

bio1NF0RM

The integrated informatics news source

copyright

**THIS NEWSLETTER IS COPYRIGHTED INTELLECTUAL PROPERTY.
BY OPENING AND VIEWING THIS ADOBE PDF FILE YOU AGREE THAT:**

- You may print and retain one paper copy. Additional copies may be printed only by specific arrangement with GenomeWeb LLC.
- You may permanently retain this Acrobat file. You may forward, copy or otherwise distribute this file only by specific arrangement with GenomeWeb LLC.
- If you have received this file under a GenomeWeb site license, your rights to print, forward, copy or otherwise distribute this file are governed by the provisions of that site license.



www.genomeweb.com
The Pulse of the New Biology

If you have questions about your use of this newsletter please contact

Allan Nixon at anixon@genomeweb.com



volume 7, number 25
June 23, 2003

in this issue

page 2

- People in the News
- Downloads & Upgrades

page 3

- **NewbieWatch:** Molecular Mining's software finds a new home with Predictive Patterns.

page 5

- FDA collaborates with NCI on a cancer bioinformatics infrastructure.

page 7

- *Bioinform* lists recent NSF awards in bioinformatics.

page 8

- **Bioinformatics Briefs:** NCBI removes 350-kb limit on GenBank sequences; New Broad Institute touts computational biology.

FDA Delves into Microarray Database Projects as First Step in Guidance Process

THE US FOOD AND DRUG ADMINISTRATION is rolling up its sleeves and diving elbow-deep into the messy world of microarray data as it girds for a potential wave of genomics-derived information submitted as part of the drug approval process.

At a June 10 meeting of the FDA's pharmacology/toxicology subcommittee on pharmacogenomics, the agency took its first steps toward tackling the questions of whether, when, and how it will accept data from microarray experiments submitted under investigational new drug (INDs) or new drug applications (NDAs).

"[FDA reviewers] haven't seen microarray data," said William Matthes, associate director at Pfizer's Kalamazoo Genomics Center of Excellence, who attended the subcommittee meeting. "They're concerned about how to handle it, they're concerned about its analysis, they basically are concerned about what does it look like, what does one do with it?"

According to another attendee, Kurt Jarnagin, who is vice president of biological sciences and chemical genomics at Iconix Pharmaceuticals, "The agency believes that in time, [microarray data] will become a standard part of any submission, either IND or NDA, and it needs to prepare itself for that day. They can't suddenly start getting inundated with data and expect to respond to that."

As a first step in familiarizing itself with the nuances of microarray data, the FDA's Office of Testing and Research has embarked continued on page 4

At Beyond Genome Conference, a Call for Bioinformatics to Move Beyond 'Pretty Pictures'

ATUL BUTTE is annoyed with the "dummy mentality" in bioinformatics — especially in the microarray analysis area. "You can tell what algorithm [someone used] by the colors of the figures in their paper," he quipped during the first day of the bioinformatics and genome research track of the Beyond Genome conference in San Diego last week.

Butte, an endocrinologist at Children's Hospital, Boston, devoted his talk to bringing bioinformatics out of this intellectual rut, and his senti-

ments seem to have been heard loud and clear by some of the other speakers: The recurring theme was the development of more sophisticated algorithms that can extract biological meaning from data — in the case of pharma, biological meaning with drug discovery relevance — rather than just making pretty pictures.

Butte focused on the construction of relevance networks, networks that show the weakness or strength of a relationship between genes. He noted that the reams continued on page 4



www.genomeweb.com
The Pulse of the New Biology

PEOPLE IN THE NEWS

10a0010c91a010t910c
101091010tt01C10t00

Visualization software developer **OmniViz** has appointed company founder **Jeffrey Saffer** as president. Saffer, who will continue his role as CTO, replaces former CEO and president **Paul Kelly**.

HPCWire, an online news service, has reported that **IDC** life

science information technology and high-performance computing analyst **Debra Goldfarb** has left to take a "strategic planning" position at **IBM** "at a senior level." **Vernon Turner**, group vice president of global enterprise server solutions at **IDC**, will take over Goldfarb's duties.

DOWNLOADS & UPGRADES

10a0010c91a0
101091010tt0

Gene Logic has launched version 2.0 of its **Genesis** data integration, management, and analysis software for its GeneExpress database. Genesis 2.0 will be commercially available during the fourth quarter of 2003 to existing customers subscribed to the BioExpress System and/or the ToxExpress System, as part of their current subscription agreement.

GenePilot, a new microarray analysis software suite, is available from **TG Services** at www.genepilot.com. The suite includes hierarchical clustering, k-means clustering, self-organizing maps, and significance analysis of microarrays, as well as a Gene Ontology display in every result screen. GenePilot is free for academic users.

GenBank release 136.0 is now available from **NCBI** at <ftp://ftp.ncbi.nih.gov>. The release contains 32,528,249,295 base pairs and 25,592,865 entries, up by 1,428,984,840 base pairs and 1,564,929 sequence records from the April 135.0 release. Uncompressed, the 136.0 flat files require about 107 GB for the sequence files only. The ASN.1 version requires around 95 GB.

Release 75 of the **EMBL Nucleotide Sequence Database** is available from <ftp://ftp.ebi.ac.uk/pub/databases/embl/release/> (UK), <ftp://bio-mirror.net/biomir->

ror/embl/release/ (US), and other mirror sites. Release 74 is about 17.2 GB compressed and 112 GB uncompressed and contains 25,214,767 sequences comprising 32,195,012,823 nucleotides, an increase of about 10 percent over release 74 in March.

Callident, a newly launched Linux cluster consulting company, will release **BioBrew**, an open source Linux cluster distribution based on the NPACI ROCKS software that is enhanced for bioinformaticists and life scientists, at ClusterWorld in San Jose, June 24-26. BioBrew includes all the software to build a cluster, along with the NCBI toolkit, Blast, mpiBlast, Hmmer, ClustalW, Gro-macs, Wise, and EMBOSS.

BioJava 1.30 is available at <http://www.biojava.org/download/>. New features include packed storage of sequence data in memory, better support for the OBDA (open bio database access) standards, and improvements to the parsers for output from Blast and Fasta. In addition, a BioJava tutorial/cook-book called "BioJava in Anger" is available at http://www.biojava.org/docs/bj_in_anger/index.htm.

An updated version of the **Dicty-Base** model organism database for *Dictyostelium* is available at <http://dictybase.org>.

bio1NF0RM

The Global Weekly of Bioinformatics

www.bioinform.com

ISSN 1094-205x

Bernadette Toner, *Editor*
btoner@genomeweb.com

Gwen Libsohn, *Production Manager*
glibsohn@genomeweb.com

Elena Coronado, *Production Designer*
ecoronado@genomeweb.com

Dennis P. Waters, PhD
Chairman and Publisher

Marian Moser Jones
Editorial Director

SUBSCRIPTION INFORMATION

BioInform is published weekly (50 times annually) by GenomeWeb LLC

Subscription rate: \$895.
To subscribe contact Allan Nixon
+1.212.651.5623
anixon@genomeweb.com

Or subscribe at our website at
<http://www.bioinform.com>

REPRINT ORDERS

Gwen Libsohn
+1.212.651.5630
glibsohn@genomeweb.com

ADVERTISING PLACEMENT

+1.212.269.4747
sales@genomeweb.com

GENOMEWEB, LLC

PO Box 998, Peck Slip Station
New York, NY 10272-0998 USA
Phone: +1.212.269.4747
Fax: +1.212.269.3686

With a Slimmed-Down Strategy, Predictive Patterns Picks up Where MMC Left off

IT DIDN'T TAKE LONG for Molecular Mining's software to find a new home following the company's closure in March. Tom Radcliffe and Mark Chatterley, the former director and manager of software development at Molecular Mining, respectively, jumped at the opportunity to launch a new bioinformatics company around the GeneLinker microarray analysis platform.

Radcliffe, who joined Molecular Mining "exactly a year and a day before we closed" and "rapidly fell in love" with the software, said the entrepreneurial urge came on strong after the company ceased operations. Realizing that "there's a value proposition there if we can deliver GeneLinker to individual researchers at a cost they can afford," he and Chatterley wasted no time negotiating a non-exclusive license to sell, support, and develop new versions of the software from Parteq Innovations, the Queens University tech transfer company that retained rights to the technology following the closure of Molecular Mining.

Despite the familiar faces, the software is in quite a different environment than its former home: It will be living on the web. The two-person company does not plan to hire a sales force, but is marketing the software solely through its website.

Radcliffe said that a key part of this strategy is ensuring a high ranking in search engines; and sure enough, the young firm's website appears at the top of the hit list for a Google search on "gene expression analysis software" [typing in the phrase without quotes yields a second-place spot, just below the web page for Michael Eisen's lab.]

"Web-based selling is obviously going to be considerably cheaper in terms of cost of sales than running a large sales force," Radcliffe said. "We will see whether or not it's successful."

Even if Google doesn't turn out to be the most effective marketing tool, Radcliffe said that Predictive Patterns' narrowness of focus should help it survive in a highly competitive marketplace where its predecessor failed. "MMC had this three-legged business model [software, services, and research collaborations], and for a small company it was difficult to cover that kind of broad perspective," Radcliffe said. "We want to stay

focused strictly on the software, and that's probably our biggest difference."

Unlike many bioinformatics software companies, Predictive Patterns has adopted a simple and transparent pricing structure that is "substantially lower" than the pricing model for the software under MMC.

The entry-level GeneLinker Gold 3.1 is \$995 for a single-user license, while the high-end GeneLinker Platinum 2.1 is \$4,995. Current MMC customers can purchase a service contract for GeneLinker Gold and Platinum for \$495 and \$2,495, respectively. All fees are for a perpetual license and include a year's worth of upgrades.

The company does not offer an academic discount, but claims that its regular prices are "far below" its competitors' academic discounts.

In another departure from Molecular Mining, Predictive Patterns has discontinued a version of GeneLinker platinum that was bundled with an IBM workstation. "People who are running high-end analyses often already have high-end hardware, and I didn't see the need for that," said Radcliffe.

The company's website has been live since June 2, and "we're starting to get fairly regular traffic in downloads,"

Radcliffe said, with interest both from former MMC customers as well as potential new customers.

As far as new hires go, Radcliffe said that he and Chatterley are currently talking with several "senior members" of the former MMC development team, but the internally funded firm plans to "grow as resources permit."

Future enhancements to the GeneLinker products include more scripting capability so that users can automate their analyses, as well as expanding the use of the technology into a broader application domain.

"One of the things that I spent quite a bit of time on at MMC was working on various proteomics problems," said Radcliffe. "The software, as it stands, can operate on proteomics data just fine... So making people aware of the fact that they can use software in other areas is one direction we'd like to take."

— BT

"We want to stay focused strictly on the software, and that's probably our biggest difference."

FDA...

continued from page 1

upon two separate gene expression database projects. One, in collaboration with Iconix, will introduce FDA reviewers to the basics of microarray data via the company's Drug-Matrix toxicogenomics database. A second project, with Schering-Plough and Affymetrix services provider Expression Analysis, will create an internal "mock submission" database for gene expression data.

The outcome of the two database projects will shape a draft

guidance document the FDA is preparing on the submission of microarray data. *BioInform's* sister publication, *BioArray News*, reported last week that Janet Woodcock, director of the agency's Center for Drug Evaluation and Research, expects the draft guidance to be prepared by August.

GETTING PAST THE FEAR FACTOR

The FDA has had access to the DrugMatrix database since March, when it began a collaboration with Iconix to gain hands-on experience with toxicogenomics data and tools. The agency is boning up on the

database as part of an effort to correlate the content and format of gene expression microarray data with standard toxicology and pharmacology study results. Iconix is training FDA reviewers on quality control and quality assurance for microarray data generation, as well as the analysis of data across multiple microarray product platforms, and the validation of biomarkers from integrated chemogenomic datasets.

The database contains findings from approximately 600 compounds, across multiple doses and multiple times. Gene expression data is linked to information on

Beyond Genome...

continued from page 1

of new information being constantly added to GenBank and other databases make these networks extremely dynamic. In other words, he said, "You're never done with microarray analysis." But he noted that microarray data has its limits. "Not all pathways will be reverse-engineered by microarray analysis."

Philip Xiang, director of bioinformatics at Roche Molecular systems, discussed his own approach to making gene expression analysis more sophisticated. He described how he combines information on gene expression and gene ontology trees. Xiang said he has also written a program that calculates average pairwise separation among any two genes.

Similarly, Jeffrey Sachs of Merck Research Labs said in his talk that he is "elucidating biological pathways by integrating gene annotations and gene expression data." Sachs has developed an approach wherein he uses 10 key variables in the gene annotation as nodes of comparison for 5,000 genes under

study. He looks at how the expression predicts annotation, and how expression and annotation predict other variables, including the effect of a compound on the genes. "We were thrilled that we were able to completely blindly make a prediction about a compound that people familiar with the datasets didn't know about," he said.

Yixin Wang at Johnson & Johnson's molecular diagnostics unit described the way his group grappled with the issue of how to treat numeric vs. qualitative descriptors in comparing microarray data. They developed a new algorithm, borrowing the fuzzy logic method from field engineering, to be able to navigate between quantitative and qualitative data to identify significant expression patterns in the data and explore transcriptional regulatory networks, he said. This particular application of fuzzy logic involved converting the gene expression value on the gene chip to a category — low, medium, or high — and grouping genes with similar values into related "triplets." By comparing the predicted relationship between members of a triplet to known relationships, they could test out the

robustness of the algorithm, he said. Already, the group has used this approach to extract an eight-gene signature that predicts survival time in colorectal cancer with around 80 percent sensitivity and specificity — an improvement over existing methods, he said.

For Brian Moldover, from Aventis, the gold mining came not in gene expression analysis, but in searching the genome for splice variants in pharmaceutically relevant protein families. "The interesting genes are heavily patented," he explained. "However, new variants with specific functions are still patentable." Furthermore, he noted, 15 percent of heritable diseases involve mutations that involve splicing. Aventis is finding new splice variants two ways: computationally, and using deep sequencing and deep cloning. The company uses Celera's genome browser, but NCBI's sequence, and has found two targets from this work, including two splice variants of hRasGRP4 that are involved in asthma, and another one that is "a target for a multibillion dollar therapeutic agent on the market now, and is very well studied," he said.

— MMJ

pharmacology, histopathology, clinical chemistry, and toxicology related to those compounds, to provide a "contextual reference set" for FDA reviewers to compare new findings with known results, Jarnagin said. "It gives the opportunity to ask specific questions," he added. "For example, 'Is it the case that the change of any oncogene can cause cancer?' You can look at the database and see that that's not true. There are dozens of drugs that elevate [expression of] one or several oncogenes, yet have been used in patients for years and years with no evidence of additional oncology."

The FDA's use of such a reference database for the evaluation of gene expression data could alleviate much of industry's lingering anxiety about submitting genomics-driven data, Jarnagin suggested. "There's a lack of trust that FDA will actually respond to the whole picture, and not respond to one gene," Jarnagin said. Drug companies "have to get

over the fear factor that 'If I submit this big experiment, some gene's going to change that the FDA's going to go nuts about and kill my compound.'"

BRINGING IT IN HOUSE

The goals of the planned internal gene expression database are a bit different than FDA's project with Iconix. In this effort, FDA, Expression Analysis, and Schering-Plough will build a framework to support the "mock submission" of data from a drug project Schering opted to discontinue. "We're taking that data, which includes microarray data, histology data, clinical chemistry data, and phenotype data, and helping FDA to understand the appropriate format, content, and context of microarray-based submissions," said Steve McPhail, CEO of Expression Analysis.

Pilot submission to the database is expected to begin in June, and the project is scheduled for completion

in October, McPhail said. A final summary report on the project is planned for November.

The project will address a laundry list of issues, including laboratory infrastructure, sample processing and array QC/QA issues, and experimental design and replication, but informatics-related questions make up the majority of topics. Data management issues such as format and file structures, linkage mechanisms between microarray data and other datasets, statistical analysis systems and software, and inference and modeling methods will all be examined as part of the project, McPhail said.

Expression Analysis will use Affy's MAS 5.0 software to analyze the data, but "we may use other methods as well," McPhail said. While the company has two years' experience processing Affymetrix data, "the linkage mechanisms are not something we've worked on in the past," he noted, so Expression

FDA Moves into Cancer Bioinformatics as Part of NCI Oncology Partnership

THE US FOOD and Drug Administration is taking bioinformatics seriously. In addition to its gene expression database efforts (see story, p. 1), a recently announced interagency collaboration with the National Cancer Institute will bring the FDA face-to-face with even more bioinformatics data.

Under the joint program, the two agencies plan to share knowledge and resources to speed the development of new cancer drugs. "Molecularly targeted drugs and other novel agents offer great promise, but they also present new challenges that require more collaboration between those involved in their discovery and development," said Andrew von Eschenbach, NCI director.

Information sharing will be the key to this collaboration, which will require the creation of a cancer bioinformatics infrastructure. The shared platform will be broad in scope, with the mission of improving data collection, integration, and analysis for preclinical, preapproval, as well as post-approval research for cancer therapies.

An Oncology Task Force made up of senior staff from both agencies will oversee the details of the collaboration. The task force will address the details of the bioinformatics infrastructure "over the next several months,"

according to an NCI spokeswoman. The FDA declined to comment on the project.

According to the NCI spokeswoman, NCI's existing bioinformatics infrastructure, which includes the caCORE and caBIO modules, "will serve as a key component of our collaboration with the FDA."

The two agencies have previously established a working relationship regarding data standards, particularly on common international biomedical standards like Health Level 7 (HL7) and the Clinical Data Interchange Standards Consortium (CDISC). "NCI will map its existing caCORE elements to these standards where they differ and work with the FDA and standards bodies to extend their coverage when required to address cancer research elements not already present within the existing standards," the NCI spokeswoman said.

In addition to the existing informatics tools and standards, "new messaging standards for adverse event reporting" will be developed as part of the joint project.

"If science is applied across the continuum of discovery, development, and delivery, safe, more efficacious drugs will be available to patients faster. Bioinformatics is an important part of that science," said the spokeswoman.

— BT

Analysis is turning to its sister company, regulatory informatics firm Constella Group, to handle the integration between microarray data and other clinical information.

Initially, the project will follow CDER's current guidance recommendations for regulatory submissions in electronic format, with the goal of identifying areas that need to be modified or redefined. This guidance stipulates that datasets be submitted as a SAS transport file of less than 25 MB per file, with data variable names of no more than eight characters, data elements defined in data definition tables, and variable names and codes consistent across studies.

The submitted array data will include raw data files after image analysis. In addition, a summary report will be provided to describe normalization, data processing, and statistical analysis steps. It is expected that these guidelines will be extended to improve compatibility with microarray data as the project progresses.

PARALLEL EFFORT

The FDA's database activities are not without precedent. A project spearheaded by the International Life Sciences Institute consortium and the European Bioinformatics Institute has been developing a centralized, public gene expression database for over a year. It is built on the EBI's ArrayExpress gene expression database, with the intention of linking toxicogenomics data from multiple platforms. Data input is currently ongoing, and the complete database is expected to come online by the first quarter of 2004.

"The intent of the ILSI effort was to establish some public offering that could be helpful in developing standards," said Pfizer's Mattes, who is on the ILSI database working group. Building on the MIAME

(minimum information about a microarray experiment) guidelines, the ILSI/EBI project has drafted a revised version of the standard called MIAME/Tox that aims to establish some consensus on the

"There is no standard yet for analysis."

minimal descriptors for array-based toxicogenomics experiments (available at <http://www.ilsil.org/committees/hesi/genomics/MIAME1.1ToxCirc-DRAFT-rev3.DOC>).

Judging by the near-universal acceptance of the MIAME standard in the microarray world, it's likely that MIAME/Tox will gain broad support within the toxicogenomics community. However, it is still in draft form, and has not been endorsed by anyone yet, least of all the FDA. CDER's Office of Information Management coordinates all of its standardization efforts, but according to Mattes, "there needs to be some communication between that group and anything going on in terms of a toxicogenomics database."

Indeed, the reigning CDISC-based guidance at CDER poses a number of differences from the proposed MIAME/Tox standard. MIAME/Tox proposes a more restrictive vocabulary, for example, with a field proposed for each clinical chemistry test. MIAME/Tox also collects information on *in vitro* experiments, while the standing CDER guidelines don't require it, and MIAME/Tox does not collect information on drug plasma levels, whereas this is currently done under the CDER guidelines.

But MIAME — along with its

accompanying data format, MAGE — is only the first piece in a much larger set of standards that need to be developed for a fully functional toxicogenomics data submission platform. In addition to a dearth of standards for experimental design, normalization, and a "universal" RNA, "there is no standard yet for analysis," said Mattes. "So, if somebody says, 'I've identified the regulated transcripts after this particular treatment,' what's the best way [to verify that analysis]? It's a huge question."

While the ILSI database project initially set out to address these standardization issues, Mattes said the group is far from a solution. "We have discussed and compared analysis, but resolved them? That's a definite no," he said.

THE RISKS OF RISK ASSESSMENT

The ILSI/EBI group has made some headway into the very issues that FDA plans to address with its own database, but there has been no formal involvement between the two groups so far, Mattes said. However, he added, "This may be the time for it. I'm sure, coming out of the [subcommittee] meeting, it would be a time when FDA would be interested in doing that, and I know we would be too," he said.

FDA could save itself some duplication of effort — and perhaps a lot of heartache — by communicating with the ILSI group. While the goals of the two projects are slightly different, the ILSI project did set out with the intention of creating a mock submission database for regulatory-bound gene expression data. However, Mattes said, after a bit of discussion on the subject, "we decided that the data we had developed was not appropriate for a mock submission."

Why was it unacceptable? "It didn't address risk assessment,"

Mattes said — a point that will likely impact the FDA's own database effort. The question of risk assessment lies at the crux of the entire regulatory process, and is one that the agency has yet to address regarding microarray data, Mattes said. So far, "genomics has been used in a predictive role, in the sense that genomics data from a short-term animal study or an *in vitro* study is used to anticipate longer-term treatment adverse events," he said. "In that case, the issue isn't risk assessment, the issue is prediction."

Microarray data may help flag potential problems in regulated assays, Mattes said, "but it has never substituted for a standard, regulated study, and I don't think anyone has anticipated that it would. In which case, then, you ask yourself, 'If I'm going to submit all

my standard, regulated studies anyway, why would I need to submit genomics data?'"

The lack of standards in the field is another sticking point, Mattes said. "It obviously gets in the way of the FDA saying, 'Submit this data.' If we can't say what the quality factors are for this data, and how we can analyze it, it's too soon to submit it."

Despite his doubts, Mattes is following the "just in case" strategy of many of his colleagues — and the FDA itself — in crafting the means by which microarray data may be submitted to the agency if and when it becomes a requirement. Though the project may still be a bit "premature," Mattes said, "working through the mock submission is a way to enlighten the agency, and enlighten the sponsors, on some of the issues that we've got to confront."

Jarnagin reiterated the ambivalence in the community over the issue. "On the question of whether the agency should encourage submission, and whether the agency should prepare itself to accept submission, I saw unanimity among the [subcommittee] panel, and the answer was yes," he said. "As to whether this becomes part of the regulatory decision today, it seemed that the gestalt of the panel was probably not today; but looking out into the future, at some point it will probably become more common. Whether it will become routine, I didn't see unanimity in the panel and I don't have unanimity in my mind either."

FDA's decision to get its hands dirty and grapple with microarray data first-hand may be the impetus that drives genomics data into the regulatory process. A number of

BioInform's Funding Update: NSF Awards in Bioinformatics through June 14, 2003

Bioinformatic Data Mining for AIDS Resistance

Genes. Start date: July 1, 2003. Expires: Dec. 31, 2003. Expected total amount: \$99,960.

Principal investigator: Walter Messier. Sponsor: Evolutionary Genomics.

SBIR Phase I project to develop data-mining software for the identification of genes and genetic mechanisms that contribute to the resistance of primates to the development of full-blown AIDS. It is hoped that the identified genes and corresponding gene products will lead to development of new therapies for HIV-infected humans.

Algorithms for the Simulation of Short- and Long-Time Dynamics of Proteins.

Start date: July 1, 2003. Expires: June 30, 2006. Expected total amount: \$405,000. Principal investigator: John Straub. Sponsor: Boston University.

Supports development of computational methods to model protein dynamics on both long and short time scales. Methods will include Monte Carlo simulations, path-integral methods, molecular dynamics, and quantum-chemical methods.

Development and Application of Generalized Ensemble Algorithms for Proteins.

Start date: Sept. 1, 2003. Expires: August 31, 2006. Expected total amount: \$328,000. Principal investigator: Ulrich Hansmann. Sponsor: Michigan Technological University.

Supports the development of generalized ensemble algorithms to the study of proteins. New means for computationally enhancing the frequency of transitions between interesting low-energy states will be developed with the goal of increasing simulation efficiency by an order of magnitude. The work will concentrate on understanding how misfolded structures of the b-amyloid peptide form and aggregate.

US-Polish-Czech Workshop on Modeling Interactions in Biomolecules.

Start date: June 15, 2003. Expires: May 31, 2004. Expected total amount: \$24,960. Principal investigator: Jerzy Leszczynski. Sponsor: Jackson State University.

Workshop to bring together students and researchers from the United States, the Czech Republic, and Poland to discuss current methods of computational chemistry and their application to biological systems.

circular arguments proliferate in the field around the fact that industry isn't submitting data to the FDA yet precisely *because* the agency is unfamiliar with it. With the FDA now taking an interest in the standard-setting process, "it certainly benefits the entire scientific community to move forward with standards in normalization and analysis," Mattes said. "And that gets to the point that you almost need that in place before

you can answer the question of how this data can be used in a risk-assessment standpoint."

Mattes added, however, that a great deal of work remains before the FDA is able to determine when and how microarray data should be included in the regulatory process. "I think we're trying to work out the nuts and bolts before we get there," he said.

McPhail agreed that it is too

soon to jump to any conclusions regarding the future of microarray data in the regulatory process. "The agency is just trying to get [its] arms around format, content, and context at this point and time, so it's probably too soon to tell what impact this will have on the future of microarray testing in support of INDs and NDAs," he said.

— BT

BIOINFORMATICS BRIEFS

NCBI TO REMOVE 350-KB GENBANK SEQUENCE LENGTH LIMIT

NCBI said last week that it would lift the current 350-kilobase limit on the sequence length of GenBank records as of June 2004.

The limit, which was originally put in place "as an aid to users of sequence analysis software, some of which might not be capable of processing megabase-scale sequences," was deemed unnecessary at the May 2003 collaborative meeting among representatives of GenBank, EMBL, and DDBJ.

According to NCBI, significant exceptions to the 350-kb limit have existed for several years, including high-throughput genomic sequences generated by the Human Genome Project and assemblies of whole-genome shotgun data. "Given these exceptions, and the technological advances which have made large-scale sequencing practical for an increasing number of researchers, the collaboration has decided that the 350 kbp limit must be removed," the NCBI said.

As of June 2004, the length of database sequences will be limited "only by the natural structures of an organism's genome." As an example, the NCBI noted that a single record might be used to represent all of human chromosome 1, which is around 245 Mb in length.

According to the NCBI, software developers for "some of the larger commercial sequence analysis packages" were asked what timeframe would be appropriate for this change, with answers ranging from "immediately" to "one year," so the one-year timeframe was selected to provide enough time for developers to upgrade their software to megabase scale.

NCBI has made sample records with very large sequences available at <ftp://ftp.ncbi.nih.gov/genbank/LargeSeqs> so that developers can begin to test their software modifications.

COMPUTATIONAL BIOLOGY TO PLAY MAJOR ROLE IN NEW BROAD INSTITUTE

Computational biology is destined to play a central role in research conducted at the Broad Institute, the biomedical research powerhouse that was announced last week as a collaboration between MIT, Harvard, and the Whitehead Institute.

The institute, which will focus on the development and application of genomics-based tools and technologies for the advancement of biomedical research, has identified computational biology as "increasingly central in converting the explosion in biological information into useful biomedical knowledge," according to a statement released by the partners.

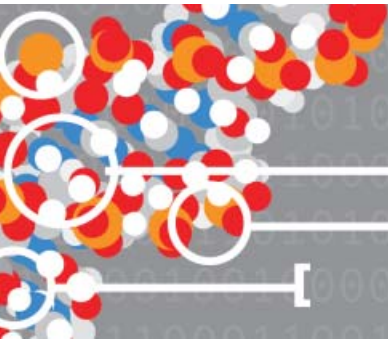
The Broad (pronounced "code") Institute received a founding gift of \$100 million over 10 years from Los Angeles philanthropists Eli and Edythe Broad, and plans to raise an additional \$200 million in private support, along with federal research grants, to support its work over the next decade.

The institute will begin operation in a new facility in the Kendall Square area of Cambridge later this year, but has not identified a site yet. Eric Lander, the director of the Whitehead Institute/MIT Center for Genome Research, will be the director of the Broad Institute.

The institute expects to employ 12 core faculty members and around 30 associated faculty members from MIT, Harvard, and the Whitehead. The initial core faculty will include Lander; Stuart Schreiber of Harvard University; David Altshuler of the Harvard Medical School and the Whitehead; and Todd Golub of the Dana-Farber Cancer Institute and the Whitehead.

It was not immediately clear who would head the institute's computational biology activities.

The Broad Institute said it expects at least 15 associated faculty members to be appointed before it is launched later this year.



bio1NFO,RM

The integrated informatics news source



3 easy ways to order!

1. Print and fax completed order form to: +1-212-269-3686
2. Print out and mail completed order form to:
GenomeWeb LLC
 PO Box 998
 Peck Slip Station
 New York, NY 10272-0998
3. If you're a first-time subscriber, **sign up online** for our introductory offer!

subscription form

yes! Sign me up for a full year of *BioInform!* That's 50 weekly issues plus access to the *BioInform* website (includes current issue and complete archive access) — the most comprehensive news resource in the bioinformatics industry — for \$895!

Please indicate your preferred method of payment:

- My check or money order is enclosed.
- Please bill me.
- Charge my credit card (see below).

Name _____

Job Title _____

Company _____

Address _____

City _____ State/Province _____

Zip/Post Code _____ Country _____

Phone _____ Fax _____

E-mail _____ @ _____

URL _____

Please indicate your preferred method of payment:

- Visa
- Mastercard
- American Express

Number _____ Exp. _____

Signature _____



www.genomeweb.com
The Pulse of the New Biology