



March 14, 2008 | Vol. 12 No. 11

SciWit, NaCTeM Tailor Text-Mining Tools For Varying Needs of Biomedical Research

By Vivien Marx

As new methods for analyzing textual data continue to emerge, a new company and an academic venture have each taken steps to tailor that growing toolkit to the varied needs of the biomedical community.

Seeking to fill a business-to-business niche, former OmniViz principals Jeffrey Saffer and Vicki Burnett recently founded SciWit, a Boulder, Colo.-based startup that offers customized information-mining solutions based on a range of methods, including natural language processing, statistical analysis, and parsing techniques.

Meanwhile, the UK's National Center for Text Mining at the the University of Manchester is taking the open access route to collaboratively create a software toolbox for text miners from publicly available tools.

Both efforts exemplify a trend that has emerged within the text-mining community in recent years: a growing awareness that there is no single out-of-the-box package for handling every text-mining task, and that the best approach is often a combination of different tools tailored to fit a particular problem.

SciWit's Saffer told *BioInform* that the company takes a consulting-based approach in which it first spends time with a client and then puts together a computationally layered text-mining process customized for the client's needs.

"What we have as a fundamental philosophy is to really understand the problem and create the tools that solve that problem," he said. "I don't like being boxed in to say, 'This is a tool kit we have.'"

Jun'ichi Tsujii, NaCTeM's scientific director, said that SciWit appears to be part of a new trend in which text-mining companies are moving toward becoming solution providers, instead of selling a fixed set of tools.

Tsujii, who also has a joint appointment at the University of Manchester and the University of Tokyo, said that this is in line with demand from life science customers, and that NaCTeM is working with several pharmaceutical companies that choose to "combine components provided by companies or provided by us or other research groups to then construct their own system."

SciWit is targeting pharmaceutical firms for its customized offering, as well as biotech and chemical companies, consulting firms, and other information-analytics providers who might want to embed some of the company's customized algorithms into their own solutions. Initial customers include software and service providers, Saffer said.

Saffer said that a typical SciWit project might be to help detect biomarkers from a wealth of information such as gene-expression or proteomics data. For the first phase of the company's development, however, the emphasis is on text analytics based on quantitative linguistics.

Burnett and Saffer are both former bench scientists from Pacific Northwest National Laboratory and previously founded data-visualization company Omniviz as a PNNL spin-out. They left six months after

Omniviz was acquired by life science software firm BioWisdom last year. [[BioInform 02-02-07](#)]

The firm is privately held and has seven employees including four PhD-level scientists. The company considers its familiarity with scientific problem solving as an advantage, said Burnett. "One of the reasons we can do this is because of our history of being those end-users."

Saffer, who describes himself as "a molecular biologist who became an informaticist out of necessity," said that as he pored over newly sequenced microbial genomes while head of the molecular biosciences department at PNNL, he felt stymied by the task of needing to understand a few thousand new genes all at once.

"I was very uncomfortable that we were still doing data analysis with the statistical analysis of one fact at a time and missing the big picture," he said. As part of his job, he said he acquainted himself with information-mining capabilities at PNNL, which were built mainly for the intelligence community. Out of that effort grew the technology behind OmniViz, and now SciWit.

The company's services are based on four main technologies. These include TopicalitySleuth to find topical terms most relevant for a client in a given set of documents. It delivers a list of key themes in a document, quantifying their importance. EmergenceSleuth, meanwhile, is set up to discover both emerging and disappearing concepts in texts in order to track scientific trends in the literature; ConceptSleuth quantifies complex business analytics in a document collection; and MarkerSleuth is a pattern-recognition and visualization tool for numeric data.

These products are processes more than algorithms, explained Burnett. "We don't mean one little equation; it is an entire approach, step by step, a quantitative approach, and in most steps an automated quantitative approach," she said.

The firm also does custom algorithm development to integrate its solutions into a client's workflow. As Burnett explained, a customer may already have natural language-processing capabilities and might tap SciWit to create the next step in its text-mining pipeline.

Competitors in the text-mining market include Linguamatics, Temis, and Connexor, but Saffer said that SciWit's customization approach sets it apart from these firms, which sell pre-designed components. And unlike the Omniviz product, which enables exploratory analysis, the SciWit method gives "definitive, actionable answers," he said.

"We asked SciWit to teach our computers how to read ... and I think they did an excellent job of that."

Unlike the market positioning with OmniViz, SciWit's target customers are not individual researchers but rather other information analytics businesses. The firm is seeking customers who say, "We can't find a company that sells a widget to solve this," said Saffer, who described the firm as mainly a B-to-B provider helping others to fine-tune their computational problem-solving in text mining. "Customization is a very important part of what we do," he said.

One ConceptSleuth customer is Michael Orlando, who runs Denver-based consulting firm Economic Advisors. In a former capacity he had, in his words, "a computationally intensive characterization problem," which was part of a business analytics contract. It required quantifying textual content across a number of dimensions. "We were trying to assess perceptions of firms," said Orlando, whose background is in engineering and economics.

Typically that is done manually, he said, developing rules and breaking down a text according to certain criteria. His company had applied basic content-analysis techniques with a team of people picking up on subtle patterns and scoring English-language documents along predefined dimensions. Rather than use human scorers, his firm wanted to bring the project to "commercially viable scale," he said.

"We asked SciWit to teach our computers how to read ... and I think they did an excellent job of that," he said. "They came up with a much more precise way of getting at what we wanted in an automated fashion."

What he found valuable was how SciWit formally reframed the company's challenge for automated content analysis. His company saw its own value in developing, interpreting, and delivering recommendations based on analytics to its clients, so it outsourced this project because it did not wish to scale up its own computational-development arm. "Being the one able to technically code that up on a regular basis isn't something we saw in our market space," he said.

In this project, explained Saffer, the quantitative linguistic approach quantified the strength of broad concepts, for example brand awareness. The framework first involves figuring out "signatures for concepts," which is an application of one algorithm on top of an anchor vocabulary. Then a second proprietary algorithm is used to perform an actual measurement.

Meeting Pharma's and Biotech's Needs?

Buoying SciWit's business model is the fact that the biopharmaceutical industry has become much more comfortable outsourcing and partnering its text mining tasks, according to William Hayes, director of Library & Literature Informatics at Biogen Idec.

But SciWit will still likely have to convince these potential customers that it offers advantages beyond anything they are capable of doing on their own.

SciWit's approach sounds "like a lot of approaches I have already seen that are fairly common in the text mining community," Hayes told *BioInform*. For example, finding hidden concepts in text for which no reference vocabulary exists is currently feasible, he said.

"You can take a corpus and extract concepts without knowing anything *a priori*," he said. What is required is looking at patterns and how often they are represented. "If two or three words show up following each other fairly often, that's a concept you want to follow."

In addition, methods for determining the rate at which literature mentions are growing are well established, he said. "The technology is out there and a whole bunch of companies have it."

What has been difficult, he said, is noise.

For example, at his former employer, AstraZeneca, Hayes ran the tex-mining initiative. In one project, he and his team sought to find new concepts in the literature about estrogen-sensitive tumors. However, they only wanted to find genes and proteins related to tumors that escape estrogen-sensitivity and become untreatable. One drawback of many text-mining algorithms, he said, is that they may yield proteins that are not related to estrogen sensitivity or insensitivity but are just being discussed in this context. "The trick is filtering out the noise," he said.

Hayes said he believes that SciWit's B-2-B positioning is surprising due to the principals' background with OmniViz, which was geared toward end-users, but sees promise in the company's highly focused niche market. "The business model makes a lot of sense," he said. "I like the concept."

However, it might be a challenge for the firm to get in the door with pharmaceutical companies, said Hayes. Even his department has a problem of "getting in front of people [within Biogen Idec] often enough and ... letting them know of alternative workflows in order to get our technology taken up," he said.

The Open Approach

On the other side of the Atlantic, meantime, the UK's NaCTeM is taking a different approach to the challenges of developing effective text-mining workflows: It's providing a public repository of integrated tools.

NaCTeM director Sophia Ananiadou said that the center, which provides text-mining services free of charge to members of higher-education institutions, is committed to open-source and open-access text mining.

"With open source you have more possibilities of selecting and integrating different tools," she said. She said she believes open access allows greater flexibility than proprietary software, and better matches the varied needs of text mining in biology.

"If tools are open access, you can mix and match" — for example, using different taggers or parsers — "and get the best output," she said. "If you don't allow that, you are a bit stuck in one specific solution."

One challenge lies in the text-mining approach chosen. "Initially text mining sees text as a bag of words with no structure of sentences," said SciWit's Tsujii. The dominant technology approach to text mining is still based on that view of a text — the counting the frequency of specific nouns or co-occurrences of two words — he said.

So a search for protein interactions would first identify proteins that co-occur in sentences. "Lots of software just enumerates the names of proteins without any specific interaction among them," he said. That leads to the problematic noise that text mining can face.

Deep parsing or full parsing technology is a new approach that gives more semantic-oriented information, said Tsujii. Recognizing this, NACTeM has begun collaborating with several academic groups to develop software that performs semantic or deep parsing, which uncovers implicit sentence structure, he said.

And given the special requirements of the life-science community, targeted text mining is appropriate, he said. With 18 million abstracts in Medline, if scientists obtain noisy results in information extraction, scientists cannot check the results, so text-mining tools must be sophisticated, he said. There is no one solution that fits all text mining needs in the life sciences, with its vastly differing subspecialties and ontologies, he said.

The Manchester center provides services and consulting in text mining, such as concept extraction and information extraction, for the entire UK academic community and it also maintains an inventory of "[best breed](#)" software, in addition to [other software tools](#).

One of the major remits of the center is to coordinate resource-building and develop software tools, its scientists said. Because they believe in an approach of mixing and matching text-mining tools, they are trying to build a publicly available repository for resources and tools, including privately owned tools.

Tsujii agreed, saying that the market "is huge and the problem is huge as well, so we don't really compete with specific companies. We want to coordinate all the efforts in this field to deliver the best services to individual users."

For its part, NACTeM offers tools with an eye to interoperability and for which workflow software is important, for example the Unstructured Information Management Architecture, or UIMA, formerly associated with IBM and now an open project that runs in OASIS and Apache, and protocols such as SOAP for XML-based message exchange. "We want to be able to select the most important tools for a specific task," said Ananiadou. Users can mix and match the tools they need.

"Academics make things freely available, so the idea of UIMA is to expose your resources to the outside world, but companies are proprietary and may not wish to share," said Ananiadou.

PGxL Prepares to Launch Decision-Support Software for Warfarin Dosing Within a Year

By Bernadette Toner

PGx Laboratories, a genetic testing lab in Louisville, Ky., is developing a software program that uses genotypic data and physical traits to help physicians adequately dose the anticoagulant warfarin.

The software, which the company expects to roll out within nine to 12 months, is envisioned as a tool that will help doctors navigate the tricky process of warfarin dosing, which currently requires constant careful monitoring due to the narrow therapeutic window of the drug — too low of a dose and it won't reduce clotting, but too high and the patient faces the risk of hemorrhage.

Variations in the CYP2C9 and VKORC1 genes have been clinically linked to warfarin response, and the US

Food and Drug Administration last summer asked drug makers to update the drug's label to note that patients with these variations should be started at lower initial doses. However, this information has not made it any easier for doctors to determine a safe yet effective maintenance dose.

PGxL is hoping that its software, called PerMIT:Warfarin, or Personalized Medicine Interface Tool, will address this problem by analyzing a patient's genotypic information — specifically by looking for mutations in the CYP2C9 and VKORC1 genes — with physical characteristics such as age, gender, weight, and height.

"The real challenge for warfarin — whether or not the genotype is known — is the physician being able to determine as quickly as possible what the optimal maintenance dose is for a particular patient," Kristen Reynolds, vice president of laboratory operations at PGxL, told *BioInform*.

Time is of the essence, she said, because "during the time while the physician is titrating that dose through what is essentially an educated trial-and-error process, the patient is at an increased risk of having an adverse drug reaction to warfarin, be they bleeding events or thrombotic events."

Mark Linder, associate director at PGxL and assistant director of chemistry and toxicology at the University of Louisville Hospital, said that the variables that contribute to warfarin dose response are "sufficiently diverse that you can't come up with a categorical approach to the problem," in which, for example, everyone with a particular genotype would be prescribed the same dose.

"Really, the only way to reconcile all this information simultaneously was with some sort of informatics tool," he said.

Linder stressed that the company doesn't envision the PerMIT:Warfarin as a replacement for the current standard of care, in which a physician administers a relatively low dose of the drug and then monitors the patient's clotting rate via the INR, or International Normalized Ratio, in order to determine whether to increase the dose.

"We still expect that the INR would be measured, but what the software does is frame it in the right context for that individual so that it's not misinterpreted."

"We still expect that the INR would be measured, but what the software does is frame it in the right context for that individual so that it's not misinterpreted," he said.

This strategy is in line with the FDA's recommendation to relabel warfarin, which states that therapy still "be initiated with a dose of 2 to 5 mg per day with dosage adjustments based on the results of [prothrombin time/international normalized ratio] determinations."

It also gibes with a brochure that the Critical Path Institute and the American Medical Association released this week to its members intended to inform them about using genotypic data when prescribing warfarin. However, the brochure stresses in boldface type that "careful monitoring of INR is still required for optimal dose

adjustment" even if they use genotyping to select a starting dose.

"Most of the time you should be able to get a benefit by knowing the genetic makeup of an individual, but that doesn't mean that that's all you need to know," said Ray Woolsey, director of the Critical Path Institute. "You still need to take into account the other factors like gender, body weight, age, and drug interactions," he said.

"A genetic test will help you better approximate the first dose, but it doesn't tell you what the first dose is," Woolsey added. "You can't cookbook it."

Computer algorithms "are going to be essential" in guiding dosing "because some of these variants have more weight than others," he said.

The key to the PGxL software, according to Linder, is that it is intended for ongoing patient care. Unlike a doctor, "it doesn't forget," he explained. "If you really put yourself into a practice setting, where you're trying to manage literally hundreds of patients, you may have done the genotyping at one point in time and know their genotype, but you would really find it difficult to continuously apply that information to your ongoing

decisions," he said.

PerMIT: Warfarin provides "a format for the ongoing utilization of that information," he said. "So every single time the physician sees that patient and decides to make some dosage-based decision, he can use the tool to put that decision in that person's genetic or genomic context."

PGxL is not the only group developing a computational approach to warfarin dosing that uses genetic data. The Barnes-Jewish Hospital at Washington University Medical Center, for example, has developed a free online server called WarfarinDosing.org that estimates therapeutic dose in patients based on clinical factors and CYP2C9 and VKORC1 genotypes.

In addition, the FDA is working with the Harvard Partners Healthcare System to develop a dosing algorithm as part of the 500-patient CROWN, or CReating an Optimal Warfarin Nomogram, trial.

But Linder said that the software PGxL envisions will be more of a decision-support tool than a simple dosing algorithm. He said that the dose-estimation method that underlies the software, which he and his colleagues published in the July 2007 issue of [Clinical Chemistry](#), "only provides you with a target of where you're likely to be headed. What the overall tool does is give you guidance as you head in that direction."

In addition, he said that the software was developed as an "operational framework" that can be customized for other drugs, though he declined to disclose what other drugs the company may also be looking at, or its timeframe for doing so.

PGxL believes that the software would predominantly be used for warfarin-naïve patients, which represents around 500,000 to 800,000 individuals per year in the US.

Linder said that the software would also be useful for existing warfarin patients, "particularly those that have demonstrated that they are difficult to stabilize," but noted that the primary target market would be new patients.

Iris Biotech Views Its Informatics Platform As Competitive Edge in Genomic Dx Market

By Bernadette Toner

Iris Biotechnologies, a developer of microfluidics-based molecular diagnostics, is currently beta-testing a comprehensive database and software system that it expects to give it a boost over its rivals in the competitive genomics-based breast cancer testing market.

The platform, called BioWindows, is an artificial intelligence system that acquires and analyzes genomic hybridization information from the company's microfluidic chips and then compares that data with a repository of hybridization profiles, patient profiles, reference information, clinical information, and other data in order to provide appropriate treatment scenarios for individual patients.

The key to the system is an interactive survey that allows patients to enter their personal information and medical history through a secure, online interface. The company plans to collect as many of 1 million such profiles in order to refine the platform in an iterative fashion.

"The informatics side is extremely important because when you talk about the human genome or how best to treat patients or develop new drugs, it's all about getting the right information and being able to piece together, in our case, a patient's personal medical history as well as family medical history, with lifestyle, with exposure, and other relevant information," Simon Chin, CEO of Iris, told *BioInform*.

In the BioWindows survey, "we ask hundreds of questions, and that information goes into the database, and then we combine this information with the information from the chip so it makes a really powerful platform," he said.

Chin said the company is not disclosing the number of patients who have already submitted data into the

system, but he said that it is aiming for a million profiles eventually. "We don't need a million people for what we plan to do, but the more we have participating, the more refined the information becomes," he said.

Iris plans to partner with breast cancer advocacy groups, research groups, and physicians in order to encourage further participation in the BioWindows database, Chin said.

The system will ultimately be able to recommend specific cancer treatment regimens based on the collection of profiles in the database, Chin said. "The biggest question that people will have is related to chemotherapy. There are more than 30 chemotherapy agents in use, and which combination of those will actually work for the patient? Today, doctors don't know, so it's a trial-and-error process," he said.

Chin noted that the survey tracks information such as the treatments that are given to patients, as well as whether those patients survived or not — information that can be used as part of a predictive tool once it is aggregated with similar data from other patients.

"Let's take, for example, an Irish woman aged 40 to 45. Say she drinks a fair amount but she doesn't smoke. And let's say she's been exposed to various diseases and a certain amount of stress and a whole host of other questions that come into play," Chin said. "If you subgroup these people, then a pattern would emerge that would show that maybe for this group of people one particular chemotherapy might work better."

"We ask hundreds of questions, and that information goes into the database, and then we combine this information with the information from the chip so it makes a really powerful platform."

The technology underlying BioWindows is protected under a patent awarded in 2006, US No. 7,062,076, entitled "Artificial intelligence system for genetic analysis."

But BioWindows is only half of the company's business. Iris has also developed a proprietary silicon biochip platform for analyzing gene-expression patterns that it believes offers a number of advantages over microarrays or PCR in diagnostic settings.

Chin last week told *BioInform* sister publication *BioArray News* that the company plans to submit its first assay on the platform — a 100-gene breast cancer diagnostic called BreastCancerChip — to the US Food and Drug Administration for 510(k) clearance as an *in vitro* diagnostic by the second half of this year.

In a prospectus filed with the US Securities and Exchange commission in December, the company acknowledged that it faces competition in this market from Genomic Health, which markets the PCR-based Oncotype DX test for breast cancer recurrence, and Agendia, which recently received FDA approval for its microarray-based MammaPrint breast cancer test.

The company said in the filing that its management "believes that research-based microarrays lack the necessary test sensitivity, gene marker specificity and speed for the best and most cost-efficient medical decision-making and patient care."

PCR, meanwhile, "is a practical system to use for clinical diagnostic if it is used to detect only a few genes," but would be "too expensive to use to detect 100 or more genes in a clinical setting to obtain the gene profile necessary for predicting the best treatment regimen for an individual patient."

The company is also differentiating itself from its competitors in its application of the technology. Chin told *BioArray News* that while companies like Genomic Health and Agendia are focused solely on predicting probable recurrence of breast cancer, "what we are doing with 100-plus genes is looking at how best to treat the patients from the beginning."

The company believes that the market for its test is worth around \$2.5 billion in the US, based on an estimate of around 2 million biopsies each year.

Chin founded Iris in 1999. In December 2007, the firm filed a prospectus with the US Securities and Exchange Commission to go public.

According to the filing, the firm incurred an accumulated net loss of around \$5 million between the time it was founded and Sept. 30, 2007, and generated no revenues during that time. Iris had five full-time employees and 10 part-time staffers as of Dec. 3, 2007.

As of Sept. 30, 2007, the company had \$251,000 in cash. It expects to spend around \$1.5 million this year to get the BreastCancerChip approved in the US.

Iris said in the filing that it plans to price its shares at \$2.25 each. Based on 10.7 million shares currently outstanding, it could raise as much as \$24 million in an IPO.

A company spokeswoman said that Iris expects to begin trading publicly on the over-the-counter bulletin board under the ticker symbol IRSB.OB "in the near future."

The company also has several other diagnostics in the pipeline, including a CardioChip for detecting and treating heart disease, and a NeuroChip, designed to diagnose degenerative neurological disorders such as Alzheimer's disease and Parkinson's disease.

Chin said that the company is collecting information on these disease areas through the BioWindows survey so that it will become part of the predictive platform in the future.

Bioinformatics Tool-Related Papers of Note, February 2008

Cai JJ. [PGEToolbox: A Matlab Toolbox for Population Genetics and Evolution](#). [*J Hered.* 2008 Feb 29 (e-pub ahead of print)]: Describes PGEToolbox, a Matlab-based open-source software package for data analysis in population genetics. Available [here](#).

Fourment M, Gillings MR. [A comparison of common programming languages used in bioinformatics](#). [*BMC Bioinformatics* 2008, 9:82]: Describes the comparison of memory usage and speed of execution for three standard bioinformatics methods — the Sellers algorithm, the Neighbor-Joining tree construction algorithm, and an algorithm for parsing Blast file outputs — implemented in six different programming languages: C, C++, C#, Java, Perl, and Python. According to the abstract, implementations in C and C++ were fastest and used the least memory, but programs in these languages contained more lines of code. "Java and C# appeared to be a compromise between the flexibility of Perl and Python and the fast performance of C and C++," the abstract states.

Ge D, Zhang K, Need AC, Martin O, Fellay J, Urban TJ, Telenti A, Goldstein DB. [WGAVIEWER: A Software for Genomic Annotation of Whole Genome Association Studies](#). [*Genome Res.* 2008 Mar 3 (e-pub ahead of print)]: Introduces WGAVIEWER, a Java-based interface to help researchers annotate, visualize, and interpret the set of P values emerging from a whole-genome association study. According to the paper's abstract, WGAVIEWER is able to highlight possible functional mechanisms in an automatic manner and can help in generating hypotheses concerning the possible biological basis of observed associations.

Hamilton NA, Teasdale RD. [Visualizing and clustering high throughput sub-cellular localization imaging](#). [*BMC Bioinformatics* 2008, 9:81]: Describes a new method, called iCluster, for visualizing, clustering, and comparing large sub-cellular localization image sets. For each member of an image set, iCluster "generates statistics that have been found to be useful in distinguishing sub-cellular localization," the paper abstract states. The statistics are mapped into two or three dimensions and the complete image set is then visualized using these coordinates. "The result is images that are statistically similar are spatially close in the visualization allowing for easy comparison of images that are similar and distinguishment of dissimilar images into distinct clusters."

Kalaev M, Smoot M, Ideker T, Sharan R. [NetworkBLAST: comparative analysis of protein networks](#). [*Bioinformatics* 2008 24(4):594-596]: Discusses the NetworkBLAST web server, which identifies protein complexes in protein-protein interaction networks. The software can analyze a single network or two networks from different species, according to the authors. Available [here](#).

Lai W, Choudhary V, Park PJ. [CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms](#). [*Bioinformatics*. 2008 Feb 22 (e-pub ahead of print)]: Describes CGHweb, a web-based visualization tool that applies several algorithms to a single array-CGH profile entered by the user and generates a heatmap panel of the segmented profiles for each method as well as a consensus profile. Available [here](#).

Li R, Li Y, Kristiansen K, Wang J. [SOAP: short oligonucleotide alignment program](#). [*Bioinformatics* 2008 24(5):713-714]: Describes SOAP, a program for gapped and ungapped alignment of short oligonucleotides onto reference sequences. The program is designed to handle short reads generated by the Illumina Genome Analyzer and is compatible with numerous applications, including single-read or pair-end resequencing, small RNA discovery, and mRNA tag sequence mapping. Available [here](#).

Liu Q, Olman V, Liu H, Ye X, Qiu S, Xu Y. [RNACluster: An integrated tool for RNA secondary structure comparison and clustering](#). [*J Comput Chem*. 2008 Feb 13 (e-pub ahead of print)]: Describes RNACluster, a platform that compares different distances between RNA secondary structures and performs cluster identification to gain useful information of RNA structure ensembles. The software "provides a user-friendly graphical interface to allow a user to compare different structural distances, analyze the structure ensembles, and visualize predicted structural clusters," according to the abstract.

Morris JS, Clark BN, Gutstein HB. [Pinnacle: a fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data](#). [*Bioinformatics* 2008 24(4):529-536]: Describes a new method for spot detection and quantification in 2D gel analysis called Pinnacle. The spot definition "is based on simple, straightforward criteria rather than complex arbitrary definitions, and results in no missing data," according to the authors.

Rohde C, Zhang Y, Jurkowski TP, Stamerjohanns H, Reinhardt R, Jeltsch A. [Bisulfite sequencing Data Presentation and Compilation \(BDPC\) web server--a useful tool for DNA methylation analysis](#). [*Nucleic Acids Res*. 2008 Feb 22 (e-pub ahead of print)]: Describes the Bisulfite sequencing Data Presentation and Compilation, or BDPC, web interface, which automatically analyzes bisulfite datasets prepared using the BiQ Analyzer. Available [here](#).

Zhang S, Kumar K, Jiang X, Wallqvist A, Reifman J. [DOVIS: an implementation for high-throughput virtual screening using AutoDock](#). [*BMC Bioinformatics* 2008, 9:126]: Discusses an application called DOcking-based VIrtual Screening, or DOVIS, that uses the AutoDock docking engine and runs in parallel on a Linux cluster. According to the abstract, DOVIS can efficiently dock millions of small molecules to a receptor, screening 500 to 1,000 compounds per processor per day.

ISCB Software-Sharing Policy, ABI SOLiD, Ohio Bioinformatics Consortium, CLC Bio, DNASTar, Agilent, Accelrys

ISCB Seeks Comments on Revised Software-Sharing Policy

The International Society for Computational Biology is soliciting comments from its members on its updated policy statement on bioinformatics software availability.

The ISCB board of directors posted a statement on the society's [blog](#) this week outlining the revised statement, which it said has been "revised from the original 2002 statement, incorporating feedback from the ISCB membership."

The original policy drew criticism for failing to reflect the views of the broader ISCB membership because it was issued without a poll, vote, or comment period [[BioInform 08-12-02](#)].

The revised statement, based on feedback on the original statement gathered at ISMB 2007, recommends that researchers, funding organizations, and publishers must "uphold the core principle of sharing methods and results;" grantors and publishers should "require statements of software availability in grant proposals and research reports;" and, "executable versions of the software should be freely available for research use to individuals at academic institutions."

The comment period closes April 15.

ABI Releases Human Genome Sequence Generated on SOLiD

Applied Biosystems this week said that it has completed sequencing a human genome using its next-generation SOLiD technology, and that it has deposited the sequence with the National Center for Biotechnology Information.

"The availability of this sequence data in the public domain is expected to help scientists gain a greater understanding of human genetic variation and potentially help them to explain differences in individual susceptibility and response to treatment for disease," the company said in a statement.

ABI scientists resequenced the genome of an anonymous African male of the Yoruba people of Ibadan, Nigeria, who participated in the International HapMap Project. They generated 36 gigabases of sequence data in seven runs on the SOLiD system, resulting in 12X coverage of the genome.

The company said the researchers were able to use this data to identify millions of SNPs greater than 99.94 percent sequencing accuracy. In addition, ABI said its scientists were able to analyze regions of structural variation using 100-fold physical coverage.

ABI said that it expects the public availability of the data "will help drive innovation and speed the development of new bioinformatics tools."

In addition to the full human dataset, ABI said it has also released "subsets" of sequence data through Genbank which "can be accessed by independent academic and commercial software developers to further enable the development of analytical tools."

The company has released several analysis tools through the [SOLiD System Software Development Community](#) to help researchers analyze the data.

At NCBI, the human sequence data is available [here](#) or by the project name, "SOLiD Human HapMap Sample NA18507 Whole Genome Sequence," under accession number SRA000272.

Ohio Bioinformatics Consortium Awarded \$4.5M in Scholarship Funding

The Ohio Consortium for Bioinformatics will receive \$4.475 million for student scholarships under the state's "Choose Ohio First" program, which is aimed at attracting and graduating more than 2,000 students in science and technology fields in the state over the next five years.

The bioinformatics initiative was one of seven awards totaling more than \$22.7 million announced this week.

The consortium includes Ohio University, 11 other colleges and universities in the state, the Ohio Supercomputer Center, and the Ralph Regula School of Computational Science. It hopes to attract and graduate an estimated 345 students over a five-year period.

Partners in the consortium will contribute an additional \$4.6 million to develop programs, expand offerings, and cover other related costs, the consortium said.

University of Copenhagen Takes Site License for CLC Bio's Software

The University of Copenhagen has bought a site license for CLC bio's Combined Workbench and Educational Suite software platforms, CLC bio said this week.

Under the agreement, the department of biology at the university has licensed access to the software for five years for "several hundred seats," the company said.

The university's genomics projects include retrieving DNA from fossils, gene regulation and SNP cancer research, primate genome evolution, bacterial genome sequencing, and other programs.

Financial terms of the agreement were not released.

BGI LifeTech to Sell DNASTar's Software in China, Hong Kong

The Beijing Genomics Institute's BGI LifeTech segment will sell DNASTar's bioinformatics software in China and Hong Kong, DNASTar said this week.

Under the agreement, BGI LifeTech has acquired the rights to sell the Lasergene, ArrayStar, Seqman Genome Assembler, and GenVision products.

The company's software platforms are used in analyzing sequence data, microarray gene expression data, and genomic visualizations.

Financial terms of the agreement were not released.

Agilent to Embed Accelrys' Pipeline Pilot in OpenLAB Software Under OEM Deal

Accelrys and Agilent said this week that they have signed an original equipment manufacturer agreement that will allow Agilent to distribute an embedded version of Accelrys' SciTegic Pipeline Pilot workflow software with its own OpenLAB enterprise content management software.

Agilent expects to have a commercial version of the Accelrys software, called Embedded Pipeline Pilot, available in the first half of the year.

The OEM deal builds upon a resale agreement the companies signed last fall under which Accelrys has been reselling Agilent's OpenLAB, Kalabie electronic lab notebook, and GeneSpring gene-expression analysis

software, while Agilent has been selling Accelrys' Accord cheminformatics solutions and Pipeline Pilot [[BioInform 09-28-07](#)].

Financial terms of the OEM agreement were not provided.

HUGR, *Pongo pygmaeus* draft assembly, Rosetta Resolver, Reactome 24, PrimerPlex, RefSeq 28

Yissum, the technology transfer arm of the **Hebrew University of Jerusalem**, has launched the [Hebrew University Genetic Resource](#) platform, a database for genetic association studies of common diseases that includes 15,000 DNA samples from Ashkenazi Jews and represents 16 different diseases.

Among the 16 diseases represented in the database are diabetes types I and II, several cancers, neurological diseases, psychiatric diseases, hypertension, and asthma, Yissum said. The database includes more than 500 samples for most diseases and a common panel of over 5,000 healthy controls. Each sample contains phenotypic information, including family history, disease characterization, drug treatments, efficacy, and adverse events. Scientists can order genotyping of any SNP of interest on any of the samples and solely own the results.

The Genome Bioinformatics Group at the **University of California, Santa Cruz**, has released the draft assembly of the Sumatran orangutan genome, *Pongo pygmaeus*, through the UCSC Genome Browser. The assembly is the July 2007 draft and was provided by the Genome Sequencing Center at Washington University School of Medicine in St. Louis. Bulk downloads of the sequence and annotation data are available via the Genome Browser [FTP server](#) or [Downloads page](#).

Rosetta Biosoftware said this week that the current version of the **Rosetta Resolver** gene expression data analysis system, version 7.1, now supports **Affymetrix Gene 1.0 ST (Sense Target) arrays** for transcriptome analysis. The company said the Resolver system supports import and data analysis of these arrays as well as interaction with third-party programs, such as RMA, gcRMA, and Affymetrix PLIER.

Version 24 of the **Reactome Knowledgebase** is available [here](#).

The release includes the beta version of a new pathway visualization tool, as well as a number of new pathway topics, including signaling by VEGF, metabolism of nitric oxide, and microRNA biogenesis.

Premier Biosoft International has released **PrimerPlex**, a tool for designing custom oligos for the Luminex xMAP platform, including the Luminex 100 and Luminex 200, and Bio-Plex Suspension Array system from Bio-Rad. PrimerPlex allows users to design custom capture probes for the bead-based detection system, the company said.

The **National Center for Biotechnology Information** has released [RefSeq Release 28](#).

This release includes genomic, transcript, and protein data available as of March 9, 2008, and includes 7,914,560 records, 5,195,116 proteins, and sequences from 5,059 different organisms.

Sergey Ilyin, Kevin Keenan

Sergey Ilyin, previously bioinformatics group leader at **Johnson & Johnson**, has left J&J to serve as chief scientific officer of **Serigene**, a developer of a PCR-based diagnostic testing platform.

Ilyin holds a BS/MS in biology from **St. Petersburg State University** and a PhD in molecular biology from the **University of Delaware**.

Kinematik has promoted **Kevin Keenan** to vice president of software development. Keenan was one of the original founders of KineMatik in 1999. Since that time he has served as chief architect for the company's flagship eNovator product suite.

Prior to founding KineMatik, Keenan helped develop knowledge management solutions at a number of firms, including **Ericsson**, **St. Georges Bank**, and various departments of the Irish government.

Copyright Notice - Subscription Terms and Conditions

GenomeWeb Application-Focus Newsletters are copyrighted intellectual property. It is a violation of US and international copyright law to forward, copy or otherwise distribute a newsletter email bulletin or PDF file to non-subscribers or other unauthorized persons. Violators will be subject to statutory damages.

Individual Subscriptions

This newsletter subscription is for a single individual user. Passwords and user logins may not be shared. You may print and retain one copy of each issue of this newsletter during the term of your subscription. Copying, photocopying, forwarding or duplicating this newsletter in any form prohibited. If multiple individuals need to access this newsletter online, you need an affordable, multi-user site license. Contact Allan Nixon at 1-212-651-5623 or anixon@genomeweb.com.

Web Postings and Reprints

Individual articles from this newsletter may not be posted on any website or redistributed in any print or electronic form except by specific arrangement with GenomeWeb LLC. Contact reprints@genomeweb.com for further information.

Site Licenses

If you have received this file under a GenomeWeb site license, your rights to forward, copy or otherwise distribute this file are governed by the provisions of that site license. Contact your site license administrator for details.

© Copyright 2008 GenomeWeb Daily News. All rights Reserved.