

CELL-BASED ASSAYS

LARGE-SCALE LIFE SCIENCE STUDIES

Fully integrated data analysis

Jens Hoefkens, Claudio Schmid, Genedata, Basel, Switzerland

Today's life science research organisations employ a variety of high-throughput experimental technologies to tackle problems ranging from bacterial strain optimisation to plant trait development or the study of drug safety and efficacy. Microarray-based technologies inspect the effect on the whole genome, high-content screening technologies enable the analysis of live cells under a variety of conditions, and mass spectroscopy-based technologies study complete protein and metabolite profiles of all kinds of samples. Each technology is capable of generating billions of data points for a single experimental project. In most organisations, these data will be generated and stored in different formats, repositories, and locations within a highly specialised but largely unconnected infrastructure that includes laboratory information management systems, shared file systems or project databases. In such a heterogeneous environment, the bottleneck for gaining comprehensive insight into information-rich data is due to the lack of time and resources, along with a supporting infrastructure for adequate data organisation, data analysis and data mining. We present how numerous collaborations with customers and international research consortia have enabled Genedata to develop novel software technology that supports those processes on a large-scale.

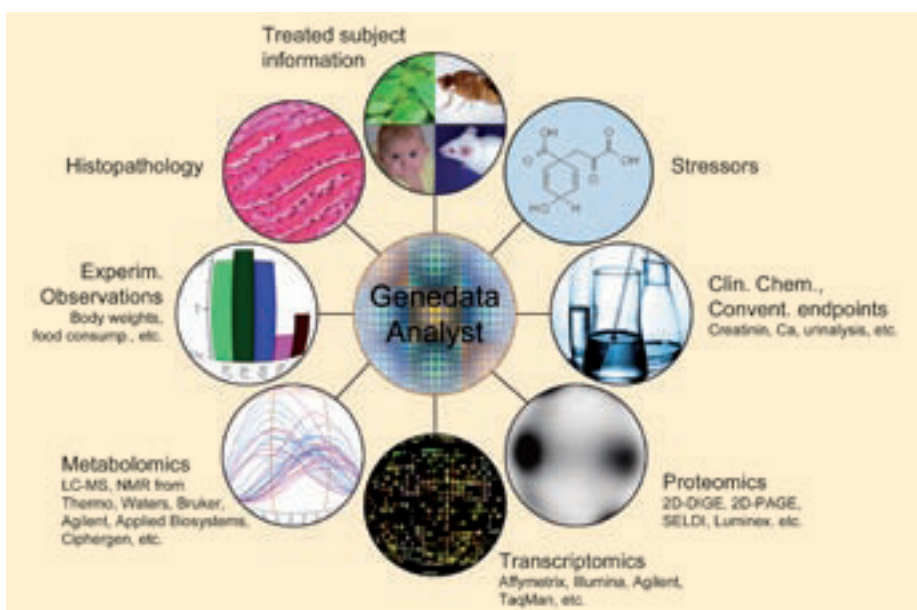


Fig. 1: Different types of data streams converging into Genedata Analyst®

Large-scale life science studies frequently collect data from a wide variety of different observations, samples, and assay types to gain a comprehensive picture of the organism's response to a compound or other external influences. Experiments are performed by separate research teams spread over multiple geographic locations, often using incompatible, vendor-specific software solutions that further complicate an integrated approach. With the advent of systems biology however, it is now generally accepted that the whole is indeed more than the sum of its parts. In an effort to maximise often substantial investments into high-throughput experimental technologies, the life science industry is seeing a paradigm shift towards an integrated data-mining approach.

Really integrating results

Yet such a shift does not come without challenges. Software systems must be able to integrate data from a variety of public and proprietary databases, and they must be capable of mining large data sets derived from disparate experimental sources. Moreover, widely used software systems should be easy to use and accessible to both end-users and expert statisticians alike. Addressing these seemingly insurmountable challenges, Genedata has worked with a number of its partners from the life science industry to develop Genedata Analyst®. Designed from the ground up as a data mining and data integration platform, Analyst includes Genedata's proprietary ExpressMap™ technology to easily integrate experimental design information with data from a wide range of different assay technologies (see Fig. 1).

Analyst's open and scalable architecture, rigorous statistical analysis of billions of data points, and numerous Application Programming Interfaces (APIs) for loading data, publishing results and integrating proprietary statistical algorithms enable tight integration into corporate IT environments. Both scientists and expert statisticians benefit from the powerful prediction algorithms available using Analyst's highly intuitive data visualisations. Unlike competing approaches that focus on integrating diverse data into a single universal data warehouse, Analyst accesses those data sources

CELL-BASED ASSAYS

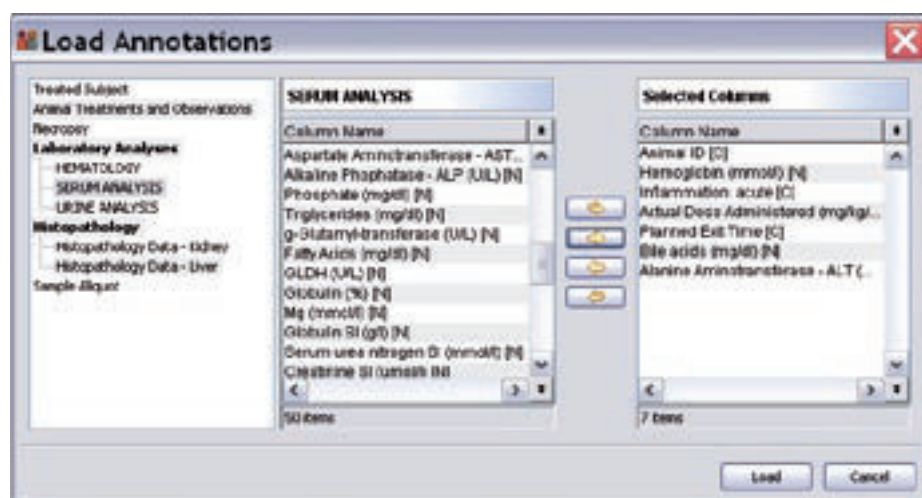


Fig. 2: Dialogue Box for loading and structuring annotation and metadata into Analyst

and utilises available information in support of large-scale data mining. As a result, Analyst can easily be integrated into existing IT infrastructures, which reduces costs and the risk of failure.

Beating the odds

Despite huge investments in target and lead compound discovery processes, the number of new drugs hitting the market has stagnated. Of those compounds that enter into preclinical development, about 90% currently fail to reach the market. Safety concerns are of paramount importance. Since the removal of a compound due to toxicity may take place at a late stage in drug development, failure costs can be very high. It has been estimated that it costs in the region of US\$1.4bn – spent over an average of 10 years – to bring a new drug to market (*DiMasi et al.* (2007) *Manage. Decis. Econ.* 28, 469-479). It is therefore crucial that compound profiling is performed at the earliest stage possible in order to save the long-term costs originating from unsuccessful candidates, and to free up development resources for more promising compounds that do not generate adverse side effects in patients.

Combining information from a variety of molecular profiling technologies with data from traditional laboratory analyses, Systems Toxicology has emerged as a promising new approach towards more informed decisions on promoting or stalling development candidates. While integrated analysis of rel-

evant data and joint interpretation in a biological or pharmacological context promises a significant reduction in drug development costs, its success hinges on the availability of rigorous study designs, structured annotation and data mining tools capable of utilizing diverse data streams (see Fig. 2).

The Genedata Expressionist[®] system is the first commercially available and professionally supported bio-IT infrastructure supporting all necessary steps for the quality assessment of raw input data, organisation of all data according to study design, and integrated data analysis and interpretation. Analyst is the unifying point providing the means to study all data within a joint analysis environment. Technological peculiarities of the different omics platforms (microarrays, 2D PAGE and MS) are addressed in domain-specific, highly-specialised Refiner modules (Refiner Array, Refiner MS, Refiner 2D-Gel), while Analyst provides the integrated analysis of all available data in a single platform.

Case study

During the recent InnoMed PredTox project (www.innomed-predtox.com) Genedata assessed to what extent a combination of different data types would improve the prediction of treatment time and dosing responses based on combining transcriptomics and metabolomics data obtained in the same study. Using the Analyst data mining platform, the team was able to conduct a

cross-omics analysis and successfully demonstrate that supervised learning methods using integrated data sets performed significantly better than models derived from individual data sets.

The transcriptomics data was obtained from 45 rat liver samples (three dosages, three time points with five replicates each) using the Affymetrix Rat230_2 array, and the metabolomics data was obtained from serum of the same animals using LC-MS technology. The goal was to see if the combination of multiple omics datasets from the same animals would lead to a better classification of treatments.

The Support Vector Machine (SVM) algorithm was used to compute the classification model. Using repeated leave-one-out cross-validation, the team demonstrated that combining data from the two different technologies resulted in better class prediction than classifiers based on individual technologies. First, only transcripts were used and a misclassification rate of 11.1% was obtained. When using only metabolite peaks, the misclassification rate was 4.4%. However, when using both transcripts and metabolites, the misclassification rate was reduced to zero.

Conclusions

The necessary foundation for any efficient integrated data analysis is actually the availability of data, the relationships that are shared among the data sources and items, and the algorithms to perform high-throughput statistics and classification with extremely large datasets. By providing a data-mining platform capable of integrating diverse profiling data with study design information and utilising uniquely scalable software architecture, Analyst enables organisations to extract high-impact knowledge from previously untapped sources, and proved that when it comes to data mining, the whole is indeed more than the sum of its parts. ▼

Contact

Dr. Jochen König
Genedata AG
Maulbeerstrasse 46, CH-4016 Basel, Switzerland
jochen.koenig@genedata.com