Innovations

Industrial-strength profiling

Rosetta Inpharmatics, Inc.

Chemistry & Biology 2000, **7**:R205–R206

1074-5521/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S1074-5521(00)00031-4

In its brief history, Rosetta Inpharmatics, Inc. (Kirkland, WA) has generated a lot of data. Data from 25 000 whole genome array scans, for starters. The company believes that the sheer volume of its data bank is the key to its future success.

When analyzing DNA expression arrays, Rosetta claims that more is more because more allows you to detect patterns of co-regulated genes. "People have assumed that the root information is at the single transcript level," says Rosetta CEO Stephen Friend, "but we are looking for patterns." Once Rosetta has accumulated enough patterns arising from known perturbations, says Friend, the company will be able to deconvolute the changes arising from any uncharacterized state, such as a disease or drug treatment. From there it is a short journey to the identification of drug targets, and the assessment of whether a drug will cause unwanted side-effects.

From the clinic to the boardroom

Friend's foray into company land was prompted by his clinical training. "As a pediatric oncologist I was getting more frustrated at how few of the targets and exciting biology ever got into therapeutic applications," he says. He took a sabbatical year to look at where the inefficiencies were in the drug pipeline. "It was really an eye-opener," he says. "It became really clear the role that serendipity – or brute force – played."

He first attempted to reduce the serendipity factor by using a more intelligent approach. With Lee Hartwell and the Fred Hutchinson Cancer Research Center he set up the Seattle Project. The idea was to use model organisms – notably budding yeast – that were defective in a pathway that is also defective in cancer cells, and then screen for chemicals that selectively kill these strains. This idea – dubbed "compounds that kill in a context" by Friend – was "an idea that could've become a company," he says. For now it remains in the public sector. Tens of thousands of compounds have been analyzed in a collaboration with the National Cancer Institute.

Rosetta has found the key to interpreting mountains of data. Generate more of it

Just as Friend was getting the Seattle Project started, DNA microarrays made their big entry. By 1995 chips were being made by both the cDNA spotting method of Patrick Brown (Stanford University, Stanford, CA; later exploited by Synteni, Inc. before it was bought by Incyte Genomics, Inc. (Palo Alto, CA)) and the photolithographic synthesis of oligonucleotides developed by Affymetrix, Inc. (Santa Clara, CA). With the encouragement of genomics guru Leroy Hood (University of Washington, Seattle, WA), Friend considered a new kind of brute-force approach to drug discovery based on DNA chips.

"We were thinking about how you might use the genome as a sensor," he says. With further study Friend realized that "the ripple-down effects when you disrupt one protein are really tremendous [because] the cell is hard-wired together." Arrays could be used to detect those ripple effects.

Array companies already existed. But still Friend felt there was a gap in the market. "The companies that were out there were interested in arrays as an end," he says. "There were companies that could make arrays but there weren't people working out how to use them to impact drug discovery. It was more 'isn't this a great technology'." In contrast, he says, "we could see that there would be a large mass of data coming off the arrays and no one would know how to use it."

Jet-powered arrays

One of Rosetta's early assets was a new method of making arrays. Alan Blanchard, first at Caltech (Pasadena, CA) and then with Hood, had been working on converting garden-variety ink-jet computer printers into DNA synthesizers. The simplest version of the machine simply spits oligonucleotides onto a glass surface, as does a bubble-jet device recently unveiled by the Canon Research Center (Kanagawa, Japan). But Blanchard's device can also be used to synthesize oligonucleotides of up to 70–80 nucleotides in situ.

Ink-jet arrays avoid the complexity and expense of photolithography, and offer greater control over spot size (and thus greater precision and higher density) than conventionally spotted arrays. For the past year Rosetta has been using 3 inch by 3 inch arrays that cover the majority of transcripts from a genome. But Rosetta was not looking to become an array company, so they have transferred rights to the 'FlexJet' technology to Agilent Technologies, Inc. (Palo Alto, CA). Commercial release of the FlexJet arrays by Agilent is expected by the end of the year 2000.

Software and databases

What Rosetta does sell is a database with reference expression patterns, and Rosetta Resolver™, a software package to analyze existing and newly generated array data.

Resolver stores data in a particular architecture so that the data can be called up and queried. It has some proprietary algorithms and tools in addition to the common public-domain programs, and most importantly the different tools are integrated.

Rosetta is releasing some parts of the Resolver package for free. These tools use a Gene Expression Markup Language (GEML) to convert diverse expression data into a standard format. But the integrated storage and analysis package is not cheap. Prices range from $250 000 for a system allowing two concurrent users, to $1 million for larger systems.

The need for Resolver comes back to the volume of data. “Very few academics can afford to put a real database back end on the experiments,” says Mark Boguski, Rosetta’s senior vice-president of research and development. And yet, he says, “you need coherent data sets to make reliable predictions. That requires the assimilation of large amounts of data collected in a controlled fashion. Only then do you believe the data.” Rosetta generates coherent data sets in customized collaborations that are separate from Resolver itself.

“Having a limited number of data sets, the academics focus on the low-lying fruit,” says Boguski. Richard Hynes of the Massachusetts Institute of Technology (Cambridge, MA) admits as much, based on his recent array experiments in which he found a role for RhoC in metastasis. “We deliberately set our threshold quite high to eliminate the noise,” he says. “When we set it that high there weren’t many genes left to deal with.”

But Rosetta reported in a recent paper in *Cell* (vol. 102, p. 109) that changes as low as 1.5-fold were often the most important, and according to Boguski “you need a tremendous amount of data to see that level.”

The *Cell* paper is a test case for the Rosetta method. “We’re putting our effort into how to design experiments,” says Boguski. “Three years ago if you had asked us what is important it would have been monitoring individual transcript levels,” says Friend. Most researchers still focus on this approach: what is up, and what is down, for example, during the cell cycle. But Rosetta realized that there would never be enough easily scored phenotypes to characterize every gene.

The best way to cover all yeast genes comprehensively was to score the patterns of changes in a vast panel of mutants. (With other organisms Rosetta will have to fall back on scoring a mixture of phenotypes, but the idea of analyzing a vast collection of changes (not just the change of interest) remains.) In the *Cell* study, patterns from 300 yeast mutants and drug treatments were used to assign eight uncharacterized genes to four different pathways and determine one drug target. The drug target was clear based on the similarity of the expression patterns after either drug treatment or mutation of the gene encoding the targeted protein.

Along the way has come a pleasant surprise. “When we first thought about doing this, we thought there would be a ten to twenty year cycle of [deciphering patterns] because people would want to know the why – what the pattern meant,” says Friend. But “the important thing was to say how similar the patterns are. The pattern recognition can go on without understanding the why.”

The long-term plan for Rosetta is to partner with drug companies in specific areas of biology to provide extensive gene expression services. A major function of those services will be to detect the likelihood of a drug producing a side-effect. Whole-genome arrays should enable Rosetta to detect all the effects of a drug on a cell, including those directed against unanticipated targets that cause side-effects. Drugs could be re-designed to avoid such interactions before the drugs enter the clinic. Competition may come from a planned company called Merrimack, which is hoping for similar outputs from its protein arrays (see *Science* vol. 289, p. 1673).

A star recruit

Rosetta’s primary focus is on bioinformatics. It is a coup, then, that they managed to hire Boguski, one of the founding members of the US National Center for Biotechnology Information (NCBI). For 12 years the NCBI has managed and massaged the increasing flood of public-domain DNA sequence data.

At first, says Boguski, users of NCBI databases had it easy. “In the early days every [sequence database] entry was the result of two years of careful work. Each hit was incredibly informative.” Now, he says, “you almost always get a hit in the database but it almost never means anything.”

The challenge for both Boguski and Rosetta will be to take bioinformatics beyond its ‘my favorite gene’ era. “Bioinformatics has not caught up with looking at 1000 gene queries at a time and returning interesting and important information,” he says. That sentiment is shared by academics trying to do their own array experiments. David Eide (University of Missouri, Columbia, MO) took a few days to generate array data then “spent several months staring at it and trying to make sense of it,” he says. “The sheer mass of data can be almost paralyzing.”

Big biology

Rosetta’s approach is, at its most basic, an argument for big biology. “Absolutely that trend is going to happen,” says Friend. “Once you have a critical mass of data you can say a lot more about the biology. But unlike physics, a single lab can focus in on a very narrow area and still make a contribution.”

Rosetta, of course, plans to be more ambitious. Most of its test case experiments have been in yeast, but the company has been busy working with mammalian cells. The results of that work will be along shortly. “Stay tuned,” says Boguski, “for some landmark scientific publications in mammalian biology.”

William A Wells
1095 Market Street #516, San Francisco, CA 94103-1628, USA; E-mail: wells@biotext.com