Genedata
solutions in silico

Performing preprocessing using Genedata Expressionist delivers a tenfold acceleration in file loading times. High-throughput host-cell protein analysis simply doesn't work unless you have an efficient automated data processing and annotation workflow.

Anders Giessing, Ph.D., Senior Scientist at Novozymes A/S, Lyngby, Denmark

AUTHOR
Anders Giessing Ph.D.
Novozymes A/S
Lyngby, Denmark

INDUSTRY
Bulk recombinant enzymes

CUSTOMER SINCE
2015

ABOUT NOVOZYMES
Novozymes is the world leader in biological solutions. Together with customers, partners, and the global community, Novozymes improves industrial performance while preserving the planet's resources and helping to build better lives

GENEDATA SOLUTION
EXPRESSIONIST

Fully Automated High-Throughput Host Cell Protein Analysis of Highly Diverse Samples

Background

At Novozymes we use bacterial and fungal hosts to manufacture bulk enzymes for a variety of industries. As with all recombinant protein manufacturing processes, host cell proteins (HCPs) are a major source of contamination that can adversely affect product stability and performance. By offering proteome-wide coverage of multiple organisms, mass spectrometry (MS) based HCP analysis is ideally suited to analyze samples from different hosts and strains. However, to keep pace with the large number of samples that our lab is required to analyze, we recognized the need to automate and streamline each stage of our LC-MS approach. Implementing robotic sample preparation, a shortened HPLC separation step, and high-speed MS acquisition increased sample throughput but generated large amounts of complex data requiring time-consuming analysis and review. To overcome this decisive bottleneck, we worked together with Genedata's scientific experts to configure and implement a tailored but highly flexible workflow, while technical consultants facilitated the integration of Genedata Expressionist enterprise software into our corporate data management infrastructure. This combination ultimately delivered a fully automated, truly high-throughput (HT) HCP analysis solution that can handle a wide range of sample types and sources.

Main Challenges

Achieving HT HCP analysis regardless of sample source
Detection of low-concentration HCPs is challenging and typically requires extensive characterization of reference samples to create specific HCP databases and/or MS spectral libraries. However, the wide range of hosts and sources from which we derive our samples makes this approach impractical. As a result, we rely on curated reference genome databases for HCP identification, an approach that requires smart and flexible data analysis to handle and leverage metadata stored in our laboratory information management system (LIMS) and corporate knowledgebase.

Keeping pace with rising data volumes using finite resources
Our goal of analyzing 200 samples per day would lead to the generation of several terabytes of complex MS data per week to be processed, posing a significant strain on our IT infrastructure and scientists. Data processing requiring manual intervention ties up personnel in performing laborious and error-prone manual data conversion, transfer, and retrieval — typically using multiple software packages.

Efficient dissemination of information across the organization
The presence of HCPs strongly affects product quality and therefore their analysis is required at different stages of product and process development.

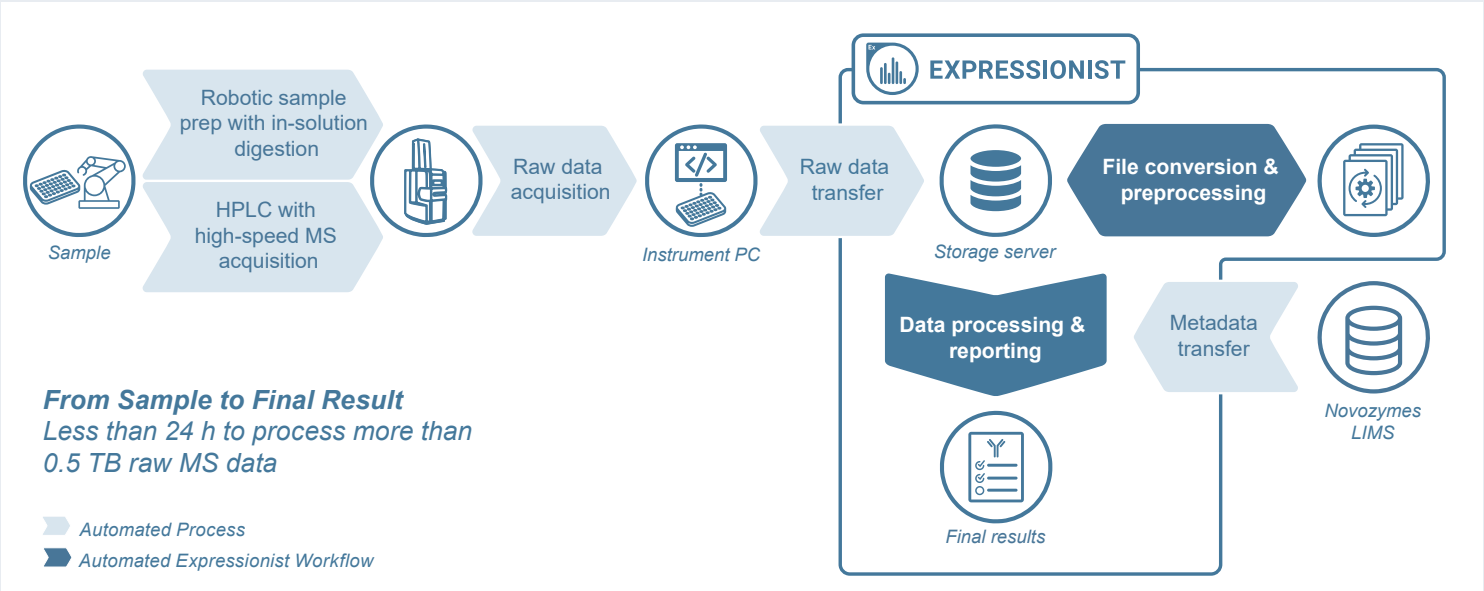
The results of our HCP identification analyses directly impact processes throughout our organization, so we needed a fast and efficient way to ingest results into our corporate data lake and provide fast and efficient ways to share relevant information to key stakeholders and decision makers.

Solution

A truly high-throughput HCP analysis data workflow
Optimizing sample preparation and accelerating MS data acquisition using cutting-edge Evosep and Bruker timsTOF instrumentation increased sample throughput but moved the bottleneck from sample preparation and MS data acquisition to data processing.

To reach the required HCP analysis throughput we eliminated all the bottlenecks in data processing, analysis, and reporting by automating the entire MS data workflow. All operations—from raw data management to generation of customized reports—are now automatically executed in a Genedata Expressionist workflow that is fully integrated into our corporate IT infrastructure through a command-line interface and dedicated plugins (Figure 1).

24/7 automated MS data preprocessing and compression
In a fully automated process, the implemented solution transfers experimental raw data to a centralized data storage server where Genedata Expressionist performs key steps such as noise reduction, smoothing, and intensity thresholding.



1 *Optimization and automation of sample- and data-processing workflows enables analysis of an entire 96-well plate in less than 24 hours; a fivefold increase in throughput compared to previous methods.*

This preprocessing step preserves all relevant signals while enabling a five- to tenfold compression of data volume; greatly mitigating strain on our IT infrastructure and accelerating downstream analysis. The subsequent dedicated proteomics-based HCP analysis workflow is designed and optimized for the specific experimental protocol—including the high duty-cycle Bruker timsTOF MS instrumentation—and configured to work continuously without human intervention.

Smart, automated, sample-specific data analysis

The seamless integration of Genedata Expressionist into our data infrastructure enables us to leverage information from our LIMS and other corporate databases. By automatically retrieving sample metadata—such as sequence identifiers—from a central Novozymes knowledgebase, Genedata Expressionist performs sample-specific data processing, analysis, and reporting. Further sample information—such as name and ID, analysis date, and the desired destination of the final report—is obtained directly from the corporate LIMS identifier. MS and MS/MS peaks are automatically detected and submitted to a proteomics search engine (Mascot, Matrix Science) together with the host organism sequence information defined in the sample metadata.

Automatic generation of customizable reports and ingestion into corporate data lake

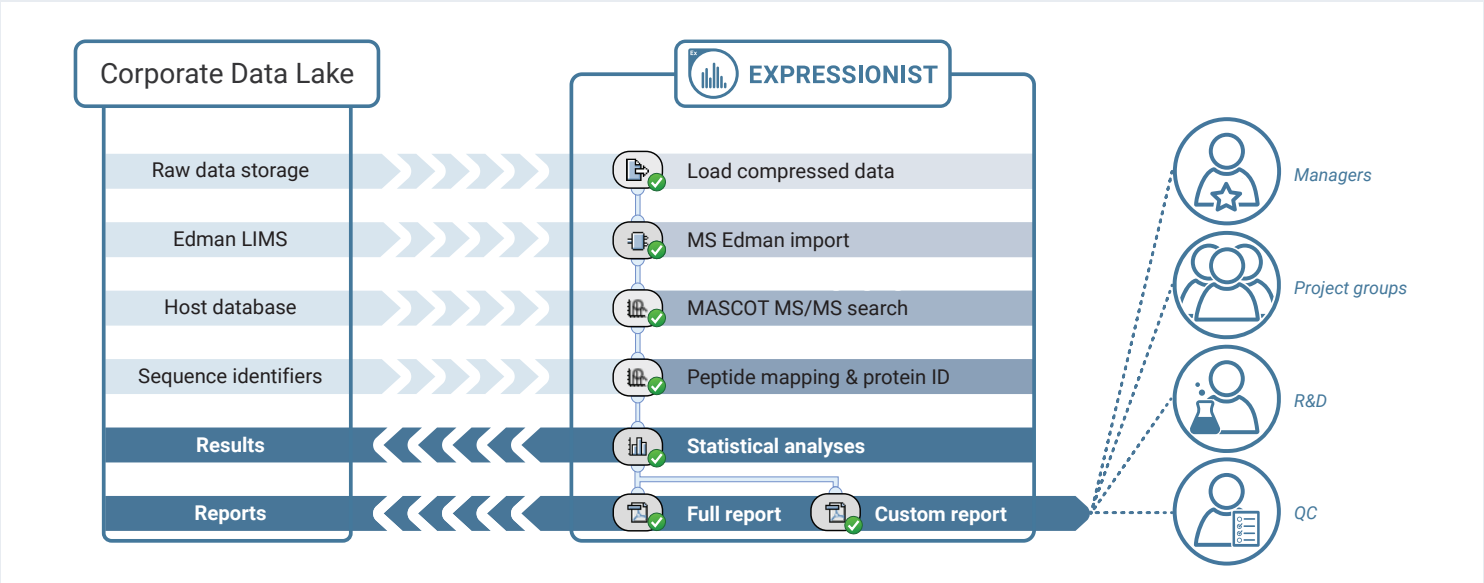
The software automatically performs statistical analyses such as Principal Components Analysis (PCA) and violin plot creation, and generates pre-configured reports containing qualitative and quantitative results on identified HCPs.

Together with detailed logs flagging any potential processing issues, these reports are sent by email to key stakeholders designated in the experiment metadata. Seamless integration of Genedata Expressionist into our corporate data lake allows us to leverage all MS data, information, and insights to inform and guide several upstream and downstream processes, such as strain characterization, expression profile modeling, and process-change monitoring.

Benefits

Keeping pace with ever-increasing MS data volumes

Automating our HCP analysis data processing using Genedata Expressionist enables us to obtain results in less than a day, compared to almost a week using our previous method. Despite increasing sample numbers, 24/7 MS data processing effectively boosted the overall duty cycle of our lab and allowed us to maintain HCP analysis turnaround times without increasing our headcount. Moreover, using a single-software platform for MS data processing eliminates the maintenance and training costs associated with the multiple software packages required for our previous method. The implemented workflow also allows a more effective use of our experts' knowledge, skills, and time by automatically generating customized reports that are typically created overnight and ready for review the following morning. Implementing a fully automated end-to-end solution on a dedicated, scalable server enabled fast and unbiased processing of the terabytes of data that our labs produce each week.



❷ ***Metadata-driven processing**: After the compressed files are loaded, information is exchanged between the Novozymes corporate data lake and the Genedata Expressionist workflow, enabling sample-specific data processing and facilitating reporting.*

Full automation delivering high-quality results

Within the Genedata Expressionist workflow, each data processing step can be precisely configured to apply parameters that are optimized for the experimental conditions and specific instrumentation (Bruker timsTOF) on a sample-by-sample basis. By combining significant time savings with sensitive and accurate HCP analysis, this smart approach to automation delivers higher-quality results than a generic "one-size-fits-all" processing method, ultimately ensuring the highest quality of our products.

In addition, automating our HCP analysis and using a single software platform for all MS data processing, analysis, and reporting not only precludes the risk of human factors influencing results, but also provides standardized and reproducible processing that significantly increases the quality of our results.

Facilitating knowledge sharing and fast decision-making

By automatically distributing customized reports containing information tailored to the respective stakeholder, the Genedata Expressionist workflow facilitates collaboration across our entire organization and allows timely interventions when quality issues arise. Data and results can be securely accessed at any time through a dedicated server enabling us to make faster and better decisions and maintain a high level of productivity even when we are working remotely. This information also enriches our corporate knowledge base and supports the decision-making processes of our internal customers in bioprocess development and quality control.

Summary

While developing a truly high-throughput MS-based HCP analysis platform, we found that optimizing and automating sample preparation and data acquisition increased throughput, but simply shifted the bottleneck from data acquisition to data processing. Together with Genedata we developed a smart, automated, metadata-driven MS data processing workflow that delivers a five-fold increase in throughput compared to our previous method without any need for human intervention.

By adopting this solution, we were able to overcome the critical MS data processing bottlenecks and meet our throughput goal of 200 samples per day. Despite ever-increasing sample numbers and limited personnel and computing infrastructure, this dramatic increase in productivity has enabled us to maintain HCP analysis turnaround times while improving data quality.

Outlook

Basing our HCP analysis on an industry-leading platform that supports all MS instruments ensures that our data processing capabilities will keep pace with new technological developments and analytical requirements. Full scalability and seamless integration into and across our corporate IT infrastructure makes Genedata Expressionist our current and future platform of choice for MS data processing, analysis, and reporting.

> ***The ability to integrate metadata—such as host, target sequence, or media conditions—is crucial to efficient analysis. We can pull in metadata on the fly from our corporate LIMS to direct the MS data processing workflow for each individual sample and facilitate downstream reporting.***
>
> **Anders Giessing**, Ph.D., Senior Scientist at Novozymes A/S, Lyngby, Denmark

GENEDATA SOLUTION

EXPRESSIONIST

Acknowledgment: Front page image reproduced by kind permission of Novozymes.