Executive Summary

Pharmaceutical target identification and validation today is an exercise in complex data mining. The amount, breadth and depth of biological data available for such mining is increasing exponentially, signaling both opportunities and challenges for the biopharma industry.

More data should lead to more insights and better decisions. However, the sheer volume of available data is overwhelming. Further, biological data findings must be considered in the context in which they were discovered and in light of their interactions and/or dependencies on other data sets and conditions. Integrating a wide variety of data sets with such understanding of their contexts is a major logistical hurdle for biologists.

To fully capitalize on the rich biological data sources available today, scientists require a technological platform that eases data integration and comparison across diverse types and sets of data regardless of their sources. The complexity of the technology supporting this platform should be hidden while it offers an easy, streamlined means of interpreting results.

Dynamic Context and Meaningful Biological Insights

In predictive fields such as oil exploration, computational finance, or climatology, data abundance poses a peculiar challenge. In these fields, relationships among data sets are rarely simple and often are not apparent without deeper investigation. So geologists study aerial photography, satellite images, rock analysis and seismographic data to attempt to locate oil basins; meteorologists examine ocean currents and surface temperatures, barometric pressure, polar ice cover and more to predict climate conditions.

The drug discovery process similarly requires assimilation and analysis of seemingly disconnected data sources with the goal of gaining insights for forming a hypothesis to validate through experimentation.

In the initial steps of drug discovery, comprised of target identification and validation, knowledge of disease biology is crucial for picking the right targets. Arguably the biggest breakthrough to that end was the completion of the human genome sequence in 2000.

However, the genome sequence is static; it does not reveal the dynamic role of the targets in a variety of cellular circumstances. Today, this data is available through genomics technologies like microarrays, which can be used to measure thousands of mRNAs or DNA or proteins at the same time. Now scientists have data at a molecular level along multiple cellular/molecular dimensions that can help them understand the
dynamic roles of targets in normal and diseased biological processes.

The data volumes in public "omics"* data repositories, like the Gene Expression Omnibus (GEO), have grown exponentially in the decade since the human genome was sequenced. A snapshot of the type of data sets in GEO reveals that omics data sets can be of varying types and that microarray technology can be used to generate functionally diverse data sets.

Omics technological advances like gene expression microarrays and aCGH offering deeper revelations about siRNAs and miRNAs have enabled us to produce diverse and mammoth volumes of data. In fact millions of data points can be produced for a whole genome single run and aggregated to even larger quantities in a very short time.

Concomitant with the explosion of data is the need for specialized data skills. Target identification has become a data mining exercise. The current challenge is how one can understand the dynamic context of the data, and analyze and interpret it to gain meaningful biological insights. It's challenging for the lead biologist working on a target to interpret and integrate the vast amount of numeric data available in his decision-making process. This leads to underuse or improper use of high throughput data sets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

Omics technological advances like gene expression microarrays and aCGH offering deeper revelations about siRNAs and miRNAs have enabled us to produce diverse and mammoth volumes of data. In fact millions of data points can be produced for a whole genome single run and aggregated to even larger quantities in a very short time.

Concomitant with the explosion of data is the need for specialized data skills. Target identification has become a data mining exercise. The current challenge is how one can understand the dynamic context of the data, and analyze and interpret it to gain meaningful biological insights. It's challenging for the lead biologist working on a target to interpret and integrate the vast amount of numeric data available in his decision-making process. This leads to underuse or improper use of high throughput data sets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.

The Challenges of Mining Today’s Trove of Biological Data

Locked within these huge volumes of data about a cell’s DNA, RNA and proteins and the relationships among them are the information and insights required to successfully identify and validate high-value therapeutic targets.
researchers must develop custom visualizations to address their specific research questions.

- **Fast-evolving domain technologies.** The microarray technology that revolutionized biological data generation is now common, and several other high-throughput technologies, like next-generation sequencing, are evolving that further push the boundaries of data generation. As they do so, the challenge of effectively mining this data to establish useful biological insights becomes increasingly critical to address.

- **Lack of a common platform.** In many modern pharmaceutical research organizations, integrated access to all the data needed to advance a discovery project is often impossible to achieve because it is spread across a variety of databases and visualization and analysis tools. Thus, the scientists maintain their own personal copies of the data, typically in the form of Excel spreadsheets, an inefficient solution.

- **Annotated integration.** Integration in the usual sense is assimilating the parts into the whole. However, this traditional notion of integration is not possible with biological data. It is much harder to numerically integrate experimental values in such a way that the resulting number represents a biologically meaningful phenomenon.

A more pragmatic approach is to integrate the results from each experiment after analyzing them separately. This approach can be extended to any molecular data type. For example, tissue microarray data can be analyzed independently by the pathologist, and the resulting inference of over-expression or under-expression of the protein of interest can be easily compared or integrated with the over-expression or under-expression inference obtained from an RNA microarray experiment.

**Making Information and Insights Obvious: An Integrated Solution**

Researchers and scientists clearly need an intelligent, intuitive solution to data collation and correlation so they can spend the majority of their time interpreting complex biological relationships and their impact on target identification and validation. A single technology platform enabling a workflow that lets scientists look at different biological data types in context and to quickly analyze their relationships would streamline hypothesis generation even as it makes those hypotheses more relevant.

We believe that an effective data integration solution enables productivity gains of at least 20%. Some of the key solution components that will drive such productivity gains are:

- **Faster insight generation.** The platform should integrate various public data sources and have the ability to combine those with commercial and private data. It should help propagate biological insights using any of the common identifiers and be easy to use, enabling scientists to more easily connect the data to reach insights.

- **More revelation of disease mechanisms.** Integrating different types of data should help researchers investigate and understand disease mechanisms more thoroughly, which assists in building the target/disease association.

- **Maximize utilization of high throughput data sets.** By providing easy access to diverse high-throughput data sets, including those available within a company’s research groups and systems, an integration platform would maximize the return on investment on generating such data sets.

- **Platform-independent and extensible.** The solution should be domain platform-independent and be extended to open platforms like caBIG or any proprietary platform.

Oncology has long been a challenging domain because of its complex biology. Our goal with Inventus is to help analyze the data and answer key questions around annotation for additional identifiers, gene ontology, mutation, expression and pathway within the solution through data propagation.

**The Cognizant Approach: Inventus**

We are creating a prototype for a biological data integration platform using oncology as a therapeutic area. Oncology has long been a challenging domain because of its complex biology. Our goal with Inventus is to help analyze the data and answer key questions around annotation for additional identifiers, gene ontology, mutation, expression and pathway within the solution through data propagation.
through data propagation. The typical analysis workflow is depicted in Figure 1.

As the next step, we investigated existing open integration platforms like Life Science Grid (LSG) and cancer Biomedical Informatics Grid (caBIG). We built Inventus using LSG because it provided the necessary minimal information technology framework. We developed the following plug-ins for Inventus:

- Web services-based pathway data from Pathway Commons.
- Access to mutation data from COSMIC.
- Web services-based gene annotation information from NCBI EntrezGene.
- Portal-based data access to BioGPS.org for gene expression.
- Cytoscape for visualizing pathways.

To demonstrate the utility of Inventus, we queried for gene TP53, a well-studied tumor suppressor gene. The default EntrezGene plug-in displays the basic functional information about TP53. The mutation plug-in displays mutation data for TP53 from COSMIC summarizing the mutations studied/identified in various cancer types (see Figure 2). The scientist is quickly able to understand that TP53 is frequently mutated in a wide variety of cancers. The BioGPS plug-in displays the gene expression data from the Novartis gene expression atlas and shows TP53 to be under-expressed (see Figure 3). The above data quickly highlights to the scientist that TP53 could be involved in cancer.

The pathway plug-in from Pathway Commons displays all the pathways that TP53 participates (see Figure 4). One can observe that TP53 is involved in cell cycle pathways. Launching the pathway in Cytoscape allows the scientist to further understand the interacting partners
of TP53 and use network analysis tools within Cytoscape to better understand TP53’s role in pathways and disease (see Figure 5). In cases of a protein like TP53, pathway analysis plays an important role in the discovery process because tumor suppressors have been difficult to reactivate using external intervention. Such proteins may not pose as targets themselves but the downstream or upstream interacting partners could be potential targets.

The above example provides a glimpse of how the solution is useful to discovery scientists, who analyze high-throughput data sets and develop a list with genes of interest to prioritize for additional individual exploration.

Seizing Target Identification Opportunities, Better Understanding Disease Mechanism

We are committed to helping pharmaceutical companies overcome the challenges posed by vast and growing quantities of biological data to seize the opportunities contained within that data. Drawing on our deep biopharma expertise and experience, we take the initiative in developing solutions to industry issues.

To learn more about our solutions for integrating biological data for true insight generation and target identification and validation, contact us at inquiry@cognizant.com.

Footnotes

* refers to the study of biological systems; includes genomics (DNA), proteomics (Proteins), metabolomics (Small Molecules) etc.

1 aCGH – array comparative genomic hybridization

2 siRNA – small inhibitory RNA

3 miRNA – micro RNA

References


LSG Sourceforge: http://sourceforge.net/projects/lsg/

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world’s leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 118,000 employees as of June 30, 2011, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.

World Headquarters
500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277
Email: inquiry@cognizant.com

European Headquarters
1 Kingdom Street
Paddington Central
London W2 6BD
Phone: +44 (0) 20 7297 7600
Fax: +44 (0) 20 7121 0102
Email: infouk@cognizant.com

India Operations Headquarters
#5/535, Old Mahabalipuram Road
Oikkuyam Petti, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060
Email: inquiryindia@cognizant.com

© Copyright 2011, Cognizant. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the express written permission from Cognizant. The information contained herein is subject to change without notice. All other trademarks mentioned herein are the property of their respective owners.

About the Authors

G.D Mahesh Kumar is a Consulting Manager in Cognizant’s Discovery Informatics Center of Excellence located in Sweden. He has over seven years of experience working in drug discovery in the pharma and bioinformatics industries. Mahesh’s core skills include bioinformatics, biological data management and research projects management. He has presented at international conferences and currently manages projects and client relationships with European pharma clients. Mahesh holds a master’s degree in biotechnology from Goa University, Goa, and has been a Project Management Institute member since 2006. He can be reached at Maheshkumar.g.d@cognizant.com.

Raghuraman Krishnamurthy is a Chief Architect in Cognizant’s Life Sciences Business Unit’s Technology Consulting Group. Raghu’s core skills include enterprise architecture, data, SOA, mobility application and convergence of technologies. He has worked with several major pharmaceuticals in envisioning and leading transformational initiatives, has presented papers at numerous conferences and was recently named a senior member of the prestigious Association of Computing Machinery (ACM). Raghu holds a master’s degree from the Indian Institute of Technology, Mumbai, and is a TOGAF-certified enterprise architect. He can be reached at Raghuraman.krishnamurthy2@cognizant.com.

Sowmyanarayan Srinivasan heads Cognizant’s Discovery Informatics Center of Excellence. He has spent over a decade focusing on building business solutions across the spectrum of discovery informatics. Sowmya has worked with leading biopharma organizations to design solutions and consult on their transformation initiatives. He has also helped to establish the Bangalore/India chapter of EPPIC Global (Enterprise Pharmaceutical Professionals from the Indian subcontinent) in early 2000. Sowmya holds a bachelor’s degree in engineering and master’s degree in business administration. He can be reached at Sowmyanarayan.srinivasan@cognizant.com.

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world’s leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 118,000 employees as of June 30, 2011, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.

World Headquarters
500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277
Email: inquiry@cognizant.com

European Headquarters
1 Kingdom Street
Paddington Central
London W2 6BD
Phone: +44 (0) 20 7297 7600
Fax: +44 (0) 20 7121 0102
Email: infouk@cognizant.com

India Operations Headquarters
#5/535, Old Mahabalipuram Road
Oikkuyam Petti, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060
Email: inquiryindia@cognizant.com

© Copyright 2011, Cognizant. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the express written permission from Cognizant. The information contained herein is subject to change without notice. All other trademarks mentioned herein are the property of their respective owners.