bioINFORM
The integrated informatics news source

May 30, 2008 | Vol. 12 No. 22

GenoLogics Shifts Product Focus to Target Growing Translational Research Market

By Vivien Marx

In a strategic shift that it has been preparing since the middle of last year, GenoLogics is moving from providing research informatics and lab-management software to becoming a translational research informatics software provider, a company official told *BioInform* this week.

Two weeks ago, the company disclosed its first comprehensive project in this area, a collaboration with the Fred Hutchinson Cancer Center to migrate its in-house biorepository and biomedical informatics software solution to the GenoLogics platform. This week, the company announced a co-development partnership with GenVault for the first module for its translational research product portfolio, a biospecimen data management system called BioVault.

Translational research informatics is a "new major focus for us," James DeGreef, company co-founder and VP of market strategy, told *BioInform*. While the company's research informatics and lab management software — such as Geneus for genomics data management, Proteus for proteomics data management, and the LIMS platform OMIX — remain core company products, GenoLogics decided to expand its scope to target a new market opportunity.

"Translational research is really driving the demand for our core products and our new products," said DeGreef.

Biomarker discovery has intensified in recent years, particularly the validation of these markers, whether for early detection, targeted therapeutics, or other applications. With the need for larger cohorts and observational studies, findings need to be validated on a larger scale. This holds true for the pharma world, too, where biomarkers are used for efficacy or safety in drug development as well to stratify a population for targeted therapeutics and clinical trials.

To meet these demands, GenoLogics is developing a multi-module biomedical informatics suite: BioVault for specimen management; BioChronicle for clinical annotations management; BioQuest for Web access; BioSphere for study management; and BioSource for electronic patient questionnaires. Although the company has officially launched the product line, most of the modules are still in production.

"There are going to be a number of modules released this year, next year and more into the future," DeGreef said.

BioVault is currently available in beta format. "We are working with early-access customers, generally customers who have our research informatics [software]. We are in major development right now," he said.

As part of this new focus, GenoLogics has hired several new employees, mainly from Canada and the US, but also from Europe, Japan, and the UK. These hires include a product manager as well as software engineers, field application scientists, and customer solutions staff to manage the early access program and larger deals.

Since it was founded in 2001, GenoLogics has focused on research institutions. But over the last year, interest from biotech and pharma companies has increased, DeGreef said, and a number of new accounts will be announced in the future, though he did not provide further details.

Translating the Data Flow

Market feedback is part of what led GenoLogics to translational research. Customers "don't want to have to build everything from the ground up and they don't want to have to deal with as many different small vendors who have a hard time integrating it all," said DeGreef.

DeGreef said that customers have provided positive feedback for GenoLogics' LIMS platform, but many were also looking for the company to help "upstream in terms of biorespository management, integrating information from clinical sources to help drive that information into the research domain, [offer] ways to query and request specimens, manage IRB [and] patient consent, [and provide options for] observational study management." That demand created the second focus for the company, he said.

This space is not void of competitors. For example, InforSense has recently accelerated its push into the translational research informatics area. Last month, the company announced a three-year collaboration with Dana-Farber Cancer Institute to develop a translational research informatics infrastructure [*BioInform* 04-18-08], and earlier this month it signed an agreement with LabVantage Solutions to integrate its software with the company's Sapphire LIMS with the intention of initially targeting biobanking management [*BioInform* 05-02-08].

> *"Translational research is really driving the demand for our core products and our new products."*

But DeGreef said that GenoLogics has more of a focus on data integration and data management than its competitors, for example pulling data and clinical information together from the hospital and the clinical trials systems, managing biorepositories, and managing lab instrument data.

"Traditionally we have provided mechanisms to get the data out into SAS or Spotfire or some of the other analytics tools, but they are not really purposed for translational research," he said.

DeGreef suggested that the company's shift in focus might bring about new partnerships, for example with analytics firms or other third-party software providers.

"There are big LIMS vendors that play into many different domains, but they are not really targeted, and then there are some smaller software vendors that might have a very focused solution around one science domain, but they wouldn't scale up," DeGreef said, "so there is sort of a sweet spot around being enterprise-enabled but having all the deep science needed for systems biology, which I think is the other side of the coin to translational medicine."

The company is also considering a move into analytics, such as building analytical data marts, providing data to statistical analysis packages, or integrating with biological modeling systems such as Ingenuity and GeneGo. "That is less of a product [area, and] more of a service [area] and working with our customers," he said.

Another aspect that played into the shift toward translational research is the search for an additional growth strategy. In recent years, GenoLogics' products have been used more for discovery work, he said, but a market survey by his company revealed that many groups are now performing validation studies on biomarkers using larger patient cohorts.

One customer, for example, is a biotech company he declined to name that uses Geneus to manage its genomics data but its main problem is finding good specimens focused on the disease area it is targeting and the clinical information around those specimens.

"That biotech wouldn't buy our new biomedical informatics system but they would be a direct beneficiary of it because it would be housed in [a] hospital or … cancer center, and [the biotech] could go to a web portal, query what specimens are there, and request them," he said.

Robustness is a big theme in translational research, and DeGreef noted that many labs have high-throughput instrumentation performing complex experiments. "You can [handle that data] on a small scale, piecemeal, without our software for sure, but with translational research and trying to validate biomarkers, there are so many confounding factors, you have to take a real enterprise approach to it," he said.

Validation is Today's Job

The Fred Hutchinson Cancer Research Center has selected GenoLogics to develop an integrated biorepository and biomedical informatics software solution for its Translational Outcomes Research group.

"That is an approach more devoted to principal investigators," DeGreef said, as opposed to "more of an enterprise approach" that another, undisclosed, customer is taking to build its biomedical informatics infrastructure.

Organizations may have biobanks at different locations and DeGreef said that as translational becomes a management focus, these groups are looking for ways to integrate these repositories.

GenoLogic's software will allow clinical and research staff to manage their specimens locally but use a web portal to integrate all the clinical information and provide "sort of a virtual biobank at the enterprise," he said.

GenoLogics believes its products can relate well to changes in the research landscape. In many cases, biobanks have in-house developed systems, perhaps built on Access or Excel. But as research teams scale-up their efforts, sharing specimens across institutes, to validate results, for example there is a renewed interest in robust systems "to be able to collaborate and share data." There is a need to use controlled vocabularies to assure consistency in collaborative efforts between several groups and to automate outside requests from external groups as well as manage the IRB-approval process

"The major driver around translational research is massive collaboration, being able to validate your findings on other people's samples, so there is a major market requirement putting together more … robust solutions that will allow people to scale up what they are trying to do," he said. In-house systems might work but may not scale well, he said.

DeGreef cited the National Institutes of Health's Clinical and Translational Science Awards as a potential driver for translational research informatics tools. NIH launched the CTSA program in 2006, and this week awarded an additional $533 million over five years to 14 medical research institutes under the initiative.

Not only will the data amounts increase as a result of this new funding, but collaboration is built in to these grants, DeGreef said.

Migrating a Growing System

Over the last decade the Fred Hutchinson Cancer Research Center has developed its own informatics system, with a grant from the Department of Defense. "We learned of the plusses and minuses of doing that," said Nicole Urban, who heads the Translational and Outcomes Research Group at the center.

That system was developed in partnership with Beth Karlan, who directs the Women's Cancer Research Institute at Cedars Sinai Medical Center. Her system predates the Hutchinson one and was developed by Oracle programmers. Karlan was going to donate it to the Hutch, "but we couldn't afford it because we didn't have Oracle programmers," said Urban. "We looked at her system and altered the specifications to meet our needs."

The problem with the "home-grown system" is that it is programmed in FoxPro, a language that that "seemed fine at the time but it probably won't be supported very much longer," Urban explained.

The fact that they have a "pretty good system" is perhaps one of the reasons GenoLogics was interested in collaborating with her cancer center, said Urban. "We actually have a comprehensive functioning system of the type that everybody wants but we needed to replace it," she said. It is "quite cumbersome" and also lacks a good querying system, because the modules of the systems are not well-linked.

For example, pulling data from blood specimens from patients who had developed resistance to chemotherapy with a focus on women with a strong family history of cancer can be done quickly but requires a database manager.

"The family history data is coming from questionnaires, and mutation status is coming from questionnaires and clinical chart review and then there is follow-up data that is in another system and then the information about the specimens is [in another system]," she said.

"We have a lot of specimens and we get a lot of requests from people to use our specimens and just trying to query the database to see if we have what they need is not easy," she said. Being able to let the investigators query the database themselves should help the center save the money it needs to pay for maintenance of the GenoLogics system, she said.

After considering another home-grown system, Urban and her colleagues realized they had staff limitations that would slow down the development process. So she and her team spoke to many vendors. "None of them seemed to quite get what we needed," she said. But as it turns out, GenoLogics had been planning to build exactly the modules that the Fred Hutchinson Center had been planning. "They can also hire a team of ten people to build this system, which we would never have been able to do," she said.

The matchmaking between GenoLogics and the Fred Hutchinson Center was set up through the Canary Foundation's founder Don Listwin, a former Cisco executive who quit his job to help accelerate research. When he asked her about her needs, Urban replied she had a database management system that "was not going to last forever." The result is a donation by the foundation of $150,000 to cover the upfront cost of the GenoLogics system, she said. The Hutch will bear the cost of the data migration and the maintenance fee of $30,000 a year.

Urban said that the Hutch would have preferred to work with an open source software vendor, which GenoLogics is not, so the maintenance contract with the firm is a bit of a concession. She said she is also unsure if the GenoLogics system will be flexible enough as the center grows.

However, she said that it so far appears that GenoLogics is building tools that will follow the development at Fred Hutchinson, such as adding new data elements to the dictionary. "Even though we don't have access to the source code, we will have tools to continue to develop the system ourselves," Urban said.

For now the system is being built according to the Fred Hutchinson Center's specifications as well as those of other clients, "but I don't think any of the other clients have asked for anything we haven't asked for," she said.

Of late, what is becoming increasingly clear, is that annotated specimens are a necessity, she explained. "Getting all that information into the database is part of the work, but getting the information back out of the database in an easy way is also important," she said. That step involves quality control and heeding patient confidentiality. "We have to be careful that it is not possible to identify an individual based on the annotation."

Next on the schedule is a data migration phase. That will probably be a year-long process, requiring a deep understanding of the data dictionary. "It's a big deal and in a way it is the scariest part of it," she said.

For now, the in-house and GenoLogics systems will be run in parallel. The center already has a backup system in place, a mirror database. "We will do the same thing with the GenoLogics system," she said, adding that the systems will run concurrently "until we're comfortable."

Carnegie Mellon Team Tackles Challenge of Homology Analysis for Multidomain Proteins

By Vivien Marx

A new homology prediction method developed by researchers at Carnegie Mellon University has raised questions about the applicability of commonly used sequence similarity tools such as Blast for analyzing the evolution of multidomain proteins.

The method, called Neighborhood Correlation, was developed specifically to deal with the challenge of multidomain proteins — proteins comprised of multiple sequence segments. While these proteins represent around 40 percent of the proteome in metazoans, they present a hurdle for current homology analysis tools because it is difficult to determine whether a common sequence is the result of shared ancestry or of domain insertion.

Neighborhood Correlation, described in a recent *PLoS Computational Biology* paper, relies on a sequence similarity network that is weighted to give gene duplication and domain insertion very "neighborhood structures," which enables the method to distinguish true homologs from domain-only matches, according to the authors.

In the paper, Dannie Durand, a computational biologist at Carnegie Mellon, and colleagues demonstrated that the method outperformed sequence similarity methods like Blast and Psi-Blast against a curated benchmark data set of sequences known to share common ancestry.

"The paper really tackles a fundamental problem that people have been hoping to avoid," said David Haussler, director of biomolecular engineering at the University of California at Santa Cruz and Howard Hughes Medical Investigator. "No one has really faced it head-on before so I commend [Durand] for [that]. … She puts her method up against some others and demonstrates better performance."

Haussler noted that current methods either filter out multidomain proteins or ignore them, which can lead to spurious results. "If the same domain is inserted into two different proteins, that doesn't make them related — it makes the little domain part of the proteins related, but it doesn't make the proteins as a whole related. And the aligner will blissfully align those two things and suggest that these two proteins might be related," he said.

Neighborhood Correlation takes a geographical view, looking at the genomic neighborhood, Durand explained to *BioInform*. "Basically we make a network … in which every dot or node is a sequence … a line between two dots means there is a meaningful Blast score," she said. "If the neighborhoods are similar, the genes are related, if they are not similar, we say the genes aren't related."

The method adds a step to a normal Blast alignment in order to look at genes in this geographic context. The scientists also developed a visualization tool for their method but have not published it yet as they are still working the "the kinks out," said Durand. She said that this visualization tool is built on Google Earth and that she plans to make it generally available for researchers to navigate their data.

The genomic network is organized such that it tells the evolutionary history of the genes in a given region, enabling scientists to ask other questions concerning homology, Durand said.

She added that Neighborhood Correlation expands the possibilities for the genomic analysis of modular gene domains by distinguishing multidomain homologs from unrelated sequence pairs that share a domain. It adds a statistical evaluation step to alignment with Blast or PSI-Blast and "gives pairwise scores … [offering] an additional parameter for triage."

By looking geographically at this network, "we are able to separate pairs that are related and pairs that aren't, where the more traditional methods haven't been able to do that," Durand said.

"Users who will find this [method] most useful are people who have to do large-scale comparison, [who] can't look at every pair individually, and [need to] make a judgment call," she said. Scientists running Blast will find "the additional cost of running the neighborhood score is very low," she said.

Building a Benchmark

To test the method, the Carnegie Mellon team created a hand-curated benchmark data set of 1,577 sequences from 20 families of known homology, covering a range of protein functional categories such as neural development, immune response, signal transduction, and enzymes. They created two sets of sequence pairs, constructing a test set of pairs with 853,465 known positive examples for homology and 40,459,204 negative ones.

The goal in building the benchmark was to have a trusted set of both known sequence pairs that are homologous and sequence pairs that do not share ancestry. The Carnegie Mellon team set out to create it so that it would test a range of homology detection problems, in single-domain as well as multidomain families.

"You're putting your head in the sand if you treat all homologies as being simple in the sense that they cover the entire extent of the protein."

The evidence for homology in the set was curated from the scientific literature, mainly by Nan Song, the study's first author.

"Although comprehensive datasets are available for testing methods for predicting homology of *individual* domains, we are unaware of any other gold-standard dataset of known multidomain families with variable domain architectures," the authors wrote in the paper.

"My prediction is that the lasting contribution of this paper will be from establishing this benchmark dataset, defining the problem rigorously, identifying it, and setting up a framework that others, I am sure, will follow, carefully testing their algorithms," said UCSC's Haussler.

In a comparison using the entire benchmark data set, the scientists state that Neighborhood Correlation "dramatically" outperformed Blast, PSI-Blast, and Domain Architecture Comparison, or DAC, a method that compares sequences based on their domains.

In a subset of the benchmark set, the kinase family, Neighborhood Correlation yielded better results than Blast and PSI-Blast and slightly worse results than DAC.

For individual protein families, Neighborhood Correlation outperformed all three methods, perfectly classifying twelve out of 20 families. Blast and DAC perfectly classified seven families, PSI-Blast perfectly classified eight.

PSI-Blast did well with single-domain families but performed poorly on complex multidomain families and sequences with promiscuous domains, which are domains that move around a great deal, the authors noted.

The scientists believe Neighborhood Correlation delivers the most reliable and consistent performance on large heterogeneous datasets, and is "particularly well suited to automated genome-scale analyses which require that a single classification threshold be suitable for the vast majority of sequence pairs in a genomic dataset," they wrote.

As an example of the limitations of sequence-similarity methods, the authors cite a homologous pair of proteins, PDGFRB and PRKG1B, and a pair that shares domains but does not share common ancestry, PDGFRB and NCAM2.

Based on pairwise alignment, both sets of proteins appear to have similar properties in terms of Blast scores, alignment length, and the number of shared domains — results that would leave such methods unable to distinguish the related pair from the unrelated pair. However, the neighborhoods of these pairs in the weighted sequence similarity network are "very different" the authors note. For example, the shared neighborhood of the kinase homologs PDGFRB and PRKG1B is 779 sequences and the pair has a Neighborhood Correlation score of 0.65, while the shared neighborhood for PDGFRB and NCAM2 is 242 sequences and the pair has an NC score of 0.29.

"Unlike sequence comparison, this clear difference in neighborhood structure can be used to recognize multidomain homology," the authors write.

"Maybe pairwise comparison is just never going to tell us [about homology] in a consistent way," Durand said. "So let's not just ask these pairs, let's ask their friends, too."

Not Always Family

Identifying homologous genes that encode proteins is a task with many pitfalls: similar sequences may be the result of convergent evolution, in which they have evolved independently to fulfill the same function. That doesn't automatically mean relatedness. Or proteins might be related but have become so dissimilar that their common ancestry is almost undetectable — a problem that scientists address by looking at protein structure. "Even after the sequences no longer have any similarity, the [protein] structures do," said Durand.

Multidomain proteins were long considered an oddity, but whole-genome analysis has revealed how prevalent they are — making up approximately 40 percent of the metazoan proteome.

Domains shuffle over the course of evolutionary history, and some move around extensively and are called promiscuous domains. "Those are the domains that are inserted [into genomes] a lot," Durand said.

These promiscuous domains can lead to significant sequence similarity but carry little information about gene homology, the scientists point out in their paper. Even in protein families with conserved domain architectures, "promiscuity can confound reliable detection of homologs," the team wrote.

Domain shuffling in general presents a can of worms when searching for an ancestral genome in comparing sequence pairs. As UCSC's Haussler explained, ideally one might like to work backwards in time to an ancestral protein. But not all of the parts of a protein share the same ancestral history. "You're putting your head in the sand if you treat all homologies as being simple in the sense that they cover the entire extent of the protein," he said.

While this is a "simple picture that most people have tried to adopt," it "ignores the fact that pieces of it can have been inserted from other places independently in one lineage or the other," he said.

Neighborhood Correlation approaches the homology challenge of multi-domain proteins by ruling out two unrelated genes that happen to have the same inserted domain because they don't help build a comparative map, said Durand.

If genes look like homologs, then the genes in the region to the left and right might also be homologous, she said. "Basically, what you are saying when you do that, is that a gene is a witness for the history of the genomic region around it," Durand said.

"I don't think the neighborhood correlation is the only way to approach this, … but the importance of the paper to my mind is that Dannie Durand has really focused us on the critical problem with these protein homologies," said Haussler.

Martijn Huynen, a researcher at the Center for Molecular and Biomolecular Informatics at the Nijmegen Centre for Molecular Life Sciences who studies the evolution of genomes and protein families, agreed that the Neighborhood Correlation approach could be useful. However he is doubtful about trying "to push these proteins into classification schemes, [such as] 'these are homologous genes and these are not homologous genes,'" he said.

His concern is that the evolutionary detail about the proteins may get lost. In some cases it may be necessary to not use the term "homology" but rather to state which gene domains are shared and which are not.

Haussler noted that while looking at network connectivity of proteins is a "reasonable way to resolve" the problems with multidomain protein homology that many people "try to shove under a rug," the approach "won't be the last word" in the field.

"There are a number of approaches and I am anticipating that the authors of these methods will come back with their own take, that this will be a lively area of investigation," he said.

Homolog predictions based on the Neighborhood Correlation method and a web-based visualization tool can be found here.

NIST Team Building Peptide Spectral Library To Improve Mass-Spec Proteomics Analysis

By Bernadette Toner

Researchers at the National Institute of Standards and Technology are building a reference library of mass spectra that they hope will be a valuable resource for improving peptide identification in proteomics experiments.

The NIST team plans to introduce an updated version of the library at next week's American Society for Mass Spectrometry conference in Denver, along with a new "hybrid" search tool that combines a spectral search algorithm called NIST MS Search that NIST developed with the National Center for Biotechnology Information's Open Mass Spectrometry Search Algorithm sequence search algorithm, known as OMSSA.

Paul Rudnick, a researcher in NIST's Mass Spectrometry Data Center, told *BioInform* that the expanded library and the hybrid search tool should be broadly available to researchers here a few weeks after ASMS.

Steve Stein, who directs the NIST Mass Spectrometry Data Center, began developing the spectral library a little over five years ago as an offshoot of NIST's Electron Ionization Library, a collection of mass spectra for small molecules that has been in use for around 25 years.

Spectral libraries are commonly used to identify molecules in chemical applications, but Rudnick noted that since proteomics is still a relatively new field, similar collections haven't been widely available for peptide spectra.

Instead, rather than matching measured MS/MS spectra directly against a spectral library of known peptides, the most commonly used peptide-identification algorithms, such as Mascot, Sequest, X!Tandem, Phenyx, and the like, match measured spectra against pre-calculated "theoretical" spectra derived from peptide sequences.

This approach is effective, but it has a number of limitations. For example, it's well known that there is very little overlap in the protein lists that these sequence-search algorithms identify — a characteristic that has led to a recent trend in the field to merge the results of multiple search algorithms into a consensus protein list.

In addition, sequence-based searching is very computationally intensive, especially for large-scale proteomics experiments, and a lot of that time is wasted by identifying the same peptides multiple times.

Finally, the "theoretical" spectra that underlie these algorithms don't account for many informative features of the experimental spectra, such as relative abundances and ratios of products with different charge states, which could be used to identify peptides more accurately.

Spectral library searching promises a more precise and efficient option than sequence searching because it directly matches an acquired MS/MS spectrum to a library of previously observed and identified peptide spectra. This method is at least twice as sensitive and up to several-hundred-fold faster than the sequence-based approach, but the challenge, Rudnick said, is building the library and ensuring that it is optimized for searching.

NIST's library is compiled from in-house MS experiments and collaborations with several large proteomics data repositories, such as PeptideAtlas.org, Tranche, the European Bioinformatics Institute's PRIDE (Proteomics Identifications) database, and the Global Proteome Machine Organization, which is collecting spectra from the proteomics community via an online deposition model.

UBC's Beavis doesn't see spectral searching as "totally replacing the existing search engines — just changing the way

The NIST library includes peptide spectra for human, mouse, rat, fly, and yeast. For human, there are nearly 224,000 spectra from around 138,000 peptides, which represents around 14.5 percent of the human proteome, Rudnick said. Coverage of the yeast proteome is the highest, at around 22 percent, while less than 1 percent of the rat proteome is represented in the current library.

Rudnick noted that this low coverage is the biggest drawback of the spectral searching approach because the algorithm can only identify spectra that are in the library.

Rudnick said that his group is looking to increase the coverage level *people think about analyzing the data."*

of the library, but noted that there are a "lot of considerations" to determine which data sets to target first.

"You need to catch the right diversity of tissue samples, of blood samples, and you want to make sure you're catching all the proteins in a developmental or disease pathway," he said. "We still don't know when and where all these proteins are expressed."

One goal of the NIST group, therefore, is "to get involved with people who really know how to do protein and peptide fractionation and who have access to the right samples" in order to build the library in a more systematic fashion, said Rudnick. In the meantime, "we're taking a really wide-reaching approach," he added.

The aim, he noted, is to create something that's "usable for most applications," including 1D or 2D separations and a range of mass-spec platforms. So far, he said, spectral searches perform better than current practice, but "now we'd like to see the library include all biologically relevant peptides, including modified forms."

For now, the NIST team is addressing that challenge via a hybrid approach that blends its NIST MS Search spectral search tool with the OMSSA sequence search algorithm. The NCBI algorithm "piggybacks our library searching" to identify peptides which are not yet in the library, Rudnick said.

An Alternative, Not a Replacement

Despite its potential promise, spectral searching won't likely ever completely replace sequence-based approaches, according to Ron Beavis, a professor at the University of British Columbia's Biomedical Research Centre and a founder of the Global Proteome Machine project.

Beavis, who developed the X!Tandem algorithm for sequence searching as well as its spectral-searching counterpart, X!Hunter, said he doesn't see spectral searching as "totally replacing the existing search engines — just changing the way people think about analyzing the data."

The main benefit of the spectra-based approach is speed, Beavis said. "What takes the most time in sequence-based searching is calculating the theoretical mass spectrum," he said, "but in the case of the library, you already have that."

In comparisons of X!Hunter to X!Tandem, the spectral library search method is typically 200 times to 500 times faster than sequence searching, and is sometimes up to 1,000 times faster, Beavis said. In addition, he noted, "You can usually find data that's further down in the noise than you could with sequence searching. It tends to dig down deeper and faster."

However, the downside to the approach is that "you can't find anything [the library] doesn't already know about," which makes it less suitable for exploratory experiments, or for organisms that are not well characterized.

In practice, Beavis said he typically runs proteomics data through X!Hunter first "to find out what it looks like." Then he uses the results from several X!Hunter runs with different parameter settings to "tune" X!Tandem for a more complete search. Because X!Hunter is so fast and can run on a laptop, that initial processing step is nearly negligible, he said.

Ultimately, Beavis said that the choice of algorithm will probably depend on the type of experiment a researcher is conducting. In a case where an investigator is looking for a particular set of known peptides, as in some clinical applications, "there's not much point in using a conventional search engine," he said. However, "if you want to do a proteome-wide study and you're looking for post-translational modifications or splice variants," it would probably be best to use a sequence-based approach.

Beavis acknowledged that even as groups like GPM and NIST rapidly expand their peptide spectra libraries in order to improve the performance of their algorithms, adoption of spectral searching in the proteomics community has been sluggish.

"I know from personal experience in this world that it takes people about five years to adopt anything, and we're still pretty early in that cycle" due to the inherently conservative nature of the proteomics community, he said.

"Because of the large investment that people make in equipment in this area, they're used to taking quite a bit of time to evaluate things," he added. "But when you're spending millions of dollars on equipment, it's probably good to be conservative."

Further information on the NIST and GPM spectral library projects, as well as several other similar initiatives, is available here.

NSF Bioinformatics Grants Awarded April 4 — May 28, 2008

Gold Standard Set for Functional Annotation of Microbial Hypothetical Proteins. Start date: May 15, 2008. Expires: April 30, 2009. Awarded amount to date: $179,752. Principal investigator: Eugene Kolker. Sponsor: Children's Hospital and Regional Medical Center.

Supports a project to create a "gold standard" reference set of annotated genes that were previously designated as hypothetical. This set will serve as a baseline for validating new approaches to functional annotation of uncharacterized genes, according to the grant abstract. The investigators note that hypothetical genes comprise around one-third of all predicted genes, "and their number is expanding rapidly with each new genome sequencing project." The goal of the project is to build a reference set of 200 previously hypothetical genes with high quality functional annotation.

Exploring Timescale Parallelization for Long-timescale Molecular Dynamics. Start date: May 15, 2008. Expires: April 30, 2009. Awarded amount to date: $103,063. Principal investigator: Srinivas Aluru. Sponsor: Iowa State University.

Supports development of timescale parallelization methods that will extend the timescales for molecular dynamics by orders of magnitude while retaining atomic level details, according to the grant abstract. The investigators will make the methods available as part of a new version of the LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) software package.

Identifying Selection Pressures on Viral Genomes. Start date: June 1, 2008. Expires: May 31, 2009. Awarded amount to date: $188,803. Principal investigator: Simon Frost. Sponsor: University of California-San Diego.

Supports development of likelihood-based statistical models of molecular evolution to study evolutionary patterns from alignments of coding sequences. The proposed models will allow the pattern of substitution to vary across the phylogenetic tree, across sites in the alignment, and to depend on the residues involved, according to the grant abstract. The models also account for recombination by permitting phylogenetic histories to change between regions of an alignment. "These evolutionary patterns will be derived from the data, rather than from *a priori* assumptions, through the innovative use of genetic algorithms to search for good-fitting models," the investigators state in the abstract.

Complex Biological Systems Across Multiple Space and Time Scales. Start date: June 1, 2008. Expires: May 31, 2011. Awarded amount to date: $1,446,022. Principal investigator: Beatrice Riviere. Sponsor: University of Pittsburgh.

Supports the development of algorithms for solving complex spatio-temporal problems that arise in biology and the application of these and other methods to neuroscience and inflammation research. In particular, the project will develop "effective discretizations of coupled biological processes, iterative solvers, and coupling/decoupling strategies," according to the grant abstract.

Statistical Methodology and Applications to Genetics, Engineering and Economics. Start date: July 1, 2008. Expires: June 30, 2011. Awarded amount to date: $587,375. Principal investigator: Tze Lai. Sponsor: Stanford University.

Supports development of an improved method for searching a parameter set for the location of one or more signals in a noisy background — a problem that occurs in gene mapping, bioinformatics, and other applications, according to the grant abstract. This problem arises in gene mapping, brain mapping, bioinformatics, and astronomy. The project plans to use parallel recursive algorithms, new ideas in change-point detection, and empirical Bayes methodology to develop asymptotically efficient estimation and prediction schemes with "manageable complexity." One goal of the proposed research is to develop a comprehensive theory of efficient parameter estimation, filtering, and smoothing in hidden Markov models on general state spaces, the abstract states.

JMP Genomics 3.2, HT-BLIS, caBIO 4.1, BIRCH 2.4, Progenesis SameSpots v3.0

SAS has launched **JMP Genomics 3.2,** which streamlines analysis workflows
for expression, exon, copy number, and genotype data. Among other changes, the product has a simplified user interface and includes new support for importing copy number CHP files from the Affymetrix Genotyping Console software and for importing expression data from Illumina BeadStudio software, the company said.

Biotique Systems has launched **HT-BLIS,** or High-Throughput BLIS, a version of the company's Biotique Local Integration System software that allows scientists to integrate, analyze, and share data generated by next-generation sequencing machines. The platform is scalable and able to receive, store, and manipulate large amounts of data in real time, Biotique said.

The **National Cancer Institute Center for Bioinformatics** has released the Cancer Bioinformatics Infrastructure Objects, or **caBIO, 4.1**. Among its new features is a refactored location model; the possibility of dynamic querying for genomic features; the possibility to access data in the Cancer Gene Index, which identifies cancer gene disease and cancer gene-compound associations to be found in PubMed abstracts; the redesign of the marker class as a UniSTS-centric object; the addition of NucleicAcidSequence to better meet caBIG modeling standards; and an interface to allow querying of microarray annotations.

The **University of Manitoba** has released **BIRCH (Biological Research Computer Hierarchy) 2.4** here. The programs and databases in a unified environment perform interconverting of file formats and other 'behind the scenes" tasks.

New features include text editors to display program output and an upgraded fasta package.

Nonlinear Dynamics has released **Progenesis SameSpots v3.0.** The latest version of the 2D gel analysis software includes such features as: flexibility in experimental groupings, statistical tagging tools, a new spot picking workflow, and concise reporting. It is also possible to add images to an existing analyzed experiment, which extends the scope of the experimental design, the company said.

Francis Collins, Carlos Brody, James Collins, Michael Eisen, Jonathan Pritchard

Francis Collins, director of the **National Human Genome Research Institute**, said this week that he will step down from his post on Aug. 1 to focus on "writing projects and other professional opportunities."

Collins has been director of NHGRI since April 1993 and led the Human Genome Project to its completion in 2003. He also guided a number of follow-up initiatives, such as the International HapMap Project, the Encyclopedia of DNA Elements, the Knockout Mouse Project, the Mammalian Gene Collection, the Cancer Genome Atlas, the Molecular Libraries Initiative, and the Human Microbiome Project. Collins also founded an intramural program in genomics within the **National Institutes of Health** in 1993.

Alan Guttmacher, deputy director of NHGRI, will be appointed acting director of the institute on Aug. 1. NIH said that it will soon begin a formal search process for a permanent NHGRI director.

The **Howard Hughes Medical Institute** has named fifty-six new biomedical scientists as HHMI investigators.

Among the new investigators are:

Carlos Brody, associate professor of molecular biology at **Princeton University; James Collins**, professor of biomedical engineering at **Boston University**; **Michael Eisen**, associate professor of molecular and cell biology at the **University of California, Berkeley**; and **Jonathan Pritchard**, professor of human genetics at the **University of Chicago**.

Brody's work has focused on computer models of neurons in the prefrontal cortex and the decision-making process in that brain region. He created a computer model using data from monkey studies in which the animals were trained to make decisions and select choices among options to simulate how a neuronal process flips from storing information in short-term memory weighing the options at hand to enabling a decision.

Collins' lab has several focus areas. He is working on a combination of computational and experimental methods to reverse engineer gene regulatory networks in microbes and higher organisms. In the field of synthetic biology, he and his lab construct synthetic gene networks for several biotechnology applications.

Eisen's lab includes geneticists, biochemists, and computer scientists who study fruit flies and other organisms such as fungi to learn how the genome specifies animal form as well as their shape. They look at the evolution of gene expression, its regulation, and its variation. Eisen also spearheaded the effort to sequence twelve *Drosophila* species and co-founded the Public Library of Science, which publishes peer-reviewed open access journals.

Pritchard studies various aspects of human genetic variation, applying mathematical and computational techniques. He uses computationally intensive approaches such as Markov chain Monte Carlo to examine large datasets to look at gene mapping of complex traits, the history of human populations, and the factors that contribute to genome variation and evolution.

EBI, Johnson & Johnson Pharmaceutical Research, Simcyp Simulator, GeneBio, Protagen, Albert Einstein College of Medicine of Yeshiva University, Montefiore Medical Center, Inserm

EBI to Lead $7M Project to Create Sustainable European Bioinformatics Infrastructure

The European Commission has awarded €4.5 million ($7 million) to a consortium led by the European Bioinformatics Institute that will integrate Europe's biological data resources into a "sustainable, integrative bioinformatics network for the life sciences," EBI said this week.

The project, called the European Life-science Infrastructure for Biological Information, or ELIXIR, involves 32 partners from 13 countries and aims to create a long-term infrastructure for biological information in Europe.

Biological databases currently depend on "insecure or short-term funding, meaning that the valuable data they contain and provide access to, are jeopardized when funding ends," EBI said in a statement.

EBI said that ELIXIR aims to address this issue by creating an infrastructure for storing, accessing, and integrating biological data that will be supported by a "secure funding mechanism."

The project, funded under the EU's 7th Framework Program, stems from a roadmap published in 2005 by the European Strategic Forum for Research Infrastructures that identified 35 large-scale European infrastructure projects, including the creation of a "shared platform for data resources in the life sciences."

EBI submitted the proposal for ELIXIR in response to a restricted call to coordinators of these 35 projects.

The activities of ELIXIR are separated into two phases over three years: a preparatory phase, in which EBI will consult with the "stewards and users of Europe's current biological resources to define the requirements, attributes and structure of the network;" and the implementation phase, in which EBI will report its findings from the preparatory phase and will seek "sustainable financial support" from national funding bodies.

According to the EBI, "key issues" that it seeks to define in the community consultation phase include:

- how best to integrate "core and specialist" data resources, including the development of standards for emerging fields;
- how the network can effectively connect with other interdisciplinary research areas;
- how to meet the needs of related European industries; and
- how to train life science researchers so that they can effectively use the information made available through the infrastructure.

EBI officials were not able to comment further on the project before deadline.

Further information on ELIXIR is available here.

Johnson & Johnson PRD Licenses Simcyp Simulator

The drug development arm of Johnson & Johnson, J&JPRD, has joined the Simcyp Consortium, which provides access to Simcyp's Population-based ADME Simulator for pharmacokinetics modeling.

The simulator is a software platform to model drug absorption and disposition in virtual populations to help assess potential drug-drug interactions and to identify individuals potentially at extreme risk for adverse reactions from a given drug, the company said.

Sheffield, UK-based Simcyp also announced that it will be providing its consortium members access to Simulator's population-based ADME Simulator and Paediatric module, geared toward pharmacokinetic modeling in infants, neonates, and children.

According to Simcyp, prior to signing the deal for three server licenses, 11 J&JPRD ADME scientists from multiple sites evaluated the simulation software, including the pediatric module. The contract between Simcyp and J&JPRD, a division of Janssen Pharmaceutica, will run for three years.

Financial terms of the agreement were not provided.

GeneBio to Distribute Protagen's Modiro

Geneva Bioinformatics, GeneBio, said this week that it will distribute Protagen AG's software tool Modiro for the rapid and automated detection of post-translational modifications in MS/MS datasets.

GeneBio and Protagen plan to "collaborate closely" to link Modiro's functionalities to GeneBio's Phenyx software platform for MS data analysis, GeneBio said.

GeneBio will have a global distribution range for the package.

CTSA Award for Albert Einstein College of Medicine of Yeshiva University and Montefiore Medical Center

The National Institutes of Health has awarded Albert Einstein College of Medicine of Yeshiva University and Montefiore Medical Center a Clinical and Translational Science Award totaling $22 million over five years.

The award is one of 14 CTSA grants totaling $533 million that the NIH awarded this week.

The grant will support the new Einstein-Montefiore Institute for Clinical and Translational Research.

Among other goals, the award will support development of a centralized infrastructure that will integrate the information systems and research databases at the two institutions, Brian Currie, co-director of the institute, vice president and medical director for research at Montefiore, and assistant dean for clinical research at Einstein, said in a statement.

Researchers Propose ISBN-like Identifiers for Biobanks

Writing on behalf of a collection of European research networks, a pair of scientists proposed in last week's *Journal of the American Medical Association* that the international research community develop an identification scheme for biobanks that would be similar to the International Standard Book Number, or ISBN, code used in book publishing.

The authors note that biorepositories are "key tools in biological research, especially in genetics, genomics, and postgenomics research," and particularly genome-wide association studies. Since these studies require large numbers of samples to ensure statistical rigor, researchers would benefit from data residing in other biobanks around the world. Currently, however, tracing these resources is "difficult because there are no standardized operational tools," according to the *JAMA* commentary.

Francine Kauffmann and Anne Cambon-Thomsen, both of France's Inserm, explain in the article that "the easiest way to trace the various uses of a given collection could be a common identifier attributed at the early stages of the collection," but note that "no such structured identifier seems to exist for biobanks, whereas it is of common use in other domains."

Kauffmann and Cambon-Thomsen propose that each biobank be assigned a unique, structured numerical code, similar to the ISBN, which comprises multi-digit "elements" that identify the language or geographical region, the publisher, and the specific edition of the publication.

In the case of biobanks, the elements of the code could identify the host of the biobank, its home country, and the individual collection, the authors write.

"Identification through ISBN for books is quite sophisticated and allows, for example, easy identification of a book, various volumes of a book, a book with several volumes, and the translation of a book, all of which represent various publication elements of one book," Kauffmann and Cambon-Thomsen write. Using a similar approach for biobanks, "it would be easy to identify a biological collection, the various surveys in a longitudinal study, or an international program made of collections built up in various countries."

In practice, they add, a unique identifier for biorepositories "would allow individuals to trace uses and results for various research partners or other interested parties; would provide a way to trace what is done with the information and the samples that individuals have provided for science; and would allow genetic data to be put in context with other dimensions of research to be considered for potential secondary users."

The authors acknowledge that there are a number of practical issues associated with adopting an ISBN-like system that have yet to be addressed, including "who would maintain the system, who would provide funding for the system, and how uptake would be ensured."

Several large-scale European research projects support the proposal and helped prepare the JAMA manuscript, including GA2LEN (Global Allergy and Asthma European Network) and PHOEBE (Promoting the Harmonization of Epidemiological Biobanks in Europe), which are funded under the European Union's Sixth Framework Program; and GEN2PHEN (Genotype to Phenotype Databases: a Holistic Solution) and BBMRI (Biobanking and Biomolecular Resources Research Infrastructure'), which are funded through the EU's Seventh Framework Program.