South African Journal of Information Management
[DoE accredited]

Student Work Vol.3(2) September 2001

Data mining and search techniques in the biotechnology and Web environment: a comparison

B.W. Koester
Postgraduate Diploma in Information Management, Rand Afrikaans University
bwk@usb.sun.ac.za

Contents

1. Introduction

The biotechnology industry is a rapidly growing industry and rather alien to most people that is not versed in biology and chemistry. It is however a very important industry and according to Webster (2001) will succeed the information age. This in itself is a very serious statement and may, in fact, overturn many of society's current outlooks on life and the way business is conducted. Forward-looking economists are determined that the technology and information era is destined to be replaced by the bio-era and that all sectors of the world's economy will become biotechnology driven which could spark a revolution in the global economy. Bio-data mining will become very important due to the enormous amounts of data and information that is generated in this industry. It could be worthwhile to investigate data mining techniques that has been developed for other industries and apply this in the biotechnology industry or vice versa. This will force researchers to think very creatively on how to tackle this challenge. De Bono (1987) says that many new ideas come about when new information gathered by observation or experiment forces a reappraisal of the old ideas. De Bono adds that new information is probably the surest road to new ideas, but it is still unreliable, for mostly the new information is explained by an old theory and fashioned to support that theory.

Although new information can lead to new ideas, these can also come about without any new information at all. It is perfectly possible to look at all the old information and come up with a very worthwhile new way of putting it together. An example is Albert Einstein. He did no experiments, gathered no new information, before he created the theory of relativity.

Since he did no experiments he contributed nothing except a new way of looking at information that had been available to everyone else. The experiments confirming the theory came afterwards. What Einstein did was to look at all the existing information which everyone else was content to fit into the Newtonian structure, and to put it together in a completely new way. It is frightening (or exciting) to contemplate how many new ideas are lying dormant in already collected information that is now put together in one way and could be rearranged in a better way (De Bono 1987). This implies that one should not try and re-invent the wheel but try and look for solutions in other areas, to cross-pollinate or to approach these issues differently. Some of the biggest challenges today on the Internet and in the biotechnology industry are the large quantities of information to be searched through, should one seek some important information. This is supported by Philopkoski (2000) when he says that the sheer quantity of the biological and chemical information that needs to be generated to work towards drug development is so huge that it could not fit on any computer available today, not even IBM's vaunted Blue Gene. Similarly, Brin and Page (2000) say that the Web creates new challenges for information retrieval. The amount of information on the Web is growing rapidly, as well as the number of new users inexperienced in the art of Web research. Accordingly, by the year 2000, a comprehensive index of the Web would have contained over a billion documents. Therefore, Brin and Page (2000) also say that to create a search engine which scales to today's Web presents many challenges. Fast crawling technology is needed to gather the Web documents and keep them up to date. The indexing system must process hundreds of gigabytes of data efficiently and queries must be handled quickly, at a rate of hundreds to thousands per second. There is also the problem of searching for quality information with the current search engine technology. Getting back to the large amount of biological data that exist in many separate databases and Web sites throughout the world, it was recommended in 1993 that integrating databases together was vital to the success of the human genome project (Shoop *et al*. 2000). Since that time, the data in separate biological databases have increased dramatically, meaning that something urgently needs to be done about this information glut.

top

2. Web searching

Google has been used in this research as an example of a Web search engine due to its satisfying search results. According to Brin and Page (2000), search engines index tens to hundreds of millions of Web pages that involve pages of a comparable number of distinct terms. These search engines answer tens of millions of queries every day. Google is a scalable search engine and its primary goal is to provide high quality search results over a rapidly growing Web. This search engine employs techniques to improve the search quality that include page rank, anchor text and proximity information (Brin and Page 2000). In addition to the above, it is also a complete architecture for gathering Web pages, indexing them, and performing search queries over them. Some new developments in Google include query caching, smart disk allocation and subindices. One promising area of research is using proxy caches to build search databases, since they are demand driven. Other features that are being explored are relevance feedback and clustering (Brin and Page 2000). This illustrates that a Web search engine is a very rich environment for research ideas, which can be applied elsewhere, such as in the biotechnology industry.

Another area of investigation is knowledge management. Greening (2000) says that these systems seek to identify and leverage patterns in natural language documents. A more specific term is 'text analysis', since the vast majority of documents operate on text. The first step is associating words and context with high-level concepts. This can be done in a directed way by training a system with documents that have been tagged by a human with the relevant concepts. The system then builds a pattern matcher that decides how strongly the document relates to the concept (Greening 2000).

Clustering, or sometimes called segmentation, identifies people who share common characteristics and averages those characteristics to form a 'characteristic vector' or 'centroid'. Clustering systems usually let one specify how many clusters to identify within a group of profiles, and then try to find the set of clusters that best represents the most profiles. This technique is used directly by some vendors to provide reports on general characteristics of different visitor groups. These techniques require training and suffer from drift on Web sites with dynamic Web pages (Greening 2000).

Another method is estimation and prediction. Estimation guesses an unknown value, such as income, when one knows other things about a person. Prediction guesses a future value, such as the probability of buying a car next year, when a person has not done it yet, or the expected number of stocks that a person will trade in the coming year. The same algorithms can perform estimation and prediction. This method is related to data mining which is also applied in bioinformatics.

Then there are also decision trees. A decision tree is essentially a flow chart of questions or data points that ultimately leads to a decision. These tree systems are incorporated in product-selection systems offered by many vendors. They are used in situations where a visitor comes to a Web site with a particular need. But once the decision has been made, the answers to the questions contribute little to targeting or personalizing for that visitor in the future. Before one could do any comparisons or make any suggestions, one should first have a closer look at some aspects in the biotechnology industry.

top

3. Biotechnology

The biotechnology industry focuses to a large extent on health and agricultural aspects but it is also spreading to other disciplines such as manufacturing. It may be worthwhile to have a closer look at certain technologies and combinations of technologies in the biotechnology industry before discussing data mining. The term 'biochips' cropped up many times during the literature scan. These so-called chips are used in the analysis of genes and these devices may soon facilitate screening of potential drugs. Persidis (1999) adds that, clinically, the immediate goal is to enable biochips to serve point-of-care diagnosis. Biochips represent a marriage of modern medical knowledge with an idea familiar to computer users, that of the chip, a miniaturized processor. Like a computer chip, a biochip is ultraminiaturized and performs highly complex tasks. Often, it has been manufactured by methods developed for computer-chip manufacture - notably, photolithography, which serves to etch intricate, minute patterns of channels and islands on a solid surface (Persides 1999). Beyond that, the analogies tend to break down. A computer chip executes logical operations on strings of electronic zeros and ones. A biochip performs biochemical reactions. A computer chip is silicon-based while a biochip may be constructed on a glass slide or even within a porous gel. A computer chip can serve a vast range of purposes, depending on the instructions imposed from outside. Current biochips are each designed expressly to perform a specific function, such as the identification of mutations in a given gene or the detection of a bacterial antigen in a water sample. A biochip constructed to fulfil any one such purpose cannot be programmed for any other. Since the potential applications are vast, both for research and for clinical use, the potential markets for biochips will be huge, a powerful driving force for their continued development, which will lead to the creation of even more data. Another fascinating development in the biotechnology industry is the linking of two rather disparate fields, namely bioinformatics and geographical information systems (GIS). These two disciplines have much in common, most notably digital maps, large databases and research involving visualization, pattern recognition and analysis. NewsRx.com reports that in general, researchers use GIS techniques and tools to find and track large patterns, for example, geographic distribution of cancer and other diseases in human, animal, and plant populations. Researchers in bioinformatics generally look at very small patterns, such as those in DNA structure that might predispose an organism to developing cancer. The potential payoff in related fields such as those looking at climate change, emerging and resurgent infectious diseases and environmental health is enormous. Bioinformatics has focused on modelling from the level of the molecules up to the whole organism, while GIS has created tools to model from the level of the ecosystem down. Both fields rely heavily on mining, managing, accessing and analysing large amounts of data, and the disciplines share many of the same challenges for data management and representation. Integrating bioinformatics with GIS will be phenomenally useful in predicting public health outcomes. There may be lessons in these developments that could be applied elsewhere, such as the Web environment.

top

4. Data manipulation in biotechnology

Eckman *et al.* (2001) say that to identify and characterize regions of functional interest in genomic sequence require full, flexible query access to an integrated, up-to-date view of all related information, irrespective of where it is stored. This can be within an organization or across the Internet and its format (traditional database, flat file, Web site and results of runtime analysis). To add to the difficulty, many of the most interesting data sources are not easily queried, nor are their results easily parsed, for example, annotations in flat-file sequence databases such as GenBank and SwissProt, Web sites such as GeneCards and alignments resulting from BLAST searches. Many data sources do not represent biological objects optimally for the kinds of queries that investigators typically want to pose. Wide-ranging multi-source queries often return unmanageably large result sets, requiring non-traditional approaches to exclude extraneous data. One system, the so-called Target Informatics Net (TINet), is a readily extensible data integration system developed at GlaxoSmith-Kline (GSK). This is based on the Object Protocol Model (OPM) multi-database middle ware system of Gene Logic Inc. Data sources currently integrated include the Mouse Genome Database (MGD) and Gene Expression Database (GXD), GenBank, SwissProt, PubMed, GeneCards, the results of runtime BLAST and PROSITE searches and GSK propriety relational databases. Special purpose class methods are used to filter and augment query results including regular expression pattern matching over BLAST HSP alignments and retrieving partial sequences derived from primary structure annotations. All data sources and methods are accessible through an SQL-like query language or a GUI, so that when new investigations arise no additional programming beyond query specification is required. The power and flexibility of this approach are illustrated in such integrated queries as:

- 'find homologs in genomic sequence to all novel genes cloned and reported in the scientific literature within the past three months that are linked to the MeSH term 'neoplasms';
- 'using a neuropeptide precursor query sequence, return only HSPs where the target genomic sequences conserve the G[KR][KR] motif at the appropriate points in the HSP alignment'; and
- 'of the human genomic sequences annotated with exon boundaries in GenBank, return only those with valid putative donor/acceptor sites and start/stop codons' (Eckman *et al*. 2001).

These queries are advanced and one could imagine the possibilities should one be able to filter in this manner on the World-Wide Web.

A number of approaches exist to integrate systems as described by Shoop *et al*. (2000). One example is the so-called linked, **indexed data approach.** While not truly integrated systems, the advantage of these linked solutions is that they provide a single entry point of access for users and they are relatively easy to implement. In practice, these systems have been extremely useful and popular with the user community, thus demonstrating the need for integrated data sources. There are a number of disadvantages of these systems such as:

- The indices and WWW links require maintenance and are prone to errors.
- Browsing or keyword searches on indices are the only available data analysis methods for users (no *ad hoc* queries).
- They require a large amount of manual work by the user to integrate the data for further analysis.

Another approach is **loose integration with views on each data source**. This approach organizes heterogeneous databases into a multi database system without a common schema, but with a common query mechanism. The advantages of these systems are as follows.

- It provides users with a single query interface.
- It transparently and automatically accesses the underlying heterogeneous data sources.
- The data sources are updated by each of their curators and the updates are obtained when queries are executed.
- It provides a single integrated result to users.

The disadvantage of these systems is that response times for executing distributed queries across the Internet are slow and this does not permit interactive use.

Another approach is **the tight integration with views**. These systems share the advantage of loose integration systems. In addition, users have a single integrated schema representation of the data sources. The disadvantages of this type of system are the effort needed to create and maintain the global schema and the response time to execute queries.

Loose integration with materialized data is another approach. This approach organizes heterogeneous databases into a data warehouse without a common schema and requires all data to be periodically loaded into a central location. There are some disadvantages such as the following:

- The only way to connect the heterogeneous sources is by exact matches on identifiers shared by each data source.
- There must be a mechanism to keep the materialized data up to date.

Lastly, there is the **data warehouse: tight integration with materialized data.** A data warehouse is constructed with a common schema and data are periodically loaded into a central location. Some disadvantages of this type of tight integration include the following:

- A single global schema is used to combine data based on semantics, thus providing the richest integration possible.
- A single interface is provided.
- Data access to a single materialized database is faster than to several distant individual databases.

The disadvantages of this type of system are the following.

- There is extra work involved with creating and maintaining a global schema.

- The materialized data must be kept up to date.

Other benefits of fully integrating multiple databases are the following:

- They provide a means of detecting anomalies and omissions in each of the underlying constituent databases.
- They provide a more complete set of information than can be found in each individual database.

Apart from integration technologies there is also data mining. Thearling (2001) defines data mining as the automated extraction of hidden predictive information from databases. This allows users to analyse large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning twists thrown in. Like statistics, data mining is not a business solution, it is just a technology. Customer relations management (CRM), on the other hand, involves turning information in a database into business decisions. For example, consider a catalogue retailer who needs to decide who to send a new catalogue to. The information incorporated into the customer relationship management process is the historical database of previous mailings and the features associated with the (potential) customers. These can include age, zip code, their response in the past, etc. The software would use this information to build a model of customer behaviour that could be used to predict which customers would be likely to respond to the new catalogue. By using this information, a marketing manager can target only the customers who are most likely to respond. To put it differently, data mining extracts information from a database that the user did not know existed. Relationships between variables and customer behaviours that are non-intuitive are the jewels that data mining hopes to figure out.

A data mining method that uses indexing terms ('keywords') from the published literature linked to specific genes within a cluster or group of interest was developed in the biotechnology industry. The method takes advantage of the hierarchical nature of medical subject headings used to index citations in the MEDLINE database, and the registry numbers applied to enzymes (Masys *et al.* 2001).

Masys *et al.* (2001) describe another type of data mining to categorize the characteristics of known genes within a group of interest. The information content of published literature is linked to these genes. Commercially available microarray interpretation software generally allows searching for results associated with specific genes by words included in a gene definition or description field, and may also provide HTML hyperlinks to specific citations in MEDLINE. A Web-based question answering utility for gene expression that exploits data linkages is contained in GenCard and PubMed database retrievals. Other methods are developed to link keyword data mining to a variety of statistical clustering approaches and to automatically update these linkages as new articles relevant to specific genes are published. It is evident that using the controlled terminology keywords of the published literature associated with groups of genes, and the organization of those keywords in biological concept hierarchies, is a useful 'cluster mining' approach that complements purely mathematical approaches to gene microarray data analysis.

top

5. Discussion

The measurement of the simultaneous expression values of thousands of genes creates a large amount of data whose interpretation by inspection may be likened to 'attempting to drink from a fire hose' (Masys *et al.* 2001). This observation is akin to what is happening on the Internet, intranets and portals. More and more information is created every second and it is almost impossible to try and make sense of this chaos without the help of advanced search tools. Eckman *et al*. (2001) add by saying that as the post-genomic era is entered, the sheer volume of data and number of techniques available in the identification and characterization of regions of functional interest in genomic sequence is increasing too fast to be managed by traditional methods. Investigators must deal with the enormous influx of genomic sequence data from human and other organisms. To derive the greatest advantage from this data requires full query-based access to all the most up-to-date information available, with flexibility to customize queries easily to meet the needs of a variety of individual investigators and gene families. Wide-ranging multi-source queries, particularly those containing joins on the results of BLAST searches, often return unmanageable large result sets, requiring non-traditional methods to identify and exclude extraneous data. Such filtering often requires more sophisticated conditions than the SQL query language can express, for example, regular expression pattern matching. It would have been helpful to investigate the internal workings of Web search engines to determine how these technologies could be applied to the Web or vice versa. What is rather disturbing is the lack on exact technical details on recent search engines, according to Brin and Page (2000). These details are closely guarded and therefore it is not easy to make comparisons on the search techniques applied in the Web environment and the biotechnology industry.

There is great pressure on pharmaceutical companies to deliver new and personalized drugs and to fulfil the hyperbole and hope surrounding genetic research. The life sciences department at IBM believes that a scalable approach to data manipulation is the answer to relieve this pressure (Philopkoski 2000). The company is working on a product called 'DiscoveryLink' a virtual database that will allow scientists to mine information from different types of files, from graphic to database to text, to find genetic or protein information.

This study is not complete and it is difficult to make a satisfying comparison or suggestion at this stage. However techniques exist in bioinformatics that enable researchers to search relevant information and data, as has been illustrated. Similarly, in the Web environment, search engines exist and data mining techniques are employed to extract unknown values. The author suspects that projects had been conducted in the past to investigate the similarities of search and data manipulation techniques in different disciplines but in this study did not find satisfying sources to confirm this belief.

top

6. Conclusion

To obtain new insights and knowledge, the huge amounts of information generated by high throughput experiments in the biotechnology industry need to be transformed into executive summaries which are brief enough for creative studies by a human researcher. The amounts of the genomics data, which are already too large to be studied by human researchers in detail element by element, will continue to increase. Novel high throughput technologies generating new types of data will keep appearing and the ability to utilize these data will depend on the success of integration of data from different sources and the development of tools able to sift through all this data. This is in parallel with what is happening on the Web and elsewhere where huge amounts of data are created which have to be managed in some way.

Novel techniques had been developed to navigate through mountains of data in different disciplines. Cross pollination regarding the application of these techniques into different disciplines most surely has been done in the past and will be in the present and the future.

As De Bono explains in his book, *The use of lateral thinking*, mankind is surrounded with solutions, it is just a matter of looking differently to be able to notice them.

top

7. References

Brin, S. and Page, L. 2000. The anatomy of a large-scale hypertextual Web search engine. [Online]. Available at http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.
De Bono, E. 1987. *The use of lateral thinking*. 17-18. Great Britain. Penguin Books Ltd.

Eckman, B.A., Kosky, A.S. and Laroco, L.A. Jr. 2001. Extending traditional query-based integration approaches for functional characterization of post-genomic data. *Bioinformatics*, 17(7):587-601.

Greening, D.R. 2000. Data mining on the Web: there's gold in that mountain of data. [Online]. Available at http://www.Webtechniques.com/archives/2000/01/greening/.

Masys, D.R., Welsh, J.B., Fink, J.L., Gribskov, M., Klacansky, I. and Corbeil, J. 2001. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17(4):319-326.
Mazumder, R., Kolaskar, A. and Seto, D. 2000. GeneOrder: comparing the order of genes in small genomes. *Bioinformatics*, 17(2):162-166.

Persidis, A. 1999. Biochips: an evolving clinical technology. [Online]. Available at http://www.hosppract.com/genetics/9911mmc.htm.

Philipkoski, K. 2000. Genetic data glut looms. [Online]. Available at http://www.wired.com/news/technology/0,1282,39656,00.html.

Protein chips offer powerful method for probing protein function. [Online]. Available at http://www.stcsm.gov.cn/fuwuzhinan/fl/la/review/200010618-2.htm.

Shoop, E., Silverstein, K.A.T., Johnson, J.E. and Retzel, E.F. 2000. MetaFam: a unified classification of protein families. II. Schema and query capabilities. *Bioinformatics*, 17 (3):262-271.

Thearling, K. 2001. Data mining and CRM. [Online]. Available at http://www3.shore.net/~kht/index.shtml.

Webster, J. 2001. Bio-era to succeed age of information. *Engineering News*, 21(27):55.

Disclaimer

Articles published in SAJIM are the opinions of the authors and do not necessarily reflect the opinion of the Editor, Board, Publisher, Webmaster or the Rand Afrikaans University. The user hereby waives any claim he/she/they may have or acquire against the publisher, its suppliers, licensees and sub licensees and indemnifies all said persons from any claims, lawsuits, proceedings, costs, special, incidental, consequential or indirect damages, including damages for loss of profits, loss of business or downtime arising out of or relating to the user's use of the Website.

ISSN 1560-683X

Published by InterWord Communications for the Centre for Research in Web-based Applications,
Rand Afrikaans University