

White Paper



Leveraging Research Informatics to Accelerate Drug Discovery

Anirban Ghosh, Siddharth Sawhney, Srikanth Srinivasan, R Arun Kumar, Subhro Mallik and Ipsita Nanda

The main drivers of research innovation in the Life Science industry are patent expiries, pricing pressures, evolving therapeutic needs and the advent of biologics for drug development. The need to find new drugs in the fastest way possible has become more urgent. Geographic spread of research endeavors, new research methods, overflow of scientific research data, and the need for collaborative research practices are defining today's research environment. In spite of having at their disposal advanced research techniques such as Genomics, Proteomics, Marker Based Assessment, and Microarray Technology, organizations find it challenging to realize optimal research outcomes. This paper presents various research challenges and ways to maximize the value of Research Informatics investments.

Introduction

In recent years the pharmaceutical industry has declined in performance ⁽¹⁾, with replenishment of the product pipeline becoming the main criterion for drug discovery research transformation. With US\$ 60bn worth of products going off patent by 2011, life science companies must identify novel and innovative methods to compensate for falling research productivity ⁽²⁾. According to Jean-Pierre Garnier, CEO GSK, “broad transformations of the organization are necessary first steps...only the very best players will be able to meet the challenge and rebuild their R&D engines” ⁽³⁾.

The emerging challenges faced by researchers have outpaced efforts to address them. Research labs are generating data faster than can be integrated. Life science companies are adopting omics-based scientific methods to gain information and knowledge on target validation. While the pharmaceutical industry has been adept at optimizing the drug development process, it has rarely implemented different structures ⁽⁴⁾ to make the discovery process more efficient ⁽⁵⁾. Therefore, scientists must use their creativity to constantly innovate ⁽⁶⁾ and align informatics and data management needs to meet an integrated cross-disciplinary discovery process ^(7, 8, 9).

Pharmaceutical companies are entering the biopharmaceutical space by investing in vaccine programs and RNA and protein-based therapies for complex chronic diseases ⁽¹⁰⁾. While some of the chemistry processes may still be in use, many novel biological methods are being introduced to validate biologics hypotheses. Biologics-based research scientists need to manage biological entities through registration, inventory and applied research systems ⁽¹¹⁾.

Trends in discovery research processes

- Adopting biological discovery methods or marker-based assessment to unravel new therapeutic solutions – build emerging core processes.
- Connecting molecular biology results with clinical research outcomes to conduct root cause analysis of a disease and its response to therapy – feedback loop processes from distinct disciplines.
- Instituting standard methods and activity procedures in most of the innovation-led project operations – follow standardized workflows.
- Using computational techniques such as predictive eBiology to harness large volumes of heterogeneous data in order to devise better experimental strategies ⁽¹²⁾.
- Enabling each research program to be supported by heterogeneous scientific and informatics groups located in different geographies – multi-geography talent pool collaboration in real-time or offline ⁽¹³⁾.
- Using cross-disciplinary scientific research processes for drug discovery – reduce ambiguity in terms and terminologies within chemistry or biology.
- Sliding R&D productivity in the past decade ⁽³⁾ not “miraculously cured” by the sequencing of the human genome and the industrialization of techniques employed in the early discovery process.
- Collaborating with external partners ⁽¹⁴⁾ and even competitors for scientific discovery or solving a common research problem ⁽¹⁵⁾ – reduce cycle times and enhance predictability.

Challenges in discovery research

Scientific research generates data by registering biological or chemical entities and testing their biological, physical or chemical character or their pharmacological action. Information related to registration and assay workflows forms the basis of all scientific innovation in any disease program. However, there are several challenges including process complexity, data indecipherability and questionable technology efficacy. The following table categorizes the most critical issues:

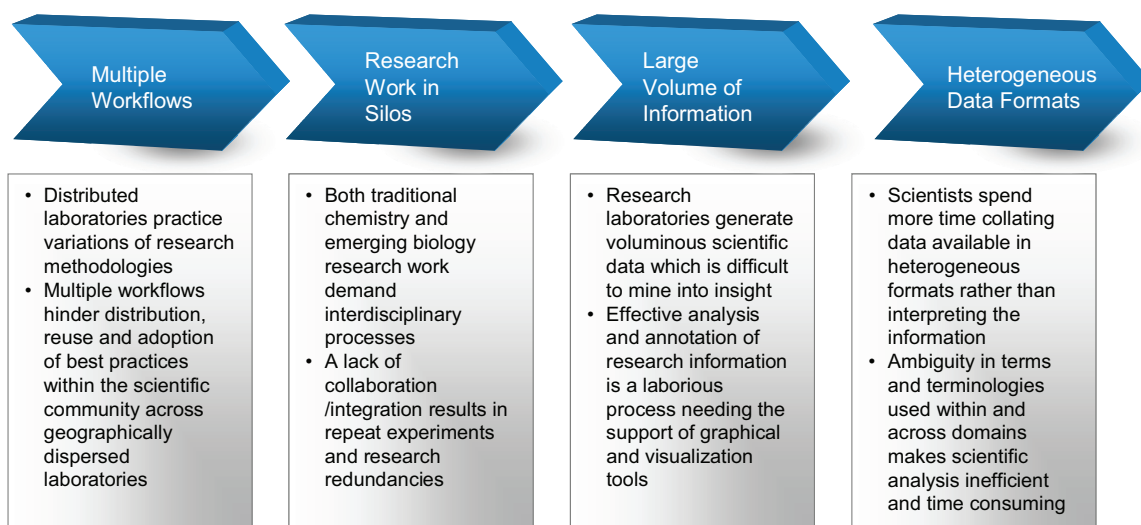


Fig. 1: Broad classification of challenges in discovery research

Multiple workflows

The drug discovery process is complex and inter-disciplinary in nature and its activities are supported by a portfolio of tools and applications collectively named Research Informatics. The current software portfolio supporting chemistry or biology functions includes packages, products, frameworks, and custom applications that consume a large proportion of the R&D budget for license fees or maintenance costs.

Multinational pharmaceutical companies have globally distributed laboratories engaged in related or similar research activities. Mergers and acquisitions lead to redundancy and lack of harmony in laboratory workflows. Laboratories belonging to multitudinous disciplines of biology and chemistry practice variations of the assays or registration workflows. Some of their activities are centered on hypotheses requiring a composite set of processes to be orchestrated together. Many of these laboratory processes are dependent on the instruments used for analysis. Multiple workflows hinder distribution, reuse and adoption of best practices within the global scientific community. For example, the most critical step in lead discovery is High Throughput Screening which varies with the target objective, creating multiple workflows across laboratories. Laboratories that focus on the lead compound study adhere to chemistry assays, which are inhibitory and plate-based. On the other hand, biologists and pharmacologists who conduct target specificity or toxicity studies, rely on marker-based cell line screening. The experimental routines conducted using different types of instruments and systems generate different data types. The screening results for different hypotheses are rolled into a normalized inhibitory concentration of the ligand-target complex. Sometimes, scientists find it difficult to correlate the results of chemistry screening with those of cell biology-based screening on account of differences in data type and value. Therefore, the standardization of similarly practiced screening operations enables reuse of a single protocol and ensures consistent information transaction.

Research work in silos

Systems biology and chemical screening centers of excellence supporting mainstream programs are located in places with easy access to infrastructure and a low-cost talent pool. Traditional chemistry and emerging biology research must be integrated to create an inter-disciplinary activity. A lack of collaboration between biology and chemistry processes leads to repeat and redundant work and inconsistent results. Integrating chemistry data within the context of biochemical processes or integrating genetic data with biological pathway information can facilitate better understanding of disease. Hence there is a need for biology and chemistry experiments to “cross-over” in the interest of disease research. A chemistry-biology interface would be required in structural biology, enzyme function, structure function, or protein folding research in order to state the disease problem better.

Take the example of non-steroidal anti-inflammatory drug research, in which structural biologists explore a novel idea to fulfill unmet therapeutic need in inflammatory arthropathies like psoriatic arthritis. By reading research findings published in various medical journals, they identify the target protein/enzyme causing the inflammation as Cyclooxygenase II. This is passed on to chemists for chemical characterization and identification of chemical information like structure

(1,6-dimethylnaphtho[1,2-g][1]benzofuran-10,11-dione), molecular weight (276.286080 g/mol) etc. A biological safety and toxicity assessment through the screening of cell based assays (IC50 data, dose response curve info) is conducted to qualify the targeted benefit based hypothesis further. If the scientists in the above scenario work in silos, they will take months to complete the entire procedure, since each group will follow up, analyse, consult with experts or iron out operational bottlenecks on their own. On the other hand, if there is seamless collaboration within and outside the research groups, the hypothesis can be cross-checked and the target's role and mechanism of action validated with ease.

Large volume of information

Research laboratories generate large volume of scientific data to draw insights and inferences. Recent experimental techniques such as the omics methods and computational simulation generate terabytes of raw data in its every run. Effective analysis and annotation of the basic reads or data sets can be laborious without the help of parsers and graphical and visualization tools. Additionally, there are concerns about the security of data stored and exchanged across laboratories and the possible theft of intellectual property.

“The new targets being identified with the advent of the post-genomic era are causing a massive explosion in the number of data points per unit time,” says John Helfrich, program manager, drug discovery and development at NuGenesis Technologies Corporation. “Typical large pharma today are generating 20 terabytes of data daily. That’s probably going up to 100 terabytes per day in the next year or so.”⁽¹⁶⁾

Heterogeneous data formats

The primary entities in drug discovery research are Diseases, Pathways, Proteins (along with their interactions) and Genes. These are the foundation stones on which new molecule research is built. The biggest obstacle to the integration of research information is that data is usually available in heterogeneous formats and stored in silos, and hence cannot be shared easily. In addition, frequent duplication of information or ambiguity in terms adds to the difficulty of making timely informed decisions.

Registration of biologics or compounds; analytical instruments, library of compounds and biologics; knowledge assets in the form of literature reports – all of these make up heterogeneous-format data silos. Thematic knowledge bases or reporting dashboards often lack the capability to update, correct, or append various sets of raw data. Hence, scientists first spend time generating raw data sets and then some more interpreting pieces of data to create knowledge assets. This lack of integration across research entities makes scientific analysis inefficient and time consuming.

Let us say that a scientist is looking to characterize and validate Janus Kinase 3(JAK3) as a therapeutic protein target for autosomal SCID. The properties of the target JAK3 are best understood with the availability of its crystal structure and its binding with a ligand like 1,2,3,4-TETRAHYDROGEN-STAUROSPORINE. The crystal structure obtained from the Protein Data Bank public repository helps to conduct a structural analysis and annotate key structural constructs that may influence sequence specific binding studies. Information on the UniProt protein sequence, its interaction with other proteins in the pathway from KEGG, and its homology to other functional proteins from the Entrez Gene provides insights on its involvement in cytokine receptor mediated intracellular signal transduction. Additional evidence from PubMed literature shows that JAK3 deficient mice had profoundly reduced thymocytes and severe Bcell and Tcell lymphopenia similar to SCID conditions. Also, internal screening assays show JAK3 proteins to be highly preferential targets for the lead compound 1,2,3,4-TETRAHYDROGEN-STAUROSPORINE. Historical information contained in handwritten notes or electronic files such as word documents, spreadsheets and power point slides are also referenced to support the hypothesis. Due to lack of automation of information transaction activities, the scientist must painstakingly assemble and synthesize data from these disparate sources before he or she can derive meaningful insight.

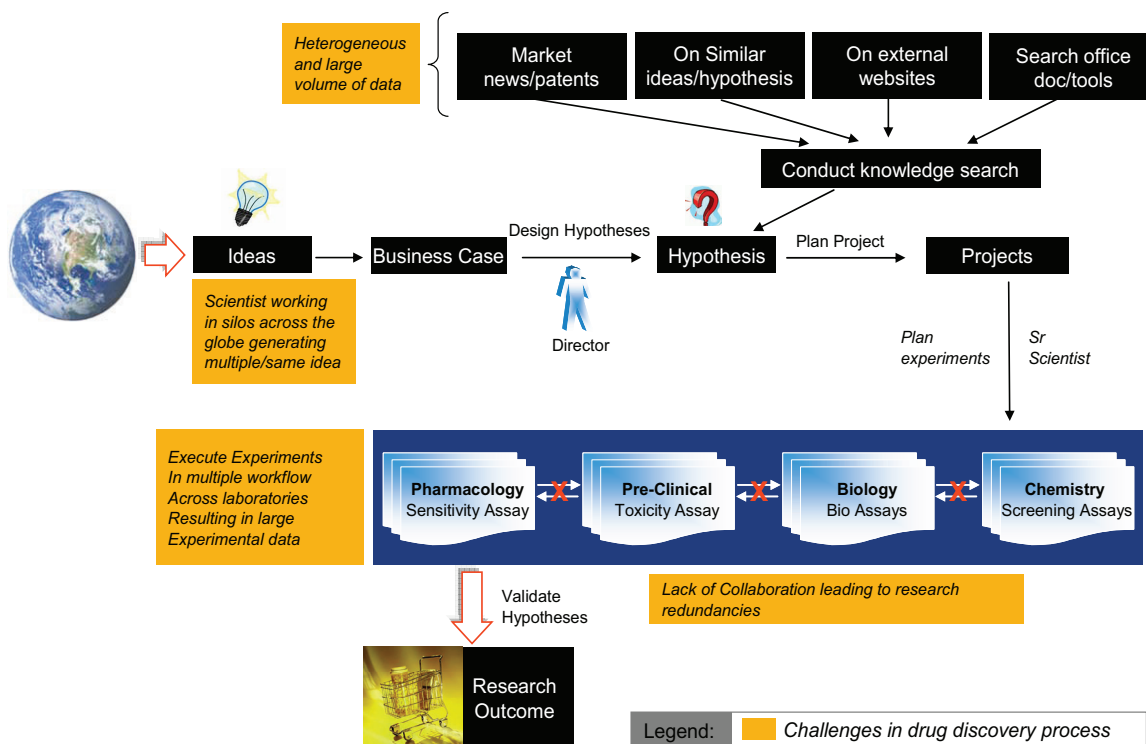


Fig.2: Research challenges during the drug discovery process

Ways to overcome research challenges

The use of techniques such as data semantics, visual analytics, collaboration and workflow streamlining in drug research can increase research effectiveness, improve predictability, foster team work among scientists and thereby produce better research outcomes.

Streamline process workflows

Pharmaceutical companies continuously optimize their processes and workflows in existing fields of research and adopt new ones to drive innovation in discovery. Procuring new products before studying their alignment with research processes only adds to license fees and maintenance costs, without adding value to knowledge capabilities. One way to optimize costs is to establish a tight linkage between processes and applications. The objective is to select processes that will improve competitiveness, prioritize business activities and enable IT solutions.

The rationalization of applications by way of retention, re-engineering or retirement should not depend solely upon technology trends, implementation costs or ease of procurement but rather on expected business value. Each of the research activities should be broken down into discrete tasks and mapped to an application portfolio for value engineering. Pharmaceutical companies should harmonize research workflows, reduce redundant processes and leverage an optimal composition of applications to process information.

- Put together regular biology and chemistry methods as standard routines
- Upload standard workflows onto an electronic system for all laboratories of an organization to follow
- Define workflow boundaries that can help join different disciplines of science

Collaborative research

The therapeutic product portfolio is being increasingly crowded by large rather than small molecule-based screening and optimization. Therefore, erstwhile chemistry and new biology research work closely together in inter-disciplinary projects. Research scientists from different disciplines need to actively understand and address various facets of the disease problem together. An example of collaboration is when results for cell line-based screening assays against a class of inhibitor compounds are jointly interpreted by a biologist and a pharmacologist.

Currently, there is only a moderate level of collaboration. Scientists are not known to freely share and exchange concepts or findings from their experiments or computations. Most often, interchange of ideas and information sharing happens via handwritten notes, whiteboard or electronic mail. For research collaboration to yield meaningful benefits, it must be viewed as an imperative.

- Make collaboration an attractive proposition to scientists, and include it in their key performance indicators
- Technology-enable collaboration by leveraging RSS, Wiki and Online Portals
- Consider using a multi-site collaboration platform that will streamline document workflow, and offline/online content sharing in standard templates
- Promote open collaboration with academia and establish pre-competitive collaboration with industry while safeguarding intellectual property

Visual analytics for large data sets

Scientists evaluate a hypothesis by gathering large volumes of multi-dimensional data for inspection. While raw alphanumeric data can be cumbersome to handle, a pictorial rendition can facilitate analysis. Even as scientists slice and dice through mountains of 2-D graphical data, they often need other types of graphical presentation for drill down analysis. Sometimes, two sets of data are compared and contrasted keeping one of two parameters constant.

Pharmaceutical companies either develop bespoke applications for molecule and data visualization or buy third party applications. While addressing complex questions, scientists first parse data to present it in appropriate views. Naturally, they prefer easy to use intuitive tools which guide them towards standard data views for inspection. Other visualization enablers provide fresh vistas to tackle multi-dimensional data views. Data visualization is undergoing a paradigm shift with unified molecule, data, graph, document and video visualization tools endowed with strong analytical capability being designed based on gaming consoles. In summary, visual inspection, analysis, editing, and annotating are now possible through various technology-enabled solutions which can dig into the data glut.

- Scientists should be equipped with visual analytics tools that include pre-defined protocols for contextual data mining and filtering
- Choose technologies that support vector-based rendering of millions of data points to show depth, perspective, and performance of 3D image, network, molecule, or data distribution
- Visually represent semantically joined concepts over a life-sized surface display

Semantics for data interoperability

Clearly, aggregation of information across the discovery value chain is pivotal to creating an integrated discovery engine, a holy grail for leading pharmaceutical companies worldwide. Still more important is the ability to carry out cross-functional search. Effective integration infrastructure enhances the ability to carry out cross-functional search on biological and chemical information categories, crashing time-to-market for new drugs. Re-wiring existing data assets in the context of domain ontology is an effective way to achieve inter-operability.

Currently, life science companies are wading through biological and chemical semantics to create a web of standard ontology. This will help scientists link a compound to a product, relate clinical protocols with an indication, associate a protocol with an experiment, determine synonym company identification with a generic name of a compound, or connect pharmacovigilance signals to genes in a pathway. Presently, product companies are building ontology-driven search capabilities and creating a connected graph of terms and concepts. Industry, academia and vendors are creating life science specific ontology according to W3C standards. The semantic web presents the possibility of a virtual electronic whiteboard displaying visual links between all documents and content supporting the concept under discussion. Cytoscape is an open source initiative that helps visualize pathway vocabulary (BioPax3) that scientists can review or modify. Pharmaceutical companies are also individually building enterprise vocabulary services to unlock the knowledge hidden in heterogeneous data sets.

Building data service methods around key domain entities is a good way to broker information across multiple points of access. These services can fetch data from diverse scientific silos in the context of the research investigation. Using semantic technologies, researchers and program directors can discover relationships that enable them to make better and faster decisions about disease targets and drug compounds. Data inter-operability with ontologically linked data sets reduces the time needed to assimilate research findings.

- To enable semantic search, structured tabular data needs to be converted into relationship models as per a commonly agreed upon domain ontology
- Web Ontology Language and Resource Description Framework are technologies that unify and integrate data into machine-interpretable concepts

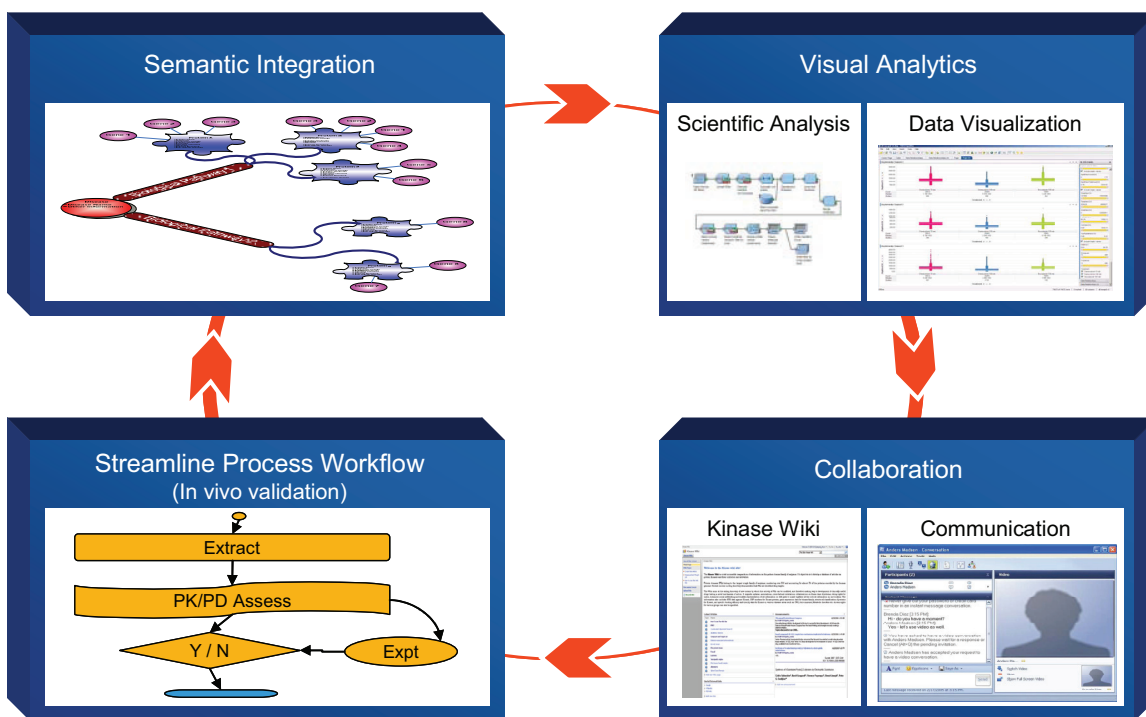


Fig.3: A schematic representation of a research informatics system

Building a Research Informatics Ecosystem

Companies looking to overhaul their research capabilities must recognize that this is a major transformation effort. As outlined in the earlier section, there are complex inter-dependencies between data and visualization as well as collaboration and workflow. Besides these, the focus areas critical to realization of transformational value are:

Business Value Creation

Some companies embark on this journey without adequately estimating the potential value generation from the investment. Without clearly articulating the levers of value generation (like scientist productivity or faster cycle times) and capabilities required to accomplish the same, it is difficult to train focus on the right areas or produce quick wins to generate momentum for the program. There are also serious risks in “boiling the ocean” on initiatives like semantic inter-operability and starting massive technology programs that are not aligned with the way scientists do research.

Adoption and accountability within the scientist community

Adoption by the scientific community is critical to the success of this effort. Some of the new practices will run counter to widely held beliefs and research practices. It is therefore important to identify early adopters and champions as well as design solutions that align well with scientists’ research practices.

Strategic partnerships with external parties

In recent years, the research community has made good progress in the areas of external collaboration and open networks. They can reap significant opportunities by harnessing this trend and integrating it within the research process. This will require new ways of thinking about how information is shared within and outside the organization, alignment of goals/incentives to encourage collaboration as well access controls for protection of IP and sensitive data. There is also an opportunity to redefine what is central to the organization from a research standpoint and what can be outsourced to partners specializing in different disciplines.

Management of the overall program

It is crucial to manage all the different threads in close alignment in order to give researchers the tools and processes to improve research outcomes. Momentum- building quick wins, clearly stated vision and goals and an integrated capability view linked to research outcomes are needed to keep the program on track. Regular supervision by leadership teams from research, research informatics and IT will enable adaptation to the changing business and IT landscape.

Conclusion

In the face of major challenges and pressure to replenish the revenue pipeline, research organizations have an opportunity to re-emerge as the growth engines of industry. This will need significant commitment from the leadership and a strong vision for the future supported by the ability to acquire and deliver value in a phased manner. Pharmaceutical organizations, patients, payers and governments alike will welcome faster and cost-effective discovery of innovative therapies for present day medical challenges.

References

1. Market watch: Pharma industry strategic performance, 2008–2013E. Michael Goodman. *Nature Reviews Drug Discovery*, 8, 348, May 2009.
2. Why has R&D productivity declined in the pharmaceutical industry? R R Ruffalo, *Expert Opin Drug Discovery*, 1, 99-102, 2006.
3. Rebuilding the R&D Engine in Big Pharma. Jean-Pierre Garnier. *Harvard Business Review*, May 2008.
4. Optimizing the discovery organization for innovation. Frank Sams-Dodd. *Drug Discovery Today*, 10(15), August 2005.
5. R&D Efficiency, Tuft Center for the Study of Drug Development. Tufts Univ., Outlook 2009.
6. Mission possible: managing innovation in drug discovery. Xiaotian Zhong, George B Moseley. *Nature Biotechnology*, 25, 945 – 946, 2007.
7. Integrating scientific data for drug discovery and development using the Life Sciences Grid. Ernst R Dow, James B Hughes, Susie M Stephens, Vaibhav A Narayan, Richard W Bishop. *Expert Opin Drug Discovery*, 4(6), 687-699, 2009.
8. Advanced biological and chemical discovery (ABCD): centralizing discovery knowledge in an inherently decentralized world. Dimitris K. Agrafiotis et. al. *J. Chem. Inf. Model*, 47 (6), 1999-2014, 2007.
9. Tripos Discovery Benchware 360 product. http://tripos.com/tripos_resources/fileroot/pdfs/FuturePharma_WYETH%20Podcast.pdf
10. The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. A Muzzi, V Massignani, R Rappuoli. *Drug Discovery Today*, 12, 11/12, June 2007.
11. Developing a cell line registration and analysis system, <http://www.infosys.com/industries/life-sciences/case-studies/biological-therapeutic.pdf>
12. High-throughput electronic biology: mining information for drug discovery. William Loging, Lee Harland & Bryn Williams-Jones. *Nature Reviews Drug Discovery*, 6, 220-230 March 2007.
13. Practitioner's Perspective: Connect and Bolster Pharma R&D Model. Mandar Ghanekar, Anirban Ghosh, Kamal Biswas. SETLabs Briefings, September 2008.
14. Can open-source R&D reinvigorate drug research? By Bernard Munos. *Nature Reviews Drug Discovery*, 5, 723-729, September 2006.
15. Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. Michael R. Barnes, Lee Harland, Steven M. Foord, Matthew D. Hall, Ian Dix, Scott Thomas, Bryn I. Williams-Jones & Cory R. Brouwer. *Nature Reviews Drug Discovery*, 8, 701-708, September 2009.
16. Drug Discovery and Biotechnology Trends: Laboratory Automation: Bursting Through the Bottlenecks. Peter Gwynne and Gary Heebner. <http://www.sciencemag.org/products/ddbtjan.dtl>.

About Authors

Anirban Ghosh, PhD is a Principal at Infosys Consulting. He has made significant contributions in scientific innovation and research informatics through international publications, patents and solution implementations at top pharmaceutical and applied life sciences companies. He can be contacted at ghosh_anirban@infosys.com

Siddharth Sawhney is a Senior Project Manager with the Life Sciences practice at Infosys. He leads the effort for design and development of informatics solutions for discovery research. He has experience in software engineering, process consulting, and program management. He can be reached at siddharths@infosys.com

Srikanth Srinivasan is the Managing Partner for the Insurance, Healthcare and Life Sciences practice at Infosys Consulting. He leads advisory projects and transformation programs at Life Sciences clients. He has experience in helping companies realize value from business transformation efforts. He can be reached at s_srikanth@infosys.com

R Arun Kumar is an Associate Vice-President, who heads the Global Life Sciences practice at Infosys and is responsible for the growth and expansion in the Life Sciences domain. Arun has more than 16 years of professional experience in the areas of business-technology alignment, IT and BPO services, global sourcing, strategy & marketing, software product development, wireless and consumer goods. His career spans multiple continents and he has worked in leadership roles in established trans-nationals as well as in start-ups. Arun lives in the San Francisco Bay Area and can be reached at R_ArunKumar@infosys.com

Subhro Mallik is Group Engagement Manager in the Life Sciences practice at Infosys. He manages key customer relationships particularly with the Big Pharma in US East Coast area. He has experience in helping companies realize value from business transformation efforts. Subhro can be reached at SubhroMallik@infosys.com

Ipsita Nanda is a Business Analyst at Infosys. She is a business management graduate having experience in biotechnology and life sciences, particularly in the area of molecular biology. She can be contacted at ipsita_nanda@infosys.com



For more information, contact askus@infosys.com

About Infosys

Many of the world's most successful organizations rely on Infosys to deliver measurable business value. Infosys provides business consulting, technology, engineering and outsourcing services to help clients in over 30 countries build tomorrow's enterprise.

For more information about Infosys (NASDAQ:INFY), visit www.infosys.com.