Knowledge assembly for the life sciences

▼ 'Knowledge assembly' technology renders knowledge derived from structured and unstructured sources into a machine-readable format. Using knowledge assembly, a computer-based environment is being developed in which scientists can perform logical analyses of a large and diverse set of pharmaceutically relevant knowledge. This environment will enable scientists to apply all available knowledge to decision-making processes in drug discovery, including target prioritization, assessment of compound liabilities, and clinical trial design.

Eric Neumann
Beyond Genomics
40 Bear Hill Road
Waltham
MA 02451, USA
tel: +1 781 890 1199
fax: +1 781 895 1119
eneumann@
beyondgenomics.com
Jeffrey Thomas
Genstruct
150 Cambridge Park Drive
10th Floor
Cambridge
MA 02140, USA
jthomas@genstruct.com

Introduction

Despite the technology revolution, methods for capturing, assembling and sharing knowledge in the life sciences are human-based and fundamentally unchanged from the time of Aristotle. Although information technologies have brought tremendous efficiencies in the sharing of life-sciences information, they are really just fast conduits. Synthesis, interpretation and application of knowledge occurs exclusively in the minds of the scientist. Owing to the exponential growth of life-sciences knowledge, there is an increasing need for computer-based systems to support the logical interpretation and association of life-sciences knowledge.

In pharmaceutical companies, knowledge tends to be codified by organizational group and scientific domain. However, informed decision-making requires the synthesis of knowledge over a wide variety of domains. For example, deciding whether to advance a compound into clinical development must consider knowledge of the efficacy of the compound in animal models, its chemical properties, its mechanism-of-action, and its potential toxicity or pharmacokinetic properties. Assembly of this knowledge is currently a manual activity and is unlikely to be comprehensive.

Current emphasis on data

There have been numerous debates about the challenges in life-science informatics. The two main issues usually presented are: (1) that data are generated at an unprecedented rate, requiring systems with massively large storage and throughput capacities; and (2) the heterogeneity and complexity of today's data warrant new approaches in the handling of multiple data-types that will support software inter-operability. Although these are important issues, this data-centric approach misses the point of where the real bottleneck is for most biotechnology and pharmaceutical companies: the consolidation and usage of diverse knowledge for maximum value realization.

> *'The real bottleneck is...the consolidation and usage of diverse knowledge for maximum value realization.'*

The representation and handling of complex data has been the center of many discussions in both industry and academia. Many 'bio-standards' groups exist (OMG, I3C, MGED, Biopathways, Bio-Ontologies, Open-Bio, CDISC), which aim to provide life-science and clinical informaticists with data standards that improve the development of inter-operable technologies. To date, implementations of such standards have proven elusive, even where specifications have been produced.

As far as usable data is concerned, there is still the challenge of distilling all the HTS data, microarray images or mass spectrometry files (which are indeed quite large); only the relevant items contained within them need be extracted, resulting in significant compression. This reduced set of data is then analyzed and mined for patterns. However, it is in the interpretation of the results in the context of their links to other information that their true significance and value is realized.

Ontologies

Heterogeneous data are often integrated to bring potentially related data into proximity through a common form that can be queried in various combinations. Several approaches exist, including schema merging (datawarehouse), federating databases (multiple databases that are linked via an interconnected query mediator), and common model indexing (a relational graph database with unique data identifiers for multiple databases). However, the meaning or semantics of why certain data elements are linked to each other is

1359-6446/02/$ – see front matter ©2002 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(02)02432-7

not always made explicit (e.g. the use of foreign keys in relational databases, or wrapper multiplexing). Assuming that this is what most scientists really want to understand and use when thinking about and searching for data, information systems need to build in such semantics, rather than selecting one point of view and obscuring other relationships.

This is where ontologies fit in brilliantly, because they enable domain experts (i.e. scientists) to specify to any degree of resolution, and how data, terminology (i.e. controlled vocabularies) and concepts relate to each other (http://www.smi.stanford.edu/projects/helix/psb98). Ontologies handle not only taxonomies, as is evidenced by the Gene Ontology database (http://www.geneontology.org/), but also complex graphs with nodes that represent different concepts, and edges that represent multiple relationships. They also introduce their suite of concept-relations, various attributes and restrictions, going well beyond what is possible using a purely object-orientated approach. Indeed, they have already been used to organize biochemical pathway information [1].

Finally, the use of ontologies will enable publishing, searching, and use of scientific text to a degree far beyond current applications [2]. Text-mining has become a very active area of interest in the life sciences because scientists need to find and access the knowledge contained within articles and link it to their own databases. Similarly, the goal of the semantic web (http://www.w3c.org/2001/sw/) is to enable semantic links to be embedded directly into the text of an article as part of the publishing process. For this to be possible, sets of ontologies will need to be defined for each of the different areas covered by literature.

Knowledge systems

Most software applications have knowledge encoded into the software, but have no understanding of the knowledge *per se*. Knowledge assembly technology draws on a long and rich history of research aimed at instructing computers to operate on knowledge much as humans do. Such research has in the past led to the development of tools like programming languages and data-mining algorithms. Knowledge Assembly techniques are used extensively in finance, for example, in mortgage approval or in fraud detection, as well as the defense sector (e.g. C4I: Consultation, Command, Control, Communications, and Intelligence).

As information is gathered from multiple sources to be interpreted and cataloged, it is assembled into semantic blocks or nuggets of knowledge that have significance and value for individuals or groups of researchers. Gene definitions can become functionally linked to inherited diseases, and their placement within a biochemical pathway can also be interwoven. We assert that this is a key process in acquiring and organizing applied knowledge, and hence have named it 'knowledge assembly'. It supports the central tenet of scientific inquiry, that is, hypothesis-driven research, and it is our belief that knowledge assembly can significantly advance informatics-based methodologies.

Just as The Human Genome Project was about the assembly of many isolated sequences into a single linear construct, knowledge assembly can too be viewed as the linking of many different but related facts and concepts rendered as a graph of ideas. The difference is mainly in the algorithms used to assemble the pieces, and the meaning behind the links (matches) generated.

Knowledge assembly in the life sciences

Any discussion of knowledge assembly must address the meaning of the word 'knowledge' and how knowledge differs from data. For our purposes, knowledge can be considered as the kind of information presented in the Results and Conclusions sections of a paper: conclusions supported by the data. Life-science knowledge is based upon data. Data are often presented in the results section as well, often in the form of figures, graphs and tables; the interpretation and relevance of such forms involves applying and recording knowledge. Knowledge, especially new knowledge, can be controversial, because interpretations might be accepted gradually by others through the scientific peer review process. Often what one scientist considers to be 'knowledge' is regarded by another as conjecture, or even as incorrect. Consequently, the formation and testing of hypotheses needs to be an integral part of knowledge assembly if it is to support the scientific method.

A knowledge assembly environment in the life sciences must therefore have the following key properties:

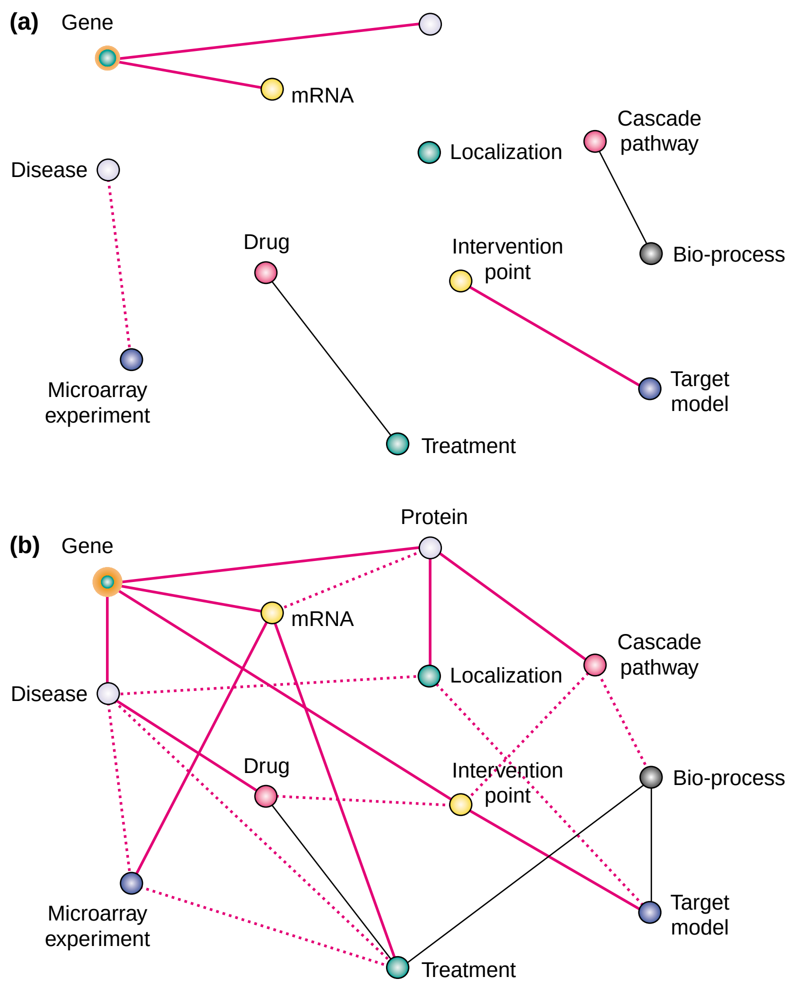
- individual scientists must be able to control the assembly of knowledge relevant to them;
- knowledge must carry its pedigree so that it can be evaluated based on its source;
- knowledge sources should include not only structured databases [e.g. Genbank, Chemical Abstract Service (http://www.cas.org)] but also unstructured sources such as literature and document management systems;
- assembled knowledge must be structured such that it can be reasoned upon.

Knowledge for life-science research can be assembled from several sources and in various combinations. Specifically these include:

- scientific literature;
- structured annotated databases;
- images and figures;
- experimental observations;
- HTS assay databases;
- ontologies and controlled vocabularies;
- local annotations by scientists;
- formal biological models.

Once the relevant information is extracted from these resources, it can then be enclosed within logical expressions using case frames [3–5]. Different data-mining algorithms and scripts are used to extract the insightful findings from various sources and insert them into the correct knowledge context.

Much of knowledge assembly can be achieved through a kind of type-validated graph merge [6], where attention is given to the types of nodes and edges that represent various data and conceptual items (Fig. 1). Nodes usually correspond to data objects or related concepts, and edges represent the relationships between specific objects. By



DDT: InfoBiotech supplement

Figure 1. Knowledge assembly of information from multiple distinct sources. (a) Comprehensive representation via a process similar to a graph merge. (b) The resulting representation makes the relationships between entities explicit and machine-readable, and provides a scaffold to support hypothesis generation and testing. The different coloured lines indicate relationships based on different evidence and strengths: bold lines represent directly determined; broken lines represent inferred relationships.

linking axioms or rules to each molecular object, various relationships can be proposed and tested, such as the hypothesis: 'genes with common roles tend to be transcriptionally regulated by containing similar regulatory elements'.

In addition to assembling knowledge from annotated and published sources, information from experimental data must also be considered for assembly. This is certainly one source for hypothesis generation and testing. Bioinformaticists typically process and analyze large amounts of information by applying specific algorithms, such as sequence analysis and microarray pattern analysis. The direct results of these analyses are mined primarily by the bioinformaticists, who develop their own supplementary scripts for extracting and filtering the relevant information from the data. However, very little has emerged in terms of intelligent 'support-ware' for gathering, annotating and uploading such enriched information in practical ways to facilitate subsequent scientific interpretations. The constant need for the development of task-specific 'perl scripts' highlights this deficiency.

Knowledge assembly includes a means for defining and managing various goal-specific scripts or 'pragmatics' through a formal language. These are then used in conjunction with algorithms that have been adapted with 'goal-wrappers' so that they can be appropriately invoked based on an immediate knowledge assembly task. This forms the basis of an agent-based approach to bioinformatics that combines tool inter-operability with knowledge capture. Much of this can take advantage of the Defense Advanced Research Projects Agency (DARPA) (http://www.darpa.mil) Agent Markup Language (DAML) (http://www.daml.org/), and its corresponding web-services specification (DAML-S).

Summary

As part of knowledge assembly, all these different technologies provide a convergence for current life-science knowledge with hypothesis-driven experimental analysis. It shifts the focus from bioinformatics analyses to interpretation and decision-making, using the expressiveness of logic systems to support the representation of hypotheses and interpretations, even if they are tentative. As hypotheses evolve into commonly accepted principles, their role as knowledge units becomes more valuable and therefore needs to be shared between researchers. Knowledge assembly supports the dynamic processes within scientific research, and transforms what has until now been a mainly manual categorization of facts and hypotheses into a scalable, machine-facilitated discovery process.

References

1 Karp, P.D. (2001) Pathway databases: a case study in computational symbolic theories. *Science* 293, 2040–2044

2 Mack, R. and Hehenberger, M. (2002) Text-based knowledge discovery: search and mining of life-science documents. *Drug Discov. Today* 7 (Suppl.), S89–S98

3 Minsky, M. (1974) A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306

4 Chandra, D.N. (1991) *Exploration and Innovation in Design*, Springer–Verlag

5 Russell, S.J. and Norvig, P. (1995) *Artificial Intelligence: Modern Approach*, Prentice Hall

6 Bollobas, B. (1998) *Modern Graph Theory*, Springer–Verlag