CURRENT STATUS AND FUTURE PERSPECTIVES OF BIOINFORMATICS IN TANZANIA

Sylvester L. Lyantagaye
Department of Molecular Biology and Biotechnology, College of Natural and Applied Sciences, University of Dar es Salaam, P.O. Box 35179, Dar es Salaam, Tanzania
E-mail: lyantagaye@gmail.com, lyantagaye@mbb.udsm.ac.tz

ABSTRACT

The main bottleneck in advancing genomics in present times is the lack of expertise in using bioinformatics tools and approaches for data mining in raw DNA sequences generated by modern high throughput technologies such as next generation sequencing. Although bioinformatics has been making major progress and contributing to the development in the rest of the world, it has still not yet fully integrated the tertiary education and research sector in Tanzania. This review aims to introduce a summary of recent achievements, trends and success stories of application of bioinformatics in biotechnology. The applications of bioinformatics in the fields such as molecular biology, biotechnology, medicine and agriculture, the global trend of bioinformatics, accessibility bioinformatics products in Tanzania, bioinformatics training initiatives in Tanzania, the future prospects of bioinformatics use in biotechnology globally and Tanzania in particular are reviewed. The paper is of interest and importance to rouse public awareness of the new opportunities that could be brought about by bioinformatics to address many research problems relevant to Tanzania and sub-Sahara Africa.

Keywords: Bioinformatics, Biotechnology, Genomics, Tanzania.

INTRODUCTION

Bioinformatics is conceptualising biology in terms of molecules (in the sense of physical chemistry) and applying "informatics techniques" (derived from disciplines such as applied mathematics, computer science and statistics) to understand and organise the information associated with these molecules, on a large scale (Luscombe *et al*. 2001). In short, bioinformatics is a management information system for molecular biology and has many practical applications. Bioinformatics is the discipline that focuses on the conversion of experimental data into biological knowledge and hypothesis (Bayat 2002). Bioinformatics is essential both, in deciphering genomic, transcriptomic, and proteomic data generated by high-throughput experimental technologies, and in organizing data gathered from traditional biology and medicine. Bioinformatics involves retrieval, storage, processing, analysis, and management of biological information through computational techniques. It uses mathematics, biology, and computer science to understand the biological importance of an extensive variety of omics data (Reportlinker 2013). Bioinformatics technologies are used in various pharmaceutical and biotechnology sectors. The medical sector is driven by the increasing use of bioinformatics for the drug discovery and development process. Students graduating in biology or biochemistry today ought to know their way around the Linux command line. They should also know bioinformatics tools such as what “tab” output is and how to bring it into a spreadsheet or to a database, etc.

The evolution of bioinformatics, which started with sequence analysis (has led to high-throughput whole genome or transcriptome sequencing today) is now

being directed towards recently emerging areas of integrative and translational genomics, and ultimately personalized medicine. This review aims to introduce scientists from Tanzania and elsewhere to the use of bioinformatics in their research and to present some of the bioinformatics resources available in Tanzania.

Considerable efforts are required to provide the necessary infrastructure for high-performance computing, sophisticated algorithms, advanced data management capabilities, and-most importantly-well trained and educated personnel to design, maintain and use these environments. With the presence of the STM-1 SEACOM Fiber-optic cable for broadband Internet connection in Tanzania, and given modern computer infrastructure and reliable power supply, Tanzanians could have relatively easy access to the products of bioinformatics. However, future use of this technology in Tanzania hinges on the availability of bioinformatics knowledge in the public domain.

High ignorance in the country of what bioinformatics could be useful for, suggests a need for awareness campaigns through journals and other publication media. The few bioinformaticians in the country trained oversees when they come back their knowledge of bioinformatics becomes redundant because their employers did not know what they could be useful for. It is crucial that local academic institutions take the responsibility to contribute to the production of a critical mass of bioinformaticians. This review also outlines the most promising trends in bioinformatics, potential and the current status of Bioinformatics in Tanzania, which may play a major role in the sensitization of the authorities in both public and private sectors to seriously consider investing in bioinformatics capacity building in the country.

APPLICATION OF BIOINFORMATICS

DNAs for thousands of organisms have been sequenced since the Phage Φ-X174 was sequenced in 1977 (Sanger *et al*. 1977, www.kegg.jp). Bioinformatics analyses are used to determine genes that encode polypeptides (proteins), RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species.

Today, computer algorithms implemented in programs such as BLAST (NCBI BLAST home) are used daily to search genomes of thousands of organisms, containing billions of nucleotides. These algorithms compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, in order to identify sequences that are related, but not identical. Another aspect of bioinformatics in sequence analysis is annotation, which involves computational gene finding to search for protein-coding genes, RNA genes, and other functional sequences within a genome (Korf 2004, Ratsch *et al*. 2007), but the programs available for analysis of genomic DNA are constantly changing and improving.

Other uses are genome assembly (Gascuel 2007), determination by gene expression of the genes implicated in a disorder (Wagner *et al*. 2010, Meiri *et al*. 2010), expression data can be used to infer gene regulation: one might compare microarray data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single-cell organism, one might compare stages of the cell cycle, along with various stress conditions (heat shock, starvation, etc.) (Liu *et al*. 2009, Wang *et al*. 2010). One can then apply clustering algorithms to expression data to determine which genes are co-expressed. For example, the upstream regions (promoters) of co-

expressed genes can be searched for over-represented regulatory elements (Rakhshandehroo *et al*. 2009).

Protein microarrays and high Throughput Mass Spectrometry (HT MS) are very much involved in making sense of Bioinformatics data (Simara *et al*. 2009,Yang *et al*. 2010), analysis of mutations in cancer (Kim *et al*. 2004, Lassmann *et al*. 2007), protein structure prediction (Chab *et al*. 2010, Peng and Xu 2010), establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms (Miller *et al*. 2004, Su *et al*. 2005), the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the complex connections of these cellular processes (Hucka *et al*. 2010, Larsen *et al*. 2010) and a variety of methods have been developed to tackle the protein-protein docking problem (Wang *et al*. 2007, Marshall and Vakser 2005, Uchikoga and Hirokawa 2010).

Basic bioinformatics services are classified by the European Bioinformatics Institute (EBI) into three categories: Sequence Search Services (SSS), Multiple Sequence Alignment (MSA) and Biological Sequence Analysis (BSA) (McWilliam *et al*. 2009). The availability of these service-oriented bioinformatics resources freely on the web demonstrate the applicability of web based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow. A variety of interfaces have been developed, e.g., Simple Object Access Protocol (SOAP) and Representational State Transfer (REST) for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world. The main advantages derived from the fact that end users do not have to deal with software and database maintenance overheads. A variety of bioinformatics systems are being developed for specific purposes, e.g., eBioKit project has been running for a number of years with the aim to help in sub Saharan Africa where bioinformatics is still in juvenile stages. eBioKit contains all the major open source databases and tools which can be used for teaching and for small research projects (Fuxelius *et al*. 2010). Another system that contains many tools that the eBioKit has but no databases is the BioLinux project (Field *et al*. 2006).

GLOBAL TREND OF BIOINFORMATICS

For more than a century, vast progress has been made in genetics and molecular biology. New high-throughput experimental techniques continue to emerge rapidly. The automation of DNA sequencing set the stage for the Human Genome Project in 1990 (Collins *et al*. 1998), which has led to genomics (the branch of genetics that studies organisms in terms of their full DNA sequences) and a range of related disciplines such as transcriptomics (the study of the complete gene expression state), proteomics (the study of the full set of proteins encoded by a genome), and metabolomics (the study of comprehensive metabolite profiles), sometimes all these domains are collectively referred to as genomics. Genomics has already, and will have a major impact on the advancement of society in many ways (GMIS 2008). By understanding genes, or the blue print, of how humans, plants, microorganisms and animals work, we can find solutions to some of the most pressing challenges facing us today in health, environment and agriculture. There is much potential for genomics to solve many of the major challenges we face today, and many

which are still unknown. Genomics has greatly accelerated fundamental research in molecular biology as it enables the measurement of molecular processes globally and from different points of view. This led to a range of applications in the biomedical sector and increasingly affects patient care (Collins *et al*. 2003).

One of the bottlenecks that prevent large-scale implementation of genomics in health care is the problem related to the management, analysis and interpretation of large amounts of heterogeneous data that are measured for (patient) samples (van Kampen and Horrevoets 2006), hence the emergence of bioinformatics.

In the recent years, the continuously changing market of bioinformatics applications enjoys a wide acceptance owing to declining new drugs' manufacturing, non-existence of potential drugs and expiry of patents. In 2004-05, this market with a US $1.4 billion has been competing with unusual and innovative tools facilitating critical Research and Development (R&D) for the bioinformatics companies. The "World Bioinformatics Market (2005-2010)" (RNCOS 2006) indicates that the bioinformatics market was at stability where 20% of the applications discovered were based on genomics and proteomics, which boost growth of bioinformatics tools. As per new research report "Global Bioinformatics Market Outlook", the market for bioinformatics will surge during 2011-2013 to a value of around US$ 6.2 Billion (RNCOS 2010). With the increasing R&D investment by companies on bioinformatics and the regulatory support in various countries, it is anticipate that the bioinformatics market will post high growth rate in majority of the countries. According to Reportlinker (2013), the global bioinformatics market was valued at $2.9 billion in 2012 and is poised to reach $7.5 billion by 2017 at a CAGR of 20.9%". The growth of the bioinformatics market is driven by decrease in cost of DNA sequencing, increasing government initiatives and funding, and growing use of bioinformatics in drug discovery and biomarkers development processes. It is expected that the market will offer opportunities for bioinformatics solutions manufacturers with the introduction and adoption of upcoming technologies such as nanopore sequencing and cloud computing. However, factors such as scarcity of skilled personnel to ensure proper use of bioinformatics tools and lack of integration of a wide variety of data generated through various bioinformatics platforms are hindering the growth of the market. Manufacturers of bioinformatics solutions will face further challenges with regard to industry consolidation and management of high volume data".

Reportlinker (2013) further shows that North America accounted for the largest market share of the bioinformatics market, followed by Europe, in 2012. However, Asian and Latin American countries represent emerging markets, owing to a rise in research outsourcing by pharmaceutical giants, increasing number of Contract Research Organizations (CROs), rise in public and private sector investment, and growing industry -academia partnerships. The major players in the bioinformatics market are Accelrys, Inc. (U.S.), Affymetrix, Inc. (U.S.), Life Technologies Corporation (U.S.), Illumina, Inc. (U.S.), and CLC bio. (Denmark).

IS BIOINFORMATICS ACCESSIBLE IN TANZANIA?

Public bioinformatics resources, such as databanks and software tools that are crucial for biotechnology projects, are today available via the Internet (Stajich and Lapp 2006, http://www.expasy.org/, http://www.ncbi.nlm.nih.gov). Scientists

need only a computer and an Internet connection of a certain quality to use them. In 2010, the STM-1 SEACOM undersea Fibre-optic Cable arrived on the Tanzanian coastal shores. To the country IT users connection to the SEACOM cable means an increase of bandwidth capacity. For the University of Dar es Salaam (UDSM) for instance, the increase is from 10 mega-bits to 155 mega bits per second (UDSM 2010), and this has resulted in tremendous improvement in the accessibility of the internet resources. With such resources the situation of a Tanzanian biologist is closer to that of an academic biologist in an industrialized country.

Modern bioinformatics research does not necessarily require more resources than any other field of Computer Science; almost all processes can be efficiently designed and modeled on a personal computer or workstation (Stevens 2006). If this basic infrastructure is sufficiently provided, biomathematics can be recommended to universities in Tanzania as an up-to-date and promising research subject that does not require excessive resources.

However, the step from theoretical bioinformatics to applied bioinformatics requires a supportive research and development climate that generates local need for such research. Like in many other developing African countries, little is known about bioinformatics in Tanzania.

Biotechnology industry in Tanzania is still very poor and hardly any bioinformatics is vividly talked about. There are several biotechnology research works around the country but no serious investments have been made for bioinformatics. Research groups most likely to apply bioinformatics are the Tanzania Genome Network (TGN) member groups and the Genome Science Centre at Sokoine University of Agriculture (SUA). The Gates Foundation through the Grand Challenges in Global Health funds the Genome Science Centre at SUA. The laboratory is well equipped with high-throughput facilities for genomics and good computing facilities but there is a lack of bioinformaticians.

While the government through the National Commission for Science and Technology (COSTEC) understands the potential of R&D in Biotechnology to the nation's economic growth (http://www.costech.or.tz/?s=biotechnology) , there is a dare need for awareness campaign on the significance of bioinformatics in biotechnology. It is the role of universities and other higher learning institution to design programmes, which also involves Bioinformatics in their curricula. With the recent development in registering several new public and private higher learning institutions, if all joins the initiative that the UDSM is taking a leading role in Bioinformatics training, the government and the private sector might get interested to invest. For the first time in the country "Introduction to Bioinformatics" has been incorporated as a core course in the reviewed curricula for biotechnology programmes of the UDSM commencing 2010/11 academic year (ARIS 2010). A new taught Masters of Biochemistry programme, which will also have a bioinformatics component, is being designed at the college of Natural and Applied Sciences UDSM (Not published).

COSTECH in collaboration with the Global Biodiversity Information Facility (GBIF) and other national and international partners established an online information tool, Tanzania Biodiversity Information Facility (TanBIF) in 2007, to support policy and decision making process on biodiversity and its related issues. TanBIF is an extensive, decentralized system of national biodiversity information units that intends to provide free and universal access to data and information

regarding Tanzania's biodiversity. It is a national node of the Global Biodiversity Information Facility (GBIF). The informatics tool will enable analysis and modeling of primary biodiversity data to support (biodiversity related) decision-making activities such as land use planning, design of protected areas risk assessment, among others. The tool will allow users to perform meaningful analysis by integrating and using biodiversity data (including occurrence data, species-level data/ecosystem data) in combination with other types of data (geospatial, climatic, demographic, economic datasets). This project is sponsored by the Department of Environment, Royal Danish Ministry of Foreign Affairs and Capacity Enhancement Programme for Developing Countries (CEPDEC) (http://www.tanbif.or.tz).

BIOINFORMATICS TRAINING INITIATIVES AND SUCCESS STORIES FROM TANZANIA

Bioinformatics is increasingly included in the undergraduate curriculum for biology students, globally (Barker *et al*. 2013). Teaching bioinformatics is made difficult, however, by the constraints of typical university computer classrooms. Some areas of basic bioinformatics may be taught using such classrooms, where all that is required is an Internet connection and Web browser [e.g. BLAST (Altschul *et al*. 1997) searches at the NCBI (NCBI BLAST home)]. More in-depth teaching requires the re-creation of a bioinformatics research environment, consisting of a Linux or UNIX operating system, standard GNU utilities (http://www.gnu.org), specialist bioinformatics software, and sequence databases (Barker *et al*. 2013).

It was not until 2009 the UDSM started teaching an introduction to bioinformatics course (BN 205) for BSc students and became the only training institution in the country to teach a formal degree course (MBB, 2009). From 14th to 18th November 2011, UDSM also successfully organized and hosted a Bioinformatics short course, which was sponsored by the network called Southern Africa Biochemistry and Informatics of Natural Products (SABINA). The target groups were postgraduate students, academic staff, and researchers in biological sciences and medical fields. The course also provided opportunity to postgraduate research students to ask for assistance from instructor for analysis of DNA sequences generated from their research works.

Although UDSM and some other Tanzanian institutions have for many years been conducting genomic studies, there have been limited collaboration leading to uncoordinated and duplication of research, and under-utilization of resources (expertise, equipment and other laboratory facilities). To address these issues and in preparation for future genomic revolution, researchers from different Tanzanian institutions decided to establish the TGN. TGN was established in 2011 with the aim to work together to consolidate existing and develop resources in genetic/genomic research in each institution to minimize duplication and ensure expertise and resources are utilized in an optimal manner. The TGN will facilitate collaboration between members and support development of centralized state-of-the-art facilities and resource centers to conduct genetic/genomic research (Ishengoma *et al*. 2012). TGN is made of 12 biomedical research institutions and universities in Tanzania each with one or more contact persons. The members include:

- Muhimbili University of Health and Allied Sciences (MUHAS)
- Ifakara Health Institute (IHI)
- National Institute for Medical Research (NIMR)
- University of Dar es Salaam (UDSM)

- Sokoine University of Agriculture (SUA)
- Ministry of Health and Social Welfare (MoH&SW)/National Health Laboratory Quality Assurance and Training Centre
- Hubert Kairuki Memorial University (HKMU)
- Nelson Mandela African Institute of Science and Technology (NM-AIST)
- African Malaria Network Trust (AMANET)
- Kilimanjaro Christian Medical Centre (KCMC)
- African Academy of Public Health (AAPH/ MDH)/ Harvard School of Public Health

In 2011, three TGN member institutions (MDH, MUHAS and UDSM) joined Human Heredity and Health in Africa Bioinformatics Network (H3ABioNet). H3ABioNet is an NIH-funded (2013-2017) Pan African Bioinformatics network comprising 32 Bioinformatics research groups distributed amongst 15 African countries and 2 partner institutions based in the USA. The main goal of H3ABioNet is to create a sustainable Pan African Bioinformatics Network to support H3Africa researchers and their projects through Bioinformatics capacity development on the African continent. The specific aims of H3ABioNet will be achieved through the following main activities: Training and Education, Research and Tool Development, User Support and Communication, Infrastructure Development, and Node Assessment and Accreditation (http://www.h3abionet.org/). The involvement with H3ABioNet is revolutionizing bioinformatics in Tanzania through knowledge transfer and infrastructure improvements. Through H3ABioNet, researchers and technical staff from MDH, MUHAS and UDSM have been attending specialized training courses with the aim that they will teach others at their home institutions. H3ABioNet has a plan of a series of courses throughout the funding period (http://www.h3abionet.org/); no doubt the project will leave Tanzania in a better position with regards to bioinformatics capacity.

H3ABioNet is also funding eBioKit system infrastructure and training to facilitate in teaching bioinformatics across the network. Again, UDSM was the first in Tanzania to acquire an eBioKit system in October 2013. eBioKit is a compilation of online open source databases and software hosted on the same server. All of the ensemble servers and the biomart server are hosted by different instances of Apache and some of the other server applications just behind a proxy. Due to the coding of the ensembl servers, they can't be behind a proxy server. For that reason, it's only possible to access the servers by their IP and port number (Fuxelius *et al*. 2010).

CHALLENGES OF TEACHING BIOINFORMATICS IN TANZANIA AND HOW TO SOLVE THEM

In most cases the available computer laboratories contain few functional computers to train a class, so classes huddle around few working computers. Moreover, the nations' energy supply would often be cut off, making teaching bioinformatics on a computer rather difficult. But also, there is a serious shortage of skilled personnel to teach bioinformatics in the country.

However, even if the power supply is intermittent and the Internet connections run at dial-up speed, it is still possible to conduct bioinformatics activities. Determined bioinformaticians can start a study with just a computer and open-source software downloaded through the Internet. Alternatively, acquiring an US$3000 wealthy eBioKit system will make a big difference. Meanwhile, awareness

campaigns should be constantly stages to appeal to the government and private sector to invest in bioinformatics.

Many organizations use computer clusters to maximize processing time, increase database storage and implement faster data storing and retrieving techniques (Barker *et al*. 2013). The major advantages of using computer clusters are clear when an organization requires large scale processing like in bioinformatics teaching and research. When used this way, computer clusters offer: Cost efficiency: the cluster technique is cost effective for the amount of power and processing speed being produced. It is more efficient and much cheaper compared to other solutions like setting up mainframe computers. Processing speed: multiple high speed computers work together to provided unified processing, and thus faster processing overall. Improved network infrastructure: different Local Area Network (LAN) topologies are implemented to form a computer cluster. These networks create a highly efficient and effective infrastructure that prevents bottlenecks. Flexibility: unlike mainframe computers, computer clusters can be upgraded to enhance the existing specifications or add extra components to the system. High availability of resources: If any single component fails in a computer cluster, the other machines continue to provide uninterrupted processing (Bader and Pennington 2001, Barker *et al*. 2013). Computer clustering, therefore, could be handy in bioinformatics teaching and training institutions in Tanzania by pooling together small computational resources available in various units.

FUTURE PERSPECTIVES OF BIOINFORMATICS

Any country intending to remain up-to-date in the biomedical, biotechnological and agricultural sectors, cannot disregard Bioinformatics. In addition to this general trend, developing countries may also want to manage their own specific data on indigenous biological species, on local epidemiology and biodiversity programmes. These tasks clearly require that statisticians and informatics experts become advanced users of bioinformatics software and develop a capability to solve problems locally. This process does not require large resources in it but will allow developing countries to further investigate their own biological resources. To facilitate this process biomathematics/bio-computing should be introduced to universities, and the establishment of small software groups and companies should be encouraged.

Whereas the 1990s to 2000s have been characterized by genome projects that have called for massive data processing solutions, the challenge remains in the understanding of the results. At present, advanced bioinformatics is concentrated in a few research centres and private companies around the world that have the capacity to employ personnel with highly specialized training. In spite of the fact that bioinformatics methods are freely accessible, there is clearly a gap between the developing and the industrialized world, which must be consciously narrowed. Bioinformatics is indeed the enabling technology for several fields of biomedical and agricultural research. The use of bioinformatics spreads freely through the Internet and it helps developing countries to catch up with industrialized countries. All this is based, however, on the principle that information resources worldwide remain freely accessible. If this should change in the future, it might widen the North-South gap in biotechnology.

CONCLUSION

There is unlimited opportunities provided by the next generation sequencing technologies which has made affordable genetic data generation for science and medicine even in developing countries. However, the

bottleneck in advancing these technologies consists in a lack of expertise in using bioinformatics tools and approaches for data mining in raw DNA sequences. There is no need of very large investments to overcome this limitation, but the need is in establishing of education programs, training courses and workshops for students, university staff members and medical practitioners and scientists. Universities and other higher learning institutions in the country need to face the fact that the world has changed. The paper is meant to rouse public awareness of these new opportunities to address many research problems relevant to Tanzania and sub-Sahara Africa.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ 1997 Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **25**:3389-3402.

ARIS 2010 The Academic Registration Information System (ARIS).https://aris2.udsm.ac.tz/

Bader DA and Pennington R 2001 "Cluster Computing: Applications". *Int. J. High Perform. Comput. Appl.* **15**:181-185.

Barker D, Ferrier DE, Holland PW, Mitchell JB, Plaisier H, Ritchie MG and Smart SD 20134273π: Bioinformatics Education on Low Cost ARM Hardware. *BMC Bioinformatics* **14**: 243.

Bayat A 2002 Science, Medicine and the Future: Bioinformatics. *BMJ* **324**: 1018-1022.

Chubb D, Jefferys BR, Sternberg MJE and Kelley LA 2010 Sequencing Delivers Diminishing Returns for Homology Detection: Implications for Mapping the Protein Universe. *Bioinformatics* **26**: 2664-2671.

Collins FS, Morgan M and Patrinos A 2003 The Human Genome Project: Lessons from Large-scale Biology. *Science* **300**: 286-90.

Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R and Walters L 1998 New Goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**: 682-689.

COSTECH: http://www.costech.or.tz/?s=biotechnology, 21 November 2013.

ExPasy: http://www.expasy.org/

Field D, Tiwari B, Booth T, Houten S, Swan D, Bertrand N and Thurston M2006 Open Software for Biologists: From Famine to Feast. Developing and Deploying Specialized Computing Systems for Specific Research Communities is Achievable, Cost Effective and Has Wide-ranging Benefits. *Nat. Biotechnol.* **24**: 801-803.

Fuxelius H, Bongcam E and Jaufeerally Y 2010 The Contribution of the eBioKit to Bioinformatics Education in Southern Africa. *EMBnet J*. Available at: <http://journal.embnet.org/index.php/embnetjournal/article/view/173/395>. 21 Nov. 2013.

Gascuel O and Steel M 2007 Reconstructing Evolution: New Mathematical and Computational Advances. Oxford University Press, New York.

Genome Management Information System (GMIS) 2008 Genomics and Its Impact on Science and Society, Human Genome Program, U.S. Department of Energy: A 2008 Primer.

GNU operating system: http://www.gnu.org

Hucka M, Bergmann F, Hoops S, Keating S, Sahle S, Schaff JC, Smith LP and Wilkinson D 2010 The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core (Release 1 Candidate). *Nature Proceedings* 1-167.

Ishengoma DS, Makani J, All TGN members 2013 Tanzania Genome Network: Setting up a Framework and Foundation for Genomic Research in Tanzania. The 8th Scientific Meeting of the African Society of Human Genetics (19th - 21st May 2013, at La Palm Beach Hotel, Accra, Ghana).

KEGG: http://www.kegg.jp

Kim IJ, Kang HC and Park JG 2004 Microarray Applications in Cancer Research. *Cancer Res. Treat.* **36**: 207–213.

Korf I 2004. "Gene Finding in Novel Genomes". *BMC Bioinformatics* **5**: 59-67.

Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D and Nelson KE 2012 Cloud BioLinux: Pre-configured and On-demand Bioinformatics Computing for the Genomics Community. *BMC Bioinformatics* **13**: 1-8

Larsen M, Yamada KM and Musselmann K 2010 Systems Analysis of Salivary Gland Development and Disease. *WIREs Syst. Biol. Med.* **2**: 670-682.

Lassmann S, Weis R, Makowiec F, Roth J, Danciu M, Hopt U and Werner M 2007 Array CGH Identifies Distinct DNA Copy Number Profiles of Oncogenes and Tumor Suppressor Genes in Chromosomal- and Microsatellite-unstable Sporadic Colorectal Carcinomas. *J. Mol. Med.* **85**: 293-304.

Liu X, Long F, Peng H, Aerni SJ, Jiang M, Sa´nchez-Blanco A, Murray JI, Preston F, Mericle B, Batzoglou S, Myers EW and Kim SK 2009 Analysis of Cell Fate from Single-Cell Gene Expression Profiles in *C. elegans*. *Cell* **139**: 623–633.

Luscombe NM, Greenbaum D and Gerstein M 2001What is Bioinformatics? An Introduction and Overview. Yearbook of Medical Informatics. Yale University, New Haven, USA.

Marshall G, Vakser L 2005 Protein-Protein Docking Methods. *Protein Reviews* **3**: 115-146.

MBB: http://www.mbb.udsm.ac.tz/mbb_files/molecular_biology&biotechnology1.html, 22 November 2013.

Mcwilliam H, Valentin F, Goujon M, Li WW, Narayanasamy M, Martin J, Miyar T and Lopez R 2009 Web services at the European Bioinformatics Institute-2009. *Nucl. Acids Res.* **3**7 (Web Server issue): W6–W10.

Meiri E, Levy A, Benjamin H, Ben-David M, Cohen L, Dov A, Dromi N, Elyakim E, Yerushalmi N, Zion O, Lithwick-Yanai G and Sitbon E 2010 Discovery of microRNAs and Other Small RNAs in Solid Tumors. *Nucleic Acids Res.* **38**:6234-6246.

Miller W, Makova KD, Nekrutenko A, Hardison RC 2004 Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5**:15-56.

NCBI: http://www.ncbi.nlm.nih.gov

NCBI BLAST home.http://blast.ncbi.nlm.nih.gov/Blast.cgi

NEPAD: http://www.nepad.org

Pauling L, Corey RB and Branson HR 1951The Structure of Proteins: Two Hydrogen-bonded Helical Configurations of the Polypeptide Chain. *Proc. Natl. Acad. Sci. USA* **37**: 205-211.

Peng J and Xu J 2010 Low-homology Protein Threading. *Bioinformatics* **26**: i294-i300.

Rakhshandehroo M, Hooiveld G, Müller M and Kersten S 2009 Comparative Analysis of Gene Regulation by the Transcription Factor PPARα Between Mouse and Human. *PLoS ONE* **4**: e6796.

Rätsch G, Sonnenburg S, Srinivasan J, Witte H, Müller KR, Sommer RJ, Schölkopf B 2007 Improving the *C. elegans* Genome Annotation Using Machine Learning. *PLoS Comput. Biol.* **3**: e20.

Reportlinker 2013 Genomics Industry: Bioinformatics Market by Sector (Molecular Medicine, Agriculture, Research & Forensic), Segment (Sequencing Platforms, Knowledge Management Tools & Data Analysis Services) & Application (Genomics, Proteomics & Drug Design), Global Forecasts to 2017. SOURCE Reportlinker. http://www.reportlinker.com, 15 November 2013.

RNCOS 2010 Global Bioinformatics Market Outlook. Publish Date: September 2010

RNCOS 2006 Bioinformatics Market Update, Publish Date: Jul 2006. http://www.rncos.com/Report/IM045.htm

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchinson CA and Slocombe PM 1977 Nucleotide Sequence of Bacteriophage φX174 DNA. *Nature* **265**: 687-695.

Simara P, Koutna I, Stejskal S, Krontorad P, Rucka Z, Peterkova M and Kozubek M 2009 Combination of mRNA and Protein Microarray Analysis in Complex Cell profiling. *Neoplasma* **56**: 141-149.

Stajich JE and Lapp H 2006 Open Source tools and Toolkits for Bioinformatics: Significance, and Where are We? *Brief Bioinform*. **7**: 287-296.

Stevens R. 2006 Trends in Cyber-infrastructure for Bioinformatics and Computational Biology. *CTWatch Q.* **2**. http://www.ctwatch.org/quarterly/articles/2006/08/trends-in-cyberinfrastructure-for-bioinformatics-and-computational-biology/, 5 March 2014.

Su Z, Olman V, Mao F and Xu Y 2005 Comparative Genomics Analysis of NtcA regulons in Cyanobacteria: Regulation of Nitrogen Assimilation and its Coupling to Photosynthesis. *Nucleic Acids Res.* **33**: 5156-5171.

TANBIF: http://www.tanbif.or.tz

Uchikoga N and Hirokawa T 2010 Analysis of Protein-protein Docking Decoys Using Interaction Fingerprints: Application to the Reconstruction of CaM-Ligand Complexes. *BMC Bioinformatics* **11**: 236-246.

UDSM (University of Dar-es-salaam) 2010 Connection of the University to The STM-1 SEACOM Fibre-optic Cable. *Press Release* Issued by the Public Relations Office 04/08/2010.

van Kampen AHC and Horrevoets AJG 2006 The Role of Bioinformatics in Genomic Medicine. In: Pasterkamp G and de Kleijn DPV (Eds) *Cardiovascular Research, New Technologies, Methods, and Applications* Springer, New York

Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T and Blanchette M 2010 Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human. *PLoS Comput. Biol.* **6**: 1-12.

Wang C, Bradley P, Baker D 2007 Protein-protein Docking with Backbone Flexibility. *J. Mol. Biol.* **373**: 503-19.

Wang H, Liu Y, Briesemann M and Yan J 2010 Computational Analysis of Gene Regulation in Animal Sleep Deprivation. *Physiol. Genomics* **42**: 427-436.

Yang HY, Kwon J, Cho EJ, Choi HI, Park C, Park HR, Park SH, Chung KJ, Ryoo ZY, Cho KO and Lee TH 2010 Proteomic Analysis of Protein Expression Affected by Peroxiredoxin V Knock-Down in Hypoxic Kidney. *J. Proteome Res.* **9**: 4003-4015.