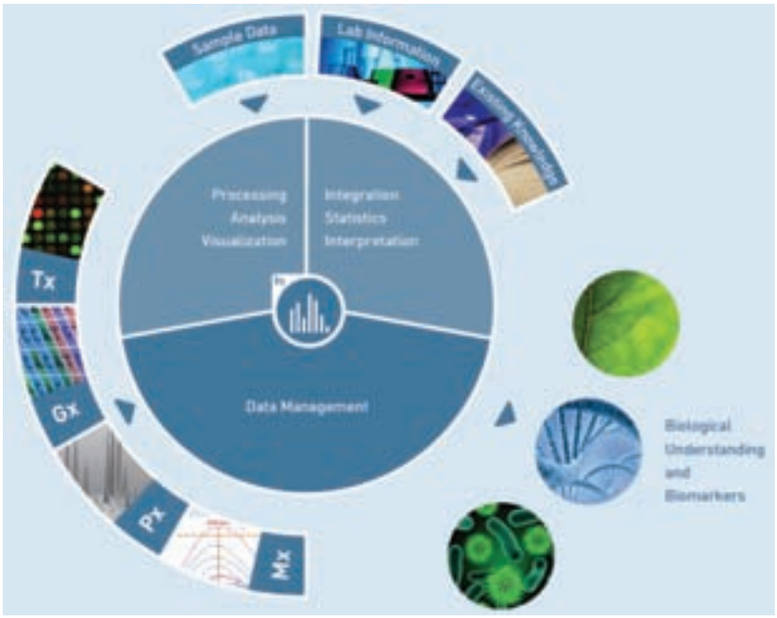

Figure 1
Combining, managing and integrating molecular profiling data from many different sources to identify biomarkers and derive actionable intelligence

Underlying many multivariate methods is the assumption of a linear relationship between data and results. However, when looking at molecular profiling data, the relationship between measurements and clinical outcome is often highly non-linear. Therefore, the applicability of multivariate statistics (and especially linear models) can be limited. While the theory of analysing non-linear data relationships is quite advanced, practical application has long been limited by the lack of computing resources and efficient algorithms. However, with algorithmic improvements and ample computing resources available to most researchers, previously impractical approaches are now well within reach and non-linear methods have become a standard tool for the analysis of complex biological data sets. The remainder of this section will briefly highlight some of those methods and corresponding applications.

Pearson's Correlation is arguably the most widely used method for analysing large biological data sets. It can be used, for example, to directly compare gene expression signatures and

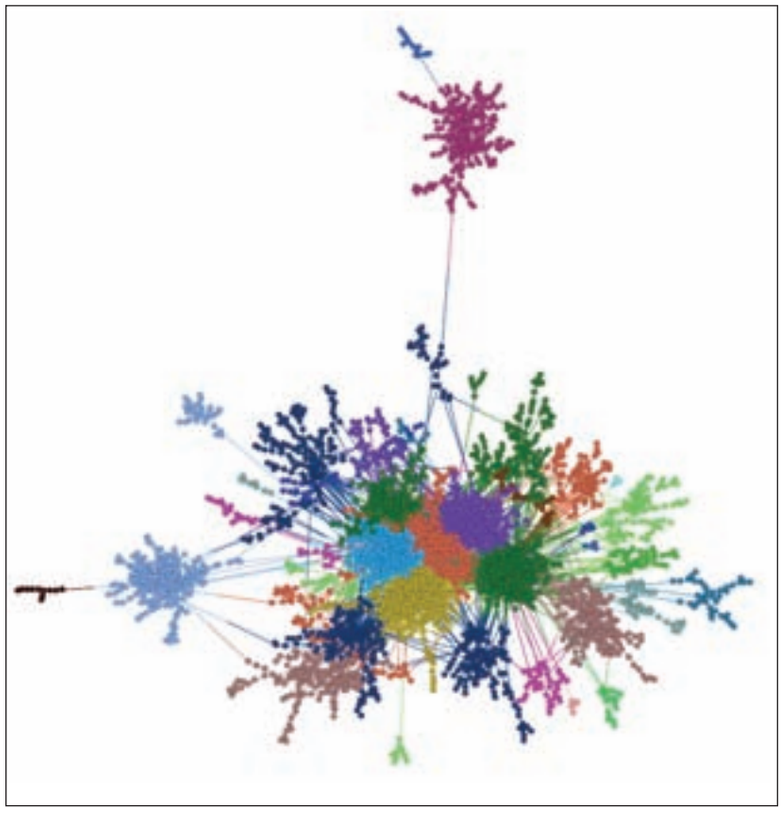

Figure 2
Gene interaction network derived from a million gene expression values and considering more than one billion gene-gene interactions using Genedata Analyst™. Inferred substructures are coloured and agree with known gene functions

is the core of popular methods such as hierarchical clustering and linear regression. It assumes, however, a linear relationship between data, and its applicability to non-linear data is limited. An alternative approach capable of handling non-linear relationships is the Mutual Information. It has been called the 'Correlation for the 21st century' by Terry Speed for its potential to take a similarly prominent role in statistical applications. While the idea itself is almost 50 years old, lack of efficient computational methods has until recently limited its practical use. However, with new algorithms and more powerful computers, widespread application of the method has become practical.

The inference of interaction networks from molecular profiling data is an important Mutual Information application. Scale-free networks promise to revolutionise the understanding of large biological data sets with capabilities to uncover not just individual biomarkers but relationships. This helps researchers determine cause and effect within large molecular profiling data (**Figure 2**). And, as many of the deduced interactions have been validated using targeted experiments, the methods are expected to lead to a better understanding of molecular pathways and protein-protein interactions.

Similar to Mutual Information itself, computation of scale-free networks was until recently considered all but impractical because of the algorithmic complexity. Methods using Support Vector Machines and Bayesian Networks are another example of advanced statistical methods with many applications in translational medicine. They can be trained to learn from existing data (eg to distinguish responders from non-responders) to subsequently predict how new patients will respond to treatment. With their ability to handle diverse data from different sources (eg different omics data and clinical endpoints) and deal with nonlinear dependencies between data, Bayesian methods have proven to be of great value in many clinical applications. Although their ability to predict endpoints is not always easily understood in biological terms, machine learning methods are nonetheless an important and extremely powerful asset for translational applications. These methods have been successfully used in medical applications ranging from leukaemia diagnostics to predicting toxicity of chemical compounds.

Usability

Software usability has many dimensions. For statistical analysis software, the most important aspects include:

- Accessibility of methods/
- Ease of use.
- Interpretability of results for scientists with limited statistical background.

The previous section highlighted some of the successful statistical methods for analysing complex translational data sets. And, as almost all the presented methods are freely available as part of academic proof-of-concept implementations, most researchers, at least in principle, can immediately use these methods. While dedicated biostatisticians can work with multiple command line tools and programming languages, biologists and medical practitioners arguably prefer an integrated package that combines all relevant methods in one easy-to-use graphical user interfaces. Selecting methods to be included in such a package requires the software vendor to strike a careful balance between flexibility and ease of use. Ultimately, end-users are better served by a careful selection of methods and a subset of options with sensible default values for obscure parameters.

While easy access to statistical methods is

important, the most important usability aspect for statistical software is the visualisation of data and results. Enabling non-expert statisticians to use advanced statistical tools requires helping users to interpret and understand the results of often complicated computations. And although access to detailed statistical results such as ANOVA tables is important for in-depth statistical analysis, many end-users prefer dedicated interactive visualisations aimed at showing statistical results in an intuitive manner (**Figure 3**). This includes the ability to:

- Interact with plots.
- Make selections and share them among graphs.
- Overlay and dynamically rearrange data using experimental designs and covariates.
- Create publication-ready images for presentations and reports.

Ultimately, most statistical results are presented, shared, and discussed in presentations and documents. As such, integration with productivity software such as Microsoft PowerPoint, Excel and Word is an extremely important part of enabling users to effectively communicate results.

Scalability

Software scalability can be defined along many different dimensions. Faced with the challenges of ever-growing data sets, lifescience researchers most often associate scalability with the ability to analyse and visualise large data sets quickly and efficiently. One aspect often overlooked in this context is data volume associated with metadata such as functional annotation, pathways and interaction data. Often, the size of external metadata easily exceed the size of the molecular profiling data, yet they are an important input required for interpretation and understanding of statistical results.

In the previous section we looked at the importance of data visualisation in the ease-of-use and data interpretation context. While visualisation of small data sets is a well understood problem, the ability to render and manipulate gigabytes of complex multidimensional data in real time poses unique challenges to the underlying software system in terms of scalability.

The main problem with designing scalable solutions is that it is very difficult to add scalability after the fact. Even for systems that work well with medium-sized data sets, if scalability has not been designed into the software architecture, it will almost certainly become a problem when data set sizes increase. And with the exponential growth of data, data analysis software designed to handle yesterday's large data sets is often unable to adequately handle the demands of today's applications in clinical research, molecular diagnostics and translational medicine.

It is an interesting twist that scalability is actually less of a problem in software targeted at expert users, who have traditionally relied on batch jobs and long-running computations. However, data analysis software targeted at biologists and non-expert users has to be able to manipulate and analyse data in real time to enable users to benefit

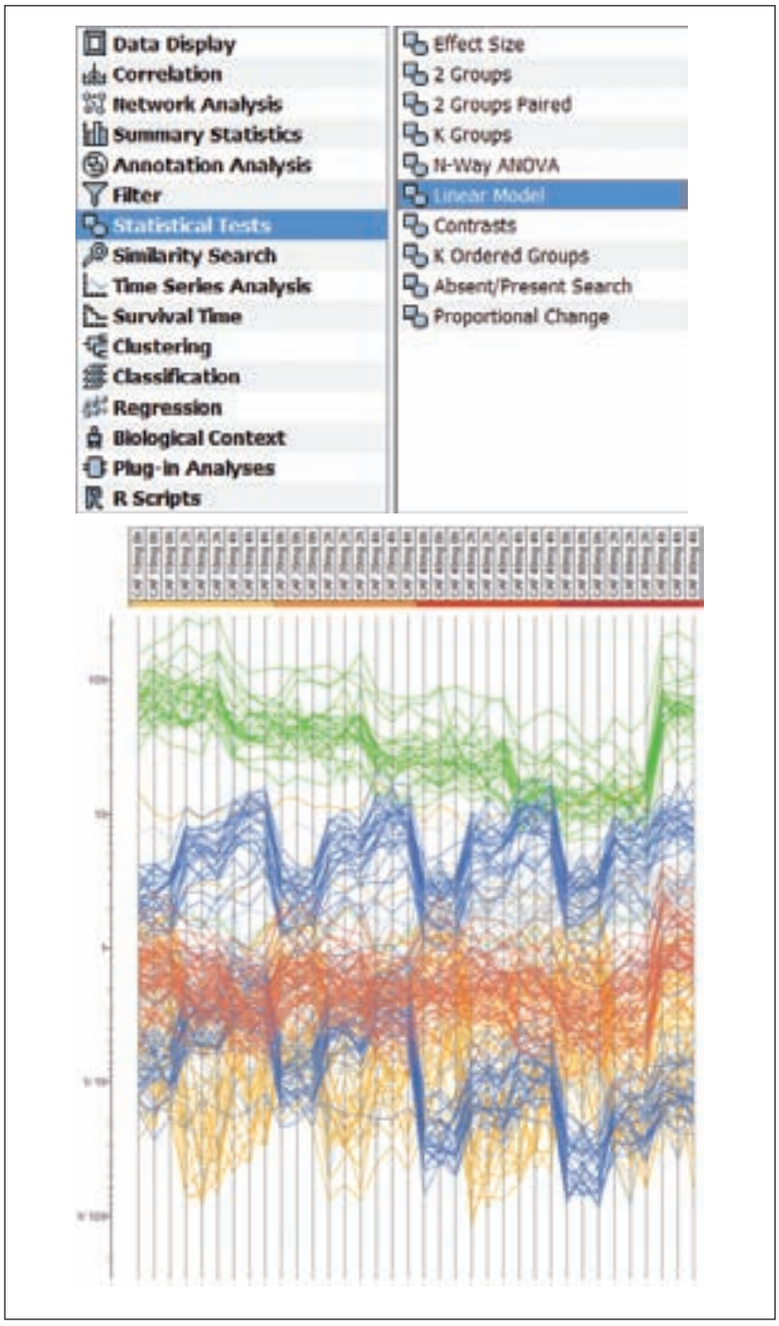

Figure 3
Example from Genedata Analyst showing access to sophisticated statistical analyses with easy-to-understand settings and intuitive visualisation of statistical results

from advanced statistical methods. Software that requires a minute or even just seconds to respond to a user action is perceived as onerous and complicated. To be responsive, software should respond in one second to any kind of user activity. Obviously, this poses unique challenges on the data visualisation and user interface design that must be included in the system architecture from the ground up.

Openness

Last but not least, software vendors must accept the fact that when it comes to data analysis software, different users have different requirements. Vendors must embrace openness and customer customisation. Users require access to existing in-house tools and databases and the variations between different customers, markets and user groups are too large to be addressed by a single solution. Instead, software should be configurable and customisable with APIs allowing users to integrate existing tools and databases.

Open source software promises the ultimate in customisation and flexibility by providing users with access to source code. Few users, however, ever make use of source code. And, those users who do quickly learn that open source software rarely guarantees API stability, and they are tasked with constantly tracking changes to critical pieces of their infrastructure. Commercial software vendors can help to reduce this risk by making API stability an explicit goal of the software design guaranteeing future support for public APIs. Having that guarantee enables customers to build and customise an infrastructure around an open commercial software solution and reduces the long-term risks associated with building such an infrastructure.

Summary

The healthcare industry can use data analysis software to stem the rising tide of data into actionable intelligence and fact-based decisions. Changes in the organisational structure of many healthcare companies pose a significant challenge to software vendors. Successful solutions must strike a balance between flexibility and ease of use, present sophisticated statistical methods in intuitive and simple terms and scale to handle billions of data points.

Going forward, successful translational medicine initiatives will depend on an organisation's ability to include biologists and medical practitioners in the analysis of molecular profiling data. By combining sophisticated statistical tools with interactive and intuitive visualisations in an open and scalable system, software systems are positioned uniquely to address the challenges of deriving knowledge and intelligence from ever-growing and increasingly-complex biological data sets. DDW

Dr Jens Hoefkens *is head of the Genedata Expressionist® business unit of Genedata. Since joining Genedata in 2002, he has been instrumental in establishing Genedata Expressionist as the leading platform for biomarker discovery and omics-based lifescience research. With a product vision for 'integrated data analysis', he managed the initial merging of transcriptomics and proteomics product lines. Leading all ongoing business and development activities for Genedata Expressionist, he also spearheads the development of Genedata Analyst™, an integrated statistical and data analysis platform with advanced visualisation capabilities. Having led Genedata USA Professional Services practice, Dr Hoefkens has a deep understanding of customer requirements, which helps to advance Genedata solutions and enrich the customer experience in lifesciences research. He earned a dual PhD in Mathematics and Physics from Michigan State University.*

Transforming data into better healthcare

The healthcare industry is drowning in data for which it lacks established data analysis processes. This situation exists despite the fact that appropriate and efficient data analysis processes are keys to better decision-making and enhanced healthcare therapies, which can potentially result in multi-billion dollar savings. This article examines software solutions that enable researchers, who do not have a PhD in statistics, to better understand the ever-increasing data collected on patients and their diseases.

By Dr Jens Hoefkens

Translational medicine holds much promise for the healthcare industry, particularly for the continued ability of pharmaceutical companies to create safe and effective drugs. As an extension of evidence-based medicine, translation medicine relies on a data-driven approach. This approach requires the integration of many different data sources, the analysis of which is alternatively described as finding the needle in the haystack or drinking from the fire hose.

Clinical trials usually involve hundreds of patients from different cohorts. Previously, such clinical trials focused on a relatively small number of phenotypic parameters (ie patient measurements) such as body weight, blood pressure, cholesterol levels, etc. With new and affordable molecular profiling technologies such as next-generation sequencing and mass spectrometry, clinical trials now generate literally mountains of molecular profiling data for each patient.

Going forward, research and development organisations will be forced to turn the available data into actionable intelligence and find ways to make real-time decisions based on available patient data. Here we discuss how software solutions enable researchers to perform advanced statistical analysis on an ever-growing sea of data and show how statistical analysis can be combined with data visualisation for improved decision making, without requiring a PhD in statistics.

The process of clinical trial data analysis has long been established. Carefully planned experiments combined with well-defined clinical outcome and endpoint measurements have been part of the drug development process and have led to thousands of drugs. While clinical trials have always involved large numbers of patients, the number of data points per patient has historically been relatively small, combining a few dozen patient data with clinical outcome measurements.

The arrival of personalised medicine, however, has a dramatic impact on this scenario. Advances in molecular profiling technologies enable clinical researchers to monitor gene and protein expression levels for every trial subject. They can make baseline comparisons of metabolite profiles, and study the correlation between genomic variations and phenotype data. Given the human genome includes about 30,000 genes, the human proteome has between 20,000 and 25,000 non-redundant proteins, and there are about 10 million annotated single nucleotide polymorphisms (ie isolated gene variations or SNPs), it is easy to see how the healthcare industry is facing an unprecedented challenge in managing, analysing and interpreting this rising tide of data.

Fortunately, this does not necessarily require the invention of a new statistical methodology. Instead, existing statistical tools can be put into the hands of scientists. Most statistical methods to

analyse large (clinical) data sets are well established. Additionally, recent advances on the algorithmic side, combined with ever-increasing computational processing power, have made the application of statistics straightforward. However, there is a major change in data analysis software requirements: Statistical methods must be accessible to scientists and researchers.

The old paradigm in which clinical data was collected and sent to a dedicated team of statisticians who would then combine clinical data with preclinical data and other information, has proven to be inflexible and slow. Therefore, statistical tool development is shifting from tool provisioning to usability and ease of use.

As clinical data analysis is regulated and requires submission of data and methods to regulatory agencies (eg FDA, EMEA), there is little competitive advantage for doing in-house method development. Most lifescience companies rely on well-tested and established commercial software packages to explore and analyse clinical and translational data sets. That said, for any number of historical reasons, different organisations use different data management solutions and semantics to describe data. Given the fragmented and disconnected processes rooted in legacy systems and solutions, the ideal statistical analysis software package provides tools and Application Programming Interfaces (APIs) that enable the rapid connection and adaption to in-house systems.

Last but not least, the ability to deal with immense data sets requires tools that can scale to support the necessary analysis. With clinical trials easily generating a million data points per patient, simple spreadsheet applications can no longer analyse data. In fact, any tool not designed from the ground up that can scale to billions of data points will quickly drown in a torrent of data.

The following examines implications of the different challenges in more detail, and discusses how software tools can be designed to address individual challenges. We will focus on statistical methods, usability and scalability and how these affect the ability manage an ever-expanding sea of data.

Statistical methods

Multivariate statistical methods are the workhorse for analysing molecular profiling data. The fact that these methods have a long and successful track record, combined with their ability to study multiple observables across many different patients makes them the premier choice for clinical trials analysis. Additionally, established experimental designs and power analyses enable users to plan clinical trials so that hypotheses and endpoints can be validated with a pre-defined level of confidence. This important category of statistical methods includes Pearson's Correlation, Student's t-Test, ANOVA and MANOVA, Linear Models, Discriminant Analyses, Clustering, Self-Organising Maps, and the Principal Components Analysis (PCA).

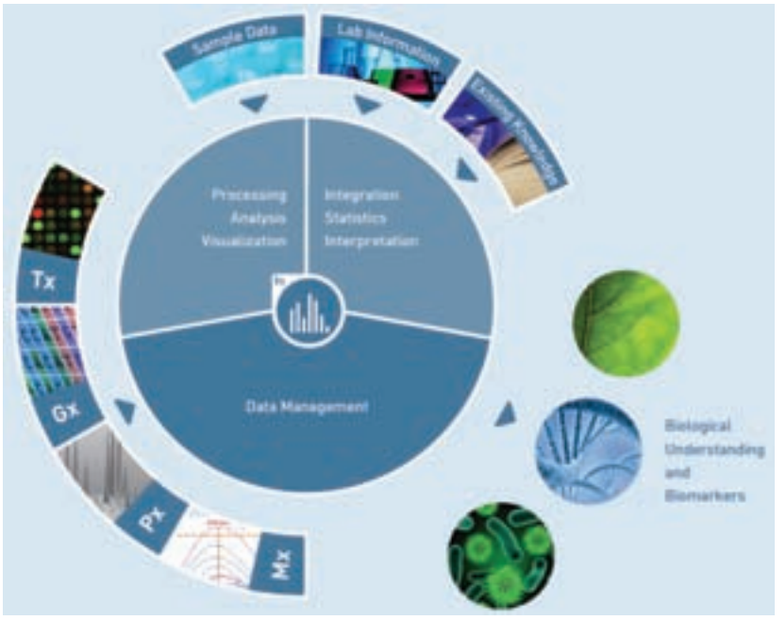

Figure 1
Combining, managing and integrating molecular profiling data from many different sources to identify biomarkers and derive actionable intelligence

Underlying many multivariate methods is the assumption of a linear relationship between data and results. However, when looking at molecular profiling data, the relationship between measurements and clinical outcome is often highly non-linear. Therefore, the applicability of multivariate statistics (and especially linear models) can be limited. While the theory of analysing non-linear data relationships is quite advanced, practical application has long been limited by the lack of computing resources and efficient algorithms. However, with algorithmic improvements and ample computing resources available to most researchers, previously impractical approaches are now well within reach and non-linear methods have become a standard tool for the analysis of complex biological data sets. The remainder of this section will briefly highlight some of those methods and corresponding applications.

Pearson's Correlation is arguably the most widely used method for analysing large biological data sets. It can be used, for example, to directly compare gene expression signatures and

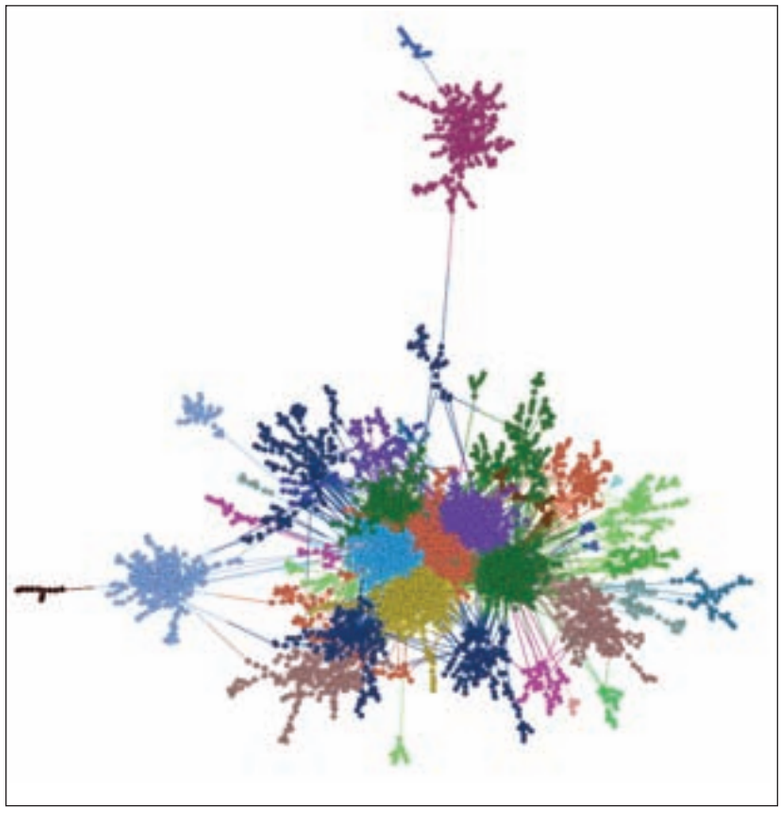

Figure 2
Gene interaction network derived from a million gene expression values and considering more than one billion gene-gene interactions using Genedata Analyst™. Inferred substructures are coloured and agree with known gene functions

is the core of popular methods such as hierarchical clustering and linear regression. It assumes, however, a linear relationship between data, and its applicability to non-linear data is limited. An alternative approach capable of handling non-linear relationships is the Mutual Information. It has been called the 'Correlation for the 21st century' by Terry Speed for its potential to take a similarly prominent role in statistical applications. While the idea itself is almost 50 years old, lack of efficient computational methods has until recently limited its practical use. However, with new algorithms and more powerful computers, widespread application of the method has become practical.

The inference of interaction networks from molecular profiling data is an important Mutual Information application. Scale-free networks promise to revolutionise the understanding of large biological data sets with capabilities to uncover not just individual biomarkers but relationships. This helps researchers determine cause and effect within large molecular profiling data (**Figure 2**). And, as many of the deduced interactions have been validated using targeted experiments, the methods are expected to lead to a better understanding of molecular pathways and protein-protein interactions.

Similar to Mutual Information itself, computation of scale-free networks was until recently considered all but impractical because of the algorithmic complexity. Methods using Support Vector Machines and Bayesian Networks are another example of advanced statistical methods with many applications in translational medicine. They can be trained to learn from existing data (eg to distinguish responders from non-responders) to subsequently predict how new patients will respond to treatment. With their ability to handle diverse data from different sources (eg different omics data and clinical endpoints) and deal with nonlinear dependencies between data, Bayesian methods have proven to be of great value in many clinical applications. Although their ability to predict endpoints is not always easily understood in biological terms, machine learning methods are nonetheless an important and extremely powerful asset for translational applications. These methods have been successfully used in medical applications ranging from leukaemia diagnostics to predicting toxicity of chemical compounds.

Usability

Software usability has many dimensions. For statistical analysis software, the most important aspects include:

- Accessibility of methods/
- Ease of use.
- Interpretability of results for scientists with limited statistical background.

The previous section highlighted some of the successful statistical methods for analysing complex translational data sets. And, as almost all the presented methods are freely available as part of academic proof-of-concept implementations, most researchers, at least in principle, can immediately use these methods. While dedicated biostatisticians can work with multiple command line tools and programming languages, biologists and medical practitioners arguably prefer an integrated package that combines all relevant methods in one easy-to-use graphical user interfaces. Selecting methods to be included in such a package requires the software vendor to strike a careful balance between flexibility and ease of use. Ultimately, end-users are better served by a careful selection of methods and a subset of options with sensible default values for obscure parameters.

While easy access to statistical methods is

important, the most important usability aspect for statistical software is the visualisation of data and results. Enabling non-expert statisticians to use advanced statistical tools requires helping users to interpret and understand the results of often complicated computations. And although access to detailed statistical results such as ANOVA tables is important for in-depth statistical analysis, many end-users prefer dedicated interactive visualisations aimed at showing statistical results in an intuitive manner (**Figure 3**). This includes the ability to:

- Interact with plots.
- Make selections and share them among graphs.
- Overlay and dynamically rearrange data using experimental designs and covariates.
- Create publication-ready images for presentations and reports.

Ultimately, most statistical results are presented, shared, and discussed in presentations and documents. As such, integration with productivity software such as Microsoft PowerPoint, Excel and Word is an extremely important part of enabling users to effectively communicate results.

Scalability

Software scalability can be defined along many different dimensions. Faced with the challenges of ever-growing data sets, lifescience researchers most often associate scalability with the ability to analyse and visualise large data sets quickly and efficiently. One aspect often overlooked in this context is data volume associated with metadata such as functional annotation, pathways and interaction data. Often, the size of external metadata easily exceed the size of the molecular profiling data, yet they are an important input required for interpretation and understanding of statistical results.

In the previous section we looked at the importance of data visualisation in the ease-of-use and data interpretation context. While visualisation of small data sets is a well understood problem, the ability to render and manipulate gigabytes of complex multidimensional data in real time poses unique challenges to the underlying software system in terms of scalability.

The main problem with designing scalable solutions is that it is very difficult to add scalability after the fact. Even for systems that work well with medium-sized data sets, if scalability has not been designed into the software architecture, it will almost certainly become a problem when data set sizes increase. And with the exponential growth of data, data analysis software designed to handle yesterday's large data sets is often unable to adequately handle the demands of today's applications in clinical research, molecular diagnostics and translational medicine.

It is an interesting twist that scalability is actually less of a problem in software targeted at expert users, who have traditionally relied on batch jobs and long-running computations. However, data analysis software targeted at biologists and non-expert users has to be able to manipulate and analyse data in real time to enable users to benefit

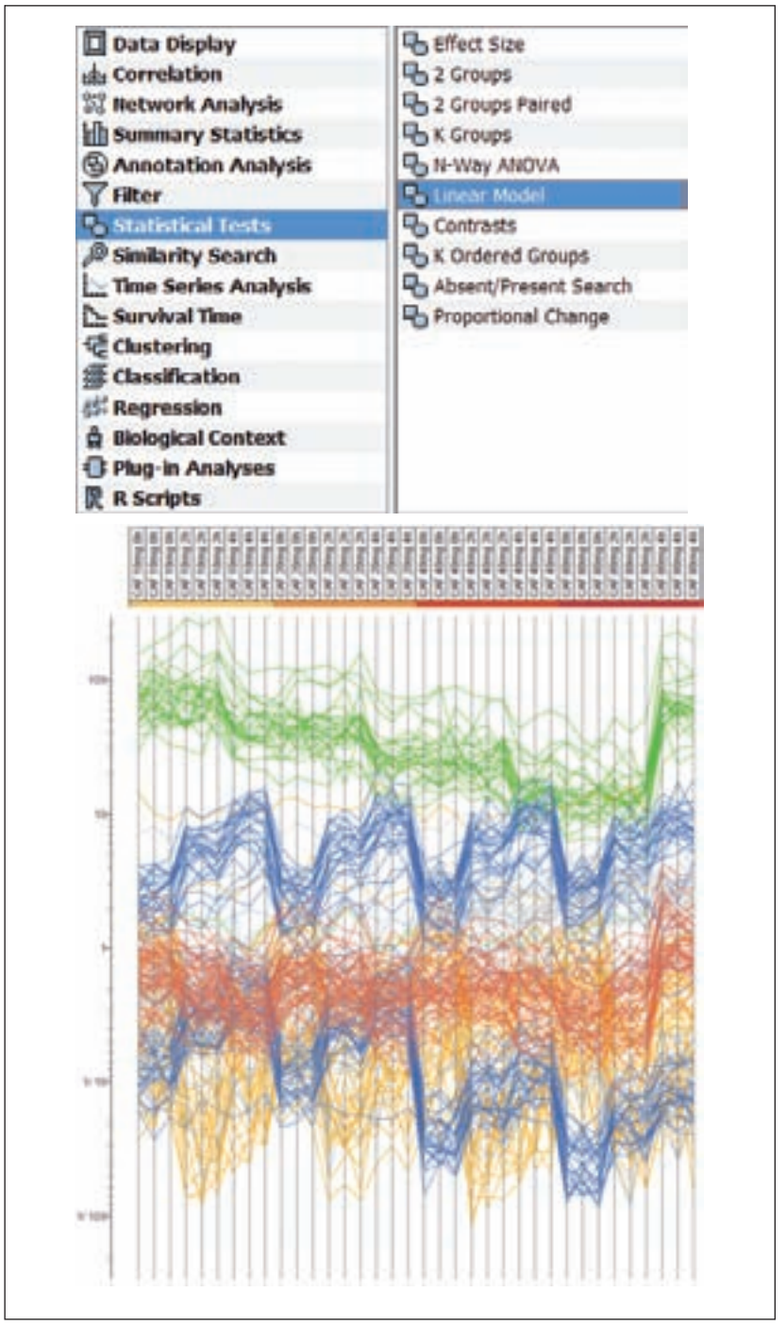

Figure 3
Example from Genedata Analyst showing access to sophisticated statistical analyses with easy-to-understand settings and intuitive visualisation of statistical results

from advanced statistical methods. Software that requires a minute or even just seconds to respond to a user action is perceived as onerous and complicated. To be responsive, software should respond in one second to any kind of user activity. Obviously, this poses unique challenges on the data visualisation and user interface design that must be included in the system architecture from the ground up.

Openness

Last but not least, software vendors must accept the fact that when it comes to data analysis software, different users have different requirements. Vendors must embrace openness and customer customisation. Users require access to existing in-house tools and databases and the variations between different customers, markets and user groups are too large to be addressed by a single solution. Instead, software should be configurable and customisable with APIs allowing users to integrate existing tools and databases.

Open source software promises the ultimate in customisation and flexibility by providing users with access to source code. Few users, however, ever make use of source code. And, those users who do quickly learn that open source software rarely guarantees API stability, and they are tasked with constantly tracking changes to critical pieces of their infrastructure. Commercial software vendors can help to reduce this risk by making API stability an explicit goal of the software design guaranteeing future support for public APIs. Having that guarantee enables customers to build and customise an infrastructure around an open commercial software solution and reduces the long-term risks associated with building such an infrastructure.

Summary

The healthcare industry can use data analysis software to stem the rising tide of data into actionable intelligence and fact-based decisions. Changes in the organisational structure of many healthcare companies pose a significant challenge to software vendors. Successful solutions must strike a balance between flexibility and ease of use, present sophisticated statistical methods in intuitive and simple terms and scale to handle billions of data points.

Going forward, successful translational medicine initiatives will depend on an organisation's ability to include biologists and medical practitioners in the analysis of molecular profiling data. By combining sophisticated statistical tools with interactive and intuitive visualisations in an open and scalable system, software systems are positioned uniquely to address the challenges of deriving knowledge and intelligence from ever-growing and increasingly-complex biological data sets. DDW

Dr Jens Hoefkens *is head of the Genedata Expressionist® business unit of Genedata. Since joining Genedata in 2002, he has been instrumental in establishing Genedata Expressionist as the leading platform for biomarker discovery and omics-based lifescience research. With a product vision for 'integrated data analysis', he managed the initial merging of transcriptomics and proteomics product lines. Leading all ongoing business and development activities for Genedata Expressionist, he also spearheads the development of Genedata Analyst™, an integrated statistical and data analysis platform with advanced visualisation capabilities. Having led Genedata USA Professional Services practice, Dr Hoefkens has a deep understanding of customer requirements, which helps to advance Genedata solutions and enrich the customer experience in lifesciences research. He earned a dual PhD in Mathematics and Physics from Michigan State University.*