Computational Knowledge Integration in Biopharmaceutical Research

Computational knowledge integration in biopharmaceutical research

David Ficenec, Mark Osborne, Joel Pradines, Dan Richards, Ramon Felciano, Raymond J. Cho, Richard O. Chen, Ted Liefeld, James Owen, Alan Ruttenberg, Christian Reich, Joseph Horvath and Tim Clark

Date received (in revised form): 23rd June 2003

David Ficenec
has led several projects in bioinformatics and knowledge engineering at Millennium. He was formerly with the Washington University Genome Center.

Mark Osborne
is a molecular and computational biologist who developed *J. Mill.* and the Knowledge Intermediary program at Millennium.

Joel Pradines
was a postdoctoral fellow at the Applied Biodynamics Lab, Boston University. His work applies mathematics and statistics to complex biological systems such as intracellular networks.

Dan Richards
leads computational biology at Ingenuity. He previously led Ingenuity's efforts in natural language processing.

Ramon Felciano
is Ingenuity's CTO and co-founder of the Stanford University Medical Media & Information Technologies Center. He founded Digital Alchemy.

Raymond J. Cho
has managed efforts in R&D and business development at Ingenuity. He previously led projects in collaborative technology innovation with Affymetrix and Stanford University.

Richard O. Chen
oversees product strategy and management at Ingenuity, Inc. He previously ran knowledge engineering and acquisition operations at Ingenuity.

Keywords: *ontology, knowledge base, pharmaceutical, drug discovery, microarray, biochemical pathway*

Tim Clark,
12 Dana Street,
Brookline, MA 02445, USA

Tel: +1 617 947 7098
E-mail: tim.clark@acm.org

Abstract

An initiative to increase biopharmaceutical research productivity by capturing, sharing and computationally integrating proprietary scientific discoveries with public knowledge is described. This initiative involves both organisational process change and multiple interoperating software systems. The software components rely on mutually supporting integration techniques. These include a richly structured ontology, statistical analysis of experimental data against stored conclusions, natural language processing of public literature, secure document repositories with lightweight metadata, web services integration, enterprise web portals and relational databases. This approach has already begun to increase scientific productivity in our enterprise by creating an organisational memory (OM) of internal research findings, accessible on the web. Through bringing together these components it has also been possible to construct a very large and expanding repository of biological pathway information linked to this repository of findings which is extremely useful in analysis of DNA microarray data. This repository, in turn, enables our research paradigm to be shifted towards more comprehensive systems-based understandings of drug action.

INTRODUCTION

Private companies undertaking pharmaceutical research normally take stringent measures to protect their research results as private intellectual property (IP), until they achieve a marketable compound. An unintended consequence of treating research as IP is that many of the research teams within a single company cannot effectively share their research results with each other, because they do not contribute to the normal public academic discourse of peer-reviewed journals and its accompanying widely-available repositories of scientific information. This public system, to which they do not routinely contribute, is the only system for widespread knowledge dissemination available internally.

Pharmaceutical company researchers are predominately consumers, rather than suppliers, of content for the published literature and its knowledge bases. Consequently, while their peers in academia and among commercial competitors are blocked from access to important proprietary discoveries, so are their (non-competing) colleagues elsewhere in the same company. Likewise the company may lack an effective historical memory of its research, outside laboratory notebooks and personal computer files, if it does not attempt to fill the gap left by opting out of public discourse. We call this gap the 'IP shadow', and we believe it creates scaling limits to productivity in private pharmaceutical research — possibly even creating negative economies of scale experienced as an 'innovation deficit'.[1,2]

The authors believe most pharmaceutical company researchers can, from their own experience, readily supply examples of wasted effort, lost opportunities and diminished productivity at their companies, owing to inability to

Ted Liefeld
is a lead software architect focusing on scalable interoperability. He developed Millenium software integration architecture, and was formerly with The Math Works.

James Owen
is a molecular and computational biologist currently working in the Knowledge Management practice at Millennium.

Alan Ruttenberg
is the PARIS software project manager. Previous work includes parallel computing, computer vision, knowledge representation and user interface.

Christian Reich
trained in oncology and molecular biology and has worked extensively in computational biology at Millennium and the EBI.

Joe Horvath
heads Millennium's Knowledge Management practice. He was formerly an executive consultant with IBM Business Consulting.

Tim Clark
led development of Millennium's bioinformatics, cheminformatics and knowledge engineering programmes from their inception. He was previously at the NCBI, where he helped develop GenBank.

Organisational memory

Framework

communicate, integrate and share research findings across intra-company organisational and geographical boundaries. A systematic approach to resolving these issues may become increasingly vital in the context of 'extended enterprises', which now appear to be the mode in the pharmaceutical industry.[3]

This paper describes a comprehensive initiative, *MyBiology*, designed to resolve the problem of 'IP shadow' using computational knowledge management, lightweight process re-engineering and web deployment. At the same time, this initiative attempts to:

- create a more sophisticated infrastructure than is currently available to the public researcher;
- build in process design and process adoption support from the beginning, realising that adoption and use are critical success factors; and
- develop advanced capabilities for computational systems biology as part of the infrastructure.

MyBiology deals, in part, with a version of the 'organisational memory' (OM) problem in knowledge management. Previous OM efforts have been primarily concerned with 'business process', that is, with better recollection and dissemination of *how* the enterprise does its work. Ontology construction has been seen as central to achieving this goal, which in turn is presumed to influence enterprise productivity and excellence of product.[4–6]

Scientific knowledge in a biopharmaceutical enterprise, however, *is itself a product*. Therefore, managing this knowledge is actually part of the production process, and failure to manage it properly contributes to wrong decisions and attrition in the pharmaceutical development pipeline.

MyBiology is a collaboration between Millennium Pharmaceuticals and Ingenuity Systems and is implemented at Millennium across multiple research sites on two continents. Initially focused mainly on results relatively far 'upstream' in the research and development process, it is rapidly being extended to have productivity impact all the way to the clinic. The system is described below and results and prospects after one year of the programme are assessed.

THE PUBLIC KNOWLEDGE MANAGEMENT INFRASTRUCTURE

Public academic researchers share information and collaborate on a large scale through the well-organised system of peer-reviewed academic journals. Over 4,600 of these journals are continuously indexed and organised into MEDLINE®[7] by the US National Library of Medicine using a formal ontology, the Medical Subject Headings (MeSH®).[8] This constitutes a knowledge base (KB). Many other databases have been linked to this fundamental resource, including whole genomes for various organisms of clinical interest. These resources are readily searchable on the web via PubMed®,[9] providing a critical means of information sharing for the scientific community. Furthermore, there are numerous computational methods available to analyse research data, such as DNA sequences, against the organised findings of other researchers.[10]

Taken as a whole, these facilities constitute an open, public framework for scientific knowledge management. The functionality of this framework must at least be matched, if not bettered, within private research organisations, to achieve parity with public efforts in sharing scientific knowledge.

The public framework is composed of:

- a peer-reviewed publication system;
- a knowledge base (MEDLINE);
- an ontology by which the KB is organised (MeSH);

- methods for searching the KB (PubMed);
- computational methods for analysing new research data against the KB;

Infrastructure

- portals that collect search methods and computational tools in a single site (NCBI Entrez[11]);
- a transparent electronic framework providing information accessibility, identity and interoperability (internet, W3C web protocols, registry systems).[12,13]

As private companies choose to replicate these elements, they must decide whether to adopt the public technology or to improve upon it. They also encounter the challenge of creating 'semi-permeable' links back to the public system.

Semi-permeable links

THE *MYBIOLOGY* INITIATIVE

MyBiology was designed to eliminate the 'IP shadow' at Millennium and within Millennium collaborations by:

- providing one or more curated internal electronic journals of key research results;
- extracting computable findings from the internal journal into a KB founded on a 'deep' ontology;

Ontology

- integrating internal findings with external public research using the same ontology;
- analysing transcriptional profiling data against biochemical pathways described in the KB using novel algorithms;
- providing both text-based KB query and graphical interface to pathway analysis.

MyBiology integrates Millennium's proprietary information with information from public literature to help scientists build an evolving picture of complex biological systems. As shown in Figure 1, *MyBiology* consists of several related components. Millennium scientists are an integral part of *MyBiology*, both as contributors and customers.

The major components of *MyBiology* today are:

- an ontology;
- Millennium's research scientists;
- knowledge intermediaries (KiMs);
- *Journal of Millennium Science* (*J. Mill.*);
- external and internal databases;
- MyTargetValidation (MyTV);
- the Life Science Knowledge Base (LSKB) and query interface;
- the public scientific literature;
- Pathway Analysis Research Information System (PARIS).

An *ontology* is critical to organisation of a KB, just as unique object identifiers and foreign key relationships are fundamental for various kinds of databases. Recent work on scientific ontologies has attempted to provide specialised annotation and organisation for genomics and genome-related research databases (eg the Gene Ontology Consortium, GO[14]) and computational service classification and organisation[15] via fundamental expressive formalisms relevant for both purposes and suitable for deployment on the internet.[16]

For our basic ontology, we selected a system from Ingenuity Systems, Inc., together with annotated content, to which significant proprietary and in-licensed content were added. The Ingenuity ontology has many advantages, among which are the ability to distinguish

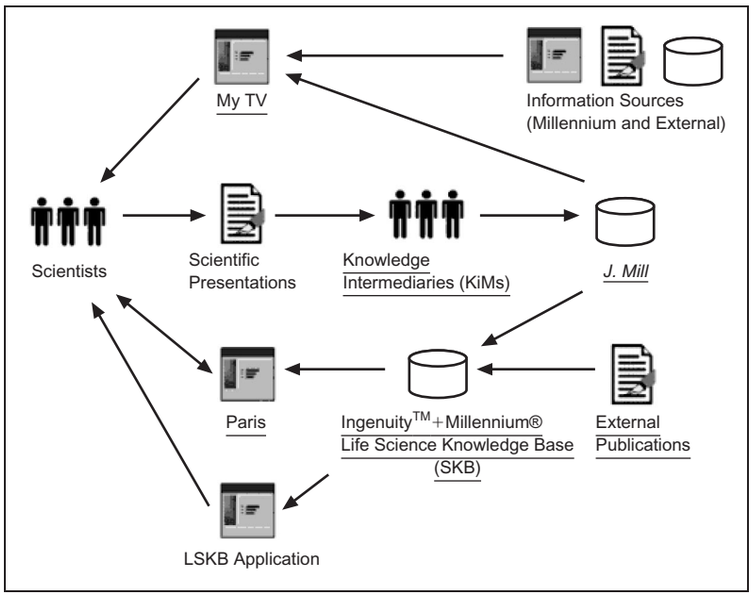

Figure 1: The *MyBiology* system at a high level component view

Repository

between objects and processes. The resulting richly-connected KB means it is possible, for example, to extract metabolic and cell-signalling pathways automatically and combine them with other information sources for computational analysis.

Experimental findings

Research scientists are both providers and consumers of information. They present their experimental methods, results and interpretations at regular research team meetings throughout the company. Finding ways to capture the information in these presentations and to make it readily available to scientists across the company is a critical component of the *MyBiology* effort.

Encoding

Knowledge intermediaries work with researchers to capture and encode the information found in research scientists' presentations. KiMs are PhD-trained biologists and chemists with additional training in computational science. The system operates effectively with a very small number of KiMs supporting more than 1,000 researchers.

Computation

The Journal of Millennium Science is an electronic journal constructed upon the Documentum ECM Platform[17] using customised interfaces. Scientific presentations, typically in PowerPoint, are stored with metadata in *J. Mill*. Metadata include an abstract, author and publication information and unique identifiers for targets, compounds or other reagents referenced in the presentation.

MyTargetValidation is a scientific portal to pertinent information on a given biopharmaceutical target (gene/protein). It integrates a large number of internal and external databases using a Web Services Architecture.[18] It also enables direct access to assay results in lead optimisation for molecules active against a target, through integration with an internal drug discovery database of compounds, assay results and basic computed compound properties.

The *Life Science Knowledge Base* is a highly-structured KB that combines selected findings from the public literature with internal Millennium research results in computable form. Ingenuity employs a large number of 'content scientists', to read articles from 33 highly cited scientific journals, curate the results and interpretations, and add them to the KB. They also use natural language processing (NLP) to add information from MEDLINE abstracts dating back to 1980. The KB contains more than 1 million findings, approximately 800,000 derived by content scientists.

Lifesciences is a web-based application from Ingenuity that provides textual query access to the LSKB including a comprehensive overview of all findings related to a given gene, its products and biological relationships.

The *Pathway Analysis Research Information System*[19] analyses the results of a transcriptional profiling experiment against the biological pathway information in LSKB. It enables deeper interpretation of cellular processes affected in an experiment, by interpolating non-transcriptionally regulated components of relevant pathways.

Several stages of scientific research in support of pharmaceutical discovery and development are supported by these components. For example, target identification and validation are enabled

PARIS

by MyTV and *J. Mill.* as individual researchers evaluate results from genome-scale experiments and seek to better understand the role that individual genes and their products play in the system under study. In characterising individual compounds, PARIS is often used better to understand the mechanism of action of the chemical, as it provides important context to understanding the large numbers of genes that are typically affected in a single experiment. The PARIS pathway browser is further used to explore the interconnections between proteins identified in an experiment.

Integration

Furthermore, there has been great value in extending *J. Mill.* into areas outside biology; several different groups of researchers have approached the KiMs about including reports and other scientific results in *J. Mill.* so that they are more broadly available and are integrated with the rest of the Millennium knowledge integration platform. PARIS is having a dramatic impact on the analysis of clinical trial data in the form of evaluating patient responses to candidate pharmaceuticals. Through its pathway mapping and interpretation, it provides analysts with context sufficient to direct further research experiments in support of the therapeutic goal.

The most fundamental goals of this initiative are to enable scientists to employ all available scientific information — public and private; to select targets and compounds most likely to succeed in clinical development; and to conduct more effective clinical trials. Ultimately, we wish to shift our research paradigm away from a dyadic compound → target view and towards a more holistic compound → pathway → disease understanding. Increases in productivity here should result in faster availability of improved therapies to patients.

MYTARGETVALIDATION

Target validation

Researchers seeking to understand the biology of a system under study seek information from many sources, including public literature and external websites with information on gene and protein families (eg NCBI[20] and GO). Also, researchers at Millennium have typically interacted with core technology groups via the web to request experiments and/or access the results. This led to a proliferation of individual unconnected websites and to redundant requests for identical experiments from different researchers. Researchers expressed enthusiasm for a single point of entry for access to multiple sources of information from both external sources and internal resources.

MyTV is a scientific portal (see Figure 2) to all available information on a given target, whether external or internal. It begins to answer the question: 'What do we know about this target?'

Access to information in MyTV is valuable in many different contexts of pharmaceutical development including target discovery, target validation, assay development, lead optimisation and analysis of clinical data. At a glance, a scientist can quickly assess what research has been performed at Millennium, by whom and in what stage of the pipeline; what types of experimental data are available; and what information is available from the public domain regarding biological targets and their annotation.

Public information is acquired through mappings to identifiers provided by Millennium's Gene Catalog system — a non-redundant catalogue of all genes constructed from both public and proprietary DNA sequences and annotations. Public sources include: SWISS-PROT,[21] LocusLink,[22,23] OMIM,[24] GO, InterPro[25] and ENZYME.[26] MyTV also links targets to Ingenuity's LifeSciences GeneView, providing access to key findings from the public literature and internal findings captured in *J. Mill.* and curated into Ingenuity's KB.

Millennium's other internal sources include:

- GeneCatalog for information on

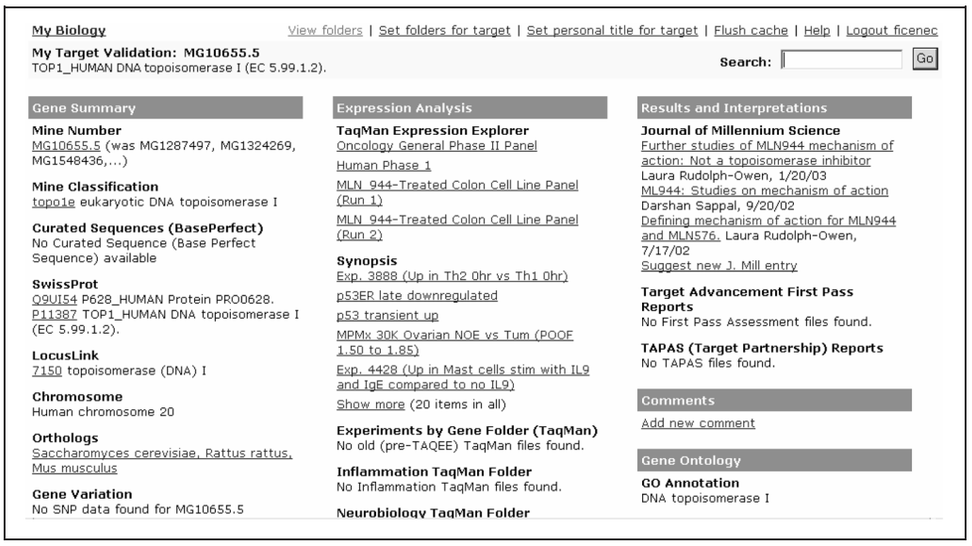

Figure 2: Example MyTV page for DNA topoisomerase I

sequences, protein families, chromosome location and orthologues.

Content aggregation

Caching

- Reagents and experimental systems that include information on expression vectors, expressed proteins, micro-arrays and experiments done in molecular pathology including TaqEE, a warehouse of TaqMan expression data.

- Scientists' analysis and interpretations are surfaced through links to Synopsis, a system for capturing interesting sets of genes and their biological context; *J. Mill.* for key scientific presentations and findings; summary reports from the Target Advancement and Assay Development groups.

Productivity

- MyDrugDiscovery, Millennium's portal for information on high-throughput screening of targets and other assay information about compounds and lead series.

Web services

MyTV uses a pull model to query multiple systems for summary information about each target. Each system to be queried implements a simple web service returning an XML[27] fragment adhering to the MyTV XML schema.[28] The system then constructs an XML file for each target. This XML content is transformed into HTML[29] by the presentation layer.

The system uses on-demand instantiation: target content is not aggregated until a scientist requests it. All requested information is then cached on the file system. Cache is periodically updated to refresh the information.

The system was initially released in a basic configuration in May 2002. Frequent releases were made, integrating new systems as requested, adding significant functionality over time, and continuously improving performance. Within four months, MyTV developed a user base of over 100 scientists. On a typical day, 20–25 individuals use the system. After about five months of use, users were surveyed to assess their perception of MyTV's impact on their work. Scientists strongly agreed that the system improved their productivity (they spend less time finding information) and quality of decision making by providing more complete and relevant information.

JOURNAL OF MILLENNIUM SCIENCE

J. Mill. is a collection of software, processes and semi-structured data. It serves as a central repository of experimental results and interpretations. *J. Mill.* is designed to maximise the connectedness of research teams across

Documentum©

Scientific communication

Repository

Metadata

GeneView

therapeutic disciplines, create an organisational memory of key scientific findings with links to supporting data and serve as a staging area for findings export to the highly structured KB. It was considered helpful to use a framework similar to one that researchers are familiar with, that of publication in a scientific journal.

J. Mill. relies on the fact that most scientific communication occurs in a group setting and is supported by electronic documents such as PowerPoint[30] presentations containing scientific results and interpretations. Since these documents form the core of the communication between scientists on a project team, if they were archived with some associated metadata, a record would be available of the important activities undertaken by the project team.

By providing a central location for scientific results, it was hoped to encourage information sharing and reuse among Millennium scientists and to prevent critical information from being lost or overlooked. *J. Mill.* is accessed via a web-based application which allows scientists to browse the repository and KiMs to publish entries in the repository. Published *J. Mill.* entries are also available through MyTV.

An entry consists of documents describing an internal Millennium presentation and curated metadata about that presentation. The documents are typically Microsoft Word[31] or PowerPoint files that have been presented within Millennium along with metadata about the presentation curated by a Millennium scientist. The metadata include information such as the author, gene and compound names and identifiers and some key scientific findings distilled from the presentations.

J. Mill. application

The *J. Mill.* application is a three-tier web application. Its user interface for browsing is deliberately reminiscent of PubMed, with which scientists are already familiar.

J. Mill. is written in Java and uses the Documentum Foundation Classes (DFCs)[32] to access Documentum's eContent Server, which stores *J. Mill.* documents and presentation metadata and manages the document life cycle. In addition, several open source frameworks and components are used, including Jakarta's Struts[33] which enforces a model–view–controller (MVC)[34] architecture within the application. The presentation layer is written using JavaServer Pages (JSPs)[35] and the Struts tag libraries to format the presentation for the user.

Authentication of users is accomplished via a web-based single-sign-on (SSO) web agent.

J. Mill. makes itself available to other applications through the publication of a simple object access protocol[36] (SOAP) interface that allows other applications to execute queries against *J. Mill.* content. The SOAP interface requires the same user authentication as the web application and uses the same SSO interface to control access.

J. Mill. was initially put into production in June 2002. Since that time, approximately 100 articles by Millennium authors have been published internally. In the future, *J. Mill.* will provide self-publication facilities to authors in order to expand its ability to collect and publish content more rapidly to our scientists.

LIFESCIENCES FOR MILLENNIUM

The Ingenuity LifeSciences Suite (Figure 3) is a user interface to the Ingenuity Pathways KB. It includes GeneView, which catalogues the functions of a given gene and gives scientists access to functional assertions between a specific gene and other genes, small molecules and cellular processes (Figure 3), as well as the specific biological experiments supporting each assertion. Most features of GeneView run dynamically off the KB; for example, statements regarding biological experiments are generated using natural language algorithms applied to structured information stored in the knowledge server. Each new piece of

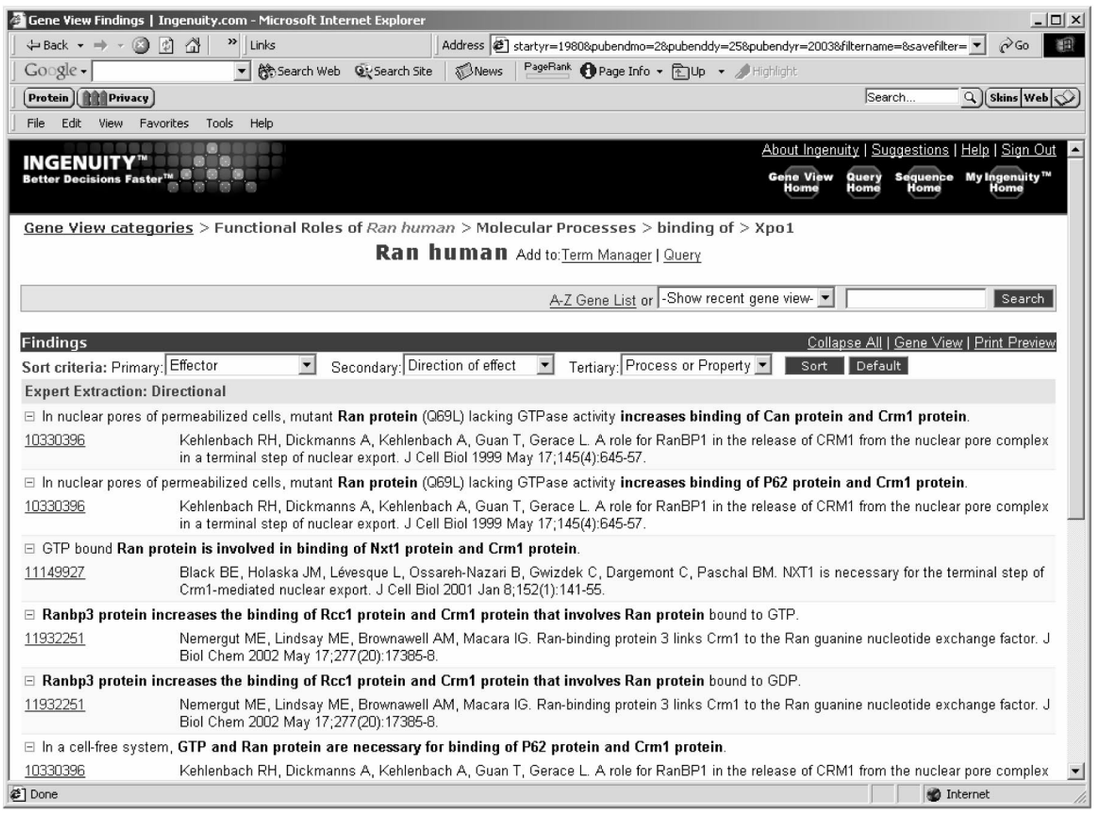

Figure 3: A GeneView page displaying relationship between *Xpo1* and *Ran* provides the user with a natural language display of the underlying relationships and properties of that object, shown here. Note that findings were also returned containing the '*Crm1*' synonym of *Xpo1*. This synonym capability resides in the Ingenuity ontology

KB Query

Findings

information is hyper-linked to other relevant findings in the KB.

Alternatively, scientists using the Query application may pose a specific question regarding relationships between biological entities; for example, 'What apoptotic effects does cytochrome c exert in neurons?' The fields in Query, which reflect the perturbations and results recorded in most biological experiments, enable scientists to construct these questions efficiently and intuitively. Query intelligently leverages the underlying ontology to empower the biologist: for example, a query about neurons returns findings regarding superior cervical ganglion cells, as the ontology recognises the ganglion cells as a type of neuron (Figure 4).

Two methods are used to capture findings. In findings capture through expert knowledge acquisition, quality-controlled operational processes enable Ingenuity-trained, PhD-level scientists to structure information from full text articles with a high degree of semantic richness. This method was applied to tens of thousands of articles from the most highly cited biological journals. The second method, semi-automated knowledge acquisition, employs algorithmic capture of knowledge from the abstracts of hundreds of journals spanning more than a decade of published research. In both approaches, every finding structured into the Ingenuity KB was quality-checked by a PhD-level scientist.

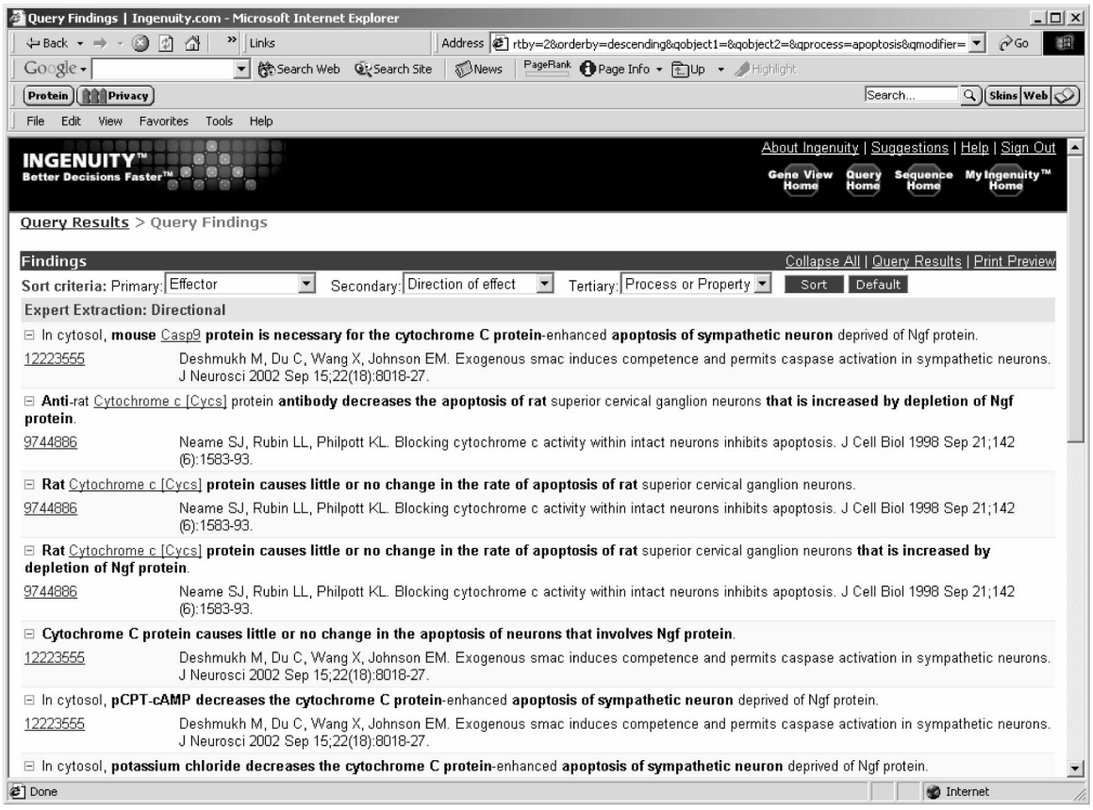

Figure 4: Query result after search for apoptotic effects of cytochrome C on neurons. Note that superior cervical ganglion neurons were recognised as a type of neuron

Frame-system

SEMANTIC INTEGRATION: THE *MYBIOLOGY* KB

To build a platform for biological computation — as well as for intelligent search and query — it was thought necessary to integrate the large corpus of literature-based findings in the Ingenuity KB with key experimental results from Millennium's research programmes into a single, computable structure. This posed multiple challenges; for example, semantic inconsistency in the literature makes it difficult to recognise when genes are identical or distinct. A need for a high degree of accuracy impedes the rate at which experimental findings can be formally represented. Finally, more sophisticated computations will probably require detailed context regarding functional relationships — for example, in what cell type one protein phosphorylates another.

Frame-based knowledge representation systems offer a versatile and powerful approach to structuring knowledge while addressing these critical needs. Generally, frame-based systems allow definition of abstract concepts and their relationships with other concepts.[37] These systems define three types of formal objects to represent knowledge in a given domain: concepts, properties and instances. *Concepts* (classes) are descriptions of particular categories of objects. *Properties* are attributes that describe the concept itself or relate one concept to another. An *instance* is a real-world example of a concept. Once information is represented in this manner, most frame-based knowledge representation systems support

basic inference capabilities such as classification ('Is this protein a type of kinase?') and declaration of axioms ('If loss-of-function mutations in a gene lead to a DNA repair phenotype, that gene participates in DNA repair'). Axioms in particular impose semantic constraints on the KB that help maintain the consistency and integrity of the data. Finally, frame-based knowledge representation systems provide basic query capabilities for retrieving stored data.

Biological ontology

Using a frame-based knowledge representation, one can build an ontology (Figure 5): a taxonomy and formal description of the concepts and relationships germane to a particular domain. Ontologies can be used in many different ways. By defining a class structure and formalising relationships between those classes, ontologies (i) allow more efficient browsing of concepts through the taxonomy, (ii) provide a standard framework for exchange of data, and (iii) provide the basis for representing domain knowledge. Building the ontology for a frame-based system is analogous to building the schema for a relational database. A KB consists of an ontology populated with instances of real data. In academia, frame-based representations of biological data have proven useful in several biological domains across a range of organisms.[38–41]

Ingenuity has developed a frame-based system with the goal of representing hundreds of thousands of functional relationships between genes for the purpose of enabling computations that provide critical path insights to scientists involved in therapeutic discovery. Toward this end, Ingenuity has constructed a proprietary ontology of more than 300,000 distinct classes including, but not restricted to, genes proteins, small molecules, cellular

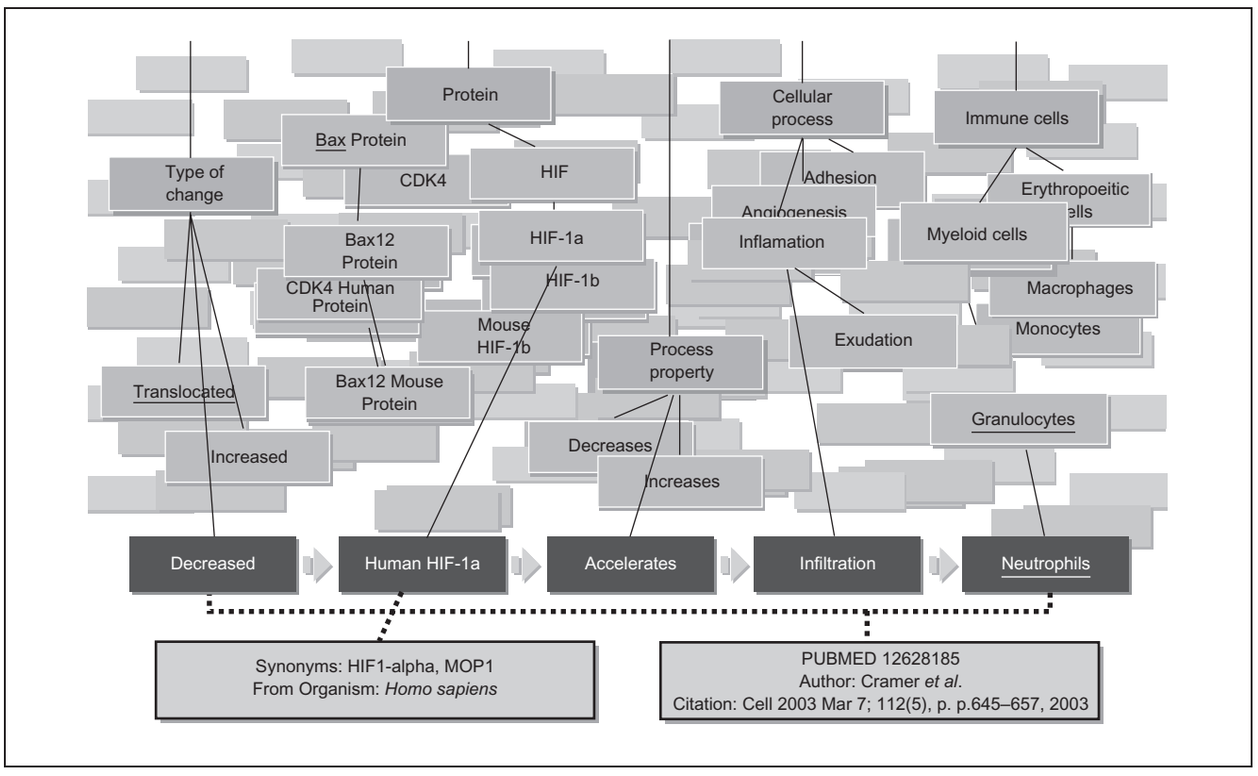

Figure 5: A conceptual model showing how a biological finding is formally represented in the Ingenuity Pathways KB using objects structured to capture meaning and context from the scientific experiment. Actual findings contain more semantic detail than displayed here, and the actual ontology is composed of more than 300,000 classes

components, cells, tissues and biological processes. Ingenuity has utilised this ontology to formally represent, in machine computable form, more than a million biological experimental findings from the public domain literature, specifically focused on human, mouse and rat genes. The Ingenuity KB can be used to structure not only knowledge from the public domain, but also a pharmaceutical organisation's proprietary scientific findings, creating a leveraged, integrated knowledge asset. This project has required scalability for both the ontology and the back-end not previously demonstrated in the smaller, academic efforts at frame-based systems described above. Ingenuity currently has operational processes capable of incorporating, per month, more than 5,000–10,000 concepts into its proprietary ontology, and more than 40,000 biological findings into the KB.

Experimental findings

Scalability

Applications have been developed to leverage the Ingenuity KB in support of a diverse array of computations, enabling the discovery process for end-user biologists — including both systems biology computations producing sophisticated pathway predictions from genomic data — and access to the underlying knowledge (LifeSciences applications, above). Also, the KB enables internal bioinformatics organisations to integrate proprietary information and use the combined resource to innovate new solutions. Both avenues were taken in supporting the *MyBiology* initiative.

Findings capture

Pathway predictions

PROCESS INTEGRATION: THE SCIENTIFIC FINDINGS CAPTURE PROCESS

If the goal of the *MyBiology* KB was semantic integration — exploiting the Ingenuity system's 'knowledge' of biological concepts and relationships to enable intelligent query and computation — then an equally important goal was process integration. By process integration is meant the ability to integrate or embed the use of the KB into Millennium's research processes. Specifically, a means of capturing and curating Millennium's internal research results for entry into the KB was needed. The challenge, of course, lay in the various forms that such results may take in a modern pharmaceutical company (notebook entries, annotated gene lists, PowerPoint presentations, etc.).

Process integration

In approaching this problem, the premise was that demands may only be made on scientists (eg to contribute scientific findings to an electronic repository) in proportion to the value they perceive from using the system. The 'bootstrapping' problem was how to get scientists to contribute as 'producers' of knowledge before there existed a critical mass of content within the repository for them to draw upon as 'consumers.' As an initial strategy, a small number of dedicated 'curation' roles were created in order to initially minimise the burden of participation upon scientists. These dedicated roles are currently being moved away from.

SCIENTIFIC FINDINGS CAPTURE

The scientific findings capture (SFC) process was developed as a systematic method for capturing presentations and their associated findings for publication Millennium-wide in both *J. Mill.* and the Ingenuity LSKB. The process is composed of several discrete steps and metrics gathering has been incorporated throughout. Initially, this work was started with the dedicated support of KiMs. KiMs served to assist individual research scientists with the process of knowledge sharing.[42] SFC KiMs were PhD scientists with knowledge of the internal and external computational tools used by Millennium scientists. SFC has several steps, which are outlined below.

Encoding the presentation

Scientists serving as KiMs are invited to laboratory and departmental meetings where scientists present the results of their work to their colleagues. These presentations are typically prepared in

PowerPoint and are excellent sources of key experimental results and interpretations. Using the *J. Mill.* web application (see above), an individual laboratory scientist's presentation is separated into components resembling a journal article. A title and abstract are prepared using background material, and hyperlinks to internal and external resources are also included within the application.

Integration

Publication

Electronic publication

After initial creation in *J. Mill.*, the completed entry is returned via e-mail to the presenting scientist for final approval. In this manner, the role of the presenting scientist is similar to that of the author of a scientific manuscript in that he/she is ultimately responsible for its contents. Upon approval, the presentation is automatically published Millennium-wide through MyTV.

Architecture

TRANSFERRING EXPERIMENTAL RESULTS FROM MILLENNIUM TO INGENUITY

Knowledge transfer

To generate a combined KB of internal Millennium content and external Ingenuity content, a transfer system was devised to facilitate the transfer of Millennium experimental results (MERs) to a private copy of the Ingenuity KB.

All scientific presentations contain potential MERs to be represented in the Ingenuity KB. Those findings determined to be useful in the KB are queued for entry. All proposed novel terms and genes are examined and validated by both teams to verify accuracy. Also, clarification of findings from the KiMs is a necessity to ensure that the translation from the written sentence in *J. Mill.* to the Ingenuity KB is performed consistently. Each MER is modelled into the Ingenuity ontology by the Ingenuity content scientists and displayed in LifeSciences within the context of the findings derived from the external literature. Further, a hyperlink to the original presentation in *J. Mill.* is also available for easy reference as well as increased searchability through Ingenuity's LifeSciences Query application.

TECHNICAL INTEGRATION: THE *MYBIOLOGY* ARCHITECTURE

A third but equally important aspect of integration in the *MyBiology* project was the technical integration of MyTV, *J. Mill.* and the *MyBiology* KB into Millennium's informatics platform within a coherent architecture, despite the rather complex technical environment.

The hardware and network for deployment were primarily Millennium's internal network but there was also a need to include some form of connectivity to the Ingenuity LSKB and Ingenuity's LifeSciences application. End-users used both Microsoft Windows and Apple Macintosh computers.

The *MyBiology* software environment included four applications (MyTV, PARIS, LifeSciences and *J. Mill.*). These applications had to be able to link to several other existing web-based applications. Each of the applications, existing and new, was targeted at different user communities and had been customised for their use.

Within this context, we also had the following requirements for the *MyBiology* architecture:

- User experience. Users must have a common user experience when using any of the *MyBiology* suite of applications, and it must be easy to navigate between applications.
- User access. The applications need to be accessible from both Microsoft and Macintosh computing platforms.
- Decoupled development. Asynchronous releases across the four development teams improved productivity.

MODEL AND DISCUSSION

Loose coupling

To make transitions between applications seamless, the *MyBiology* software architecture (Figure 6) focused on the interconnections and transitions between the applications. To maintain loose coupling between the applications and databases, connections to application databases were limited solely to the applications that owned them. Connections between the applications are permitted at two levels. Low-level connections were made between applications using web services. User interface level connections were made between applications via HTML links. At the lower level, application-to-application connections were made using web services via the SOAP interface. A single standardised SOAP interface was exposed by all applications. These SOAP connections were used to pass data between the applications and to provide URL links that the client application then used to provide links between the applications at the user interface level. These SOAP connections were also used to connect to other applications outside the *MyBiology* initiative.

SOAP

The exception to this form of connectivity is the LifeSciences application. In this case, since the application was hosted outside Millennium, the applications are connected via periodic file transfers, over a secure network, of the data to be shared with Ingenuity's KB and the links back to the Millennium applications.

The architectural approach used in *MyBiology* has been successful in its primary goals of preserving loose coupling between the applications for ease of development while simultaneously presenting a unified seamless view of the applications to the end users.

THE PATHWAY ANALYSIS AND RESOURCE INFORMATION SYSTEM

Intracellular processes

The Ingenuity KB integrates large amounts of biological knowledge. Because this information is encoded in a structured format that enables computation, the Ingenuity KB can be used to solve one of the more difficult challenges in drug discovery: the identification of the intracellular processes perturbed or activated in a biological experiment.

The PARIS database of protein functional relationship integrates data

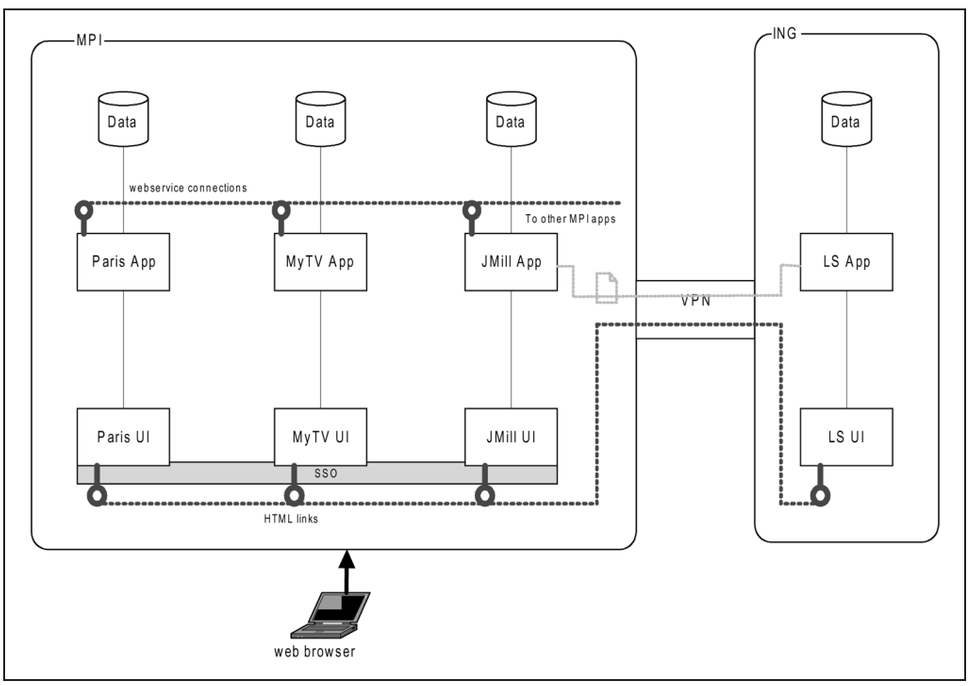

Figure 6: The *MyBiology* architectural model

Graph

Pathway database

Microarray

Neighbourhood

Regulation

from both public sources (CSNDB,[43] LIGAND[44]) and the Ingenuity KB, including Millennium scientific findings encoded in the KB. This integration has given Millennium a human pathway database much larger than any publicly-available resource.

The detection of perturbed pathways is required to understand the function of a novel gene, the mechanism of action of a compound or the different responses of patients to treatment. Transcript profiling (TP) using microarrays is a powerful technology to perform such investigations: in a single assay the transcriptional activity of the entire genome can be measured.[45]

Translating hundreds of observed mRNA changes in terms of perturbed pathways is work too complex to be optimally performed by hand, however. Help can be provided to the researcher by using computation.[46,47] Such computation combines the experimental data with large amounts of pathway knowledge and extracts the portion of the knowledge relevant to the observed transcriptional activity. Besides helping the scientist deal with large amounts of data, appropriate computation can also significantly improve the sensitivity of TP data interpretation. Indeed, it is now well established that functionally-related genes, eg pathway neighbours, have a significant tendency to be simultaneously regulated at the mRNA level.[48–50] Therefore, even if the individual gene expressions stand in the range of experimental noise, the coordinated differential expression of genes belonging to a same pathway can indicate its perturbation.

Computational interpretation of TP data requires large quantities of pathway knowledge. The coverage provided by current public sources[51,52] is relatively low. Moreover, to enable computation the data have to be consolidated into one format. The Ingenuity KB is an ideal source of pathway knowledge.

Consequently, Millennium uses Ingenuity to maintain a system called the Pathway Resource and Information System (PARIS). The knowledge contained in PARIS can be summarised in a graph where vertices are genes and edges represent their functional relationships (Figure 7). Those relationships are: membership in the same complex, protein binding, post-translational modifications, transcriptional regulation and neighbouring of enzymes in metabolic pathways. Currently the network contains about 5,000 genes and 22,000 interactions.

Millennium's computational biologists have developed algorithms that use this network to analyse TP experiments.[19] The algorithms identify the portions of the graph that are significantly regulated in a given experiment (Figure 8). For each gene the average regulation in its pathway neighbourhood is computed. This local density of transcriptional activity is then compared with those obtained when randomly reassigning the individual gene expressions.

Figure 8 shows an example of application of the PARIS system to identify pathways perturbed upon *p53* mutation in ovarian tumour samples.[53] The PARIS system has extracted a set of genes having significant density of differential expression in their functional neighbourhood when comparing transcript profiling of *p53* null samples against *p53* wild-type samples. Those genes define pathways related to cell cycle and proliferation, tumour growth factor beta signalling and cellular adhesion.

The PARIS system, based on pathway knowledge extracted from the Ingenuity KB, is used at Millennium to analyse TP experiments performed at several stages of the drug discovery process. The computationally extracted pathways can point to the cellular processes downstream of a target, which are perturbed by a compound, responsible for drug resistance or involved in the different responses of patients to treatment.

The PARIS web application contains hyperlinks to MyTV and other sources to enable scientists to rapidly learn about genes regulated in their experiments.

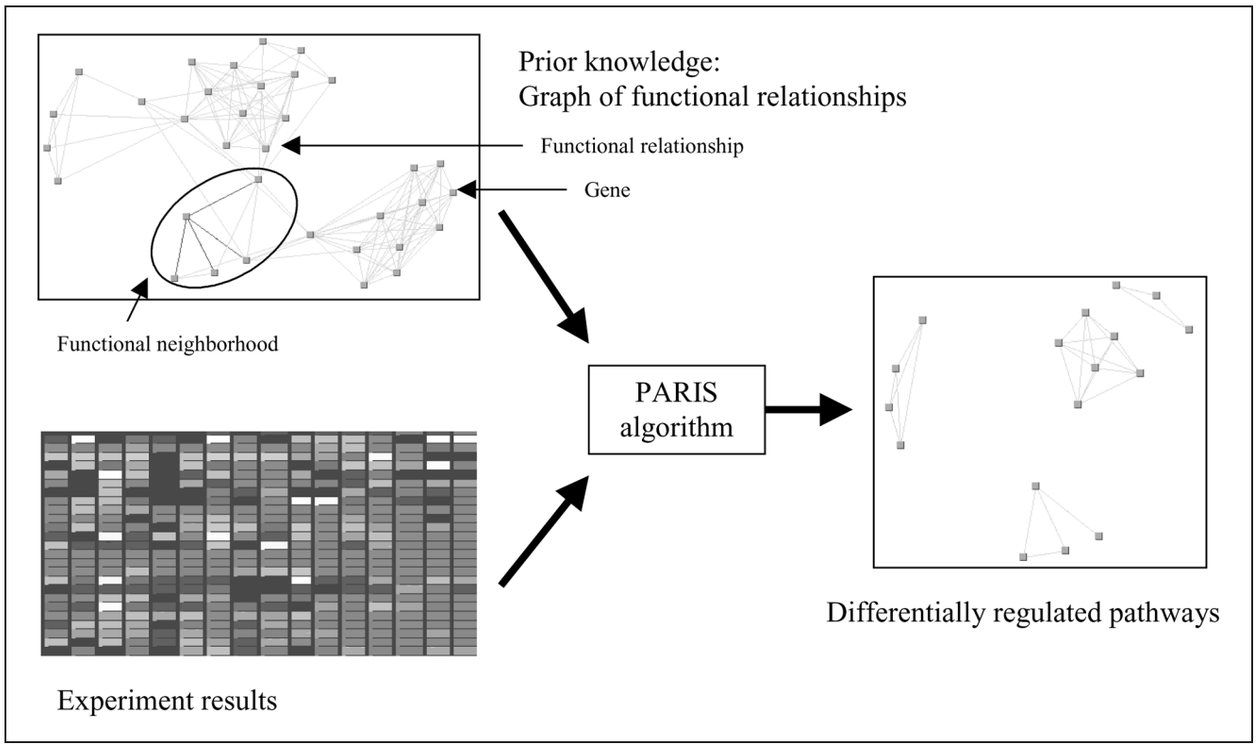

Figure 7: Interpretation of transcript profiling experiments using the PARIS system. Knowledge extracted from the Ingenuity knowledge base is represented as a simple undirected graph where vertices are genes and edges represent their interactions (eg binding). During a TP experiment analysis, each gene of the graph is scored for the density of differential expression observed in its neighbourhood. Genes at the centre of significantly regulated neighbourhoods (eg $P < 0.05$) induce a subgraph representing the regulated pathways

CONCLUSION

Decisions

Biopharmaceutical-scientific knowledge is acquired through a long and expensive series of directed experiments, in which the supporting data can become quite voluminous. New experimental data must be posited against the conclusions from previous experiments. Integration and synthesis with public knowledge must be complete, but knowledge developed inside the enterprise must remain shielded so long as it is treated as intellectual property. Keeping knowledge hidden from competitors may result in it being hidden from the broader organisation. Our goal was to resolve these problems using a suite of technologies and organisational process change.

Transcript profiling

What was achieved? The system described in this paper is currently in production at Millennium and is being actively extended. It provides not only a long-term organisational memory for key science in the biopharmaceutical enterprise, with enhanced analytical capability for experimental data, but also the fullest record of inter-linked data and information to enable more informed decisions about the project pipeline.

Experiences with this system to date show that:

(1) The semi-structured and lightweight *J. Mill.* KB is much more significant than was originally thought.

(2) The availability of public domain, *highly structured* findings provides substantial value even in the absence of internal experimental results.

(3) The ability to directly posit experimental data from transcriptional profiling upon the Ingenuity KB is extremely productive and has led to

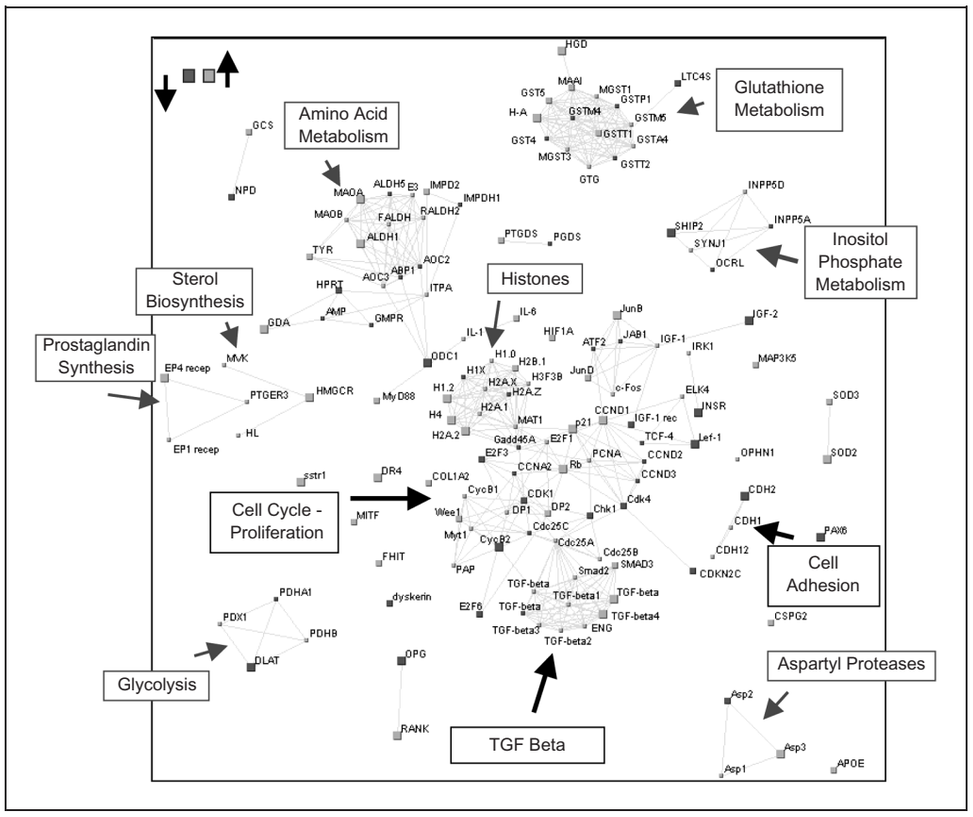

Figure 8: PARIS analysis of *p53* null *v.* wild-type clinical ovarian tumours.[53] The graph shows the pathways that were computationally extracted using the PARIS system when comparing *p53* null ovarian tumour samples with tumour samples that have not lost *p53*. Each gene is represented by a square whose size and shade reflects the differential expression between groups of tumours. Dark grey means more expression in *p53* null samples and light grey less expression. A line between two genes indicates the existence in the PARIS system of a functional relationship (eg binding) between their products. The position of the genes on the picture tries to preserve their respective distances in the full network

key insights concerning possible drug mechanisms of action.

(4) It is important to direct the content acquisition of the KB, and to align it with the research.

Semi-structured findings

J. Mill. was originally conceived mainly as a staging area for capture of the highly structured findings. As it turns out, it has become valuable in its own right as a primary repository for enterprise research findings in human-readable form. This is due both to the rich context provided by the underlying presentations and to the familiar PubMed usage pattern where scientists can skim a title and abstract before going to the original source material. While Ingenuity's highly structured KB is invaluable for computation, links to the original source material (both *J. Mill.* and public sources) provide critical additional background and context vital to correct interpretation.

Insight

The linkage from MyTV to *J. Mill.* also contributed to the effective use of *J. Mill.* and proved to be highly beneficial for gaining new insight into future experiments. One example case occurred when a scientist reviewing target information in MyTV encountered a link to related experiments in *J. Mill.* An examination of the presentation led to a change in future experimental design and a potential new research avenue.

The ability to capture *and structure*

Compound

PARIS algorithms

Clinical

Decisions

Model systems

Target

scientific findings in the KB is crucially important, because proper structure allows computation across a very large corpus of knowledge. At present our internally generated knowledge constitutes only a tiny fraction of the KB — consequently most of its computational power comes from public knowledge, as expressed in the formal language of the KB's ontology. In this context, its single most valuable application has been found to be in combination with the PARIS algorithms, in analysing transcriptional profiling experiments. In part, this is a reflection of the state of experimental technology and the power of genome-scale expression analysis. We conclude that the most useful knowledge *currently* going into the KB is therefore that concerning pathways. Accordingly we have accelerated our work in capturing pathway-related findings from public literature and databases.

Unlike browsing by humans, computational approaches can leverage the consistency, structure and quantity of information in the knowledge base. The PARIS approach has proven successful because it enables scientists to apply this prior knowledge to genome-scale expression data and quickly gain insight into biological processes relevant to the experimental context. It is expected that PARIS will be the first of many computational approaches that will be able to exploit Ingenuity's KB.

The value of computational pathways analysis is increased by its broad application across the pharmaceutical pipeline:

- In the area of target discovery, analysis of model systems and their perturbations was enabled our scientists to identify new pathways containing potential therapeutic targets. The relevance of these pathways was not anticipated and represents new knowledge about the biology of the model systems.
- For compounds that are in-licensed for further development, pathway analysis has proven to be valuable for validating and exploring the supposed mechanisms of these compounds. For one compound, the analysis suggested a different mechanism of action — one that had a more favourable competitive landscape.
- Applying pathway analysis to our clinical data has provided valuable insights into the molecular differences between responders and non-responders and suggested biologically relevant signatures for distinguishing the two populations. This is being used to guide development of follow-on compounds and biomarkers for pharmacogenomics.

Lastly, it is vitally important in such a system to direct the content acquisition and to establish control and feedback mechanisms. If the process and software systems are working, proper direction and feedback will advance the scientific understanding and this will be perceptible by the scientific staff and management. Better operational decision-making will be enabled, and this in turn should be captured.

A record of operational decisions should complement scientific knowledge and deserves an equal place in the organisational memory. There are never enough resources to do all the desired experiments. Operational knowledge in this context really concerns scientific judgments in the context of limited resources. Future versions of *MyBiology* will begin to capture such higher-level decisions affecting the entire drug discovery pipeline within the enterprise.

Acknowledgments

The authors are grateful for the support of Millennium's research scientists, and in particular Laura Rudolph-Owen, Robert Coopersmith and Harshwardhan Bal. We also thank Carole Goble and Robert Stevens of the University of Manchester Department of Computer Science, and Peter Szolovits of the MIT Laboratory for

Computer Science, for helpful and clarifying discussions in the early stages of this project; and Keith Robison, for his thoughtful comments on the manuscript.

References

1. Drewes, J. (1997), 'Strategic choices facing the pharmaceutical industry: A case for innovation', *Drug Discovery Today*, Vol. 2(2), pp. 72–78.
2. Drewes, J. (1998), 'Innovation deficit revisited: Reflections on the productivity of pharmaceutical R&D', *Drug Discovery Today*, Vol 3(11), pp. 491–494.
3. Cavalla, D. (2003), 'The extended pharmaceutical enterprise', *Drug Discovery Today*, Vol. 8(6), pp. 267–274.
4. Abecker, A. *et al.* (1998), 'Toward a technology for organizational memories', *IEEE Intelligent Systems*, Vol. 13(3), pp. 40–48.
5. Stabb, S. *et al.* (2001), 'Knowledge processes and ontologies', *IEEE Intelligent Systems*, Vol. 16(1), pp. 26–34.
6. Maedche, A. *et al.* (2003), 'Ontologies for enterprise knowledge management', *IEEE Intelligent Systems*. Vol. 18(2), pp. 26–33.
7. National Library of Medicine (2002), 'National Library of Medicine Fact Sheet: Medline' (URL: http://www.nlm.nih.gov/pubs/factsheets/medline.html).
8. National Library of Medicine (2002), 'National Library of Medicine Fact Sheet: Medical Subject Headings (MeSH®)' (URL: http://www.nlm.nih.gov/pubs/factsheets/mesh.html).
9. National Library of Medicine (2002), 'National Library of Medicine Fact Sheet: PubMed ®: Medline ® Retrieval on the World Wide Web' (URL: http://www.nlm.nih.gov/pubs/factsheets/pubmed.html).
10. Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1999), 'Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids', Cambridge University Press, Cambridge.
11. National Center for Biotechnology Information (2003), 'Entrez Search and Retrieval System' (URL: http://www.ncbi.nlm.nih.gov/Entrez/).
12. Stein, L. (2002), 'Creating a bioinformatics nation', *Nature*, Vol. 417, pp. 119–120.
13. Clark, T. (2003), 'Identity and interoperability in bioinformatics', *Brief. Bioinformatics*, Vol. 4(1), pp. 4–6.
14. The Gene Ontology Consortium. (2000), 'Gene ontology: Tool for the unification of biology', *Nature Genetics*, Vol. 25, pp. 25–29.
15. Greenwood, M., Wroe, C., Stevens, R. *et al.* (2002), 'Are bioinformaticians doing e-Business?' in Matthews, B., Hopgood, B. and Wilson, M., Eds, 'The Web and the GRID: from e-science to e-business', Proceedings of Euroweb 2002, St Anne's College, Oxford, UK, Dec. 2002, Electronic Workshops in Computer Science, British Computer Society (URL: http://www1.bcs.org.uk/DocsRepository/03700/3782/greenwoo.pdf).
16. The DAML-S Coalition (2001), 'DAML-S semantic markup for web services', 'Proceedings of the International Semantic Web Working Symposium (SWWS)', (URL: http://www.daml.org/services/daml-s/2001/05/daml-s.html.
17. Documentum, Inc. (2003), 'Documentum 5 ECM Platform: Managing Content Across the Enterprise' (URL: http://www.documentum.com/products/collateral/platform/ds_ecm_platform.pdf).
18. Champion, M. *et al.* (2002), 'Web Services Architecture' (URL: http://www.w3.org/TR/ws-arch/).
19. Pradines, J., Rudolph-Owen, L., Hunter, J. *et al.* (2003), 'Detection of activity centers in cellular pathways using transcriptional profiling', *J. Biopharm. Statistics*, in press.
20. National Center for Biotechnology Information. URL: http://www.ncbi.nlm.nih.gov
21. Boeckmann, B., Bairoch, A., Apweiler, R. *et al.* (2003), 'The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003', *Nucleic Acids Res.*, Vol. 31, pp. 365–370.
22. Pruitt, K. D. and Maglott, D. R. (2001), 'RefSeq and LocusLink: NCBI gene-centered resources', *Nucleic Acids Res.,* Vol. 29(1). pp. 137–140.
23. Pruitt, K. D., Katz, K. S., Sicotte, H. and Maglott, D. R. (2000), 'Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI', *Trends Genet.*, Vol. 16(1), pp. 44–47.
24. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) (2000) *Online Mendelian Inheritance in Man, OMIM*™. URL: http://www.ncbi.nlm.nih.gov/omim/.
25. Mulder, N. J., Apweiler, R., Attwood, T. K. *et al.* (2003), 'The InterPro Database, 2003 brings increased coverage and new features', *Nucleic Acids Res.*, Vol. 31, pp. 315–318.
26. Bairoch A. (2000), 'The ENZYME database in 2000', *Nucleic Acids Res.*, Vol. 28, pp. 304–305.
27. Bray, T., Paoli, J., Sperberg-McQueen, C. M. and Maler, E. (2000), 'Extensible Markup

Language (XML) 1.0' (2nd Edn) (URL: http://www.w3.org/TR/REC-xml).

28. W3C (2001), 'XML Schema Part 0: Primer' (URL: http://www.w3.org/TR/xmlschema-0/).
29. W3C (2002), 'XHTML™ 1.0 The Extensible HyperText Markup Language' (2nd Edn) (URL: http://www.w3.org/TR/html/).
30. Microsoft, Inc. (2002), 'Powerpoint® 2002 Product Information' (URL: http://www.microsoft.com/office/powerpoint/evaluation/default.asp).
31. Microsoft, Inc. (2002), 'Microsoft® Word 2002 Product Information' (URL: http://www.microsoft.com/office/word/evaluation/default.asp).
32. Documentum Technical White Paper (2002), 'Developing Web Services with Documentum' (URL: http://www.documentum.com/products/collateral/platform/wp_tech_web_svcs.pdf).
33. The Apache Software Foundation Jakarta Project (2003), 'The Apache Struts Web Application Framework' (URL: http://jakarta.apache.org/struts/).
34. Singh, I. *et al.* (2002), 'Designing Enterprise Applications with the J2EE™ Platform', 2nd Edn, Addison-Wesley, Boston, pp. 348–371.
35. Singh, I. *et al.* (2002), 'Designing Enterprise Applications with the J2EE™ Platform', 2nd Edn, Addison-Wesley, Boston, p. 28
36. W3C (2000), 'Simple Object Access Protocol (SOAP) 1.1' (URL: http://www.w3.org/TR/SOAP/).
37. Minsky, M. (1985), 'A Framework for Representing Knowledge', in Levesque, R. B. A. H. (Ed.), 'Readings in Knowledge Representation', Morgan Kaufmann Publishers, Los Altos, CA.
38. Sanchez, C. *et al.* (1999), 'Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database', *Nucleic Acids Res.*, Vol. 27, pp. 89–94.
39. Karp, P. D. *et al.* (2002), 'The EcoCyc Database', *Nucleic Acids Res.*, Vol. 30, pp. 56–8.
40. Karp, P. D., Riley, M., Paley, S. M. and Pellegrini-Toole, A. (2002), 'The MetaCyc Database', *Nucleic Acids Res.*, Vol. 30, pp. 59–61.
41. Chen, R. O., Felciano, R. and Altman, R. B. (1997), 'RIBOWEB: Linking structural computations to a knowledge base of published experimental data', in 'Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 84–87.
42. Sharon, J., Sasson, L., Parker, A. *et al.* (2000), 'Identifying the key people in your KM effort', *Knowledge Management Rev.*, Vol. 3(5), pp. 26–29.
43. Takai-Igarashi, T., Nadaoka, Y. and Kaminuma, T. (1998), 'A database for cell signaling networks', *J. Comp. Biol.*, Vol. 5(4), p. 747.
44. Goto, S., Okuno, Y., Hattori, M. *et al.* (2002), 'LIGAND: Database of chemical compounds and reactions in biological pathways', *Nucleic Acids Res.*, Vol. 30, pp. 402–404.
45. DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997), 'Exploring the metabolic and genetic control of gene expression on a genomic scale', *Science*, Vol. 278, pp. 680–686.
46. Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A. F. (2002), 'Discovering regulatory and signalling circuits in molecular interaction networks', *Bioinformatics*, Vol. 18, Suppl. 1, pp. S233–240.
47. Zien, A., Kuffner, R., Zimmer, R. and Lengauer, T. (2000), 'Analysis of gene expression data with pathway scores', in 'Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 407–417.
48. Grigoriev, A. (2001), 'A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*', *Nucleic Acids Res.*, Vol. 29, pp. 3513–3519.
49. Jansen, R., Greenbaum, D. and Gerstein, M. (2002), 'Relating whole-genome expression data with protein–protein interactions', *Genome Res.*, Vol. 12, pp. 37–46.
50. Miki, R., Kadota, K., Bono, H. *et al.* (2001), 'Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays', *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp. 2199–2204.
51. Goto, S., Nishioka, T. and Kanehisa, M. (1998), 'LIGAND: Chemical database for enzyme reactions', *Bioinformatic*, Vol. 14, pp. 591–599.
52. Takai-Igarashi, T. and Kaminuma, T. A. (1999), 'Pathway finding system for the cell signaling networks database', *In Silico Biology*, Vol. 1, pp. 129–146.
53. Rudolph-Owen, L., Buller, R., Kovats, S. *et al.* (2002), 'A new method to detect modulation of cellular pathways using transcriptional profiling: Consequence of *p53* and *BRCA1* expression on proteasome inhibition', Oncogenomics AACR Conference poster (URL: http://www.nhgri.nih.gov/CONF/Oncogenomics2002/abstracts.cgi?view=359/).