Data mining and knowledge discovery: Emerging fashions in science

Balaram in his recent editorial[1] succinctly highlighted the emergence of the new field of 'data mining' in the country. I would like to share some thoughts on this new field and propose that notwithstanding the fact that 'they contribute little by way of original research', data mining could become a very productive enterprise for a country such as India.

The last few decades have witnessed an unprecedented explosion of data in almost all fields, ranging from anthropology to astronomy. The People of India Project, funded by the Anthropological Society of India has contributed to the preparation of nearly 4000 maps, compilation of 21,362 photographs spread over 120 manuscript volumes[2,3]. The earth observing system (EOS) of NASA, USA is estimated to spew out 46 megabytes of data per second. The Human Genome Project is estimated to generate a mind-boggling amount of data on the nucleotide sequences in less than five years of the project period. In large part, the generation of such enormous volume of data has been made possible by an unparalleled synergy between the modern science gadgetry coupled with data holding and computing devices. On a more cottage scale, numerous small institutions, universities and colleges have been generating scores of pages of data.

Whatever the means and agencies involved in generating data, one thing is clear – there seems to be an insatiable hunger for data and even more data. It is not uncommon that in the urge to generate more and more data (and not necessarily more and more information), laboratories and institutions end up with more data than they could possibly process. Consequently, the cost effectiveness of such data gathering could be rather low, because they are rarely put to complete use. And as Jules Henri Poincaire (1902) mentioned, 'Science is built up with facts, as a house is with stones, but a collection of facts is no more a science than a heap of stones is a house'.

Herein lies the importance of mining[4] the available data and 'pulling out nuggets' from otherwise sterile digits. A case in point which has a relevance to the issue Balaram has raised[1], is in the field of biodiversity. Hundreds of institutions in the country along with the government sponsored Botanical and Zoological Survey of India have been documenting the occurrence of thousands of our plant and animal species for tens of decades. However, for want of proper organization of the data, we do not yet have in a single place a complete and comprehensive listing of what species occurs where. In fact this concern is precisely echoed in the theme raised by E. O. Wilson cited by Balaram[1] and Ganeshaiah and Uma Shaanker[5]. Only recently have there been a few initiatives, including at our centre to mine the data and provide biodiversity atlases for the country[6,7].

Data mining has assumed a global proportion, what with the extensive worldwide data transfer systems and data available in public domain sites. No longer is one required to generate data *de novo*. If you are hungry for data, just hunt with your mouse (yes, we have left the hunter dogs far behind) and the data are just a click away. It is easy to realize that data mining such as this could be accomplished in an infinitely less cost and with perhaps infinitely huge gains in terms of patterns found, insights drawn and concepts proposed.

This is best exemplified by the data spewed out by the Human Genome Project. With the DNA sequence data available in public domain, we in India seem to have the option to put our ingenuity to the challenging task of mining the sequence data in a myriad of ways – to discover hidden patterns in the organization of the base sequences, to discover useful nucleotide sequences that might be of relevance to human health and finally to conceptualize and advance the frontiers of science, instead of going through the drudgery of sequencing and more sequencing.

In fact I see an important role for countries such as India in data mining and knowledge discovery. With its huge technologically literate manpower, data retrieval, management and mining could become a major encashable industry. If managed strategically, we could be riding on the crest of knowledge generation and discovery in the world. However, before data mining and knowledge discovery could become an important activity, it might be necessary to change the mind set[8] of our students and even faculty to accept the fact that data mining also represents a legitimate way of doing science.

Data mining has already commenced in a big way by several corporate companies in the USA and in certain parts of Europe. Among the companies that head the list are pharmaceuticals and genomics laboratories. Firms specialized in handling data have sprung up and offer to mine data on a commercial basis. Research in the field of data mining and knowledge discovery (KDD) to evolve rapid and efficient ways of archiving and treating data has become a major field of study at many universities in USA, Europe and Australia. The journal *Data Mining and Knowledge Discovery*, published by the Australian Association of Statisticians is completely devoted to research in KDD.

It is not too late for our country to realize the vast potentialities of data mining and steal the march over others. Perhaps there is need to develop a national agenda and agreement on KDD in major spheres of science and technology. There is an urgent need to pool the vast array of disparate data held by numerous scientists, universities and institutions. Such data sharing could elevate the total synergy in the system and hasten the pace of discovery. During the process of exploring the path of carbon in photosynthesis, Frank Salisbury asked the Nobel Laureate, Melvin Calvin, 'Just what is the nature of creativity in science'? Melvin replied that 'The real creative trick is to get the right answer when you only have half enough data, half of what you have is wrong, and you don't know which half is wrong'.

Nowhere is this synergy so evident as in data mining in the area of biomolecules. I am sure there have been a number of discoveries or pattern findings possible based on such data banks. Not long ago a group of scientists, including three from Hyderabad discovered an intriguing pattern in the frequency distribution of even and odd numbered carbon compounds from a world database of organic compounds[9]. Our own finding that the chemical composition of seeds (as oil, carbohydrate and protein) is shaped by the dispersal mode of the species was possible due to a large database on seed chemical composition that we could access from USDA[10]. Mining data from floras, faunas and virtually any other retrievable source can be fun and often leads to simple, yet important patterns[11]. Finally, as my colleague Ganeshaiah demonstrated, you can get cracking by mining data from far and wide, from cricket to stock markets, without ever having played cricket or engaging the bulls (the stock brokers)[12,13].

Gone are the days of coal hunters and coal mining or for that matter gold hunters and gold mining. Let us unashamedly and without loss of much more time enter the age of data hunters and data miners. And why not, considering that data hunting and mining is environmentally safe?!

1. Balaram, P., *Curr. Sci*., 2000, **79**, 1511–1512.
2. Singh, K. S., *Curr. Sci*., 1993, **64**, 5–10.
3. Singh, K. S., *People of India: An introduction*, Seagull Books, Kolkata, 1992.
4. Conventionally, the term 'data mining' has been used to refer to the process of treating large sets of data to large-scale data analysis and discovering patterns in them. However, in this article I use the term to also include the process of collating widely dispersed and seemingly disparate data onto a common platform and then examine them for certain underlying patterns.
5. Ganeshaiah, K. N. and Uma Shaanker, R., *Curr. Sci*., 1998, **75**, 292–298.
6. Ganeshaiah, K. N. and Uma Shaanker, R., *GIS@Development*, 1999, **III-V**, 67–69.
7. Sen, N., *Curr. Sci*., 2000, **79**, 1046.
8. There is an inherent bias among students that their thesis should contain new (original) data. In fact my own graduate student was extremely uncomfortable with the fact that three of the four chapters in his thesis were based on data mining, while all his friends reported original data. He seriously doubted if his thesis would be approved by the external examiner. The fears of the student might not be completely unfounded. External examiners and reviewers of manuscripts tend to regard papers that do not contain significant amount of primary data as soft work and thus such papers run the risk of a high rejection rate.
9. Sarma, J. A. P., Nangia, A., Desiraju, G. R., Zass, E. and Dunitz, J. D., *Nature*, 1996, **384**, 320.
10. Lokesha, R., Hegde, S. G., Uma Shaanker, R. and Ganeshaiah, K. N., *Am. Nat*., 1992, **140**, 520–525.
11. Uma Shaanker, R., Ganeshaiah, K. N. and Ravishankar, K. V., *Curr. Sci*., 1997, **73**, 646–647.
12. Ganeshaiah, K. N., *Curr. Sci*., 1992, **63**, 345–347.
13. Ganeshaiah, K. N., *Deccan Herald*, Science and Technology Supplement, 18 May 1999.

R. UMA SHAANKER

Department of Crop Physiology,
University of Agricultural Sciences,
GKVK,
Bangalore 560 065, India
and Jawaharlal Nehru Centre for
Advanced Scientific Research, Jakkur,
Bangalore 560 065, India
e-mail: rus@vsnl.com

Towards a biodiversity information network for India

Convergence of research in modern biology and informatics is a reality now that genomics has become an accepted branch of scientific research[1]. The recent announcement of the completion of the sequencing of the *Arabidopsis* genome, while a significant milestone in modern biology research by itself, is also notable for the rapidity with which results have been made available to the community of researchers via the World Wide Web[2].

The transformation of outputs of modern biological research into easily dispersible and transferable digital data is thus now fully functional at the molecular level. At the levels of species and ecosystems, such a convergence is still incomplete. However, significant initiatives are under way in applying informatics at the species level. The Global Biodiversity Information facility (www.gbif.org) and the project Species2000 (www.sp2000.org) are notable efforts to build global gateways for taxonomic data in digital format held in collections located in different countries and continents. The challenge facing them is one of overcoming the limitations posed by the existence of widely different and occasionally incompatible digital formats used by various database system[3].

These initiatives are spearheaded by organizations located in countries that are members of the organization for Economics Cooperation and Development (OECD). At the international level, the World Conservation Monitoring Centre (www.unep-wcmc.org) and the clearing-house mechanism of the Secretariat of the Convection on Biological Diversity (www.biodiv.org/chm) provide detailed information on conservation and policy issues, but their emphasis is not on providing in-depth data on biological species.

In India, large collections of herbarium data exists. However, they are not readily available in a digital format. The Botanical Survey of India (BSI) has long pursued a project for the computerization of its accessions, but the digitized data are not yet available to public. With the passing of the Biodiversity Conservation Bill, a number of individuals and organizations connected with administering the national sovereignty over biological resources will be required to access species level information frequently. The Plant Variety Protection and Farmers' Rights Bill, when passed, will also create a need for access to varietal information.

There is thus a need to establish a digital network for species level information in India, which is the easiest and surest way to enable access by individuals and institutions connected with implementing the provisions of the above-mentioned Bills. Centralized databases such as the one under construction by the BSI are slow in coming. When such databases are constructed, the problem of incompatible formats cannot be ruled out, as seen with the OECD initiatives mentioned above. There is a clear need to avoid such bottle-necks.

One strategy would be to allow a variety of individuals and organizations to create and share or digitize and share species level data using the World Wide Web and through use of the technique of peer-to-peer file sharing. Creation of a minimum common format is essential for this purpose. It can be designed on the basis of consensus among the major national organizations such as the Botanical and Zoological Surveys of India, the National Bureau of Plant Genetic Resources and the allied bureaus, and the National Bio-resources Board. Data can be either input