



This is a promotional banner for NextGENe software. On the left, the text reads 'Relax! If you have NextGENe™ Next Generation Sequencing Software'. In the center, it says 'Software for Illumina®, Roche 454 and ABI SOLiD™ systems' and 'Click now for more details and trial software'. On the right is a yellow cartoon character with a smiling face, wearing a white shirt and a black tie.

Bracing for a Metabolomics Data Storm, Thermo, Genedata Partner on Informatics

October 3, 2008
By Vivien Marx

As metabolomics takes its place alongside genomics, transcriptomics, proteomics, and other large-scale experimental platforms, instrumentation vendors and their customers are seeking new ways to handle its ballooning data yields.

In one example, Thermo Fisher Scientific is partnering with Genedata to optimize metabolomics-specific informatics workflows for its mass spectrometers.

Thermo is no stranger to bioinformatics. The company markets its own suite of analysis tools for proteomics, which it recently updated [[BioInform 09-12-08](#)]. However, Donna Wilson, Thermo's strategic marketing specialist in metabolism and metabolomics, said that the informatics requirements for metabolomics are very different than for proteomics, even though both fields need to analyze mass-spec data.

"Metabolomics is trying to understand how small-molecule biomarkers can indicate disease [or] exposure," Wilson explained to *BioInform* via e-mail. These metabolites "are closer to the phenotype of an organism than proteins or genes."

In practice, the experimental workflow for metabolomics, which is "small-molecule oriented," is different than for proteomics, which is "large molecule-oriented," she said. Data processing and interpretation differ as well, she said, necessitating different tools and appropriate statistical analysis, even though the two fields are often trying to answer the same questions.

One metabolomics project involving Thermo and Genedata is developing methods that quantitatively and computationally assess metabolic profiles generated on Thermo's LTQ-FT mass spec.

Led by biochemist Dean Jones, who directs the Clinical Biomarkers Laboratory in Emory University's Department of Medicine, the project is currently assessing informatics tools for its metabolomics pipeline, and Jones and his colleagues are comparing results generated with Genedata's Expressionist and Refiner MS software with various open source tools, including the [XCMS module](#) within the R framework.

"The Genedata folks are perfectly comfortable with this because it validates their product as well," Jones told *BioInform*.

Jones said that the volume of metabolomics data generated in his lab is formidable. He estimated the experimental pipeline generates millions of datapoints per half second during a 10-minute analysis.

"Most of those data points are meaningless, but the difficulty is figuring out which ones are useful and not making [the decision-making method] so rigorous that it is going to throw out useful information," he said.

Jones and his colleagues have spent the last two years building a pipeline for high-throughput metabolic analysis with the

goal of answering a key question: “Would it be in fact practical to get a comprehensive profile of metabolism that one could use clinically? This is quite contrary to the usual clinical assays where we measure one metabolite at a time,” he said.

The first phase of moving toward personalized metabolomics, he said, is to not yet shoot for a complete inventory of all metabolic processes, which would involve tracking over 20,000 metabolites in a blood sample. His group is trying first to sample many metabolic pathways. “It is a broad coverage but doesn’t cover everything,” Jones said.

“We have to figure out how we can get the system rugged enough so that we can take samples from any place, put them in instruments scattered around the country and the data are going to be comparable so that they can be used in a meaningful way ... that is what we are after.”

The procedure he has worked out combines a short liquid chromatography separation with mass-to-charge analysis on Thermo’s LTQ-FT Fourier transform mass spectrometer. After analyzing the mass resolution and mass accuracy data through software tools such as Genedata’s Expressionist, he and his co-workers within minutes have been able to obtain a data table with around 2,000 mass-to-charge features, Jones explained.

“We can do 80 samples in one dataset and we have validated that it is completely reproducible [and] reliable,” he said. “Even though we don’t know what the identity of those chemicals are, we know that when we take that snapshot picture, that it’s reproducible.”

In order to ensure that clinical metabolomics samples are processed in “a very reproducible way,” Jones and his team are developing methods to standardize the process, exploring ways to get the data into a library, and “using very powerful bioinformatics and biostatistical methods to tell us how good things are and where the problems are,” he said.

Jones said the interaction with Genedata has been “a positive one” that “definitely solves a big problem.” The advantage of the Genedata software, he said, is that it “takes the dataset and has routines written to move it through in a very consistent, rapid manner, and reduce it to about 2,000 features — the same number we have reached through our other methods — which are robust. You can trust that they are real and in that sample.”

Of those 2,000 mass-to-charge features, some will be identifiable, but most will not. “The beauty of this is that it gives you relative amounts,” he said.

For example, in an experiment that looked at individuals with a sulphur deficiency in their diet, his team found 57 metabolic features that change significantly. “We know what some of them are, like methionine and direct products of cysteine metabolism, but most of them we don’t know,” he said.

‘Very Exciting Times’

Thermo’s Wilson said that projects such as those in Jones’ lab “push Thermo to be innovative in instrument development, providing researchers with the tools they need to answer important questions.”

By teaming up with Genedata, Thermo can offer its metabolomics customers a “total solution,” Wilson said. “It is crucial to have both pieces — hardware and software.”

Expressionist is able to “translate” the mass-spectral data into models and lists that can eventually point to a biological understanding,” she said.

“These are very exciting times for the metabolomics market,” said Claudio Schmid, director of Genedata’s Expressionist business, in an e-mail to *BioInform*. As mass spectrometry hardware vendors have put forward a variety of increasingly sophisticated and higher-throughput machines, they are enabling ever larger studies “with superior mass accuracy,” he said.

“With this increase in detection power and concomitant increase in both file size and number of files, metabolomics researchers require automated and scalable software systems such as Genedata Expressionist to effectively process and mine this newfound wealth of information,” he added.

Schmid said that many research areas, including drug discovery, toxicology, agrochemical, and industrial biotech are drawn to metabolomics, creating an opportunity for bioinformatics companies like Genedata.

"This has really come up pretty quickly," he noted. "Our customers across these fields are using metabolite expression levels to identify biomarkers and better classify diverse biological states in ways that have not been previously possible."

Streaming Through the Lab

Jones envisions an informatics workflow that streams the data from the mass spectrometer through the software to get a mass-to-charge data table, and then feeds that information directly into a cumulative data library.

The eventual intent "is to make [the database] publicly accessible," he said, but for now it is accessible only to direct collaborators, because data-sharing logistics have not been figured out.

"If we had this type of a database with a routine way to capture samples and profile those samples what it would mean is that for things like drug-drug interactions, which are extremely rare, we would have an exquisite way to identify them," said Jones.

One challenge, he said, involves normalization. "We don't have a good normalization procedure so we can put all of those data into a common data library, so you could feel confident you could use it," he said.

"We have to figure out how we can get the system rugged enough so that we can take samples from any place, put them in instruments scattered around the country, and the data are going to be comparable so that they can be used in a meaningful way," he said. "That is what we are after."

Another challenge is database-related. Human metabolic databases currently contain about 1,500 to 2,000 metabolites that have been characterized in human plasma, Jones said. However, he said that when he tries to match the 2,000 features his team has identified in its study to the metabolites in those resources, "on a good day it may be 15 [percent] to 18 percent, [and] on a bad day it will be 12 percent."

As a result, Jones and his colleagues are developing their own database "because we didn't trust the others as far as the curation is concerned." Of the 2,000 features he and his colleagues detect in their experiments, he finds that around 1,000 of them are present in every individual studied.

He and his team also study the results from experiments to assure they are not instrumentation artifacts but real metabolites. "My impression is they are real metabolites," said Jones. "They are just chemicals that we don't know, that have never been characterized in biology."

Unlike other teams, his group "has not sweated the fact that we don't know what these things are," he said. Rather than waiting until all the chemicals are known, which could delay the introduction of this method into the clinic, the researchers want to hone the methods and standardization so "we can get relative quantification of all of the mass to-charge features and put them into a data library."

It has not yet been determined where the library will sit physically, for now the team is building it on the Emory server. "We want to build this library so we can test it," Jones said.

Open Source and Commercial Platforms

Jones said that so far, Genedata's tools have generated very similar results as the open source tools his group is evaluating. For example, the XCMS package uses a different peak-picking algorithm than the Genedata software, but it also identified 2,000 metabolic features, he said. "That's what makes me comfortable that they are capturing the same thing."

XCMS, an open source data analysis tool developed for peak picking and alignment in metabolite studies, is written in R and is "quite widely used," Gary Siuzdak, senior director of the Scripps Center for Mass Spectrometry and developer of the software, told *BioInform*.

XCMS uses "endogenous metabolites as internal standards for aligning the data," which "has turned out to be a really neat way of solving a big problem not only in metabolomics but also in drug discovery," he said.

This approach uses endogenous metabolites to align the data with respect to all the other datasets from runs that have been performed. "It is using the ones observed in the run as internal standards," he said. "You get this non-linear alignment, which is really critical for doing good comparative analysis between datasets."

Once the alignment is completed, it statistically analyzes the different peaks. "It tells you based on a P-value which ones are most statistically significant in terms of change taking place ... so now it pulls out the interesting metabolites," he said.

The tool does the peak picking, alignment, and the statistical analysis, giving scientists the mass of the molecules and retention time of the ones changing most significantly. "Those are the ones we concentrate on for identification," Siuzdak said.

XCMS also links to the lab's [Metlin](#) database, which contains 23,000 metabolites as well as MS data to help researchers confirm the identity of a given metabolite.

Siuzdak and his colleagues recently published a paper in [Analytical Chemistry](#) describing a new version of the software called XCMS².

© Copyright 2008 GenomeWeb Daily News. All rights Reserved.