DEVELOPMENT OF SOFTWARE TOOLS AT BIOINFORMATICS CENTRE (BIC) AT THE NATIONAL UNIVERSITY OF SINGAPORE (NUS)

P. R. KOLATKAR[1,2], M. K.SAKHARKAR[2], C.R.TSE, B.K. KIONG, L. WONG, T.W. TAN,
BioInformatics Centre,
National University of Singapore
Singapore

S. SUBBIAH *
**Wistar Institute*
3601, Spruce Street,
Philadelphia, PA 19104

There is a burgeoning volume of information and data arising from the rapid research and unprecedented progress in molecular biology. This has been particularly affected by the Human Genome Project which is trying to completely sequence three billion nucleotides of the human genome (1),(1a). Other genome sequencing projects are also contributing substantially to this exponential growth in the number of DNA nucleotides and proteins sequenced. The number of journals, reports and research papers and tools required for the analysis of these sequences has also increased. For this the life sciences today needs tools in information technology and computation to prevent degeneration of this data into an inchoate accretion of unconnected facts and figures.

The recently formed BioInformatics Centre (BIC) at the National University of Singapore (NUS) provides access to various commonly used computational tools available over the World Wide Web (WWW) - using a uniform interface and easy access. We have also come up with a new database tool, BioKleisli©, which allows you to interact with various geographically scattered, heterogeneous, structurally complex and constantly evolving data sources. This paper summarises the importance of network access and database integration to biomedical research and gives a glimpse of current research conducted at BIC.

Introduction

BioInformatics deals with organising and presenting information in effective ways (2). With the globalisation of the Internet and the data deluge from the Genome sequencing projects bioinformatics is going through a period of explosive growth and development. The WWW facilitates the sharing of this treasure and has changed the nature of learning by providing increased access to resources in a variety of media.

[1] presenting author
[2] communicating authors

The emergence of such an interdisciplinary field at the confluence of Biology, Computational science, Chemistry and Mathematics will be critical for solving many problems:

- Data generated by sequencing projects require novel ways of processing and analysing.
- Better algorithms are required to fully explore the biological databanks.
- Vital clues to experimentation can be obtained by bioinformatics and biocomputational tools which save time and effort.

In this report we will elaborate on the current status of BIC and the tools developed by the centre for database integration and facilitated access to various BioInformation and BioComputing software. In particular we will elaborate on access to:

- BioKleisli
- BioPortal.
- BioAgent

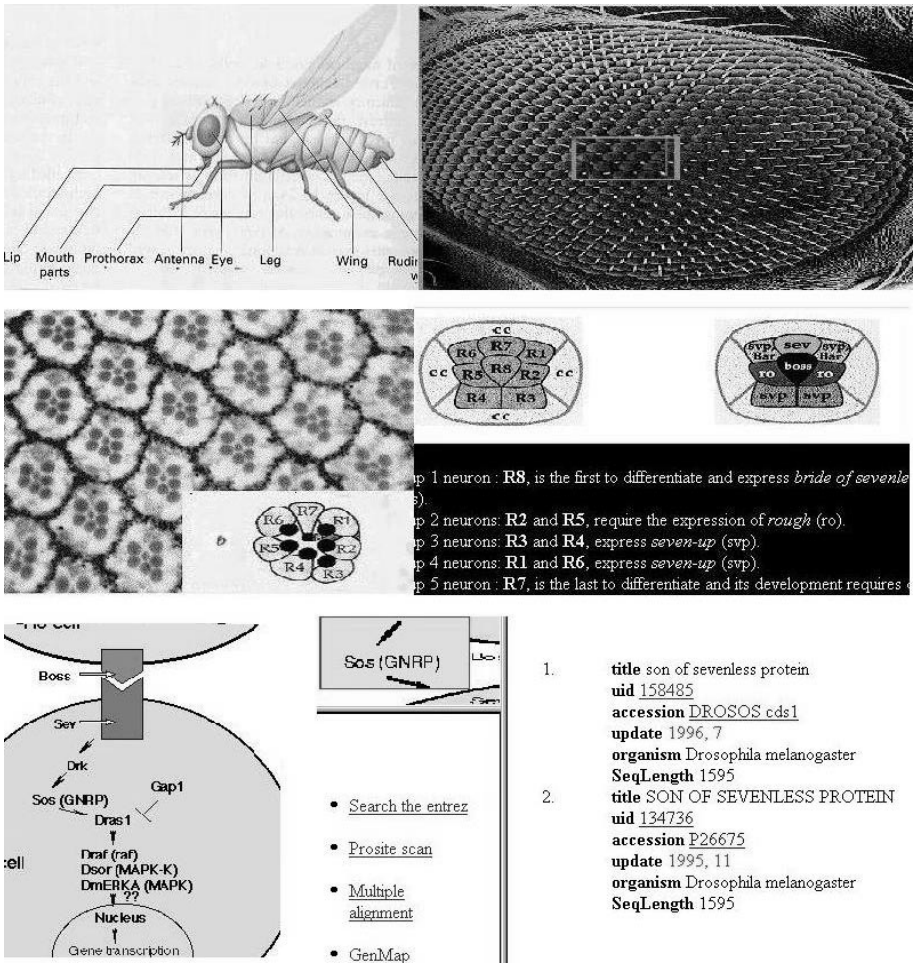
Access to International databanks and their Integration (BioKleisli)

The data collected from genome sequencing projects will be viewed as less than a success if users were unable to share their findings. With the rapid increase in volume of data catalysed by the advent of recombinant DNA technology there arose the need to allow researchers to submit their data directly to the community and do the editorial review of sequences.
Three main institutions are currently responsible for world wide acquisition, annotation and distribution of DNA sequences (3):

- The National Centre for Biotechnology Information (NCBI) in US with online submission tools called BankIt which provides a simple form based method for submitting sequences to GenBank.(3a)
- The European Molecular Biology Laboratories (EMBL) in Europe with EBI sequence data submission system.(4)(4a)
- DNA Databank of Japan (Japan) with sequence submission tool called Sakura.(5)

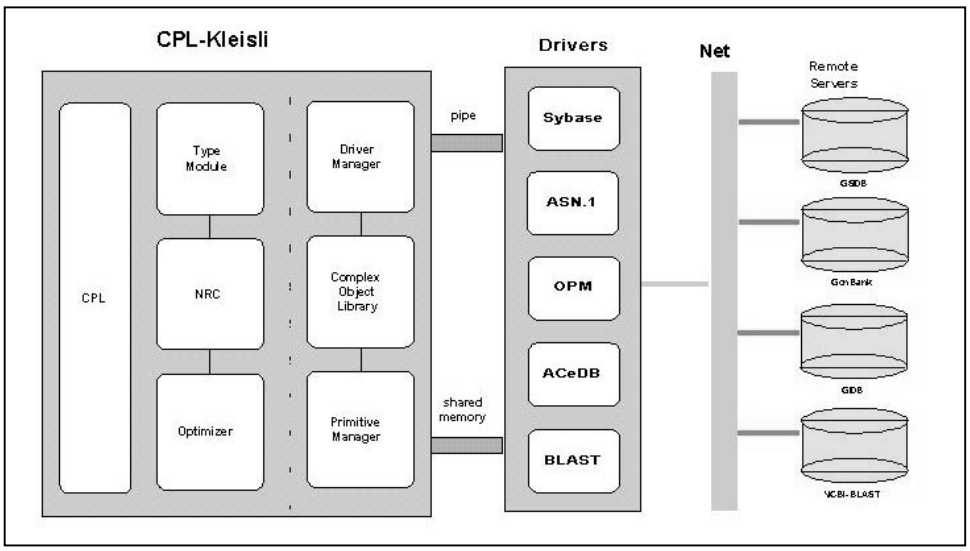
Besides these three main data banks, there are more than 100 specific databases as given in LiMB (6). This generation of databases has become so varied and dispersed that they are no longer easy to access and use. Industrial users find the current system particularly inconvenient, as they do not want to put their data in public domain but at the same time want to use the public domain databases and if possible integrate the information from public and private databases to generate new information. Thus, new ways need to be found for linking data and information across biological and

geographical divisions so that the databases become interoperable. BioKleisli © provides links between complex and heterogeneous data sources which are geographically scattered. It also allows integration of data banks, analysis software, and visualisation tools. It can be deployed for high-level data-intensive software. BioKleisli©, is developed under a joint collaboration between BIC, the Institute of Systems Science and University of Pennsylvania.



An example query from BioKleisli

This interface provides a holistic approach to using information. The diagram on the previous page shows how the power of BioKleisli can be harnessed for a real example. The user can control the journey from macro level (organism) to the level of genes involved in its development. As biological data is more meaningful in context, this interface is a step towards integrating the current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting genes or molecules and, second, to link individual components of the pathways with the gene catalogues being produced by the genome projects. Specific operations like searching ENTREZ (7) and PROSITE (8) database can be carried out. The advantage is that these tools are made available in an integrated fashion. This helps the user to carry out the main mission of understanding Drosophila at the gene level rather than hunting for the programs. This could further be extended to cross the barrier of being simply a static resource to be searched and browsed to a deductive database in the sense that additional information can be deduced dynamically from the stored information. We are currently investigating many queries for BioKleisli© which could shed light on specific biological systems. The power of BioKleisli was revealed when it managed to solve many of the so called "impossible" queries raised by US Department of Energy (DOE) for its Human Genome Project in a single day. BioKleisli © could be used in the future for computer assisted or even automated analysis of biological sequences.



A graphical representation of BioKleisli Functionality

Online Access to BioSoftware (BioPortal)

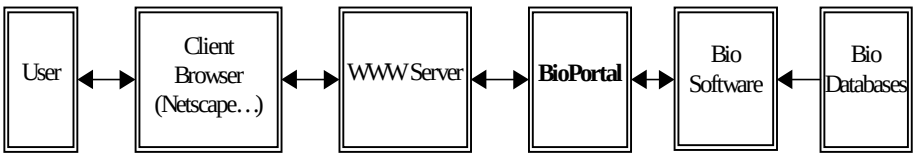
Powerful computational tools are a paramount requirement for research in molecular biology. Many of these tools are not user friendly and can be platform dependent. The Internet allows you to access biocomputing and bioinformation tools from the World Wide Web (WWW) at the click of a button from your personal computer. We at NUS have engaged the capability of WWW to provide access to the most widely used biosoftware. This is done using a set of Common Gateway Interfaces (CGI) programs in PERL and C on your specific system environment. The programs currently supported are:

- Java implementation of the GCG© package with file handling capability (9).
- PHYLIP© suite of evolutionary analysis programs (10)
- Signal Scan.(11)
- Promoter Scan. (12)
- GNUPLOT© (13)

We also support other commonly used biosoftware like ClustalW© (14) and PatScan© (15). Some 150 or so separate bioinformation programs have been done to date. We intend to add interfaces for many of these programs, particularly ones relevant to 3D structure and drug design.

The specialised features of BioPortal are:

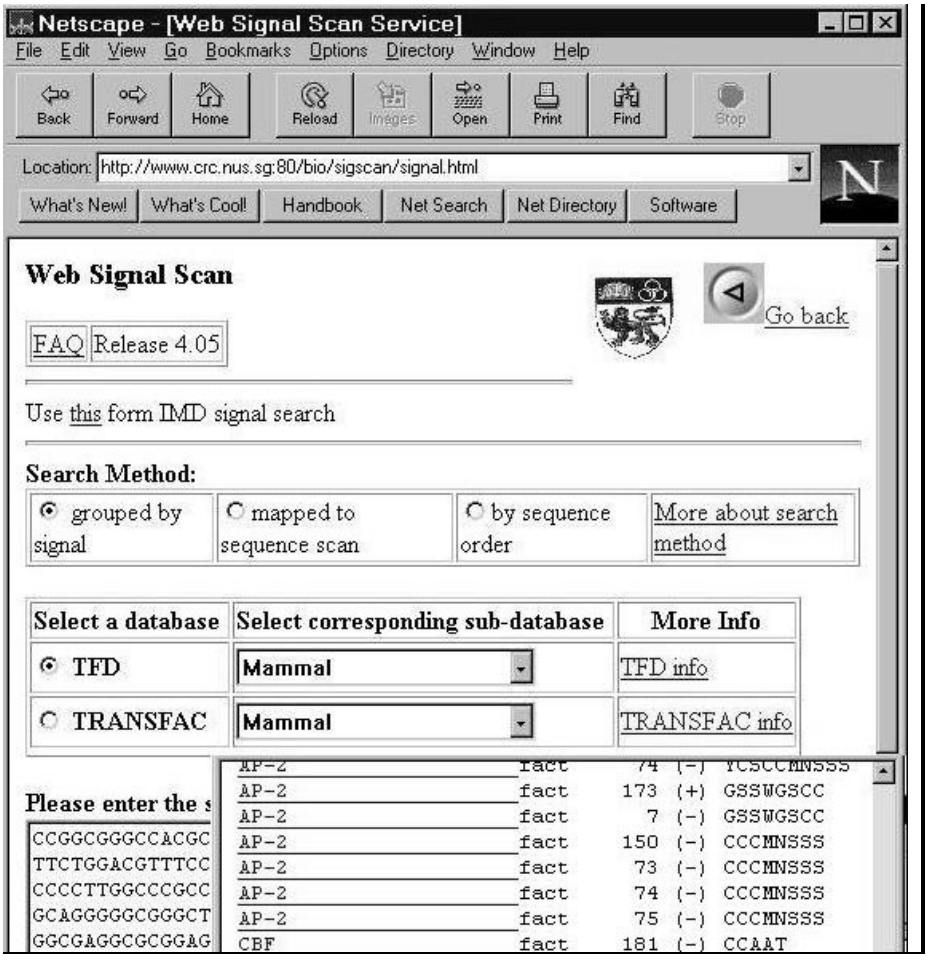
- User Friendly Interface : It follows the "form based" approach. So the user does not need to learn to run the software.
- Online Help : Online help speeds up data entry and analysis.
- Remote retrieval and analysis: Allows you to access the program from anywhere in the world.
- Conduit : Output can be piped from one program to another for further analysis.
- Hyperlinks : The results can be hyperlinked to appropriate sites to obtain further information about them.
- Improved Operation Control : A fully customisable interface provides a flexible environment.



A graphical representation of BioPortal Functionality

Other programs like WWW2GCG (16) exist which also have similar functions but our interface has certain advantages:

- File handling capability which allows the user to browse the local desktop and select for files that get auto-pasted in the selected area.
- Format independent interface allows the user to submit the sequence in all the commonly used formats.
- Hyperlinks to full sequences for the retrieved sequence list (using fetch).
- The web interface links programs where the output from one becomes the input for another e.g Mfold and PlotFold and Bestfit and DotPlot.



An example interface for the Signal Scan Program.

In the future, we will customise web interfaces for other biosoftware and if possible provide a Java implementation of the interface. We are also working towards integrating BioPortal with BioKleisli, which would allow more convenient utilisation of database and visualisation tools while facilitating links to more databases.

Biological information retrieval and indexing system (BioAgent):

The Internet has become an invaluable asset in the globalisation of computer resource sharing and speedy communication. Although, this new frontier holds vast potential, the power of the Internet has yet to be fully harnessed. Information exists on the Internet but the tools for locating the information are relatively crude. Dissemination of updated information for R & D scientists is essential. Tools now exist to post the newly arrived sequences in primary databases such as NCBI, EMBL or DDBJ to you. We have come up with an intelligent information retrieval and indexing system called BioAgent. This is based on iAgent created by Information Technology Institute (ITI), Singapore. BioAgent periodically sweeps through the user or system specified biological knowledge sources for new information and notifies the user that the information is available. BioAgent allows you to prioritise retrieved entries (based on keyword) and notifies only if the articles are in the same context. BioAgent filters out articles of your interest based on your keyword search and voting profile. We have tried to find a solution to a point re-stated by Bains , " accessing the information is not a problem but sifting the relevant items from the millions of 'hits' is a monumental task" (17). Currently, BioAgent supports bionewsgroups and archives. We are exploring the possibility for indexing BioJournals and Medline database.

An example of BioAgent:

Future Directions:
We hope to develop BioKleisli into a system which will enable the bioscientist to search and retrieve, survey and research virtually any discipline in the biological sciences, from anywhere at anytime. Projects in the pipeline are:

BioKleisli related:

1. QUICK: to develop an automated query formulating system on top of BioKleisli.
2. Object Oriented view management for BioKleisli.

Other projects:

- Virtual cell project: The primitive level of organisation of biological data has led to a decrease in the knowledge-to-data ratio. We intend to initiate and build a unified global biological hyper-infostructure presented to the scientist in a user-friendly aesthetic yet powerful virtual reality interface. From this biohyper-infostructure built on top of the Internet, access to other bodies of information including medical information, other scientific data and data from commerce will be facilitated. It will be more than a project - a bold and ambitious vision worthy to carry all disciplines of the life sciences into a new millennium.
- Genebit recognition of EST project: The Expressed Sequence Tags (EST) database is a database of short DNA sequences from expressed genes. These genes presumably code for proteins that may be of biological, medical or pharmaceutical importance. However, the EST sequences will be of greater value if there functions are known. Once known, the researchers can make use of ESTs to fish out the entire gene from the genome, clone them and sequence them in entirety, as well as produce recombinant versions of the gene for potential medical application. We are trying to answer the question “How can I guess the function of EST when traditional methods do not work?”. We propose to do this by cutting up the DNA sequence of all proteins with known 3D structure into fragments that are structurally meaningful. Each genebit will have specific associated function(s), since only full-length sequences/structures will be included. This collection of genebits will then be searched against the EST database for sequence similarity.
- HLA Database project-HLA biotechnology is increasingly important. This project involves setting up of an International databank of HLA allele typing, structure and sequence database in Singapore. This will help enable the modeling of HLA structure for design of new drugs for treating specific disease predispositions in the region and world wide.

Conclusion

Networks and integration of databases are keys to success in BioInformatics. The technology now exists for global access to databases and software. The next step is to make the databases interoperable, deductive, and design better tools to interrogate the databases. Further, the marriage of data computation (the domain of theoretical, predictive computer algorithms collectively developed over many decades) and data integration (the domain of the multitude of experimental bio-data - e.g. literature and genomic databases) into a single cohesive whole, thus seamlessly uniting BioPortal with BioKleisli, is one key step along this path. This will also increase the efficiency of research effort by reducing the serendipity and hit and miss nature of empirical research and will help provide clues to the bioscientist on their choice of experiments to carry out under limitations of funds, manpower and time. Better education in BioInformatics is required. Users have to know what is available and how to access and use the resources they are offered. Only then will we all be able to exploit the true potential of BioInformatics. We are embarking on several collaborations with other institutions to expand the portfolio of integrated and linked information systems with the goal of a globalised information infrastructure.

Acknowledgment
Bioinformatics Centre (NUS), is funded by the Economic Development Board (EDB) of Singapore. Subbiah .S acknowledges DOE grant no. DE-FG03-95ER62135 for partial support.

References:

1. Hawkins TL, McKernan KJ, Jacotot LB, MacKenzie JB, Richardson PM, Lander ES: A magnetic attraction to high-throughput genomics. Science 276 (5320): 1887-1889 (Jun 20 1997)

1a. Patrinos A, Drell DW: The Human Genome Project: view from the Department of Energy. J Am Med Womens Assoc 52 (1): 8-10 (1997)

2. University of Berkeley Museum of Paleontology (UCMP) Internet Web Site. Internet URL: http://ucmp1.berkeley.edu/subway/bioinfo.html

3. Tuli MA, Flores TP, Cameron GN: Submission of nucleotide sequence data to EMBL/GenBank/DDBJ. Mol Biotechnol 6 (1): 47-51 (Aug 1996)

3a. The National Centre for BioTechnology Information (NCBI) Internet Web site. http://www.ncbi.nlm.nih.gov/

4. The European Molecular Biology Laboratory (EMBL) Internet Web site. http://www.embl-heidelberg.de/Services/index.html

4a. Stoesser G, Sterk P, Tuli MA, Stoehr PJ, Cameron GN: The EMBL Nucleotide Sequence Database. Nucleic Acids Res 25 (1): 7-14 (Jan 1 1997)

5. The DNA Data Bank of Japan (DDBJ) Internet Web site: http://www.ddbj.nig.ac.jp/

6. Lawton JR, Martinez FA, Burks C. Overview of the LiMB database. Nucleic Acids Res 17 (15): 5885-5899 (Aug 11 1989)

7. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. Methods Enzymol 266, 141-162 (1996)

8. Bairoch A, Bucher P, Hofmann K. The PROSITE database, its status in 1997.

9. The Genetics Computer Group Internet Web site: http://www.gcg.com.

10. The PHYLIP© Internet Web site: http://evolution.genetics.washington.edu/phylip.html

11. Prestridge, D.S. SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. CABIOS 7, 203-206 (1991)

12. Prestridge, 1995 Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol 249, 923-932 (1995).

13. GNUPLOT© 1986-1993, Thomas Williams, Colin Kelley.

14. Thompson, J.D., Higgins, D.G. and Gibson, T.J. CLUSTAL W: improving the, sensitivity of progressive multiple sequence alignment through sequence weighting positions-specific gap penalties and weight matrix. Nucleic Acids Research, 22:4673-4680 (1994)

15 The Patscan is a copyright of Dr. Ross Overbeek of Argonne National Laboratory.

16. WWW2GCG has been written by Marc Colet (mcolet@ulb.ac.be), at the Belgian EMBnet Node, Brussels.

17. Bains. W Scrip Magazine. March, 18-19. (1995).

18. Davidson SB, Overton C, Tannen V, Wong L. BioKleisli: A Digital Library for Biomedical Researchers. International Journal of Digital Libraries, 1(1), (November 1997, to appear).

19. Peter D. Karp. Database links are a foundation of interoperability, TIBTech , Vol 14, 273-279 (1996)

20. Anonymous Strategies in Bioinformatics and BioComputing : a European (19) perspective. Proceedings of a meeting held in Nijmegen, The Netherlands. 26-28 Sep 1994. p8.

21. Harper R. World Wide Web resources for the biologist. Trends Genet 11 (6): 223-228 (Jun 1995).