



February 24, 2003

Oracle Hopes to Strike Bioinformatics Gold with Embedded Data-Mining Tools

Oracle currently claims to hold 85 percent of the life science research database market, but the company isn't resting on its laurels. On the contrary, the database giant is expanding the capabilities of its software in a bid to retain its edge in the increasingly competitive market. For the bioinformatics sector, the company is offering a suite of data-mining tools that are embedded within its database and that can provide a level of functionality many Oracle customers may not even know they already have.

One of the company's goals right now "is to make people aware of these things," said Pablo Tamayo, a bioinformaticist who splits his time between Oracle's data mining group and the Whitehead Institute Center for Genome Research. "Many times people don't know that this technology is already available — they get the database and all these capabilities are already there. They don't even have to pay for them."

Oracle's current database system, 9i, includes two supervised learning algorithms — naïve Bayes and decision trees — and three unsupervised learning algorithms — association rules, hierarchical k-means clustering, and a proprietary clustering algorithm called O-Cluster. Tamayo and his colleagues are planning a number of enhancements in the data-mining lineup for the next release, due out some time in mid-2003.

According to Oracle, embedded data analysis offers a number of advantages for bioinformatics customers: It eliminates the extra step of moving data out of the database for analysis; it does away with the data integrity and security risks of moving data in and out of the database for multiple analytical steps; and it permits users to create automated analytical workflows for pre-processing, data analysis, and interpretation right within the database itself.

Of course, Oracle's competitors also know of these advantages: IBM has signed several customers — including AxCell Biosciences and DeCode Genetics — for the Intelligent Miner software that accompanies its DB2 database, and data warehouse firm Teradata is also moving into the life science market with a full set of data-mining tools on hand. With competition heating up in the sector, these value-added features may ultimately determine whether Oracle is able to retain its current top spot.

Building Blocks

Mining its large customer base, Oracle is actively getting the word out to its loyal life science users about the benefits of its data-mining tools. The company added a special biology session to a data-mining customer advisory board meeting it held in late

January, and is collaborating closely with those customers who are still evaluating the features.

Tamayo stressed that Oracle is not in the business of designing end-user applications, but only “building blocks” that developers creating in-house research pipelines or commercial services can put together to do the “heavy lifting” at the heart of their finished product. “This is more for the people building the applications than for the biologists,” said Tamayo, who added that “the mixture [of building blocks] is probably different in each company or institution.”

Incellico, which uses Oracle to store over 200 million semantic relationships on biological entities in its CELL (Coded Electronic Life Library) database, began using Oracle Data Mining (ODM) at the end of 2002, according to Incellico CEO John Wilbanks.

The companies are currently collaborating on a project to add data mining capabilities to CELL, “and [we] hope to have some interesting results to report this summer,” Wilbanks reported via e-mail.

Wilbanks anticipates that Incellico customers will see two primary benefits from the data mining features. “First, users can leverage ODM on their raw data; for example, a user might leverage O-Cluster to generate patterns in their microarray data. Those patterns are loaded into CELL as relationships ... Second, ODM can operate on the relationships themselves, generating patterns and structure within CELL itself,” he said.

“The market has been telling us to get data mining into our set of solutions,” said Wilbanks, who added that the Oracle tools permitted the company to abandon a time-consuming in-house development effort to generate a proprietary data mining solution.

Dennis Mock, principal statistician for the Alliance for Cell Signaling at the University of California, San Diego, is also giving the data mining tools a try. Mock is using the tools to perform association studies on expression data and protein-protein interaction data for B-lymphocytes and myocytes in mice. “The statistical filtering is computationally intensive, so calculations performed directly on the data from the database will be desirable,” he said.

The Competition

Meanwhile, IBM has also signed on some early adopters to try out DB2 Intelligent Miner (IM) for bioinformatics. In August, AxCell Biosciences began evaluating the predictive algorithms in IM as part of a research collaboration with IBM “to assess the utility of the tool in the life sciences,” said Luping Lian, director of bioinformatics at AxCell. The company had already developed two data-mining packages on its own to speed population of its ProChart database of signal transduction pathway information. So far, according to Lian, IM shows promise as a means to predict the likelihood that ligands from public databases will bind with a given protein domain. Early results show a prediction success rate of between 50 percent and 70 percent, Lian said. If proven successful, this capability would reduce the amount of candidates — and the associated time and costs — for the wetlab screening process, Lian noted.

One advantage of the IBM technology, Lian added, is that even though IM is embedded in DB2, AxCell did not have to relocate the data it had housed in other database platforms because IBM's DiscoveryLink data federation middleware is able to access information from distributed resources.

Lian said that AxCell does not intend to make the Intelligent Miner capabilities available to users of its ProChart database. Right now, he said, the company is still evaluating whether to use the technology for its production-scale ligand mining work.

And IBM isn't the sole contender for a chunk of Oracle's market share. Teradata, a data warehouse vendor that has found success in the retail and banking markets, is dipping a toe into the life science waters. Its first customer, the Salk Institute, recently installed a Teradata system through its work with Teragenomics, a business unit of IT consulting firm Information Management Consulting [*BioInform* 01-27-03]. Salk Institute neurologist Carolee Barlow, who is using the database to support a gene expression atlas of the mouse brain, said that the embedded data-mining tools have significantly reduced the number of analytical steps her team had to carry out previously.

But Barlow pointed out that despite her satisfaction with the product, Teradata has its work cut out for it in the life science market. "In the life science industry, there's been a huge investment in Oracle and IBM, and that's where Teradata will have their trouble, because people don't already have Teradata in place. ... So the challenge for them is going to be, how can they penetrate a market when somebody else is there, even if they have a superior product?"

To Embed or Not to Embed

Despite a few initial positive reviews, the jury is still out on the benefits that in-database mining provides. "It seems to me it would be a good thing to have the tools as close as possible to the data," said John Elder, founder and CEO of data-mining consulting firm Elder Research. However, Elder did point out one disadvantage: "You wouldn't have the power you would have with external tools — kind of like what a doctor would bring on a house call vs. what you can do in your office. The [embedded data-mining tools] would be sort of a first-aid kit."

John Hotchkiss, CTO of discovery informatics firm AnVil, agreed. "We've been trying to understand what the value is of having this in the database vs. in an analytical tool that applies itself to the database, and we think there's some value there. ... What it isn't, though, for all their good intentions, is really a solution."

AnVil, whose business relies on in-house and third-party data-mining technologies, is using Oracle Data Mining as one step of its analytical pipeline, but Hotchkiss noted that use of the technology requires case-by-case evaluation. "You have to take a whole-system view of your analytical pipeline, and ask, 'Do I want that computation to happen on the database machine; would I rather have it elsewhere; would it be more network-intensive one way or the other?'" It could be either way."

Advantages of in-database mining, Hotchkiss said, include the ability to automate data processing. On the other hand, he said, one risk is that "you get in trouble with databases if you try to multipurpose them too much...you end up with a complex database maintenance issue and the ramifications are sort of hard to predict."

Ultimately, Hotchkiss is waiting to see some performance numbers. "One would think it would be more efficient to perform the analysis within the database, but I still have to see that," he said.

In the meantime, AnVil insists it doesn't perceive the entry of large database vendors into the data mining market as a threat to its own business. Citing the importance of domain experience in any serious bioinformatics-related data mining project, Hotchkiss quipped, "I don't think of Oracle as being in the data mining space as being competitive with us any more than a gardener is competitive with his shovel. If they give me a better shovel, great. I'm still going to have to grow the plants myself."

— *BT*

Copyright © 2003 GenomeWeb LLC. All Rights Reserved.