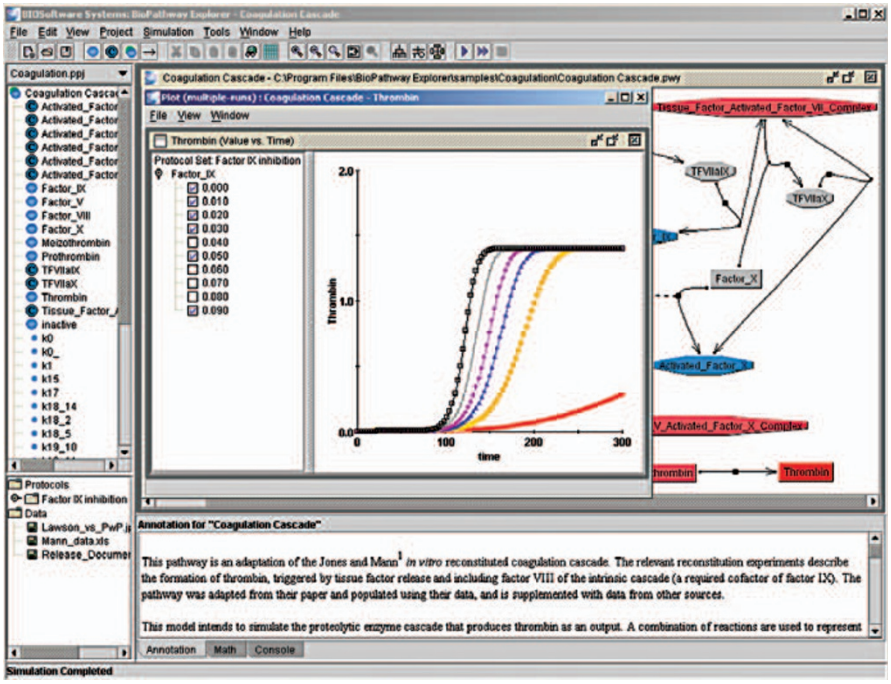
Bioinformatics—from genes to pathways

Combined with the right computational tools, genomic data can uncover unknown pathways to cellular processes. Because few researchers have the resources to develop their own bioinformatics software, companies have stepped in to meet this need. Laura Bonetta reports.

The sequence of the human genome has provided a complete 'parts list' for constructing cellular functions. Yet negotiating how all the pieces fit together is a formidable task. Of all the tools scientists have at their fingertips, DNA microarrays are perhaps the most popular. These tiny chips can monitor the expression of thousands of genes at once and help group them into functional categories—for example, genes that are expressed to a greater or lesser degree in response to a drug or at different times during development. They also allow rapid genotyping of DNA sequence variations among individuals (See **Box 1**, 'SNP CHIP Software is on It's Way)

During the past decade, large-scale expression profiling experiments have generated a deluge of data. At the same time, commercial software packages have come onto the market to help extract meaning from these data. These products can pick out relevant subsets of genes with various analytical methods, scout the literature and databases to find commonalities among gene products, and draw interactive graphs and diagrams that can be queried with a click of the mouse.



BioPathway Explorer displays experimental and simulation data. (Courtesy of BIOSoftware Systems, Inc.)

The promise of pathways

Identifying hundreds of genes whose expression is markedly higher in one sample than in another provides a large amount of potentially valuable data; the trick is to home in on the genes that are relevant to the questions being asked. One way to do that is to find a subset of genes that are functionally connected through common pathways.

Justin Lamb's group at the Broad Institute in Cambridge, Massachusetts, uses an approach that he refers to as "functional annotation". His method, based on the Kolmogorov-Smirnov statistical test, first identifies a gene expression signature that lights up when a pathway is active, say when an oncogene is ectopically expressed inside a cell. Next, it mines sample microarray data sets to see if the gene signature matches a pattern of genes differentially expressed in, for example, a particular tumor. "If we find a match, it will tell us that the oncogene is likely to be involved in the tumor," says Lamb, whose work has uncovered functional relationships between different cancer genes[1].

Lamb's group is now conducting proof-of-concept experiments to determine whether it would be feasible to derive a gene signature for every human gene. "If we need to profile every gene in 50 cell lines because their effects are exquisitely context-dependent, it is an impossible task. But if we only have to study two or three cell lines, it is doable," he explains.

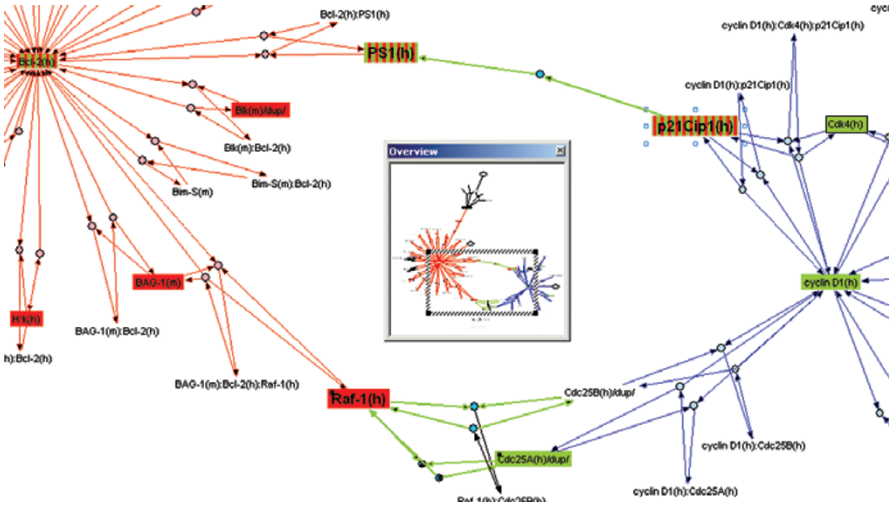
From signals to gene lists

Lamb's work is but one example of the ways in which microarray data can be mined to extract biological meaning. Whatever the approach, the first step in microarray analysis is to obtain a list of genes that are differentially expressed. Several suppliers of microarray hardware, such as Affymetrix, Agilent Technologies and

Applied Biosystems, provide software that allows the user to go easily from fluorescent signals to a list of genes.

Applied Biosystems' 1700 Chemiluminescent Microarray Analyzer, launched in April, was designed to "show exactly what you are measuring on a microarray," says Clark Mason, senior product line manager for gene expression arrays. The system carries integrated software for image analysis, quantification and normalization, as well as an Oracle database of annotations, including gene names, cross-referenced IDs, gene ontology and Celera's Panther Protein Classification System. The end product is a "meaningful list of genes that is rightfully annotated," says Mason. The software allows seamless and customizable integration with third-party software, such as Spotfire's DecisionSite and Silicon Genetics' GeneSpring, for more complex data analysis.

In a similar vein, the NetAffx Analysis Center, an online resource created by Affymetrix, allows researchers to correlate results from their GeneChip probe array experiments with biological information from both Affymetrix's own and public databases.



Vector PathBlazer finds connections between two groups of genes that are differentially expressed in a microarray experiment. (Courtesy of Invitrogen.)

Data analysis and visualization

A large number of companies offer software for analyzing microarray experiments. "In the last 2 to 3 years there have been great improvements in the accuracy of DNA chips, and they have become cheaper. Most researchers can now obtain good, reliable data," says Eric Olson, director of science at VizXLabs. "It has become possible to devel-

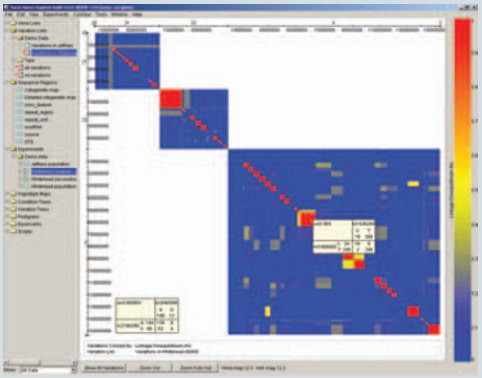
op products with built-in access to standard statistical tools for analyzing these data."

Many software packages provide users with a variety of analytical techniques (including time series and clustering analyses), gene and probe annotations by linking to internal and external databases, and tools for visualizing data and preparing figures. Some products have an even wider range of capabilities, such as integrating data from a wide variety of sources, adding a user's own statistical algorithms, and providing data and project management tools. In most cases, the software will guide a researcher from a list of genes to a first pass at a cellular pathway.

VizXLabs caters to "biologists working at the bench," says Olson. "We spent a lot of time trying to build in processes for the kind of things that biologists would want to do." Its GeneSifter product is entirely Web-based, avoiding the need for high-power hardware in-house. It uses pull-down menus from which the user can choose what kind of statistical tests to use and can set the *P* values and other parameters without having to do any of the number crunching. The result of a typical analysis is a list of annotated genes that are differentially expressed, "but you can also start to ask some questions about the list," says Olson. "[GeneSifter] can tell you whether the genes are involved in cell cycle or apoptosis, or, in a time series experiment, it can find a subset of genes that are expressed at a later time."

BOX 1 SNP CHIP SOFTWARE IS ON ITS WAY

In March 2004, Silicon Genetics released Varia, a product that provides tools for analyzing variations in DNA sequence. Varia is an example of a growing number of commercial products that can be used to build family trees as well as to conduct association and linkage studies and genotyping, in combination with powerful visualization tools.



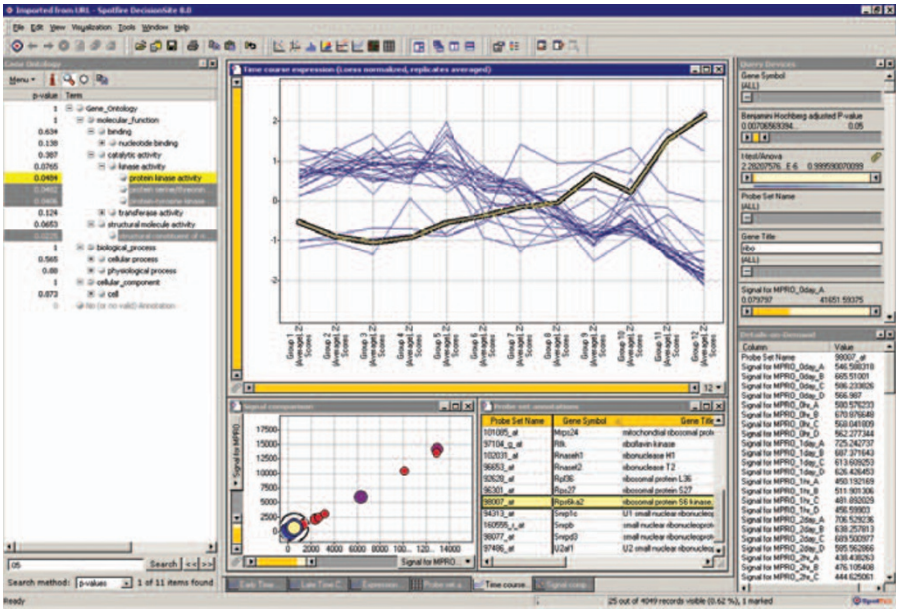
A variety of visualization tools in Varia allow researchers to identify genetic regions that are inherited in blocks. (Courtesy of Silicon Genetics.)

Single-nucleotide polymorphisms (SNPs) represent the most common source of genetic variation. Less than a year ago, Affymetrix launched its GeneChip Mapping 100K Array Set—a set of two arrays that allow the genotyping of >100,000 SNPs with a single primer. "Many labs are collecting large number of samples from large, medically relevant populations and asking sophisticated questions about them. SNPs provide the power needed to address those questions," says Steve Lincoln of Affymetrix.

Presently the number of companies offering software for analyzing SNP data is limited in comparison with those offering microarray expression analysis tools, but that may soon change. "There is a diverse set of methods for analyzing genotype data so it is harder to imagine one shrink-wrap package," says Lincoln. "But we will probably get there."

Although software products like GeneSifter are built with the nonspecialist in mind, they often have the flexibility to satisfy more sophisticated customers. For example, Silicon Genetics' flagship product, GeneSpring, provides "an interface to connect to R Bioconductor, a popular software product for conducting your own custom analysis," says Kevin Wandryk, vice-president for marketing and business development. Similarly, although these software products are designed to analyze microarray data, many of them also handle other kinds of results. "As long as you have IDs and expression information, you can analyze any kind of data," says Wandryk. "Some of our more savvy customers use GeneSpring to analyze data off their proteomics experiments." (See **Box 2**, 'Proteomics—The Next Challenge.')

For researchers conducting large-scale analyses that incorporate different kinds of data and need more customization, Spotfire provides DecisionSite, a product that can analyze data from microarray experiments, high-throughput screens, chemical reactions and proteomics experiments. "Configurable workflows guide users through the analysis process," says Matt Anstett, market manager for life science. With a click of a button, any data in a spreadsheet can be brought



A typical gene expression analysis workflow in DecisionSite. (Courtesy of Spotfire.)

into DecisionSite and visualized as graphs or diagrams that can be quickly manipulated to ask various questions. "One of the features our customers like is how quickly you can answer questions," says Ian Reid, vice-president of applications marketing. Visualizations can be annotated and shared among members of a team, and even emailed to distant collaborators, facilitating interactive discussions that can be archived in the 'library'.

Likewise, Inforsense's Knowledge Discovery Environment integrates different kinds of data for a range of analytical applications. "Extensive normalization and analysis tools are complemented by powerful integrated visualization capabilities," says Stephen Misener. "You can, for example, use brushing and linking between different clustering visualizations to compare one clustering method with another." Comparing patterns obtained with different statistical and analytical methods may help a scientist home in on a subset of genes that is most relevant to the process being studied. "If you are looking for distinguishing features of our software it would be the openness and flexibility of our platform," says Misener. "You can easily select the components and data sources you need, and even add other algorithms or external programs."

Agilent Technologies Inc. provides what they call "bridging informatics," in other words, products that integrate different kinds of data and analyses. "We want our customer to use one set of products to produce results and formulate hypotheses that can then be analyzed using another set of tools," says Francois Mandeville, business manager of informatics solutions. In addition to marketing Rosetta Resolver and Rosetta Luminator gene expression data analysis systems (both developed by Rosetta Biosoftware), Agilent provides Synapsia, "a project assessment tool that enables different team members to exchange and analyze information," says Mandeville. Synapsia imports and stores different data types, correlates gene to protein expression, connects to third-party analysis software such as DecisionSite and GeneSpring, and links to internal and external databases. "Our customers told us that there are a lot of point solutions for data analysis, but there was a need for correlating information," says Mandeville.

Databases for functional annotation

It is vital for most bioinformatics tools to link to external databases. Many software products are available that sift through private or public databases to find, for example, what pathway a particular protein participates in or find a pathway that connects two genes. Databases like GenBank and Gene Ontology (http://www.geneontology.org/) give descriptions about the gene themselves, whereas others, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/) or the Alliance for Cell Signaling (http://signaling-gateway.org), provide information about biological interactions. On the commercial front, examples of 'pathway databases' include Biobase's TRANSPATH-NetPro, a database for molecular path-

ways and cellular network modeling, and GeneGo's Metacore, which also contains disease information data that can be overlaid on the pathways.

Ingenuity pioneered the hand-curated approach. Its database was created from millions of individually modeled relationships. "I have not seen another application with the same breadth and depth of information as Ingenuity Pathways Analysis," says Daniel Siu, director of product management. Deployed through the Internet, the software can accept data from microarray, mass spectrometry or two-dimensional gel experiments. "Any results that can be translated into protein or gene IDs can be uploaded into the application to perform pathway analysis," says Siu.

"Up until recently, pathway databases would have been considered an emerging technology. A few years ago most of our customers looked at these databases but did not find sufficient information about their genes. That has clearly changed," says Affymetrix's Steve Lincoln.

From genes to pathways and back again

New software products take advantage of the knowledge contained in these databases to construct biological pathways and provide additional features. At one end of the spectrum are text-mining products, such as the one offered by Acumenta. It can search for several genes, using all their aliases, simultaneously querying over a dozen sites, including PubMed, the U.S. Patent and Trade Office, NCBI, sequence databases and international patent databases. Researchers can also ask the software to search for co-occurrences among the genes of interest. "If you start with a list of 50 genes, you can do a gene search and gather abstracts for each gene, and then rank them according to the articles that have all or most of the genes in them," says Paul Martinez, vice-president for sales and marketing. In addition, the latest version of the software provides the "ability to annotate documents that

BOX 2 PROTEOMICS—THE NEXT CHALLENGE

As gene expression studies become a regular occurrence in most biology labs, researchers are turning their attention to the next logical step: proteomics. Proteomics is by definition a catalog of every protein in a given organism or tissue for a particular set of conditions.

Protein identification and quantification is typically accomplished by protein separation using two-dimensional (2D) gels or liquid chromatography with subsequent mass-spectrometric analysis of the individual protein spots or fractions. There are many software products that assist scientists in every step of the way, but most of them are from academic labs. The commercial sector has just recently started to pursue this area. "Proteomics is a newer science," says Francois Mandeville of Agilent Technologies. "There is not yet an abundance of data and fewer tools to analyze it."

"Proteomics analysis is technically much more difficult to do than microarray analysis," says Matthias Berth of Decodon GmbH. "In microarray analysis spots turn up at exactly the same places in every experiment, but in 2D gels, the spots turn out in different positions depending on the conditions for running the gel. They are also more difficult to detect." Decodon's Delta2D image analysis software takes raw data from 2D gels and provides a catalog of all the protein spots in the gel, along with information about their relative quantities and other properties, such as isoelectric point. "You can layer different external annotations on top of the expression data," says Berth. With this information on hand and easily visualized, a researcher can more readily narrow down the number of spots to further analyze by mass spectrometry.

Applied Biosystems' 4700 Proteomics Discovery System couples their mass spectrometer with bioinformatics software, which produces a list of protein IDs and quantitative information about them. The software also integrates with Celera Discovery System Online Platform to obtain information on protein molecular functions, biological processes, known modification sites, references and comprehensive genetic information. For those who want to carry out complex statistical analysis of the protein expression data, the product is also integrated with Spotfire's DecisionSite software.

The 4700 Proteomics Discovery Systems. (Courtesy of Applied Biosystems.)

"One of the major technical hurdles in proteomics is getting the same throughput as gene expression studies," says Dale Patterson, senior marketing manager for proteomics at Applied Biosystems. One technology that may resolve this problem consists of protein chips. By linking antibodies to glass slides, all proteins in a sample can, at least in theory, be identified and quantified simultaneously.

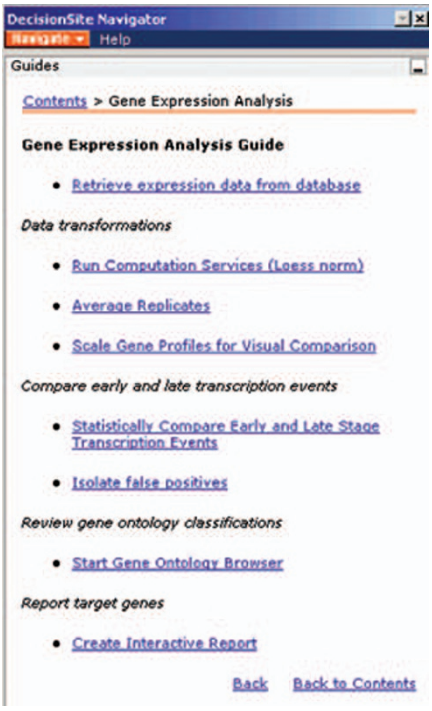
Although it will be some time before antibodies specific to each human protein are available, a small number of studies have already indicated the potential appeal of protein chips. Michael Snyder's group at Yale University constructed a yeast proteome microarray containing ~80% yeast proteins and screened it for a number of biochemical activities[2]. The authors found that, once the proteins have been prepared, proteome screening is markedly faster and cheaper than with the use of conventional methods.

are retrieved by Acumenta and share the annotations with the whole research team, compiling comments from each individual scientist," says Martinez.

Ariadne Genomics' PathwayAssist (which is also distributed by Stratagene) combines text mining with tools to identify biological relationships among genes of interest and to visually display these relationships as interactive clickable maps. PathwayAssist incorporates a text-mining technology called MedScan that can sift through scientific abstracts and articles to extract biological relationships. "MedScan can process any type of file including PubMed abstracts and search through the full text of articles," says Anton Yuryev, director of application science. PathwayAssist can also link to and download information from pathway databases, such as KEGG. In addition, the latest version of the product allows integration with GeneSpring and Iobion Informatics' ArrayAssist gene expression analysis software. Microarray gene expression data can be overlaid on a pathway to show how genes and proteins are affected under different conditions.

BIOSoftware Systems Inc.'s BioPathway Explorer is a tool for drawing pathways, which can then be edited and annotated. "One researcher described it as a whiteboard for pathways," says company representative Ned Haubein. The software allows for all relevant information about a pathway to be incorporated in one place. "You can attach images and overlay time series data on top of a pathway so that you can easily pick out a pattern, rather than having to look at a table of numbers," says Haubein. Another use for this product is to numerically simulate pathway models. Once reaction kinetics have been defined, equations describing an entire pathway are automatically generated by the software. A researcher can then use the software to find out, for example, which step in a pathway is sensitive to various manipulations and then experimentally test these results.

Integration between software for gene expression and pathway analysis can provide a powerful tool for designing experi-



Gene expression data analysis in DecisionSite. (Courtesy of Spotfire.)

ments and interpreting results. Invitrogen's PathBlazer can find all biological reactions in which a set of genes identified by microarray analysis with the companion Vector Xpression software participates. The product uses the TRANSPATH database licensed from BIOBASE and other public databases to identify and visualize reactions and find links between them, while highlighting major features.

By alternating between the Vector Xpression and Vector PathBlazer software, scientists can more quickly home in on a particular gene or set of genes. "For example, with Xpression you can find through one Affy chip experiment that 192 genes are 99% certain of being differentially expressed in thyroid cancer samples. With PathBlazer you find proteins that are part of the apoptosis and cell cycle pathways and further analysis shows which proteins are common between the two pathways. You can then go back to your microarray data and ask whether these particular proteins are differentially expressed in thyroid cancer," says David Pot, product manager for Vector Xpression.

PathBlazer can also be used to inform a scientist on how to design a microarray experiment. "If you know that an oncogene is involved in a cancer and that another gene is overexpressed, PathBlazer can find the shorter pathway between the two genes and then query a microarray experiment to see if the pathway is involved," says Pot.

The currently available bioinformatics products have allowed scientists to carry out functional genomics studies that may not have been otherwise feasible. As genomic information continues to pour out of every laboratory, computational tools to analyze, annotate and visualize these data will also continue evolving. With the right software, going from genes to pathways will soon be only a few clicks away.

1. Lamb, J. *et al. Cell* **114**, 323–334 (2003).
2. Zhu, H. *et al. Science* **293**, 2101–2105 (2001).

Laura Bonetta is a freelance writer based in the Washington, DC area (lbonetta@nasw.org)

SUPPLIERS GUIDE: COMPANIES OFFERING BIOINFORMATICS SOFTWARE FOR GENE EXPRESSION AND PATHWAY ANALYSIS

Company	Web Address
Acumenta	http://www.acumenta.com/
Affymetrix	http://www.affymetrix.com/
Agilent Technologies	http://we.home.agilent.com/
Applied Biosystems	http://www.appliedbiosystems.com
Applied Maths	http://www.applied-maths.com
Applied Precision	http://www.api.com
Ariadne Genomics Inc.	http://www.ariadnegenomics.com
Array Genetics	http://www.arraygenetics.com/
Axon Instruments	http://www.axon.com/
Bioalma	http://www.almabioinfo.com
BioBase	http://www.biobase.de
BioConductor*	http://www.bioconductor.org
BioDiscovery	http://www.biodiswwcovery.com
Bioinformatics Solutions Inc	http://www.bioinformaticssolutions.com/
Bio-Rad Laboratories	http://www.bio-rad.com/
BioSoftSolutions	http://www.biosoftsolutions.de/
BIOSoftware Systems Inc.	http://www.biosoftwareinc.com
Clondiag Chip Technologies	http://www.clondiag.com
Compugen	http://www.cgen.com
Decodon	http://www.decodon.com
GeneData	http://www.genedata.com
GeneGo	http://www.genego.com
Geneva Bioinformatics (GeneBio) S.A.	http://www.genebio.com
Genotypic Technology	http://www.genotypictech.com/
Imaging Research	http://www.imagingresearch.com
Improved Outcomes Software	http://www.improvedoutcomes.com/
InforSense	http://www.inforsense.com
Ingenuity Systems	http://www.ingenuity.com/
Insightful	http://www.lionbioscience.com/
Insilicos	http://www.insilicos.com/home.html
Invitrogen	http://www.invitrogen.com
Iobion	http://www.iobion.com
Jubilant Biosys	http://www.jubilantbiosys.com/
LION bioscience AG	http://www.lionbioscience.com
MediaCybernetics	http://www.mediacycom
Metalife AG	http://www.metalife.orbitel.bg/
MiraiBio	http://www.miraibio.com
MolMine	http://www.molmine.com
NonLinear Dynamics	http://www.nonlinear.com
Paracel	http://www.paracel.com
Partek	http://www.partek.com
Proteome Systems	http://www.proteomesystems.com
Rosetta Biosoftware	http://www.rosettabio.com/
Scanalytics	http://www.scanalytics.com
SilicoCyte	http://www.silicocyte.com
Silicon Genetics	http://www.silicongenetics.com
Stratagene	http://www.stratagene.com
TIGR*	http://www.tigr.org/software/tm4
Universal Imaging	http://www.image1.com
Visualize Inc.	http://www.visualizeinc.com/markets/genomica.html
VizXlabs	http://www.vizxlabs.com/

*Bioconductor and TIGR software is open source.