ELSEVIER

Beyond data integration

Ted Slater[1], Christopher Bouton[2] and Enoch S. Huang[2]

[1] Pfizer Worldwide Technology, 700 Chesterfield Parkway West, Chesterfield, MO 63017, USA
[2] Pfizer Global Research and Development, Cambridge Laboratories, 620 Memorial Drive, Cambridge, MA 02139, USA

Pharmaceutical R&D organizations have no shortage of experimental data or annotation information. However, the sheer volume and complexity of this information results in a paralyzing inability to make effective use of it for predicting drug efficacy and safety. Data integration efforts are legion, but even in the rare instances where they succeed, they are found to be insufficient to advance programs because interpretation of query results becomes a research project in itself. In this review, we propose a coherent, interoperable platform comprising knowledge engineering and hypothesis generation components for rapidly making determinations of confidence in mechanism and safety (among other goals) using experimental data and expert knowledge.

Introduction

In his excellent analysis of the business of biotechnology, Gary Pisano [1] observed that the biotechnology industry differs from other high-tech sectors in that (1) it is characterized by 'profound and persistent uncertainty' in R&D, related to our shallow understanding of how human biological processes and systems respond to therapeutic intervention; (2) its various specialized disciplines must work in an integrated fashion, that is the R&D process cannot be broken into separate problems to be solved independently; and (3) much of the knowledge in its various disciplines is 'intuitive or tacit, rendering the task of harnessing collective learning especially daunting.'

Pisano is completely correct, in our experience. The high-stakes nature of drug discovery and development has now reached critical levels. George Milne, former president of Pfizer Research, commented that despite dramatic investments in R&D, the overall productivity of the pharmaceutical industry has not increased over the past decade [2]. The grim statistics collected by various senior leaders in the industry are remarkably consistent: the ratio of preclinical candidates to approved product has been roughly 25 to 1, a staggering 96% attrition rate [2–4]. These punishing odds are a direct consequence of the complexities in translating the sound theoretical basis for molecular target selection into clinical proof of concept and the difficulties associated with predicting the *in vivo* and clinical safety of novel compounds. Pisano is also accurate in his assessment that our limited knowledge is highly decentralized and mostly tacit in nature. In other words, the expertise and experience required to successfully complete the process of discovering and developing a new drug is necessarily spread out among thousands of people in different departments at different geographic locations. Yet there are technologies, some new and some old (even ancient), we can use to strengthen our approach.

Data integration is not the answer

Data integration (see Glossary) has been the rallying cry of the pharmaceutical industry for many years now, the sentiment being that if we could just get all of the information we need in one place so that we can query it, we will have eliminated many of the roadblocks to our success in moving compounds through the pipeline. All of the answers will be available at the end of every query. There is no doubt that properly integrated biological information can make the drug discovery process much more efficient by providing testable hypotheses after visualization and mental processing [5,6]. These high-value hypotheses might lead to the uncovering of a causal relationship between the biological activity of an enzyme or receptor to the progression of a disease, or the discovery of a secreted biomarker that correlates with a particular adverse event. However, whether or not we have collectively

Corresponding author: Huang, E.S. (enoch.huang@pfizer.com)

1359-6446/06/$ - see front matter © 2008 Published by Elsevier Ltd. doi:10.1016/j.drudis.2008.01.008

Please cite this article in press as: Slater, T, Beyond data integration, Drug Discov Today (2008), doi:10.1016/j.drudis.2008.01.008

GLOSSARY

Data integration: the process of combining disparate data and providing a unified view of these data
Domain: a body of knowledge, such as biology
Domain knowledge: the terminology and facts of a domain without a focus on any particular task
Graph (mathematics): a set of objects called nodes or vertices connected by a set of links called edges
Inference engine: that part of a knowledge system that uses the knowledge in a knowledge base to reason its way to solutions to problems
Knowledge base: repository for the knowledge used by a knowledge system
Knowledge system: short-hand for knowledge-based system; a computer system that represents and uses knowledge to carry out a task
Ontology: a formal description of set of entities within a domain and the relationships between those entities, used to reason about the entities within that domain
Semantics: what a symbol means, separate and distinct from the symbol itself
Triple: a statement of a relationship, called the predicate, between a subject (the entity the triple is about) and the object of the triple (another entity or value)
Triple store: a data store (such as Kowari) geared towards storing and returning RDF triples in response to queries

worked out a sustainable solution for 'properly' integrating biological information remains an open question. Indeed, the remarks of Buneman and co-authors in their piece featured in Towards 2020 Science [7] are sobering: 'attempts to solve the issues of scientific data management by building large, centralized, archival repositories are both dangerous and unworkable.'

The scientific data repositories in use throughout the pharmaceutical industry today are characterized by being siloed (isolated, unconnected to any other repositories or applications), redundant (many 'new' repositories being built contain information already present somewhere else), and inaccessible (researchers typically have very limited query capability, if they have access at all). Furthermore, current bioinformatics systems do not have the ability to reason, that is, to automatically produce novel insights that were not previously captured and integrated. Consider the analogous situation related to the Human Genome Project (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml). As we undertook to sequence the entire human genome, many of us believed that when we were finished, medicine would be revolutionized. The reality is that the human genome has been completely sequenced for some years now, and while this accomplishment should in no way be diminished from the major scientific and technical achievement that it is, the fact remains that medicine has not quite been revolutionized (see Towards 2020 Science, p. 60 [7]). It is true that you can download the entire human genome (and soon even your own genome's sequence) to your iPod and take it to your physician, but alas, it is unlikely that this will have any impact on the quality of care you receive.

Semantics is required to interpret information

Genomic technologies indeed require time to mature before significant societal benefits can be realized. Nevertheless, one major reason we cannot make better use of the fully sequenced human genome is that it currently lacks semantics (but help may be on the way [8]). For your physician, there is simply not enough meaning associated with the information to make it directly useful for your treatment. Some prior interpretation of the information is necessary to endow it with meaning before it can be used to diagnose and guide treatment. Indeed, information becomes knowledge when it is interpreted, absorbed, and socialized such that it becomes part of an individual's knowledge resource base [9]. Interpretation endows information with practical meaning, which ultimately allows it to be applied to a particular purpose.

The point is that we will not make the best use of the sequence of the human genome unless we know its meaning, and by the same token, the pharmaceutical industry cannot expect to derive maximum benefit from the mere integration of its data without the corresponding semantics. Providing such interpretation in the domain (see Glossary) of pharmaceutical R&D is made very difficult because of its enormous complexity. We are only beginning to learn how to interpret thousands of empirical data points per experiment in the context of everything known about biology, chemistry, and other disciplines. This is because, as Peter Karp has pointed out [10], our conceptualizations of biology have grown in size and complexity to such an extent that even experts cannot hold them in their heads in order to reason with them. Unfortunately, this is something we simply must learn how to do, because it is increasingly clear that we need to understand everything we can about the biological context of targets and their mechanisms of action in order to truly understand disease and drug treatment. How, then, can we use technology to interpret information germane to pharmaceutical research and convert it into knowledge?

Beyond data integration

In 1975, Newell and Simon presented the Physical Symbol System Hypothesis in their Turing Award paper [11]. This hypothesis states that a physical symbol system has the necessary and sufficient means for general intelligent action; in other words, any system exhibiting general intelligence (e.g. the ability to make plans and formulate hypotheses) will necessarily be found to manipulate symbols that represent entities in the physical world, and that the ability to do so is all that is necessary for general intelligent behavior. Knowledge systems (see Glossary), which are well-established artificial intelligence (AI) systems, are physical symbol systems that comprise a reasoning engine or problem-solving process coupled with a domain-specific knowledge base (see Glossary). The reasoning engine is a system for making inferences from the information in the knowledge base, usually using a set of rules from a rule base. A knowledge base is a way to store and manipulate the physical symbols that represent entities in a domain of interest, such as the enzymes, targets, metabolites, and other elements in a pathway relevant to the domain of pharmaceutical R&D. Knowledge engineering is a practical discipline that provides the tools and techniques for creating knowledge systems.

A properly designed knowledge system provides the means to manipulate symbols representing entities in some domain of interest computationally. Such manipulation allows computers to perform various kinds of seemingly intelligent behavior within the domain, including reasoning and other useful kinds of computation. Because the pharmaceutical R&D domain is so complex,

Please cite this article in press as: Slater, T, Beyond data integration, Drug Discov Today (2008), doi:10.1016/j.drudis.2008.01.008

the benefit of such a system to researchers already overwhelmed with information is clear. However, the choices made in creating a knowledge system can have another, perhaps unexpected benefit: the right knowledge system can help solve the problem that attempts at data integration have failed to solve.

To see how this might work, consider a knowledge base representing the domain of human diabetes. Imagine that the facts, or assertions, relevant to diabetes were represented as subject-predicate-object 'triples' like

p53 decreases-expression-of insulin-receptor

This assertion, represented graphically in Figure 1, is about the transcription factor p53, and it describes how p53 down-regulates the expression of the insulin receptor. Now, further imagine that thousands of such facts, all in some way relevant to diabetes, are interconnected by having some subjects serve as objects of other triples (see Glossary), and vice versa. The resulting construct is a mathematical graph (see Glossary), a set of nodes (here representing entities like p53) and interconnecting edges or lines (here representing relationships such as decreases-expression-of) representing facts about diabetes. Figure 2 depicts a simple graph. The blue circles (nodes) represent entities in the domain, and the arrows (edges) represent the relationships between them. The fact that the edges are arrows indicating directionality, and not merely lines, indicates that the graph is a *directed* graph, just what is needed to represent a particular relationship between a subject and an object (though not necessarily the converse).

The semantics associated with the entities and their relationships are formally described, and thus controlled, by a set of ontologies: formal descriptions of the domain, including the entities in it and how they can (and cannot) interact. We can develop a simple *if-then* rule base that takes advantage of the semantics in the ontologies and the regular graph structure to implement an expert reasoning system over the graph (Figure 3). Finally, imagine that the graph of diabetes knowledge is stored in a relational database, the schema for which being relatively straightforward because of the simple node-edge architecture of graphs.

Toward a universal, interoperable knowledge architecture

Now, suppose we would like to create a knowledge base about a separate but related domain, such as dyslipidemia. What would we need to do to accomplish that? We notice that the subject-predicate-object triple format for representing assertions in the original domain is flexible and powerful and therefore perfectly serviceable for the new domain. Because the assertions about our new domain are represented as subject-predicate-object triples, we can simply duplicate our original database schema and reuse it as-is. There is no need to design a new database schema, because the kinds of information being stored (entities and relationships, or nodes and edges) are the same as before. Since our database schema has not changed and our knowledge representation is identical, we can continue to use the same algorithms for searching, traversing, manipulating, and visualizing the graph.

We will have to make some changes to accommodate the new domain, however. We will probably need to modify our ontologies and rule base to some extent in order to represent dyslipidemia-specific concepts, but there is likely to be sufficient overlap that such modifications will be minimal. We will also have to populate the new knowledge base schema with information relevant to dyslipidemia. That notwithstanding, it is probable that we will be able to reuse a significant portion of the assertions from the original diabetes knowledge base for this purpose.

So, how does this help overcome the limitations of data integration? Because each knowledge base built in this way shares

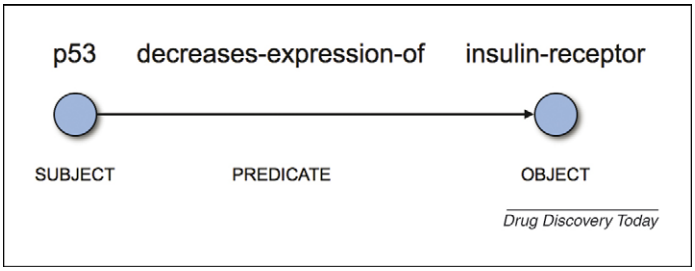

FIGURE 1

A graphical representation of an assertion, or triple. Circles represent the subject and predicate of the assertion, and an arrow from the subject to the object represents the predicate.

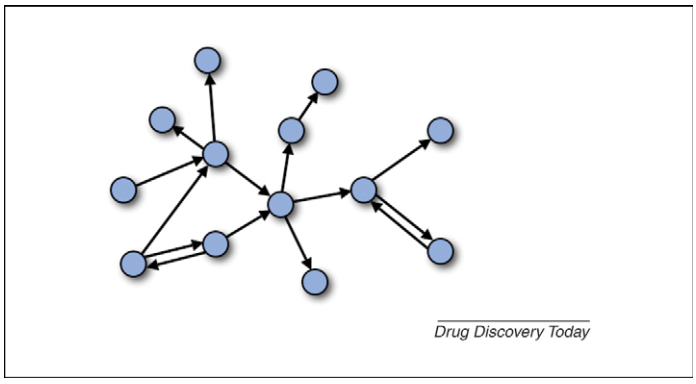

FIGURE 2

A small directed graph. The circles in the graph are 'nodes,' and the arrows are 'edges.'

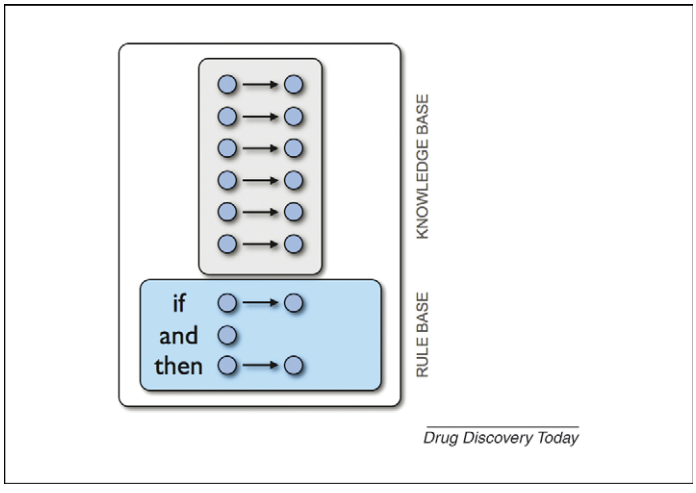

FIGURE 3

A knowledge system comprises a knowledge base and a reasoning engine. In the proposed architecture, the reasoning engine generates hypotheses by applying the appropriate rules from the rule base to the knowledge base graph.

Please cite this article in press as: Slater, T, Beyond data integration, Drug Discov Today (2008), doi:10.1016/j.drudis.2008.01.008

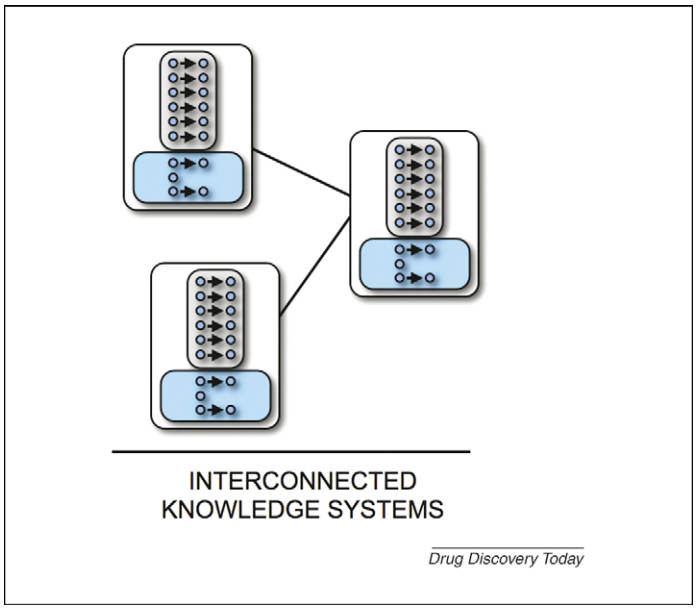

FIGURE 4

An interconnected set of knowledge systems.

exactly the same relational database schema, the same knowledge representation scheme (triples and ontologies), and the same algorithms for working with it, each knowledge base is a fungible resource. If not for the actual content, any knowledge base can be substituted for any other. If the knowledge bases are connected to each other over a network (Figure 4), such as the World Wide Web, then the same search tools, algorithms, rule bases, and visualization methods can be used across all of them. Rather than attempting to achieve data integration by 'getting all the data in one place,' the stated goal of so many doomed integration projects, success is achieved by virtue of the uniformity of the underlying information architecture and the software used to access and manipulate it. The goal is not to put all of our information in one place; the goal is interoperability between repositories. 'Boutique' knowledge bases, relevant to just a particular indication, molecule type (such as compound), or any other particular concepts, are not only accommodated but encouraged, because they keep database sizes small and therefore increase manageability, decrease computation demands, and they cannot become silos or 'data tombs' [12] because of their uniform architecture and explicit semantics (see Glossary).

Putting it all together

A practical system for implementing a physical symbol system can be built using current technologies. Such a system requires the following components:

- A knowledge representation scheme for representing facts or observations in a given domain
- A relational database schema for representing mathematical graphs
- A knowledge acquisition system for creating graph-based knowledge bases
- A set of algorithms for manipulating, searching, and visualizing graphs
- An expert system for reasoning over the knowledge base

Ideally, the knowledge representation scheme will be powerful and flexible enough to represent any fact in a given domain, while at the same time being simple enough to perform well during computation over a large body of such facts. In the complex domains of life sciences and drug discovery, the knowledge representation chosen must take the form of a common format for the integration and combination of data taken from diverse sources. Adoption of such a format is the only way to avoid the pitfalls of data integration attempted in the past. The Semantic Web (http://www.w3.org/2001/sw/) community, under the direction of Sir Tim Berners-Lee and the World-Wide Web Consortium (W3C, http://www.w3c.org/), has recommended the Resource Description Framework (RDF, http://www.w3.org/TR/rdf-syntax-grammar/) as the way to represent triples. The set of all such triples in a knowledge system comprises a graph which constitutes the system's knowledge base.

Other W3C standards, including RDF Schema (RDFS, http://www.w3.org/TR/rdf-schema/) and Web Ontology Language (OWL, http://www.w3.org/TR/owl-features/) augment RDF with sufficient expressive power to support inference (see Glossary) and other more sophisticated operations on the knowledge represented. The W3C supports the Healthcare and Life Sciences (HCLS, http://www.w3.org/2001/sw/hcls/) Interest Group, whose mission is to facilitate the development and adoption of Semantic Web technologies in those domains.

The relational database schema for representing graphs serves as a fungible resource for the creation of graph-based knowledge bases in any domain. Because of the simple and uniform way in which knowledge is represented, the database schema itself is uncomplicated, suitable for duplication and reuse for any number of knowledge bases in any domain without modification. Ideally, a special-purpose triple store (see Glossary) would be employed, but in our hands, currently available triple-store technology cannot adequately support reasoning over large graphs.

The knowledge acquisition system is necessary to flesh out the graph with assertions from the domain. The assertions can come from structured sources, such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), or they may come from unstructured sources, such as the scientific literature. Unfortunately, pre-assembled knowledge bases are very rarely available for purchase, so they must be created for the task at hand. This often requires the use of text mining and other technologies to make the process efficient in the face of huge amounts of information (see the recent review by Jensen *et al.* [13], and references therein for an excellent discussion on this topic).

A set of algorithms for manipulating, searching, and visualizing graphs is essential in order to make use of this information architecture. Fortunately, mathematical graphs have been a subject of research since Leonard Euler published the first paper in the history of graph theory (on the Seven Bridges of Königsberg) in 1736, so algorithms for searching and manipulating graphs are well studied. Such algorithms are necessary for reasoning over information represented as a graph. Visualization of large graphs using computers is still an unsolved research problem; however, good solutions for graphs of moderate size (3000 nodes or so) are readily available at reasonable cost.

Finally, an expert system (or similar system for reasoning) is required for hypothesis generation over the large space repre-

Please cite this article in press as: Slater, T, Beyond data integration, Drug Discov Today (2008), doi:10.1016/j.drudis.2008.01.008

sented in the knowledge bases. These typically take the form of a set of *if-then* rules representing knowledge from domain experts. For example, one such rule might look like, 'If protein-X increases the expression of gene-Y, and gene-Y's expression is measured to be significantly increased, then posit that the concentration or activity of protein-X is also increased.' Of course, such a rule is invalid as a logical argument; it is the 'fallacy of affirming the consequent.' However, as a heuristic for exploring a set of possibilities in the service of hypothesis generation, rules of this form turn out to be quite useful, especially when reasoning defeasibly such that conclusions and assertions can be retracted when better-supported ones are discovered.

In theory, serial application of such *if-then* rules, especially when starting at nodes representing molecular entities (such as genes) which have been empirically measured in the laboratory, can provide the basis for a graph traversal over related assertions. These traversals can represent chains of causal reasoning which ultimately provide explanations for experimental results, and in most cases are too complicated for human minds to make. We are aware of at least one company that has successfully used this methodology for the interpretation of experimental data sets, in particular molecular profiling data, from several diverse and complex disease areas across several different model systems [14]. We are also intrigued and encouraged by academic efforts in computational hypothesis evaluation [15] and hypothesis formation [16] in the context of biological systems.

Conclusion

What does knowledge engineering have to do with addressing the demoralizing attrition rates faced by drug discovery project teams? In our opinion, an R&D organization's lack of confidence in the efficacy or safety of a novel therapeutic agent in patient populations is a result of their inability to relate events and observations at the molecular level (i.e. results from *in vitro* assays against which initial drug leads are found and then optimized) to the desired endpoints at the preclinical and clinical stages of testing or vice versa (see the review by Butcher *et al.* [17] and references therein). Consider the challenges associated just with trying to nominate a 'safe' compound for clinical development. Classical *in vivo* toxicity studies require timelines and quantities of experimental compound that are fundamentally incompatible with the lead optimization process, and there are few prospective *in vitro* toxicology assays with cycle times short enough to run alongside a medicinal chemistry campaign. Molecular profiling technologies, such as DNA microarrays for toxicogenomics [18], are powerful in that they produce large-scale, systems-level molecular data in response to a toxicant, and one can even train accurate classification models using standard algorithms from these data [19]. However, because this technology still requires samples from whole animals, they cannot practically be used in lead optimization cycles, nor do multivariate signatures or classification models lend themselves well for informing compound design. Any fundamental breakthrough awaits interpretation and rationalization of the manifold molecular state changes occurring in a target organ into a sequence of comprehensible biological processes or pathways, ideally through a triggering event such as an off-target activity by the compound. Hence, there is a need for a framework to represent, populate, and assemble domain-specific knowledge [20], over which algorithms can reason and propose mechanisms of toxicity, perhaps resulting in more efficient discovery of qualified biomarkers and new *in vitro* screens. The ability to deduce causal relationships from these data might likewise be applied towards elucidating the mechanism of compounds showing desirable *in vivo* activity (e.g. improving insulin sensitivity or elevating HDL cholesterol levels).

Here we have described an idealized information architecture that obviates the recurring need to integrate and re-integrate data even as it provides the tools and techniques for powerful computational reasoning. While many of the standards and technologies we mention above are themselves being actively researched and developed (some with significant and unsolved technological problems), many are available now and appear in practical solutions. Pharmaceutical R&D is an information-based endeavor. Accordingly, we believe that significant competitive advantages will be enjoyed by those companies that are best able to represent, encode, and combine their precious internal knowledge with findings in the biomedical literature for the express purpose of automating the formation of relevant and experimentally testable hypotheses for drug discovery and development.

Acknowledgements

The authors wish to thank Dr Lee Harland and Dr Eric Neumann for their thorough and critical reviews of this manuscript.

References

1 Pisano, G.P. (2006) Can science be a business? Lessons from Biotech. *Harvard Business Review*, October issue. DOI:10.1225/R0610H
2 Milne, G.M. (2003) Pharmaceutical productivity – the imperative for new paradigms. *Annu. Rep. Med. Chem.* 38, 383–396
3 Roses, A.D. *et al.* (2005) Disease-specific target selection: a critical first step down the right road. *Drug Discov. Today* 10, 177–189
4 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715
5 Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4, 45–58
6 Loging, W. *et al.* (2007) High-throughput electronic biology: mining information for drug discovery. *Nat. Rev. Drug Discov.* 6, 220–230
7 2020 Science Group. (2005) *Towards 2020 Science*. (http://research.microsoft.com/towards2020science/downloads/T2020S_Report.pdf)
8 Giles, J. (2007) Key biology databases go wiki. *Nature* 445, 691
9 Neumann, E. and Prusak, L. (2007) Knowledge networks in the age of the Semantic Web. *Brief Bioinform.* 8, 141–149
10 Karp, and Peter, D. (2001) Pathway databases: a case study in computational symbolic theories. *Science* 293, 2040–2044
11 Newell, A. and Simon, H.A. (1975) Computer science as empirical inquiry: symbols and search. *Commun. ACM* 8, 113–126
12 Fayyad, U. and Uthurusamy, R. (2002) Evolving data mining into solutions for insights. *Commun. ACM* 45, 28–31
13 Jensen, L.J. *et al.* (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* 7, 119–129
14 Elliston, K.O. *et al.* (2005) Identification of the androgen-induced transcriptional program of human prostate cancer revealed by causal analysis. *J. Clin. Oncol.* 23, 4637

Please cite this article in press as: Slater, T, Beyond data integration, Drug Discov Today (2008), doi:10.1016/j.drudis.2008.01.008

15 Racunas, S.A. *et al.* (2004) HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics* 20 (Suppl. 1), i257–i264
16 Tran, N. *et al.* (2005) Knowledge-based framework for hypothesis formation in biochemical networks. *Bioinformatics* 21 (Suppl. 2), ii213–ii219
17 Butcher, E.C. *et al.* (2004) Systems biology in drug discovery. *Nat. Biotechnol.* 22, 1253–1259
18 Ulrich, R. and Friend, S.H. (2002) Toxicogenomics and drug discovery: will new technologies help us produce better drugs? *Nat. Rev. Drug Discov.* 1, 84–88
19 Yang, Y. *et al.* (2004) Toxicogenomics in drug discovery: from preclinical studies to clinical trials. *Chem. Biol. Interact.* 150, 71–85
20 Neumann, E. and Thomas, J. (2002) Knowledge assembly for the life sciences. *Drug Discov. Today* 7, s160–s162

Please cite this article in press as: Slater, T, Beyond data integration, Drug Discov Today (2008), doi:10.1016/j.drudis.2008.01.008