bio1NFORM
The integrated informatics news source

April 13, 2007 | Vol. 11 No. 15

New Algorithms Indicate That 'Microread' Genome Assembly May Be Within Reach

By Bernadette Toner

A number of bioinformatics projects underway indicate that *de novo* genome assembly using extremely short reads from next-generation instruments — once thought impossible — could begin generating new bacterial genomes in the next few months.

While the various platforms in the emerging next-generation sequencing market each rely on different technologies to translate DNA molecules into a machine-readable format, they all have one thing in common: read lengths that are far shorter than those used in traditional Sanger sequencing.

These short reads, which can range from 25 base pairs to around 250 base pairs as opposed to an average of 750 base pairs for capillary electrophoresis, have presented an obstacle for conventional assembly algorithms that rely on long stretches of overlapping sequences to build out the contiguous regions, or contigs, that are the foundation of *de novo* genome assembly.

This challenge is particularly acute for platforms that generate so-called “microreads” of around 30 base pairs, such as Illumina’s (formerly Solexa’s) Genetic Analyzer or Applied Biosystems’ SOLiD technology.

Until very recently, the idea of *de novo* assembly with microreads was considered impossible, and these platforms were considered better suited for resequencing projects in which the short reads are aligned to a reference genome.

Over the last six months, however, several bioinformatics developers have begun looking into the problem a bit more closely and have made significant progress in creating a new generation of assemblers that can stitch microreads into full-length genomes.

Projects are currently underway at Illumina, the Broad Institute, the British Columbia Cancer Center’s Genome Sciences Center, Stony Brook University, and elsewhere to create new assembly algorithms for microreads.

These methods are now at a crucial point in their development because they are getting their first taste of real sequence data. The Illumina Genetic Analyzer has just entered the market and ABI’s SOLiD is not yet available, so most of these algorithms were developed and tested on simulated data — microreads generated by chopping up known genomes into tiny chunks — and are only now moving into the proof-of-principle stage.

Most of these developers “have been working with theoretical in silico data and we’re just starting to see some of them being applied to real data sets,” Francisco De La Vega, senior director of computational genetics at ABI, told *BioInform* this week.

“This is really early work and most of these algorithms are untested — they’re theoretical,” he said. “They still have yet to be fine-tuned to the error profile of these reads, which is different than with [capillary electrophoresis].”

De La Vega is organizing a “birds of a feather” session at the upcoming Research in Computational Molecular Biology conference in San Francisco that will focus on emerging algorithms for next-generation sequencing applications. One of the goals of that session, he said, is for developers of short-read assembly approaches to discuss how they are testing these algorithms and what they are finding.

ABI is not planning on developing a *de novo* assembler on its own, De La Vega said. “Our position here in this marketplace is that we will deliver the core technology — the platform — and some basic tools, but I don’t think we’re going to get into the assembly business,” he said. “There are so many applications, and *de novo* genome assembly is just one of them, that it would be difficult for us to get resources to deal with all of them.”

However, he noted, the company is hoping to “create a community and nurture this community so that independent researchers in the bioinformatics area can start contributing to the field.”

Illumina, on the other hand, is working on a *de novo* assembler that is currently at the proof-of-principle stage, according to Tony Cox, principal scientist at the company’s UK computational biology group.

“Over the past few years our thinking has evolved from detecting just single base differences from a reference towards encompassing more complicated differences such as indels and copy number variants, and our X chromosome sequencing collaboration with the Sanger Institute and our new paired read protocol have both opened up many exciting possibilities in this regard,” Cox explained via e-mail.

While noting that “people tend to make an artificial distinction between resequencing (aligning to a reference) and pure *de novo* assembly (where you start from nothing),” Cox said that his team is nevertheless working on a “pure *de novo* assembly algorithm.”

In Theory …

Several bioinformatics developers began working on this problem while ABI’s and Solexa’s technologies were barely prototypes.

“We have been working for about two, three years on trying to do *de novo* assembly for short read technologies,” said Steve Skiena of Stony Brook University, who has developed an algorithm called Shorty for the *de novo* assembly of read lengths of around 20 to 30 base pairs.

“Up until very recently, a lot of it has been sort of hypothetical because people didn’t know how good the technologies were going to be, how long the read lengths were,” he said. “But now the technologies are getting mature enough and there’s a clear vision of what kind of data is going to be produced by these — or a much clearer version than before.”

Currently, Skiena said, “We’re at a point where we’re starting to work with real and simulated data from companies as opposed to our fantasies about what their data was going to be like, given published specs.”

René Warren of the BC Genome Sciences Center said that his group is at a similar point. Warren co-authored an applications note that was published in *Bioinformatics* in December describing an algorithm called SSAKE (Short Sequence Assembly by progressive K-mer search and 3'read Extension), which clusters and assembles 25-mers into longer contigs.

Warren said that he and his colleagues began developing the algorithm last fall before they had a next-gen sequencer, "and at that time, it was pretty much accepted that no one would be interested in doing *de novo* assemblies with that data because of the read length," he said.

While the *Bioinformatics* paper was published based solely on simulated data, Warren said that his lab has since begun generating sequencing data from an Illumina Genetic Analyzer and has just started running SSAKE on real data.

"Our expectations were not really high," he acknowledged, but early results appear to be promising, he said.

In one example, using a human BAC that Illumina recommends as a resequencing control, the GSC group generated 490,000 25-mers for 70X coverage of the BAC. Warren used SSAKE to assemble those reads into 13,000 contigs with an average size of 44 bases.

Using only those contigs that were longer than 75 nucleotides — around 10 percent of the total — Warren found that they covered 98.4 percent of the BAC with 96-percent sequence identity.

"I was pretty happy to see this," he said.

Other algorithms have also relied on simulated data until very recently. At the Advances in Genome Biology and Technology conference in February, Jonathan Butler of the Broad Institute presented early results from an algorithm called ALLPATHS that was developed to assemble read lengths of around 30 base pairs.

In his presentation, he discussed results based on simulated data from a reference genome that was "assigned an error pattern modeled after real [Illumina] reads," according to the abstract for his talk. At the time he said that he "expect[ed] real data soon," though he acknowledged that "this will present new challenges."

"We're at a point where we're starting to work with real and simulated data from companies as opposed to our fantasies about what their data was going to be like, given published specs."

ABI's De La Vega said that the proof of the pudding for these new assembly algorithms will be in real data, which includes errors and biases that simulated data can't account for.

"Definitely there are going to be more errors right now than what you see in CE, and that needs to be dealt [with] at the assembly level, too, because a short read with one or two errors suddenly becomes a lot more difficult to align and to assemble," he said.

Even those synthetic data sets that do account for errors do a poor job of replicating real biological data, he said. "They are assuming random error models, and in reality there is always going to be some bias in the system."

For example, he said, most sequencing platforms exhibit some bias across the length of a read, or bias related to GC content. "Those things I don't think have been taken into account right now," he noted, "but if the error profile is well understood, and those biases are understood, than you can compensate in the algorithm for that."

As real data becomes available to test these algorithms, he said, “I think that then there is going to be the need to do some tweaking to adjust for the error profiles of these new technologies.”

Paired-Read Promise

One relatively recent development in the field of next-gen sequencing is the availability of paired-read data, which “is really a critical element in being able to make *de novo* assemblies from these platforms,” De La Vega said.

“If there are CE backbone scaffolds maybe you can get away with not having the mate pairs, but if you want an assembly, I think all of these algorithms are assuming that mate pairs are going to become available,” he said.

Illumina’s Cox agreed. “Clearly, as with any assembly, the resolution of repeat regions is the tough bit, and most of the short read assembly algorithms I know rely on read pairing to help out with this,” he said, noting that the availability of Illumina’s paired read data “opens up exciting new possibilities here.”

Stony Brook’s Skiena said that his Shorty algorithm relies on the availability of mate pairs. “The fact that you’ve got two reads that are a certain distance apart from each other, where there is some expected distance and some variance there — this turns out to be a powerful thing to help you in assembly, and it turns out to be much more important in assembling short reads than in long reads,” he said.

SSAKE doesn’t use mate pair data, but Warren noted that the algorithm will primarily be used to characterize unknown genomes within metagenomics data sets via Blast searches and gene-prediction tools, rather than *de novo* assembly of individual genomes.

For pure *de novo* genome assembly using microreads, he noted, paired-end data will likely be necessary. “Until we have a better feel for sequence or base quality with these short reads, and the read length becomes a bit bigger, and we have some information to put these contigs in the context of the genome — pairing information — it’s not going to be trivial, and people will have a problem in assembling very contiguous sequences,” he said.

Despite advances in the field, “I haven’t seen yet a *de novo* assembly completely out of short reads,” De La Vega said. “I expect that’s because the data sets are just now being generated, and [the fact] that for short reads, mate pairs are necessary.”

However, he added, the availability of such an assembly may be only a few months away. “I’m confident that, based on the people I know and the data that’s been generated, that by the summer we’re going to start looking at some microbial genomes assembled *de novo* from short reads,” he said.

Skiena said even though the quality of the data from next-generation sequencers is “still a moving target,” he’s confident that *de novo* assembly is possible with microreads.

“Do I believe you can do *de novo* assembly of bacteria to an interesting state using reads of length 20? The answer is yes. Exactly what ‘interesting’ means is a separate question, but I am convinced that is doable,” he said.

“The question about higher organisms is maybe a little bit more open, but I still believe that if you work hard enough at the assembly and have high enough coverage, I think you could produce something interesting.”

Genedata, Cellomics Integrate Software to Meet Demands of 'Very Mature' HCS Market

By Bernadette Toner

Bioinformatics firm Genedata and high-content screening provider Cellomics have taken steps to further integrate their HCS software products to meet the requirements of a market that officials from both firms claim is rapidly becoming more sophisticated.

Next week at the Society for Biomolecular Science conference in Montreal, Genedata will present a new "workflow" it has developed in collaboration with Cellomics for importing data from HCS experiments into its Screener software, while Cellomics plans to announce a new HCS analysis product that can integrate more effectively with Screener and other third-party tools.

Mark Collins, senior product manager for informatics at Cellomics, a subsidiary of Thermo Fisher Scientific, said that the company's customers have been asking for tighter integration between their HCS software and the rest of their drug discovery IT infrastructures. He called this a good sign because "it shows that the technology of high-content [screening] is very mature now and people trust the results and they want to use those results in making decisions about compounds or targets."

Collins noted that Screener is one of a number of tools that drug-discovery groups currently use to manage traditional biochemical screening data, so it made sense to work with Genedata to ensure that the integration was as seamless as possible.

Collins noted that the project is "not a formal partnership," but rather an arrangement under which the firms collaborate with mutual customers to link their software products. The relationship is built upon a software integration project that the two companies carried out for Serono and disclosed late last year [*BioInform* 12-15-06].

"We've worked together with Genedata to make our software export data in a format that their software can read in," Collins said. The HCS workflow also works the other way, via a web-based link from Screener back to the image-based information in the Cellomics software, he added.

Underlying the integration capability is a new data standard that Cellomics and several collaborators are developing called MIAHA, short for Minimum Information for a High-Content Assay. The proposed standard, modeled after MIAME and its multitude of "minimum information" offspring in the bioinformatics community, is a combination of two existing standards: MIACA (Minimum Information About a Cellular Assay) and the Open Microscopy Environment.

"The OME standard is really good at describing optical stuff and things to do with microscopes and images, but it doesn't really have enough richness to describe the kind of data you get from a cellular assay," he said. "Whereas the MIACA standard has nothing about microscopy in it."

Cellomics and its collaborators, including researchers at the German Cancer Research Center, merged the two standards into a single standard to address the requirements of high-content screening. Cellomics will be hosting a special interest group meeting at SBS next week to further discuss the proposed MIAHA standard.

Collins said that the company was able to use MIAHA to create a "round trip" that starts with the Cellomics HCS software, exports that data to Screener for analysis, "and then [is] able to reach back to look at the images, which is what everybody wants to be able to do."

Kurt Zingler, head of US business for Genedata, said that the ability to link back to raw cellular images is one thing that distinguishes Screener's HCS capabilities from its traditional HTS features.

"In a typical primary assay, you may measure one value in a million compounds, but in a high-content assay, you may have 50,000 or 100,000 siRNAs or compounds that you're testing, but you may have five to 25 parameters that you're testing."

In a biochemical screen, he said, the information on the plate or well "is really just a number and there's nothing to look at." In a high-content assay, on the other hand, "the scientist at this point usually wants to go back and look … at the image or the set of images that was used to make that call."

Zingler said that Screener also includes improved capabilities for multiparametric analysis to meet the demands of the HCS market.

"Our understanding is that people are putting things together and finding the problems with them and starting to move toward systems that are more adept at handling multiparametric data," he said.

"In a typical primary assay, you may measure one value in a million compounds, but in a high-content assay, you may have 50,000 or 100,000 siRNAs or compounds that you're testing, but you may have five to 25 parameters that you're testing," he said.

"Generally what people have done with high-content assays is treat them as a standard high-throughput assay," he noted. In the example of nuclear localization, he said, "they'll pick one number out of those 20 different parameters and say, 'This is our measure of nuclear localization and what we'll use to determine a hit or to determine activation.'"

However, Zingler said, the HCS field is moving toward an understanding that it needs to take all of those parameters into account and determine which combination of parameters is most informative. "Instead of just grabbing the one thing, we'll actually rank the 20 parameters that they may be measuring and say which of these parameters is most relevant for identifying the positive reaction or the looked-for reaction," he said.

Zingler compared the approach to that used in microarray analysis, where researchers would prefer to discover single-gene or single-protein biomarkers that indicate whether a patient has a disease, "but what they're really getting out is four or five things that together really give you a picture."

Genedata is "doing the same in the high-content world to say, 'Let's make use of all those parameters and figure out what combination of those parameters actually gives us the best answer and the most predictability," he said.

Stratagene Buy Will Give Agilent Two Array Software Packages; Post-Merger Plan Unclear

By Bernadette Toner

Agilent Technologies last week announced that it plans to acquire Stratagene for $246 million in cash in a deal that will add the smaller company's consumables and molecular diagnostics to a product portfolio that is heavily weighted toward instrumentation.

While Agilent officials claimed last week that Stratagene's products and experience are "highly complementary" to its life sciences portfolio, there is one area of overlap in the firms' bioinformatics portfolios. Both companies sell microarray analysis packages that are very well established in the research community: Agilent sells the GeneSpring suite of tools that it picked up in its 2004 acquisition of Silicon Genetics, while Stratagene sells the ArrayAssist software package.

Officials from both firms told *BioInform* this week that they could not disclose details of any plans for the software business before the merger closes.

The companies expect the deal to close in around 90 days.

Unification Program

Agilent is already in the midst of an effort to "realign" its informatics portfolio in an effort to better integrate several products it picked up through previous acquisitions.

Jordan Stockton, informatics marketing manager at Agilent, told *BioInform* in January that the realignment should result in a more unified software platform [*BioInform* 01-26-07].

While Agilent officials claimed last week that Stratagene's products and experience are "highly complementary" to its life sciences portfolio, there is one area of overlap in the firms' bioinformatics portfolios. Both companies sell microarray analysis packages that are very well established in the research community.

"We've fairly recently taken a look at all of our informatics assets and sort of realigned them to really speak toward there being a single platform — basically the Agilent informatics platform — to address the entire laboratory's needs," Stockton said at the time.

Stockton declined to provide specifics on the realignment plan, but noted that GeneSpring would factor heavily in the platform.

As of January, Agilent had not yet determined whether to rebrand any of its current products as part of the realignment process.

Even though informatics contributes very little to Agilent's overall life science revenues, the company considers software to be a key part of its strategy in the marketplace.

In January, Nick Roelofs, Agilent's general manager of life sciences, included software as one of four growth initiatives for the company at the JP Morgan Healthcare Conference in San Francisco.

Describing informatics as “really fundamental in transforming how laboratories analyze data and handle data,” Roelofs estimated the informatics market at around $600 million and said that Agilent currently has around 12 percent market share and “tens of thousands of users” in the sector.

While noting that $600 million is “not a particularly high revenue opportunity for any company, and certainly not a particularly high revenue opportunity for us,” he said that informatics is “a fundamental backbone across the laboratory” and therefore an important focus for the firm.

At the time, Stockton declined to comment on whether Agilent was looking to acquire additional informatics firms to fill in any gaps in its software portfolio, but he cited partnerships with pathway informatics companies as “a big emphasis that we’ve had over the last year that continues to be even more of an emphasis” for the GeneSpring product line in particular.

In addition to ArrayAssist, Stratagene sells a pathway analysis package called PathwayArchitect, which would expand Agilent’s software offerings in that area, as the company does not yet sell a pathway informatics package.

Even in the area of microarray analysis, the two firms have taken slightly different paths. While both GeneSpring and ArrayAssist were originally designed for gene expression analysis, the companies have evolved in slightly different directions to address the requirements of emerging microarray applications.

Agilent, for example, offers software packages for array-based comparative genomic hybridization and chromatin immunoprecipitation-on-chip analysis, while Stratagene has developed versions of ArrayAssist for exon arrays and copy number analysis.

Humboldt University Team Finds Little Overlap in Eight Human Interaction Maps

Matthias Futschik,
Institute for Theoretical Biology, Humboldt University

The number of publicly available resources for human protein-protein interaction data is on the rise, but a recent study led by researchers at the Institute for Theoretical Biology at Germany's Humboldt University indicates that these resources have very few interactions in common.

The study, published in the online version of *Bioinformatics* in January, compared eight publicly available human interaction maps of three different types — manually curated, computationally predicted based on interactions between orthologous proteins in other organisms, and large-scale yeast two-hybrid scans.

The researchers note in the paper that they were surprised at how little overlap they found between these resources: Out of a total of 10,769 proteins, only 10 were found in all eight maps; and out of 57,095 interactions, none was found in six or more maps.

As part of the project, the Humboldt University team developed its own integrated human protein interaction database called the Unified Human Interactome, or UniHI, to provide access to all available human protein interaction information from one entry point.

BioInform spoke with Matthias Futschik, the lead author on the paper, this week to discuss these findings and their potential impact on biological research, and to get additional information about UniHI.

Considering that the initial comparisons between yeast protein-protein interaction maps in 2002 generated so much interest, it seems odd that no one has compared human interaction maps prior to this. Why do you suppose this is the case, and what drove you to undertake the challenge?

In contrast to yeast, the human interaction networks haven't been around for such a long time. The literature-based ones were developed in 2000 or 2001, and the orthology-based ones, which are purely computational, started in 2003, but four of the eight networks we compared were only available in 2005.

We actually started at the end of 2005 and it originated from a collaboration with Erich Wanker [of the Max Delbruck Center for Molecular Medicine], who is a coauthor on one of the interaction networks. They did yeast two-hybrid screens and they published in *Cell* [A human protein-protein interaction network: a resource for annotating the proteome. *Cell*. 2005 Sep 23;122(6):957-68] — the first human interaction network based on yeast two-hybrid screens.

So because we were in a collaboration, we started to compare it to other available networks, and at this time there was no second yeast two-hybrid, so we compared it with [the Human Protein Reference Database], and then when the other yeast two-hybrid came out, the one by [Marc] Vidal's group [Towards a proteome-scale map of the human protein-protein interaction network. *Nature*. 2005 Oct 20;437(7062):1173-8], we were surprised that not many interactions were in these three networks.

So this was the motivation for why we started to systematically examine the coherency and the concurrency between and within interaction networks. We wanted to see if our first results were typical for the current state of interaction networks, or whether they were because the yeast two-hybrid maps were maybe outliers.

But it turned out that the low number of interactions in common between networks is quite a common feature in the current stage of the human interactome.

You said you were surprised that there were so few interactions in common, but wasn't that in line with what was seen with yeast? Or have there been improvements in the yeast two-hybrid methodology over the past few years that led you to believe the results would be better in human?

Actually, we thought so, because … these techniques have improved over the last few years, so we assumed that you would catch more interactions that are known in other interaction networks.

But that's obviously not the case.

No, and there are probably many reasons. One reason is that maybe in contrast to yeast, for human proteins the interactome is probably more dynamic. So you won't find some protein modifications when you're using only yeast two-hybrid interaction screens.

You mentioned that you looked at the coherency and concurrency of these networks for this analysis. Can you explain what you mean by those terms and why they're significant for this study?

Concurrency is just how many interactions of one network that you find in another network. So it's just the overlap. The total overlap was quite small, but we saw some quite considerable tendencies. One important finding was that networks generated by the same approach have got a larger overlap. On the positive side, this means they are probably reproducible because if they weren't there would be a random overlap.

On the other side, they have got tendencies, or internal biases, in them. So we also checked if there is an enrichment for proteins of certain functions. One example was that there is a large enrichment in literature-based networks for signal transducers or regulatory proteins. So this means they are well studied in these networks.

With coherency, we took several approaches and one approach is, based on the knowledge we have about proteins, one can assume that proteins of the same function act together. So we checked if this is true for the interaction networks, and you see that they are all of them coherent to a certain degree. Of course, the literature-based ones perform better in this test, but then you must also say that the knowledge about function is derived from the literature, too. So it's not a truly independent benchmark test because it's maybe derived from the same publications. So it's a catch-22 a little bit.

So to avoid this, we also used expression data from the Gene Atlas. This is a large expression data set with different human tissues, and we checked if interacting proteins are co-expressed, because people have found that in yeast, frequently interacting proteins are co-expressed. So we checked this too, and you get quite high co-expression relative to random, and co-expression to literature-based ones and orthology-based ones, and to a lesser degree for yeast two-hybrid networks.

But again, with each of these tests, you can say on the one side it catches a feature, but on the other side, there are a lot of interactions where we know they are transient, so we know they don't need to be co-expressed. Often this co-expression only exists in the stable complexes.

Would you consider any of the three types of networks to be more or less reliable than the others? Or is it just a question of knowing what the limits of each of them are?

On the one side, one has to ask, 'What's reliable?' because in contrast to the DNA in the genome, the interactome is not stable. So under different conditions, different proteins, it heavily depends on modifications, so when you look at these interaction networks, and you see these hairballs, this is a projection of many, many conditions at the moment. That's one of the major problems I think in the field, that we have to define more conditions. So some proteins may interact under one condition — in, for example, a liver cell — but under other conditions — in a nerve cell — they would hardly interact or not even be expressed.

So the reliability depends on the conditions. Most people would say the literature-based ones are more reliable, and people use them as the gold standard. And I think that maybe they are more reliable than the two others, but they definitely have got a high false positive rate, too, in the sense that there is a heavy inspection bias. And several papers will come out, by us and by other groups, that probably will point this out — that if you are really looking hard for some interactions between some proteins, then you will find them.

So in a way, I think it's important to see the limitations of these approaches, and to know what biases and what pitfalls they have.

What does this mean for researchers who are using this data? As you mention in the paper, one goal for these networks is to use them as frameworks for modeling and simulation in computational systems biology, so what effect would the current state of this data have on these efforts?

I think it means that they first have to say where this data comes from … [and] to see if such a network is actually expressed at the same time under the same conditions. So just taking out a network and trying to do modeling with it, that won't be very reliable.

Therefore, for example, with UniHI, our database, we're trying to provide one more resource. Our philosophy is not to give a final network because this will depend on the conditions. If you're looking at a certain cell type it might look totally different from another cell type, so we want to give researchers a tool so that they can say, 'I'd like to search for protein interaction networks that are specific to the brain because I'm interested in neurodegenerative diseases.'

Our aim is to provide really more of a dynamic network and to give integrative tools for researchers to get out of this mass of interaction data the most suitable data for the task they want to solve.

So UniHI provides a sort of filtering capability to create subnetworks based on particular cell types or experimental conditions?

Yes. And our philosophy is to link up to different databases, so UniHI should be an entry gate to the human interactome. We don't say, 'Just use the literature ones or just use the yeast two-hybrid ones because they're experimentally verified and essentially tested.' We say, 'OK, here are the links.' We give them a lot of filtering possibilities, so you can straight away say, 'I don't use the orthology-based ones because they are computationally based, and I only want to have interactions that are experimentally tested.'

We also want to give real-life information about where the genes are corresponding to the protein expressed. So in some tissues, if you don't find expression you wouldn't expect the interaction of the corresponding proteins. And we also want to give filtering functions that people can say, 'I only want to have interactions that are reproduced by two experiments or by two different experimental techniques.'

On the one side, there is of course a [requirement] that the user has some insight as to what he's looking for. So we're not presenting something that's just plug and play. What we want to present is, for a researcher who is competent in the field of interaction networks and interaction network modeling, a tool that is as flexible and as powerful as possible, but also narrows down this large interaction space to some manageable format that they can use for systems biology.

So this is more of a way to manage the available information from different resources rather than a monolithic resource of all of those interactions in one place.

We don't think there is a static, ultimate interaction network, but there are lots of interaction networks out there and we want to link them and just give researchers the freedom to look for their specific ones based on other information.

The idea of UniHI was not to download everything and put it in one place, but to provide links to all of the interactions. We put together all the networks, but we provide links to the original databases so that people can go and find more information and more annotation about specific interactions.

So it's an entry gate. … You have to find your own way, but we try to give as much information as possible.

Entelos, EBI, UK BBSRC, Tripos, Provid Pharmaceuticals, Galapagos, Inpharmatica, Boehringer, GeneGo, Gene-IT, GenomeQuest

Entelos Delays Release of First Annual Financial Report as a Public Firm

Entelos said this week that it has pushed back by two weeks the release date for its preliminary financial results for the period ended Dec. 31, 2006.

The company will issue its results on April 23 instead of April 12.

The Foster City, Calif.-based company went public last April on the London Stock Exchange's Alternative Investment Market [*BioInform* 06-16-06].

“As a result of additional work required for a Delaware company to satisfy the AIM rules, the company must allow its accountants to complete an appropriate audit before releasing its first set of preliminary results,” Entelos said in a statement this week.

Entelos said that it expects its 2006 results to be in line with previous guidance.

EBI Wins Japan Partnering Award from UK BBSRC

The European Bioinformatics Institute said this week that the UK Biotechnology and Biological Sciences Research council has granted it a UK-Japan Partnering award for a project called Interfacing Standards and Ontologies in Systems Biology.

The project is a collaboration between Nicolas Le Novere of EBI, Ken Fukuda of Japan’s AIST Computational Biology Research Center, Douglas Kell of the Manchester Center for Integrative Systems Biology, and Hiroaki Kitano of the Tokyo Systems Biology Institute.

According to the BBSRC website, the grant is worth £45,000 ($89,200).

BBSRC said that its UK-Japan Partnering award program is designed “to gain access for BBSRC-supported scientists to the rapidly developing research base in Japan and to take advantage of the tremendous opportunities collaborative activity represents.”

The first awards under the program were granted in 2001. Since its inception, 28 awards have been made totaling £1.14 million ($2.2 million).

Tripos Terminates Sale of Discovery Research Business to Provid

Tripos will not sell its Discovery Research business to Provid Pharmaceuticals, according to a recent US Securities and Exchange Commission filing.

Tripos said in the filing that it told Provid on April 2 that it has terminated the stock purchase agreement the companies signed in early January, under which Provid would have acquired Discovery Research for $2 million in cash [*Bioinform* 01-05-07].

Tripos said when the deal was made that it would hinge on whether Provid could gather the necessary financing, and later warned that Provid was having trouble raising the money [*BioInform* 03-02-07].

The initial purchase agreement gave both companies the option to terminate the deal on or before March 31, and Tripos said it exercised that option last week.

Tripos sold its Discovery Informatics business in March to Vector Capital for $26.2 million, as part of its ongoing efforts to liquidate the company [*BioInform* 03-30-07].

In last week's SEC filing, Tripos reasserted its intention to sell the remaining parts of the company and to settle all debts, liabilities, and obligations to its shareholders.

The company stated it is still engaged in efforts to sell Discovery Research, but would not guarantee it would be able to make a deal "on satisfactory terms."

Galapagos Cuts Value of Inpharmatica Acquisition by $9.5M

A missed milestone by Inpharmatica has prompted Galapagos to cut around €7 million ($9.5 million) from the value of its acquisition of the company, Galapagos said this week.

Galapagos has already issued 613,000 new shares to former Inpharmatica stockholders, which Galapagos had agreed to acquire last December for as much as €19 million ($25.4 million) in stock [*BioInform* 12-08-06].

Galapagos planned to grant Inpharmatica shareholders as many as 2.2 million shares priced at €8.82 apiece based on certain milestones, cash on hand, and worth as a business. Galapagos issued 623,000 shares at the time of the agreement.

This week, Galapagos said that Inpharmatica failed to meet one of these milestones, resulting in a loss of 815,000 shares.

Galapagos said it will issue a maximum of 113,000 more shares in May to complete the acquisition, bringing the total exchange to around €12 million.

Boehringer Expands License to GeneGo's MetaCore Software

Boehringer Ingelheim has expanded a license agreement for GeneGo's MetaCore data-mining software suite that the companies signed last year to cover "global locations and multiple departments."

Under the expanded license, Boehringer will use the software, which uses a database of human disease and toxicity information, in its toxicogenomics programs.

Financial terms of the agreement were not disclosed.

Gene-IT to Change Name to GenomeQuest

Gene-IT is changing its name to GenomeQuest, after its flagship sequence search software, the company said this week.

GenomeQuest is a gene sequence search product designed for biologists and IP lawyers. It contains more than 40 million patented sequences, the company said.

Michael McManus, vice president and general manager, said in a statement that “uniting our product and company names is a natural evolution for our business.”

Blast 2.2.16, GeneDirector 3.5, UCSC Genes, Image-Pro Plus Version 6.2, Merlin version 1.1

The **National Center for Biotechnology Information** has released **Blast 2.2.16** here. The new release includes performance improvements for nucleotide searches and some bug fixes. This will be the last release for **Solaris 8** and **Tru64**, NCBI said.

BioDiscovery has released **GeneDirector 3.5**. The new version of the software includes "extended" data management support for **Affymetrix**, **Illumina**, and **Agilent** microarray platforms, the company said. GeneDirector 3.5 includes several "major upgrades" to its backend data-management engine in order to improve support for these microarray platforms, the company said.

The Genome Bioinformatics Group at the **University of California, Santa Cruz**, has released a new gene prediction set, **UCSC Genes**, via the **UCSC Genome Browser** (hg18, NCBI Build 36). The annotation set includes putative non-coding genes as well as protein-coding genes and 99.9 percent of **RefSeq** genes. It supersedes the existing Known Genes annotation on the hg18 assembly, UCSC said.

In conjunction with the release of the UCSC Genes data set, UCSC has also released companion annotation track on the hg18 assembly called **Alt Events**, which shows alternative splicing, alternative promoter, and other events that result in more than a single transcript from the same gene.

Media Cybernetics has released **Image-Pro Plus Version 6.2**. The new version of the company's scientific image processing and analysis software includes the ability to remove haze from image stacks via built-in **SharpStack** image deconvolution tools. These tools include real-time 2D deconvolution, nearest neighbor and no-neighbor deconvolution, and the inverse filter algorithm.

An alpha release of **Merlin version 1.1** is available here. The release includes support for estimation of missing genotypes and support for quantitative trait association analyses. There are also several other smaller enhancements, including support for base-pair allele labels in pedigree files.

Robert Strausberg, Craig Venter, Eric Eisenstadt, Marv Frazier, Julie Gross Adelson, Aimee Turner, David Urdal, Mark Gabrielson, Jules Blake, Robert Stanley, Patricia Rougeau

Robert Strausberg has been named deputy director of the **J. Craig Venter Institute** as part of a broad reorganization of its two research divisions — the **Institute for Genomic Research** and the **Center for the Advancement of Genomics** — into several more tightly focused units.

The JCVI will now consist of 10 distinct research groups: Genomic Medicine, Infectious Disease, Synthetic Biology & Bioenergy, Plant Genomics, Microbial & Environmental Genomics, Pathogen Functional Genomics, Applied Bioinformatics, Research Informatics, Software Engineering, and a Policy Center.

Craig Venter will remain president and chairman of the institute, and Strausberg, who had been president of TCAG, has been named deputy director. **Eric Eisenstadt**, former vice president for research of TIGR, has been named deputy vice president for research.

Marv Frazier, former vice president for research at TCAG, is now the executive vice president for research. **Julie Gross Adelson** has been named general counsel, while **Aimee Turner**, formerly CFO at TIGR, is now CFO of the JCVI.

Gene Logic has elected **David Urdal** and **Mark Gabrielson** to its board of directors.

Urdal is currently senior vice president and chief scientific officer of **Dendreon**, while Gabrielson is a co-founder, director, and CEO of **Pulmatrix**.

In addition, the company said that **Jules Blake** is retiring from the board of directors after serving on the board since 1994.

IO Informatics has tapped **Robert Stanley** as president and CEO. Stanley is a company co-founder and was formerly chief technology officer at IO Informatics.

Stanley replaces former CEO **Patricia Rougeau**, who left the firm at the end of March.

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 2007 GenomeWeb Daily News. All rights Reserved.