

Bioinformatics Tools @ NBBNet: Online Infrastructure for the Management and Analysis of Biological Data

MOHD FIRDAUS RAIH, MOHD YUNUS SHARUM, AHMAD FUAD HILMI MUHAMMAD, HAFIZA AIDA AHMAD, RAJA MURZAFERI RAJA MOKTAR, MOHD NOOR MAT ISA, NOR MUHAMMAD MAHADI & RAHMAH MOHAMED

ABSTRACT

The use of informatics tools for the management and analysis of sequences for nucleic acids and proteins has resulted better throughout capability of wet lab research work to infer biological data to functional biological information. The field of computational biological information management and analysis is generally known as bioinformatics. We discuss some tools and processes which have been developed or integrated into a data management and information presentation pipeline by the Malaysian National Biotechnology and Bioinformatics Network. Central to this is the Bioinformatics Tools @ NBBnet online infrastructure system. This infrastructure system utilizes grid computing technology. In addition, the deployment of niche databases and database shells for research applying specific datasets such as a particular protein function, protein family or genomes have been discussed.

Keywords: Bioinformatics; meta-server; niche datasets; web infrastructure; grid computing

ABSTRAK

Penggunaan teknologi maklumat dalam pengurusan dan analisis bagi jujukan-jujukan asid nukleik dan protein telah mengakibatkan peningkatan kemampuan penterjemahan data yang dijana di makmal kepada maklumat biologi yang berguna. Bidang pengurusan dan analisis maklumat biologi berkomputer juga dikenali sebagai bioinformatik. Penerbitan ini membincangkan beberapa perisian dan proses yang telah disepadukan dalam satu aliran kerja bagi analisis dan pemaparan data serta maklumat biologi oleh Rangkaian Bioteknologi dan Bioinformatik Kebangsaan Malaysia (NBBnet). Prasarana dalam talian ini berpusat sekeliling sistem Bioinformatics Tools @ NBBnet. Sistem prasarana ini mengaplikasikan teknologi pengkomputeran grid. Di




sampling itu, beberapa konsep pengurusan dan analisis data subset daripada famili protein tertentu juga dibincangkan.

Kata Kunci: Bioinformatik; meta-server; data set khusus; prasarana web; pengkomputeran grid

INTRODUCTION

Bioinformatics is a field flooded by a plethora of tools for the analysis of the ever increasing and diverse life sciences data, specifically, molecular biology data sets (Galperin 2004; Nucleic Acids Research Web Server Issue 2004). The diversity of these data sets and the tools available to analyze them has resulted in the necessity to integrate and unite this heterogeneity and at the same time break down the diversity of the data into more niche subsets. Adding to this complexity of accessibility is the fact that these resources are spread out on many different servers across the globe (Nucleic Acids Research Web Server Issue 2004). A further constraint, which even though existent worldwide, but especially magnified in developing countries, is the lack of expertise which needs to be further developed to harness many powerful software tools available for data analysis. This is especially true as many of the more powerful tools available in bioinformatics, are still accessible only via command line environments running on technical operating systems such as UNIX. Adding to this problem is the sheer volume of data being generated by modern biotechnology, which without informatics tools for management and analysis, would be almost impossible to sift through within a reasonable time frame. While more and more tools are made available with graphical user interfaces (GUIs), these at times do not address the problems at hand. GUI enabled tools at times are added costs to software which was originally free or academically licensed. Again, this presents a hindrance to developing countries with small R&D expenditure, which is typically a very small fraction of the GNP (Huete-Perez & Orozco 2001).

The data at hand, which is being referred in this text, is mainly biological sequence data and its derived information. Biological sequences are basically either nucleic acid sequences such as DNA and represented as a series of 4 letters namely A, C, G or T for each monomer of DNA termed as bases; or protein sequences which are displayed by a series of 20 letters each representing amino acids protein monomers. Genes are sequences of nucleic acids which hold information as to what type of protein will be synthesized from the encoded information. The sequence of amino acids will determine the three dimensional structure the protein takes up. The structure of a protein determines what biological function it will carry out. Examples of derived information from these basic building blocks can be notes referred to as annotation and protein structure data. Annotations are notes attached to the sequences



usually assigning biological function to a specific sequence. For example, a particular sequence of 264 amino acids may function as a particular enzyme, once this information is known for that sequence it is marked as such. Most of these data are in the format of text in flat files.

The basic management of these data is centered on assigning information of protein function to a particular sequence (Figure 1). The most basic analysis available for a biological sequence is known as sequence analysis which predominantly includes the sequence similarity searches against related sequence database. This procedure has resulted in the generation of sequence alignment. Sequence alignment is an arrangement of two or several biological sequences (e.g. protein sequences or DNA sequences) highlighting their similarity. Sequence alignment usually represents a hypothesis about common evolutionary origin of sequences involved. As discussed earlier, the function of a protein is dependent on the structure the polymer of amino acids take up in three dimensional spaces. The sequence similarity of an unknown sequence to a known sequence can be used to comparatively extrapolate the possible function of a protein, based on the assumption that there is an evolutionary relationship between similar proteins (Zuckerlandl & Pauling 1965). As such, if a protein shares a similar structure, this would mean that it may share a similar function. It is more or less known that a similarity in sequence of 30% or more for a protein sequence may result in a similarly folded protein structure (Chothia & Lesk 1986). The use of comparative approaches particularly pertaining to sequence alignments is seen as a major contributing factor to the rapid elucidation of biological information from the mountains of biological data being generated presently. This enables new information to be layered onto new data using presently available knowledge. Taking this a step further are methods of alignments which deploy Hidden Markov Models (HMMs) and neural networks to elucidate remote similarities between sequences. This rings true in biology as sequences may diverge in similarity due to the effects of evolutionary adaptation or differentiation arising from genetic variation or mutations.

Bioinformatics, besides modern DNA sequencing technology, can be seen as the savior of the numerous genome projects in the world today. Of these, one of the most important is probably the Human Genome Project (Collins et al. 2003). A genome is the whole genetic content of an organism. A genome project is a scientific endeavor that aims to sequence complete set of genes in any particular genome and thereon derive the function of these genes and the proteins which make up the functional organism. Further downstream applications may include associating interdependent functions as part of a network. This is referred to as systems biology, which can be defined as the ability to look at all the elements in a biological system such as genes, proteins, protein interactions and from there measure their relationships to one another as the system functions in response to biological or genetic

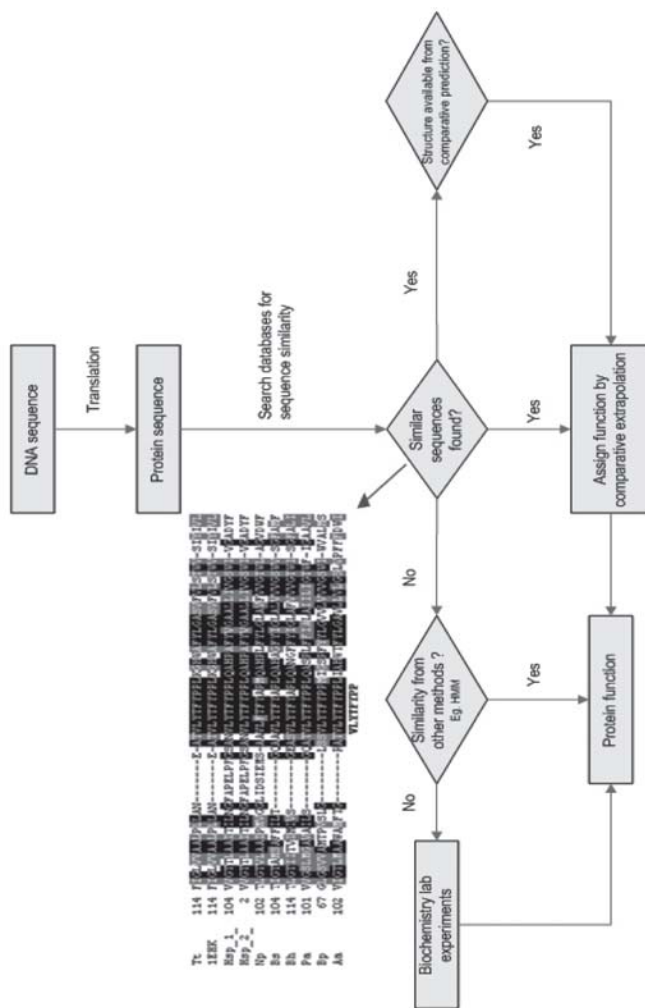


FIGURE 1. Basic processes of extracting function information from biological sequences using sequence alignments



perturbations (Hood 2003). For example, protein A influences protein B to induce the activation of protein C, which acts upon protein D. Mapping out these interactions (gene regulatory networks) and understanding the molecular mechanisms involved in these interactions will open doors for molecular structure design and protein engineering for therapeutic as well as industrial applications. DNA can therefore be said as having two types of a digital nature as well, these being the genes that encode the protein and the gene regulatory networks which specify the behavior of the genes (Hood & Galas 2003). The critical issue is how all these exponentially expanding sequence information can be converted to knowledge of the organism and how biology will change as a result (Hood & Galas 2003). This sort of endeavor is in fact an integration of technology for computation, biology and ultimately medicine (Hood 2003).

We have successfully set up a primarily web based infrastructure which includes a common gateway and database integrated meta-server for submission of sequence analysis jobs which serves the Malaysian biotechnology community on a national level. Further to this, database shells for niche datasets have been developed using MySQL interfaced with PHP. All the developments are operated as a web based service while the source codes are distributed free for academic and non-profit use (<http://genome.ukm.my/tools/>).

STRUCTURE, ORGANIZATION AND DEPLOYMENT

The web based infrastructure named 'Bioinformatics Tools @ NBBnet' is operated as a portal for data, resource management as well as a metasever environment for submission to external resources (Figure 2). This portal is operated using the MySQL database interfaced via PHP and PERL scripts. Primary CPU resources are integrated as a compute grid utilizing the Sun Grid Engine (SGE). Other software agents for management, formatting and analysis of data were also written using PERL. The Malaysian National Biotechnology and Bioinformatics Network (NBBnet), is the institutional host for this infrastructure. NBBnet is a virtual resource and capacity building network tasked at providing and managing administrative services as well as scientific computational biology services for the Malaysian biotechnology community through the use of informatics (Firdaus Raih et al. 2003).

Access to the portal is controlled via a user database, which in turn is connected to the data and resource management database system. The presentation of the Bioinformatics Tools @ NBBnet interface is operated on two fronts. One is an open to all curated database of web based resources inclusive of tools, online tutorials and courses as well as web sites of labs. The second front is accessible only to registered users, which includes a meta-server system directly attached to the users' data to ease management of data analysis and annotation. The use of a meta server, which acts as an



agent for remote data submission, is to enable users to have access to a diverse range of freely available analysis tools from a single interface.

This resource also includes an interface to other tools such as a grid enabled bioinformatics software pipeline (Firdaus Raih et al. 2004). The grid operated services, are however independent of the data attachments described above, which is typically limited to 50 entries per user and is targeted for use by users not requiring high throughput analysis capabilities. High throughput analysis of sequence data encompassing automated high-throughput editing of chromatograms (results from DNA sequencing runs), high-throughput BLAST, BLASTClust, high-throughput vector screening are submitted directly via the interface to SGE or users have an option of using the more powerful BioGrapppler interface accessible via ssh logins. BLAST an acronym for 'Basic Local Alignment Search Tool' (Altschul et al. 1990), is a sequence similarity program which can be used to compare a query sequence against those available in a database. This program is seen as the most basic and core tool in bioinformatics for sequence similarity purposes (McGinnis & Madden 2004).

BioGrapppler is an ongoing development of an integrated pipeline of proven tools operated in a grid computing environment. The primary idea is to provide users with access to a suite of bioinformatics tools, which are available via academic licensing, covering functions from sequence quality editing, sequence database searching, contiguous sequences assembly for reassembling sequences, phylogenetics study on evolutionary aspects, as well as protein structure prediction. It is operated on a grid computing infrastructure enabled via SGE, without technical hassle. BioGrapppler is coded in PERL and operable in a Linux as well as Solaris 9 environment.

The current deployment of SGE to this system is primarily for the processing of high-throughput input data. These jobs may have massive inputs, but are not necessarily very compute intensive. As such, an input manager splits large input data to a number of input files to take full advantage of the cluster grid resources available to the SGE queue master. Users do have options of splitting input into single submission units with the results specific to each submission. This service is currently available for running high throughput BLAST jobs via web interface. Currently, phred (Ewing et al. 1998; Ewing & Green 1998), CLUSTAL W (Thompson et al. 1994), phrap - CrossMatch (<http://www.phrap.org>) and MODELLER (Sali & Blundell 1993) are the applications that are operated on this grid. These applications are also available via a secure shell (ssh) connection or interfaced via the EMASGRID BioBox initiative (Firdaus Raih et al. 2004; <http://www.apbionet.org/grid/apbiobox/lsvgcApr04/index.shtml>).

Integrated into the web based infrastructure are readily deployable data managers for niche or in-house datasets built using MySQL, PHP and PERL. An operational example is the Prokaryotic Lysis Enzymes Database (ProLYSED) accessible at <http://genome.ukm.my/prolyses/>. This example is a database of

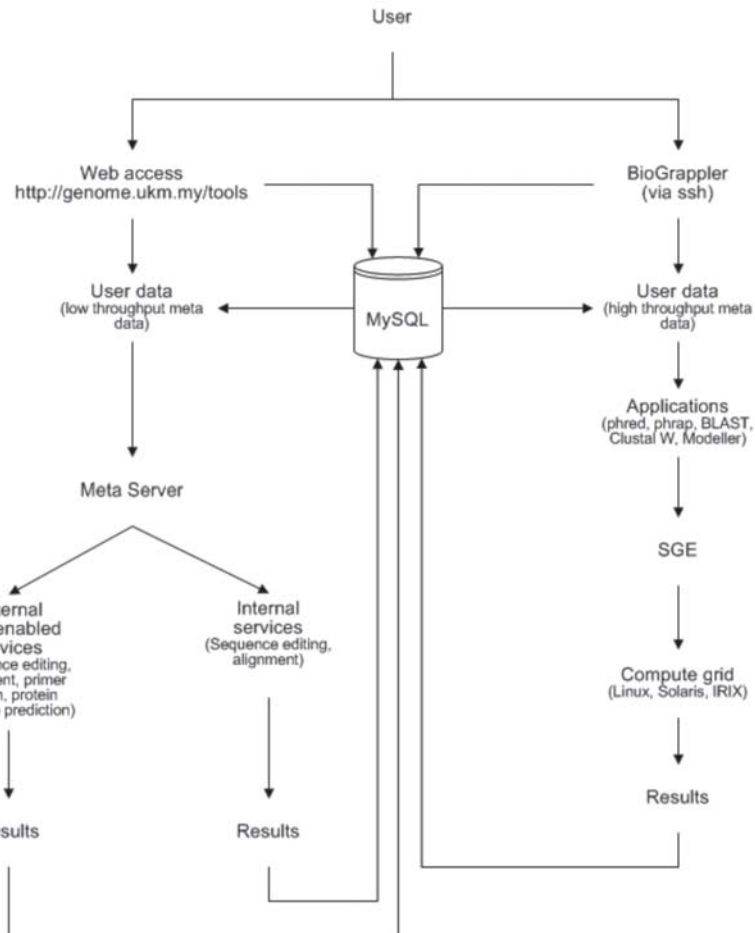


FIGURE 2. Overview of the 'Bioinformatics Tools @ NBBnet' online infrastructure system

bacterial protease systems which is integrated to a metasever system using the housed sequences as meta-data for submission to internal or external tools for on-demand analysis and cross-referencing of the stored dataset (Figure 3). The system was designed to be easily adapted to different datasets. Amongst the current deployment for this database system by Malaysian research groups are for *Eimeria tenella* genome sequencing data, Seabass (*Lates calcarifer*) expressed sequence tags (EST) data and the *Burkholderia pseudomallei* clinical investigations and functional genomics database.

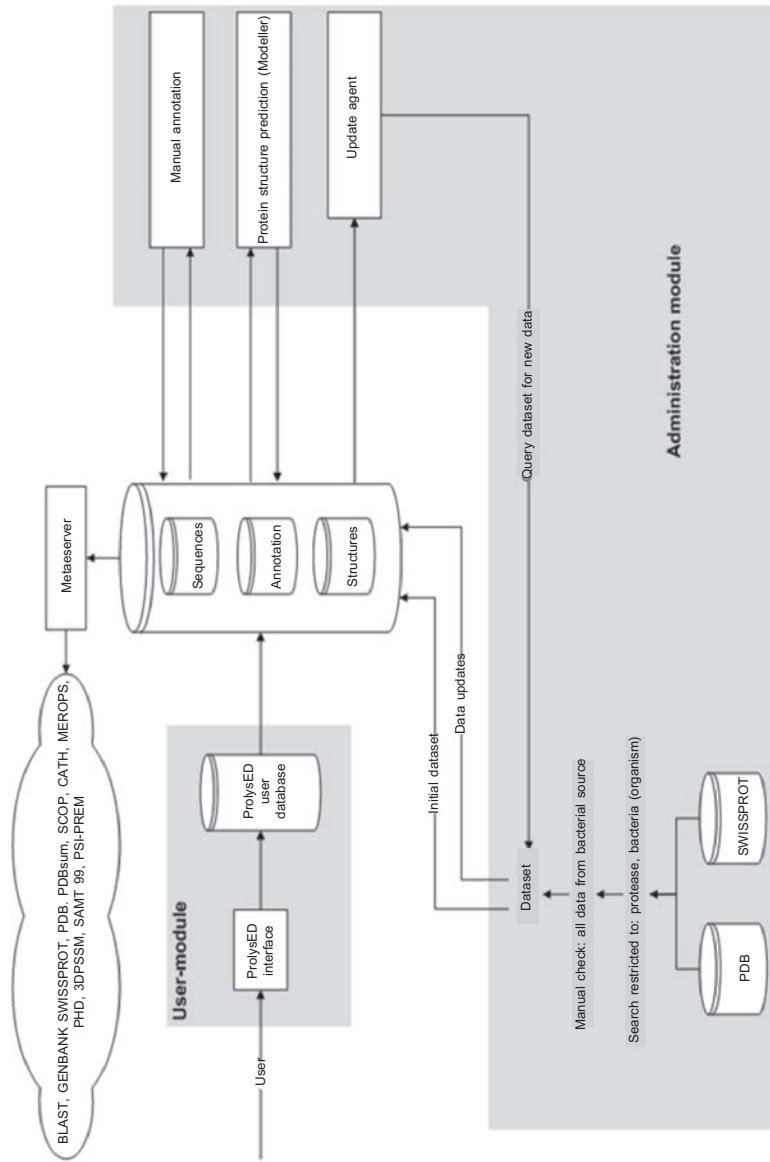


FIGURE 3. Overview of value added annotation for publicly available sequence data of niche protein families



DISCUSSION

The concept of a web based bioinformatics tool box is not new. What we have sought to do is not just to deploy a web based sequence analysis workbench which has been done successfully many times. The objective of our resource is to provide a wide range of resources which includes access to high performance technical computing tools with a high degree of user customization in setting up job runs in addition to more commonly available tools. By enabling users to extract submission data from a database, it is important to provide a single environment where users can store and selectively analyze biological sequences, with the resulting output attached to the original input.

The problem of heterogeneous formats encountered when using different bioinformatics applications is also addressed using the web interface described and BioGrapppler. In the case of the web service, the metadata is automatically reformatted and presented in the required format to the services available to the meta-server. Some processes include a series of format changes and may require the user to issue continuation commands for certain process divergences. BioGrapppler, on the other hand operates as a pipeline environment for highly deployed applications covering the basic processes needed to analyze a nucleic acid or protein sequence. As such, the input for a particular analysis may have been pre-prepared by automatically reformatting the output from earlier analysis. For example, the inputs for the phred application are in text and graphical formats (chromatograms), for which the output is a text format called FASTA. The FASTA text format is the required input format for a subsequent series of analysis for similarity search using a program such as BLAST. The output from the similarity searching can then be reformatted as input for a clustering application based on sequence similarity. Further on, the output from ClustalW can be reformatted as an input in the protein structure prediction applications such as MODELLER. These layers of assistance allow the biologist to bypass the technicalities of reformatting input files and searching for the resources to run on the web. At the same time, a correct input format and the use of the correct program will minimize the errors caused by the programs in terms of technical errors such as input rejection or erroneous output resulting from submission of an input to an incompatible set of analysis resources.

Furthermore, a resource was designed to support users ranging from a local group to a wider geographically dispersed user base, with all users having access to the same tools and resources. This is important solution to many departments and countries which may not have the capacity of having bioinformatics support at every separate research unit. This solution was also aimed at ease of use for a wide range of users, from first time users to experienced users. This particular aspect was accomplished by adding usage



guides to submission interfaces. There is a tendency for first time users to attempt using applications without understanding many aspects of the program and its parameters therefore resulting in either erroneous submissions or results. A single interface environment was used for various bioinformatics applications with options to run the applications sequentially. These approaches towards providing a seamless integration of proven solutions deployed as web infrastructure is directed towards the evolution of an environment which can use multiple data formats and standards, and more importantly does not create any new standards or formats. This online services model can be a stepping stone towards the evolution of a truly seamless database and applications model as described by Lincoln Stein (2002). At the same time, we hope that this environment is easy to use for most non-bioinformatics savvy biologists to appreciate.

CONCLUSION AND FUTURE DIRECTIONS

The operation of a centralized biological computing infrastructure is an important aspect in biotechnology research and development in developing countries. The deployment model can be seen as an easy replicable infrastructure for other developing countries or resource restricted laboratories which can harness the power of high performance computing using off-shelves heterogeneous hardware and low cost or free software solutions. While the infrastructure model was targeted for use as a national based service requiring minimal human technical support, it can also be adapted on a departmental level as a cost effective infrastructure solution. An integrative and cooperative multidisciplinary approach to biological informatics infrastructure is crucial to ensure successful and sustainable operations, especially for support of high-throughput wet laboratory generated data. The online infrastructure concept with the combined input from molecular biologists and bioinformaticians, software and systems engineers, electronics and computer engineers is testimony to this fact. While the project may have succeeded without this integrative approach, it may not have been as rapidly operational and sustainable. Furthermore, such an approach allows the contributions of specific expertise, allowing the parties involved to continuously explore their own research niches, therefore negating the needs for certain groups to involve in retraining exercises.

ACKNOWLEDGEMENTS

We thank the National Biotechnology Directorate, Ministry of Science, Technology and Innovation, Malaysia for the funding and support to NBBnet; the NBBnet R&D team at the Interim Lab, National Institute for Genomics

and Molecular Biology, Malaysia; Sun Microsystems for a hardware grant in grid computing and Asia Pacific Science and Technology Centre, Singapore for the technical assistance in grid computing.

REFERENCES

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.
- Chothia, C. & Lesk, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *European Molecular Biology Organization Journal* 5: 823-826.
- Collins, F. S., Morgan, M. & Patrinos, A. 2003. The Human Genome Project: Lessons from large-scale biology. *Science* 300(5617): 286-290.
- Ewing, B., Hillier, L., Wendl, M. C. & Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* 8: 175-185.
- Ewing, B. & Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* 8: 186-1.
- Firdaus Raih, M., Harmin, S. A., Ahmad, H.A., Isa, M. N. M., & Mahadi, N. M., Ibrahim, A. L. & Mohamed, R. 2003. NBBnet – The National Biotechnology and Bioinformatics Network: A Malaysian initiative towards a national infrastructure for bioinformatics. *Electronic Journal of Biotechnology* 6(1), (dalam talian), Available online: <http://ejbiotechnology.info/content/vol6/issue1/issues/03/>.
- Firdaus Raih, M., Harmin, S. A., Isa, M. N. M. & Mahadi, N. M., Ng, L. K., Stoelwinder, A., See, S. & Mohamed, R. 2004. EMASGRID - An NBBnet grid initiative for bioinformatics and computational biology services infrastructure in Malaysia. *Proceedings of the First International Workshop on Life Science Grid (LSGRID2004)*, 31 May-01 June 2004. Kanazawa-City, Ishikawa, Japan, 123-124.
- Galperin, M. Y. 2004. The molecular biology database collection: 2004 update. *Nucleic Acids Research* 32(Supplement 1): D3-D22.
- Hood, L. & Galas, D. 2003. The digital code of DNA. *Nature* 421: 444-448.
- Hood, L. 2003. Leroy Hood expounds the principles, practice and future of systems biology -Interview by Stephen L. Carney. *Drug Discovery Today* 8(10): 436-438.
- Huete-Perez, J. A. & Orozco, D. A. 2001. Biotech gap between the north and south. *Science* 294: 2289-2290.
- McGinnis, S. & Madden, T. L. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids. Research* 32: W20-W25.
- Nucleic Acids Research Web Server Issue. 2004. *Nucleic Acids Research* 32 (Supplement 2).
- Stein, L. 2002. Creating a bioinformatics nation. *Nature* 417: 119-120.
- Sali, A. & Blundell, T. L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* 234: 779-815.
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22(22): 4673-4680.



Zuckerlandl, E. & Pauling, L. 1965. Molecules as documents of evolutionary history.
Journal of Theoretical Biology 8: 357-366.

Mohd Firdaus Raih, Mohd Yunus Sharum, Ahmad Fuad Hilmi Muhammad, Hafiza Aida Ahmad, Raja Murzaferi Raja Moktar, Mohd Noor Mat Isa, Nor Muhammad Mahadi & Rahmah Mohamed Makmal Genomik Biovalley
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor D. E.
mfirr@pkriscc.ukm.my
hafiza@cgat.ukm.my

