

In recent years pharmaceutical research has been able to reap huge benefits from scaling-up automation. New automated techniques for screening potential drug compounds have produced enormous amounts of data. But raw data alone is not enough. It must be translated into knowledge. Then it becomes the key factor in steering future research into successful products. The data mining team at Janssen Research Foundation (JRF) has created a competitive edge for its scientists by developing systems that speed up the delivery of this knowledge to researchers. The first thing they needed however were flexible and dynamic software solutions; solutions that could easily link existing databases and effectively integrate the specific tools used in advanced stages of data mining processes. It is no surprise that SAS ended up as their partner.

Scaling-up

The odds for success in discovering effective new medicines depend mainly on the number of compounds that can be screened for their activity against disease. With the development of modern High Throughput Screening (HTS) methods, researchers at JRF are able to screen up to 100,000 chemical compounds a day. This compares with a minuscule 50 compound a day rate with the manual techniques of only a few years ago. This has resulted in an explosion of new data, which has created in itself, something of a problem. "A lot of valuable information was hidden by the sheer volume of the data. That is why the decision was made to establish a data mining team," says Dr. Michael Engels, Principal Scientist in Theoretical Medicinal Chemistry at JRF and Scientific Project Leader of the team. "The major objective of research scientists is to develop the knowledge and skills needed to create new medicines. Their expertise is not necessarily in the mastery of the underlying statistical theories and the highly specialised calculation techniques needed for the mastering of massive amounts of data."

Existing databases easily merged

The data mining team supports all discovery research activities at JRF, irrespective of the particular areas under investigation. Their first task was to bring all the data from different divisions in one general data warehouse. "The smooth interfacing between SAS/Warehouse Administrator® and our existing Oracle-databases, together with the data manipulation and trans-

formation tools in SAS, made this job a lot easier than we expected," recalls Dr. Rudi Verbeeck, Project Leader of the IT team.

"Each research division at JRF has historically developed its own database. Biologists typically use a different approach than chemists. Pharmacologists and biochemists put their own emphasis on some aspects of data as well. Merging all these different relational databases seemed like a large and complex project when we initially analysed the problem. We were extremely pleased when we discovered that SAS/Warehouse Administrator® allowed us to access existing databases directly from within the graphical user interface. In no time, we built a dynamic data warehouse that synchronises automatically with all underlying databases. With the visualization tools of SAS, this was mainly a matter of graphical configuration. Actual coding was very limited."

Most extended statistical library

With this treasure of information transparently available in the data warehouse, the team immediately started their data mining activities. Continues Dr. Engels, "Compared with every other tool we considered, SAS/Enterprise Miner® has by far the most extended library of high-powered statistical methods. Regardless of what approach our modellers want to use, from logistic regression, artificial neural networks, or decision trees, they have it all instantly available in the visual library of SAS software. Our statistical analyses can



"SAS software allows us to translate enormous amounts of raw data into knowledge. This is the basis for our continued efforts for innovation in the drug discovery research process."

Michael Engels

PhD. - Principal Scientist in Theoretical Medicinal Chemistry,
 Janssen Research Foundation
(Right on this photo)

Left: Dr. Rudi Verbeeck,
 Project Leader, IT team

reveal previously unknown and even unsuspected relationships in data. We can then use those relationships for building accurate models for the research department. The fact that the analyses are directly performed from within the SAS software makes the data mining process particularly creative and effective. It also greatly speeds up the development of models. Since SAS/Enterprise Miner® allows us to compare different statistical methods at a glance, we have much more confidence that the models we select are the most likely ones to yield the desired result."

"Data is translated into easy-to-use applications for researchers. They do not have to be concerned about the underlying SAS code or statistical theories. Researchers don't have to be code experts, but can still benefit from continuous refinement of statistical analyses."

From data mining to web tool in a single step

"We have profoundly benefited from the perfect integration between the wide variety of SAS software tools. We have come to rely on it," says Dr. Verbeeck. "We use AppDev Studio®, the SAS visual application building tool for creating new research applications. This tool permits us to plug in the original SAS code - generated in the SAS/Enterprise Miner® - directly into the application. That way, developing a web application becomes nothing more than designing the user interface. AppDev Studio® allows us to dynamically upgrade our applications to mirror the constantly changing needs of our research organisation."

Steering future research

The flexibility of SAS software allows the data mining team to think big. Dr. Engels: "We are currently working on innovative applications to improve the rate of discovery of valuable compounds in HTS. At the moment, only about 0.1 percent of the tested compounds turns out to be active. But even with our state-of-the-art screening systems, the number of compounds available for testing far exceeds the capacity. An accurate pre-selection is necessary to increase the number of hits. The information needed to improve our suc-



"SAS software supports the whole data mining process, always combining the highest degree of versatility, customizability, and flexibility with the easiest-to-use graphical user interface."

cess rate has previously been hidden in the mass of data from former screenings. With SAS/Enterprise Miner® we can extract predictive models from this data. When we plug the models into an application, JRF researchers are able to perform virtual screenings on both virtual and existing compound libraries. On the basis of these cheap and easily obtained results, they can make a much more productive first selection of compounds to be investigated. They will be able to find many more active compounds in much less time. And the more data that becomes available, the more accurate those self-learning predictive models become."

Customizable and easy-to-use

According to Dr. Engels, SAS is the leading edge data mining software. "In our comparison with other suppliers, we found packages that are very generic and customizable. But the problem was they all required a thorough programming knowledge. We also found highly specialized, pre-programmed and easy-to-use software. But they all lacked the flexibility that is absolutely essential for successful research. For our money, SAS software combines the perfect mix of versatility, customizability, and user-friendliness with the widest range of visual approach applications. Since all SAS tools are perfectly integrated with each other, our data mining processes have become a continuous fluid movement."



SAS Institute
Kasteel de Robiano
Hertenbergstraat 6
B-3080 Tervuren
Tel.: +32 (0)2 766 07 00
Fax: +32 (0)2 766 07 77

SAS Institute s.à.r.l.
Office City
6 Circuit de la Foire Internationale
BP 2507
L-1025 LUXEMBOURG
Tel.: 00-352-264.20.410
Fax: 00-352-264.20.608

www.sas.com