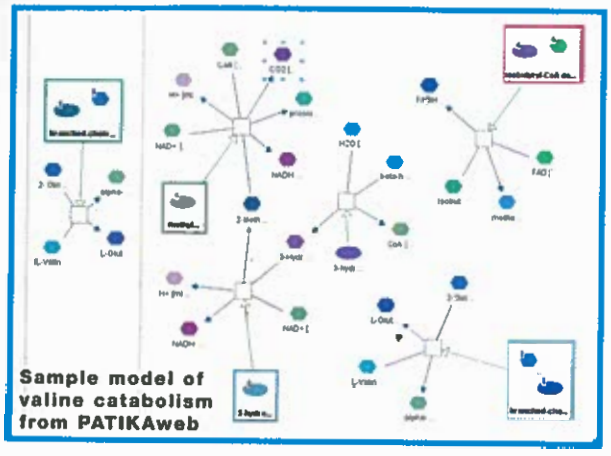
Systems Cartography

Commercial and public resources for pathway analysis abound. We talked to field leaders to learn what's what and how standards figure into navigating cellular landscapes.

By Jennifer Crebs

If you can find it, investigating the pathway less studied can make all the difference. The trick is in using a reliable map, preferably made with expertly curated data, and ideally one that isn't averse to incorporating new info from your own cellular wanderings. Problem is, systems-level experiments have generated reams of molecular details, most of which are encoded in different biological dialects. One scientist's map is another's Babel.



Sample model of valine catabolism from PATIKAweb

Such is the state of pathway informatics. We're still in the early days of high-throughput experimentation, and while the goal of understanding systems *in toto* is a noble one, there is still not a universal language — visual or otherwise — in which to trade information about the goings on of metabolites, proteins, genes, and other bio-entities.

That's not to say that standard formats for exchanging pathway data don't exist. Anyone familiar with the field's love affair with acronyms knows otherwise. SBML, CellML, and PSI-MI are just a few of the formats used to encode pathway information. The problem is one of translation. It's especially damning if you want to integrate knowledge extracted from different pathway databases or software programs. BioPAX, a recent effort to generate a data format for all pathway parts, aims to change all of this.

There is a wide range of tools available if you are interested in looking at pathways, regulatory networks, or molecular interactions. Open access databases and public informatics tools are only a Google search away — or, if you want to browse, Chris Sander's group at Memorial-Sloan Kettering has assembled a list of more than 200 resources called Pathguide. Meanwhile, commercial purveyors of proprietary databases and software products are aiming to be one-stop shops for anyone interested in visualizing and analyzing cellular phenomena.

This all leads to the question: what does what, and how will it help me? Although we can't provide a comprehensive survey of everything available in pathway informatics — that alone would take an army of expert curators — we can give you an idea of some of the tools out there, both on the commercial and public-sector sides. We'll also give you an idea of how standards figure into the mix.

NODE BY ANY OTHER NAME

Before delving into the bells and whistles of pathway systems, let's be clear on what a pathway actually is. Biological processes are, by definition, dynamic affairs. To understand them from our macro perspective, it's necessary to make inferences based on static snapshots of molecular data. Pathway diagrams representing networks of reactions are used to do this.

The grist for pathway models is generated by experimentation. Pathway databases are brimming with empirical data, which is extracted by curators (both human and machine) from the literature. The ideal visualization system would provide a model based on expertly selected results, and it would be capable of serving as a scaffold for further investigation. Anyone can read about a model; the point is to make use of it in a way that permits both the addition of new data and the generation of processes.

Researchers at Virginia Tech have looked into what scientists really want from pathway visualization systems and, in doing so, have charted the many meanings of 'pathway.' In a paper published last year in *Information Visualization*, Purvi Saraiya, Chris North, and Karen Duca say a pathway can be understood as any "user-defined network" of molecular interactions under study. Representations are diverse and can cover anything from specific events to

higher level abstractions. As the authors put it, "Some [pathways] are sketchy, while others are rigorous."

Saraiya and colleagues noted that components of a pathway are typically depicted as graphs containing nodes and edges. Depending on the complexity of a model, nodes can be used to represent a single molecule, enviromental stimulus, or to summarize an entire interconnecting pathway. Edges represent relationships or interactions, such as gene expression, inhibition, modification, and so on.

PATHS TO THE MARKET

There are a number of pathway analysis tools on the market, some more specialized than others. To get an idea of system flexibility, database creation, and clinical utility of software, *GT* spoke with a few vendors representative of the different approaches to building pathway tools.

In their study, Saraiya and the Virginia Tech crew interviewed life scientists to determine the most important requirements for pathway information systems. The team then evaluated a sample set of applications — GenMAPP, Cytoscape, GScope, Pathway Assist, PATIKA, and BioCarta — to judge how the tools fared in light of the requirements. They found, albeit in a study of limited scope, that Ariadne Genomics' PathwayAssist (now known as PathwayStudio) was preferred over the publicly available resources.

Ilya Mazo, president of Ariadne, says he was pleased with the findings, primarily because they indicated what is required from the user's perspective. "In today's incarnation of systems biology, people want to deal with knowledge rather than data," Mazo says, adding that feedback between pathway models and experimental data is key to truly understanding the biology and mechanisms behind disease.

Ariadne's PathwayStudio is a software package that helps identify relationships among proteins, small molecules, cellular process, and therapies. It provides a sort of scaffold upon which high-


Ilya Mazo, president of Ariadne

throughput data, including microarray and proteomics results, can be imported and filtered to classify molecules and draw diagrams. The system runs on Windows and works with other public and commercial databases, including BIND, KEGG, GO, PathArt, STKE Connection Maps, and Prolexys HyNet.

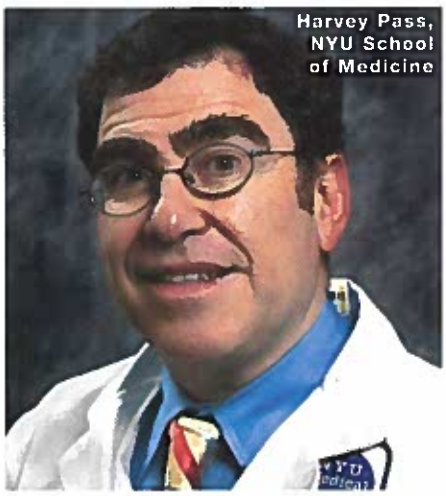
The system is fed by Ariadne's ResNet molecular networks database, which contains more than 1 million functional interaction events and is updated quarterly. This amount of available data ensures that "what you know is what humankind knows," Mazo says. The database itself is populated with data automatically extracted by Ariadne's MedScan mining technology from all of PubMed and more than 40 full-text journals.

Another pathway analysis offering from Ingenuity Systems was recently updated with the release of version 4.0 in June. Ingenuity CTO Ramon Felciano says Ingenuity Pathway Analysis, while not reviewed in the Virginia Tech paper, stacked up well with what the key requirements identified by the study's authors.

The current version of Ingenuity's software allows users to conduct and save advanced searches for genes, proteins, or molecular relationships, as well as to create reports and high-resolution images of pathway information. Jake Leschly, Ingenuity's CEO, says that the company's Web-based tool was designed with the biologist end-user in mind, and is applicable to questions from basic science to clinical work across all phases of the R&D process.

Like Ariadne's system, IPA 4.0 features the ability to analyze high-throughput microarray and proteomics data to get a handle on biological context and chart changes in biological states. The Ingenuity system can be accessed via Mac and PC platforms. Unlike PathwayStudio, however, IPA works with a manually curated database. "The biggest and clearest differentiator [of IPA] is the size and quality of our Knowledge Base," Felciano says. The Ingenuity Pathways Knowledge Base is hand built by curators around the world, he says, and features networks assembled from millions of individually modeled molecular relationships.

Harvey Pass, NYU School of Medicine's director of surgical research, used IPA in an exploratory analysis of genes associated with pleural mesothelioma, an asbestos-related cancer with a miserable survival rate. Pass used IPA to find a common pathway of genes that dif-


Harvey Pass, NYU School of Medicine

fered between progenitor and diseased mesothelial tissue. He then narrowed down that set of 35 genes to one: the gene for osteopontin. The software was a "pretty easy" to use, says Pass, who believes that Ingenuity's workflow will become a standard method for biomarker discovery.

Ingenuity is not alone in the clinical arena, as GeneGo is "heavily used in bio-

marker discovery research" as well, according to Julie Bryant, the company's VP of business development. She's not kidding. The company has licensing deals with both government health agencies (such as NCI, NHLBI, and FDA) and big pharma (like GlaxoSmithKline, Merck, and Proctor & Gamble), in addition to users in university labs.

GeneGo offers a suite of tools for looking at high-throughput data, and especially those data relating to human biology and medicinal chemistry. MetaCore, the company's flagship product, is used for drug target selection, validation, and biomarker discovery. Its content is manually provided via MetaBase, an Oracle-based database curated by 35 expert annotators who scan 200 journals to provide data on transcriptional regulation, metabolism, cell signaling and protein interactions, and disease conditions. Users can network interactions for more than 90 percent of known human proteins, the company says, claiming that this is the most comprehensive database of its kind on the market.

GeneGo's other tool, MetaDrug, focuses on predicting the metabolic activity and toxicity of novel small molecules and compounds. Mainly used by academic chemists and pharma researchers, MetaDrug is able to "virtualize metabolic pathways and provide the answer to 'What is the biological relevance of this compound?'" Bryant says, noting that the software is also used in phase III clinical trials.

OPEN ACCESS PATHS

In the public sector, pathway databases and visualization tools have proliferated at a dizzying pace. Examples of frequently used pathway and interaction databases include BioCyc, DIP, KEGG, Reactome, and SABIO-RK. On the pathway visualization and analysis front, there are CPath, Cytoscape, GenMAPP, and PATIKA, to name only a few. Open access databases and analysis systems operate symbiotically, as new analysis facilitates the creation of more data.


GeneGo's Julie Bryant

Reactome, an expert-curated database containing human biology pathways and reactions, is run through a collaboration among Cold Spring Harbor Laboratory, the European Bioinformatics Institute, and the Gene Ontology Consortium. Imre Vastrik, the project's EBI-based leader, says that Reactome's two major user groups are bioinformaticists and wet lab biologists.

Biologists with sub-field expertise contribute content to the site, which is maintained and cross-referenced by Reactome editorial staff. Human molecular events are not all Reactome offers; the free online database and open source software also feature orthologous events that are electronically inferred in 22 other species. The database's latest features include modules for specific signaling cascades, a new Web display for species annotations, and stable identifiers to track data objects over Reactome's releases.

PATIKA, an acronym for Pathway Analysis Tools for Integration and Knowledge and the Turkish word for "path," is a pathway analysis tool that relies on Reactome for at least some of its data. The Bilkent University-based project integrates information from several other

PATHWAYS AND THE STANDARDS THEY USE

Resource Name	Type(s)	URL
BioPAX		
BioCyc – BioCyc Knowledge Library	M	http://biocyc.org/
Cancer Cell Map	PP, S	http://cancer.cellmap.org/
EcoCyc – Encyclopedia of *E. coli* Genes & Metabolism	M	http://ecocyc.org/
MetaCyc – Metabolic Pathway Database	M	http://metacyc.org/
PathCase – CASE Pathways Database System	M	http://nashua.cwru.edu/PathwaysWeb/
Reactome KnowledgeBase	M, S	http://www.reactome.org/
CellML		
BioModels - BioModels Database	S	http://www.ebi.ac.uk/biomodels
CellML Model Repository	M, S	http://www.cellml.org/
PSI-MI		
BIND - Biomolecular Interaction Network Database	G, PP	http://www.bind.ca/
BioGRID - General Repository for Interaction Datasets	G, PP	http://www.thebiogrid.org/
DIP - Database of Interacting Proteins	PP	http://dip.doe-mbi.ucla.edu/
HPRD - Human Protein Reference Database	PD, PP	http://www.hprd.org/
IntAct	PP	http://www.ebi.ac.uk/intact
MINT - Molecular Interactions Database	PP	http://mint.bio.uniroma2.it/mint/
MIPS CYGD - MIPS Comprehensive Yeast Genome Database	G, M, PD, PP	http://mips.gsf.de/genre/proj/yeast/
OPHID - The Online Predicted Human Interaction Database	PP	http://ophid.utoronto.ca/ophid/
SBML		
BioCyc - BioCyc Knowledge Library	M	http://biocyc.org/
BioModels - BioModels Database	S	http://www.ebi.ac.uk/biomodels
EcoCyc - Encyclopedia of *E. coli* Genes and Metabolism	M	http://ecocyc.org/
JWS Online - Online Cellular Systems Modelling	M, S	http://jjj.biochem.sun.ac.za/
MetaCyc - Metabolic Pathway Database	M	http://metacyc.org/
Reactome KnowledgeBase	M, S	http://www.reactome.org/
SBML Model Repository - SBML Model Repository	M, S	http://sbml.org/models/
STKE - Signal Transduction Knowledge Environment	S	http://stke.sciencemag.org/cm/

Adapted from information available on "Pathguide: The Pathway Resource List" (http://www.pathguide.org/)
Legend: M: Metabolic pathways PP: Protein-protein interactions S: Signaling pathways
G: Genetic interaction networks PD: Pathway diagrams

databases to provide an environment suited to further modeling and analyzing cellular processes across data sources.

"The main component of PATIKA is a querying tool," says Uğur Doğrusöz, director of Bilkent's Center for Bioinformatics. Users can do simple field queries or more sophisticated graph theoretic queries to find the shortest path between sets of genes. At press time, the database contained nearly 42,000 biological entities and more than 7,700 interactions.

This month, Doğrusöz and his team hope to launch PATIKAweb 2.0, a user-friendly Web interface for querying and visualizing information in the PATIKA database. Registration isn't required for access, and neither is a local installation. The XML-based tool also has support for loading, mapping, and clustering microarray data on top of pathways.

Uğur Doğrusöz, of Bilkent's Center for Bioinformatics

"Basically, we present everything visually — graphs, the objects, and the links," Doğrusöz says. As he previously worked at Tom Saywer Software, the same company that provides the graph visualization backbone of PATIKAweb, Doğrusöz is confident that his team is "using visualization to its fullest potential, as far as the state of relational information visualization is concerned."

One of the main challenges facing those looking to integrate datasets, cited by both Doğrusöz and Vastrik, is having the raw data in a standard format. "If you leave it up to the community, in which everybody has different needs, it will be almost impossible to exchange data," Doğrusöz says. BioPAX, the Biological Pathway Exchange, is trying to meet this need through community-level work on the problem. Both the Reactome and PATIKA projects are involved, as are a number of individual researchers and representatives from other public pathway databases.

Vastrik points to a different problem: knowledge acquisition, or "being able transfer the knowledge from our experts' brains, in a form that is computationally important, into the databases. This is not a thing that scales very well. I mean, you just can't write a decent code to do it for you." And that's something that even the commercial developers will agree with. GT