22

An Integrated Biomedical Knowledge Extraction and Analysis Platform

Using Federated Search and Document Clustering Technology

Donald P. Taylor

Summary

High content screening (HCS) requires time-consuming and often complex iterative information retrieval and assessment approaches to optimally conduct drug discovery programs and biomedical research. Pre- and post-HCS experimentation both require the retrieval of information from public as well as proprietary literature in addition to structured information assets such as compound libraries and projects databases. Unfortunately, this information is typically scattered across a plethora of proprietary bioinformatics tools and databases and public domain sources. Consequently, single search requests must be presented to each information repository, forcing the results to be manually integrated for a meaningful result set. Furthermore, these bioinformatics tools and data repositories are becoming increasingly complex to use; typically they fail to allow for more natural query interfaces. Vivisimo has developed an enterprise software platform to bridge disparate silos of information. The platform automatically categorizes search results into descriptive folders without the use of taxonomies to drive the categorization. A new approach to information retrieval for HCS experimentation is proposed.

Key Words: Data mining; document clustering; federated search; information retrieval; metasearch; ontology; structured and unstructured data; taxonomy.

1. Introduction

1.1. Information Explosion

The genomics information explosion that started in the 1990s has evoked various methods to store, annotate, retrieve, and analyze the bourgeoning data aftermath. Information assets that have been created include gene expression data from microarray experiments, genomic sequences, proteomic identification, and data from high throughput SNP arrays ***(1)***. In addition there has also been a growth in high-throughput and high content screening (HCS) data. These data are typically highly structured but lack cross-informatics platform capabilities.

The continued growth of biological research articles, fueling the continued information expansion and the emerging biotools that created them, lie juxtaposed with the aforementioned structured information entities. For example, Entrez PubMed, developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine, is a web-based citation repository and search tool spanning more than 12 million citations across 4800 biomedical journals published in over 70 countries ***(2)***. Staying current with this expanding research repository,

From: *Methods in Molecular Biology, vol. 356:*
High Content Screening: A Powerful Approach to Systems Cell Biology and Drug Discovery
Edited by: D. L. Taylor, J. R. Haskins, and K. Giuliano © Humana Press, Inc., Totowa, NJ

in addition to the various commercial informatics tools to retrieve and analyze structured and unstructured data, becomes ever more challenging.

NCBI has extended Entrez's reach beyond journal citations to help store and retrieve the more structured information assets. These tools include: PubChem, GenBank, Entrez Protein, and Online Mendelian Inheritance in Man. In addition to these publicly available information resources, growing repositories of gene, protein, cytotoxicity, compounds, and other biological information are growing within the data-warehousing confines of the highly competitive pharmaceutical and biotechnology companies. A growing challenge has emerged for industry to blend proprietary internal data assets with the growing public assets, and to make swift analysis of these data to optimize the relatively new commitment to HCS experimentation.

Beyond NCBI, a growing legion of private corporations is developing information storage and information mining solutions to help swiftly gain insights from these growing repositories. New tools now exist to:

- Input genome sequences to identify relevant patents (Gene-IT; www.gene-it.com).
- Identify novel protein–protein interactions through canonical pathways systems (GeneGo, Ingenuity; www.genego.com).
- Rapidly create cloning vectors (Vector NTI from Invitrogen; www.invitrogen.com).

As the situation stands, not only the information, but the islands of bioinformatics tools to analyze the information remain scattered throughout the enterprise. Now the questions of where to go, how to search, and which tools are best suited to analyze the specific data continue to be difficult to answer. This chapter will explore beyond the methods by which traditional high throughput and high content data have been stored and analyzed; it will propose an information architecture to blend the structured with the unstructured knowledge that have fueled the information explosion and knowledge discovery. This new information infrastructure brokers output from various bioinformatics tools—public and proprietary—as inputs to other tools. This new platform will help promote information integration across primary functional groups in industry—from preclinical to marketing and competitive intelligence—as well as academic research enterprises, by means of a unified, single-point research architecture that transcends complex query syntax.

1.2. Current Information Retrieval Challenges

The traditional information mining and retrieval challenges have been:

- Disparate information silos (information residing in multiple physical locations.
- Information overload (too much information being returned to the user).
- Information overlook (missing critical information that has been clouded by information overload).

Researchers in academia and industry may have a wealth of information sources and bioinformatics tools at their disposal. These information sources include public, private/licensed, and internal proprietary content residing in multiple data repositories. For example, a researcher might be interested in all published and proprietary documents related to the interaction between the proteins p53 and mdm2. Relevant information research sources may include PubMed, Science Direct™, Genbank, and an internal Oracle database that stores drug discovery experimental results. The traditional method to search for information such as the pathway components p53 and mdm2 is to access each source in series. The researcher must then know where to go, how to search once they get there, and how to synthesize the results into a manageable result set for further review and study.

1.3. Current Functional Workflow—From High Content Results to New Discoveries and Insights From the Literature

HCS was originally implemented as a smaller vertical within secondary screening in drug discovery and has begun to penetrate primary screening. Given early HCS successes, HCS has

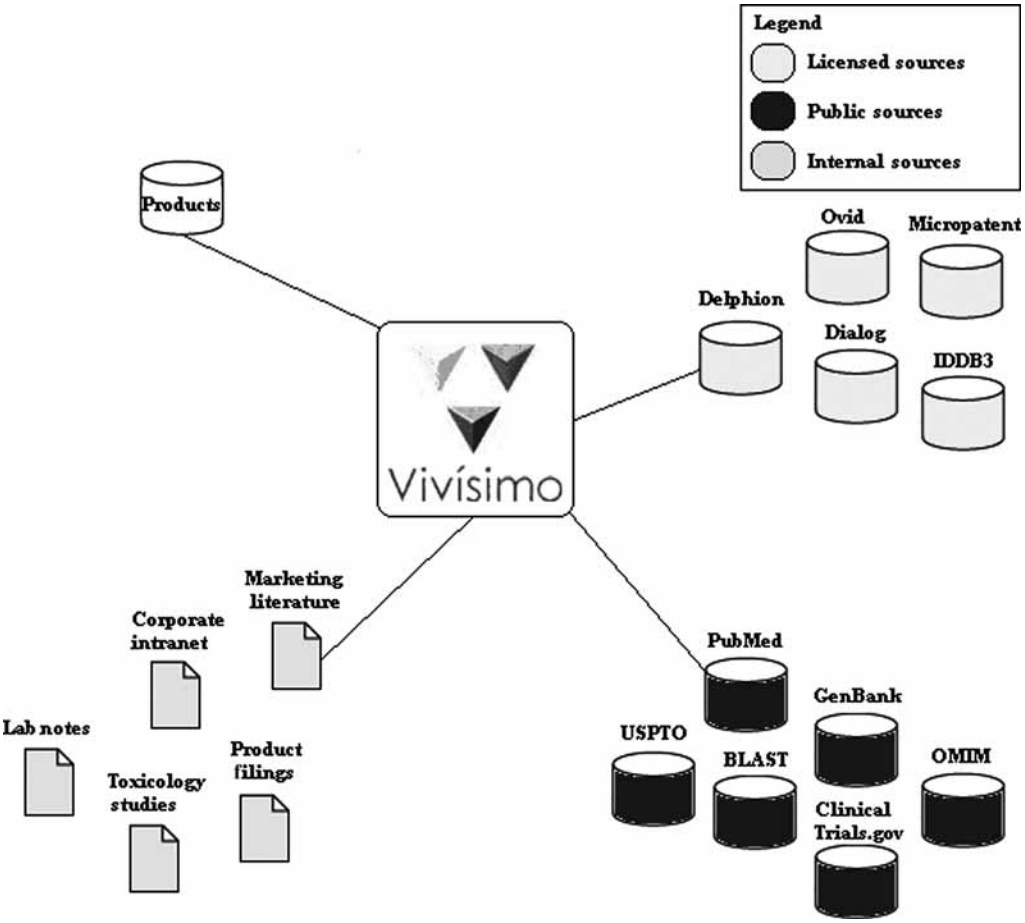

Fig. 1. Biomedical information landscape.

expanded its reach upstream and downstream of the second and primary screening continuum. HCS is now applied to target identification/validation, to lead optimization, and to toxicology *(3)*. Consequently, HCS assay design, the information collected, and the informatics tools to help derive knowledge have a greatly expanded scope. Additionally, interpreted results from an HCS experiment when combined with additional information from the literature and other information sources can lead to the next level of HCS investigations. The domain of scientific literature serves as a potential validator for HCS experimental design and results alike *(3)*. Imagine the bottleneck created when critical pre- or postexperimental knowledge is buried within over 12 million citations hosted by the PubMed search engine—in addition to the internal data stores scattered throughout the enterprise.

2. A New Approach—Integrated Research Portal Bridging Both Data and Information Tools

Vivisimo has developed a webserver-based software application (Velocity for Life Sciences™ [VVLS]) to bridge the information retrieval and information brokering challenges (*see* **Fig. 1**). With this application, researchers are able to simultaneously search across multiple information sources by means of a single query. The Vivisimo application then organizes—spontaneously—search results into descriptive, hierarchical folders allowing for the consolidation of similar documents into logical groupings. Moreover, the Vivisimo software intelligently routes queries

beyond the information silos into the myriad of bioinformatics tools available to the organization. Consequently, researchers can leverage the information assets as well as bioinformatics tools all within a simple, unified user interface.

2.1. Content Integration

Vivisimo technology provides a single search interface to access and retrieve documents from any number of internal or external sources. This capability is called metasearch, or federated search; the application accesses host search engines remotely, and the content is maintained by each host source. In this way, federated search enables users to access the most up-to-date documents across multiple information domains simultaneously. Examples of consumer-based, public metasearch engines include Clusty.com and Ixquick.com. Clusty.com federates search results from host search engines such as Yahoo, and Ask Jeeves. Clusty routes each search query to multiple search engines; each host search engine then returns results from their current document index.

The Vivisimo technology includes a federated search component. Unlike public metasearch engines such as Ixquick.com, the VVLS technology may be configured to access any number of electronic sources internal and external to the organization (*see* **Table 1** for sample external and internal sources).

The VVLS technology will allow the end user to choose any number of sources to federate per query. VVLS will access each source's search interface and will return a preconfigured number of results per source. VVLS also includes advanced duplication-prevention algorithms. VVLS presents the search results back to the user in a ranked result set, allowing the user to hyperlink directly to documents of interest. Although federated technology has overcome the challenge of disparate information repositories, hundreds (if not thousands) of search results may be returned per query. Although the host search engines' rankings have been preserved, end users may wish to delve more deeply into the search results. Thus, some mechanism must be implemented to organize the search results for greater knowledge discovery.

2.2. Document Clustering

VVLS includes a document clustering module that dynamically transforms search results into "crisp, hierarchical folders" ***(4)***. This clustering is performed on the fly without the use of taxonomy. The VVLS software is preconfigured to use specific outputs from each of the sources as inputs for the Clustering Engine. For example, one clustering method may be to use the title along with the abstract (or snippet) from each document as inputs to the Clustering Engine. The Clustering Engine will then group the documents based on their similarities; from the grouping, VVLS generates meaningful folder headings. The Clustering Engine may be configured to incorporate metadata as inputs. Metadata may also be used as clustering inputs. For example, documents from sources that share metadata (including author and date of publication) may be organized accordingly.

3. Methods—An Example Application

3.1. Intelligent Query Routing

The Vivisimo technology is capable not only of remotely administrating host search engines from information sources, but is also capable of routing queries to bioinformatics and other software tools. For example, consider the following hypothetical scenario wherein a researcher is broadly interested in the interaction between proteins p53 and mdm2 (*see* **Fig. 2**). In this example application, the query will be presented to the Vivisimo technology and will then be brokered to public information sources, iPath (systems biology tools created by GeneGo and hosted by Invitrogen), and Invitrogen products and services relevant to the query. This is only one example on how the Vivisimo tools can be applied to multiple databases and biotools at the same time.

Table 1
Sample Internal and External Information Sources

Government scientific sources
- Biomedical literature (NLM)
- Multidimensional drug screening results (NCI-DTP)
- Genomics databases (NIH-NCBI)
- Protein data bank (NSF-NIH)
- Clinical trials (NLM)
- Regulatory issues (FDA)
- Scientific project information (NIH-CRISP)

Commercial scientific sources
- Genomics (e.g., Gene cards, XenneX)
  - Gene function
  - Role in disease
- Proteomics (e.g., Proteome bioknowledge library, Incyte)
  - Organismal distribution
  - Protein structure
  - Enzyme activity
- Chemiformatics (e.g., iResearch library, ChemNavigator)
  - Compound structure–function
  - Biological activity
  - Business intelligence
  - Biomedical literature

Competitive sources
- News and press releases
- Investor information
- Patent application information

Internal sources
- Cheminformatics
  - Compound structure–activity relationship data
  - Biological activity
- High-throughput and HCS results
  - Compound structure–activity relationship data
  - Target validation
  - Dose–responses
  - Phenotype effects
- Toxicology studies
  - Drug–drug interactions
  - Metabolism
  - Animal toxicity
- Clinical development
  - Patient responses
  - Cumulative reports

Internal source continued
- Regulatory correspondence
  - Product filings
  - FDA responses
  - Warning letters
- Manufacturing
  - Formulations
  - Production specifications
  - Packaging requirements
- Institutional intranet
  - Reports
  - Memos
  - Presentations
  - Marketing materials

1. The researcher begins by presenting the query, "p53 and mdm2" to the Vivisimo software. The demonstration URL is: http://vivisimo.com/metademoHCS. The software will route the query to the aforementioned sources and return information results sets in the form of citations, products, and pathways. Initially, the researcher may be interested in the overall landscape of citations that are related to p53 and mdm2 to understand the scope. The left hand pane will dynamically cluster the returned citations to accommodate this exploration. The researcher may then discover a clustered folder titled, "serine, phosphorylation p53."
2. Each folder may be expended into subfolders by clicking on the "+" sign captioning each cluster. So for example, out of 500 returned results for "p53 and mdm2," the cluster titled, "serine, phosphorylation p53" contains 34 related citations. Further exploration within this folder illuminates the process by which p53 is phosphorylated causing the disassociation from mdm2 and the subsequent transcription of p21. This knowledge is easily derived from scanning the abstracts wherein p53 and mdm2 are indicated in bold font.

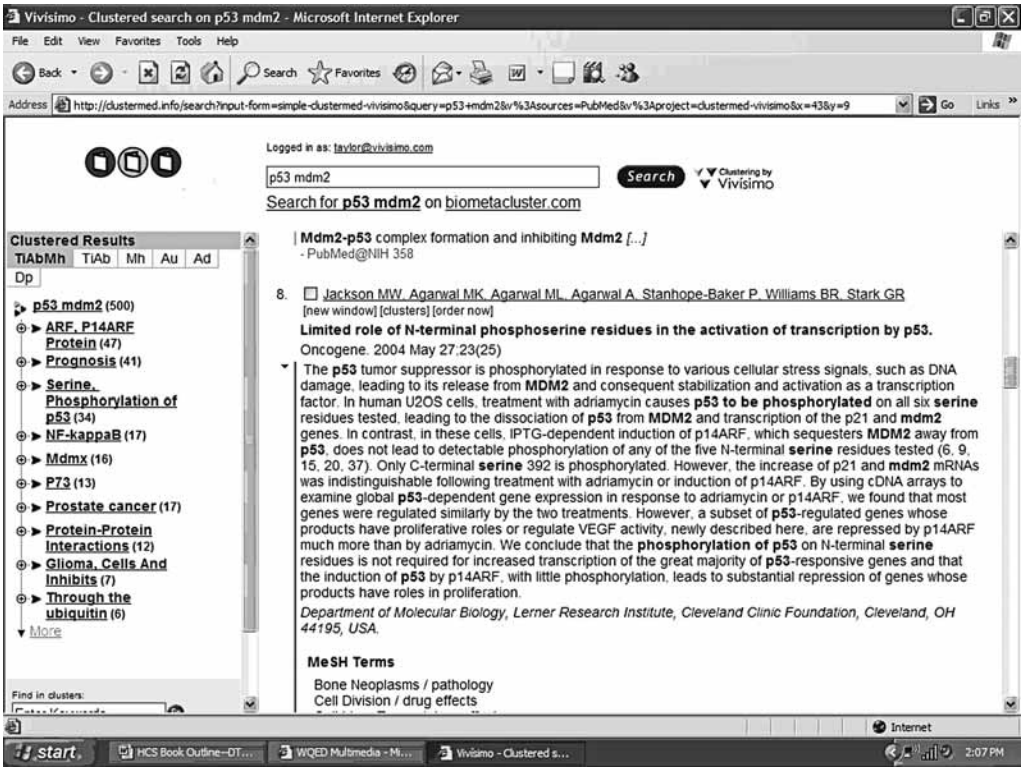

Fig. 2. Velocity for Life Sciences™ screen shot.

3. Now let us assume that p21 is not well understood by the researcher. The Vivisimo software can be requeried to append p21 to the original query (query expansion), or a new query with "p21" as the sole search string may be presented. The latter is a more traditional approach to further exploring "p21" as a topic of interest. However, let us assume that the researcher may wish to implement a systems biology search using p21 as the search string in order to identify the pathways in which p21 is implicated (as well as the relevant literature citations). Without the Vivisimo technology, the researcher must know where and how to access their systems biology platform (e.g., GeneGo, Ingenuity Pathways, and so on).
4. With the Vivisimo technology the researcher simply enters p21 and performs another search. Now the query is routed directly, in this example, to iPath by Invitrogen that is built on the GeneGo platform. Selecting the iPath tab at the top of the Vivisimo search screen will take the research directly into the list of Genes and associated pathways pertaining to p21 (*see* **Fig. 3**). Now the researcher is able to link directly to the following pathway maps:
   - ATM/ATR regulation of G1/S checkpoint.
   - IFN-γ signaling pathway.
   - Role of AP-1 in regulation of cellular metabolism.
   - Role of HDAC and calcium/calmodulin-dependent kinase in control of skeletal myogenesis.
   - ATM/ATR regulation of G2/M checkpoint.
   - Putative integrins pathway. Part 1.
   - Putative integrins pathway. Part 2.
   - TPO signaling through JAK-STAT pathway.
   - DNA damage response in inhibition of CDK during G2.
   - TGF-β receptor signaling.
   - Cell cycle regulation by Brca1.
   - Angiopoietin—Tie2 signaling.
   - AKT signaling.
   - PTEN pathway.

CDKN1A	Homo sapiens cyclin-dependent kinase inhibitor 1A (**p21**, Cip1)	• ATM/ATR regulation of G1/S checkpoint • IFN gamma signaling pathway • Role of AP-1 in regulation of cellular m... • Role of HDAC and calcium/calmodulin-depe... • ATM/ATR regulation of G2/M checkpoint • Putative integrins pathway, Part 1 • Putative integrins pathway, Part 2 • TPO signaling via JAK-STAT pathway • DNA damage response in inhibition of CDK... • TGF-beta receptor signaling • Cell cycle regulation by Brca1 • Angiopoietin - Tie2 signaling • AKT signaling • PTEN pathway
CDKN1B	Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1)	• Cdc25 regulation of cell cycle • TPO signaling via JAK-STAT pathway • Regulatory cascade of cyclin gene expres... • Cell Cycle Regulation by Brca1 • AKT signaling
CPS1	Homo sapiens carbamoyl-phosphate synthetase 1, mitochondrial	• Aspartate and asparagine metabolism • Arginine metabolism • Urea cycle • UMP biosynthesis
CYP21A2	Homo sapiens cytochrome P450, family 21, subfamily A, polypeptide 2	• Biosynthesis and metabolism of pregnenol... • Biosynthesis and metabolism of cortisone • Aldosterone biosynthesis and metabolism
CYP3A4	Homo sapiens cytochrome P450, family 3, subfamily A, polypeptide 4	• Androgens and estrogen metabolism, part ... • Androgens and estrogen metabolism, part ... • Role of VDR in regulation of genes invol...
DOK1	Homo sapiens docking protein 1, 62 kDa (downstream of tyrosine kinase 1)	• H-RAS regulation pathway • Signaling pathways for GDNF • CXCR4 signaling pathway

Fig. 3. p21-related pathways.

5. Further exploration within the Vivisimo results expose the following information and related hyperlinks:
   - Cyclin-dependent kinase inhibitor 1A (CDKN1A) is the common gene name for p21.
   - p21 is one of seven aliases for CDKN1A.
   - CDKN1A summary.
   - Nucleotide accession numbers.
   - Transcript map.
   - Fourteen links to pathway maps including: ATM/ATR regulation of G1/S checkpoint and cell cycle regulation by Brca1.

   The Vivisimo user interface displays this information, thus enabling the end user to browse the information in greater, consolidated detail.
6. The user may click on the hyperlink to expose the graphical pathway of the AMT/ATR Regulation of G1/S Checkpoint. Or, the user could link directly to Entrez Gene's entry for CDKN1A.
7. Suppose at this stage the researcher is interested in patents that correspond to the two spliced variants of CDKN1A. The researcher would select the two variants and select the check box for a Gene sequence search technology called GenomeQuest by Gene-IT. The Vivisimo technology will automatically link to Entrez Nucleotide and pull both nucleotide sequences. The sequences will be submitted

via secure HTTP to the GenomeQuest software tool. Patent titles and abstracts are returned that match the sequences to the Vivisimo user interface. Further, those titles and abstracts are fed through the Vivisimo Clustering Engine and organized into folders.

4. Conclusion

This one example demonstrates the power and flexibility of implementing federated search coupled with document clustering. The full capability of this approach can be realized by setting up these tools to search a wide range of databases, both public and private. There are now solutions available to optimize information management and mining without the burden of information overload.

References

1. http://kd.cs.uni-magdeburg.de/ws03.html. Last accessed March 2, 2006.
2. http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html. Last accessed March 2, 2006.
3. Giuliano, K. A., Haskins, J. R., and Taylor, D. L. (2003) Advances in high content screening for drug discovery. *ASSAY Drug Dev. Technol.* **1,** 565–575.
4. http://vivisimo.com. Last accessed March 2, 2006.