EMERGING BIOTECHNOLOGIES:

BIOINFORMATICS SERVICES APPLIED TO AGRICULTURE

MARTHA DELPHINO BAMBINI
Technology Transfer Analyst at Embrapa Agriculture Informatics, Campinas, SP, Brazil
martha.bambini@embrapa.br

POLIANA FERNANDA GIACHETTO
Researcher at Bioinformatics Multiuser Lab at Embrapa Agriculture Informatics, Campinas, SP, Brazil
poliana.giachetto@embrapa.br

PAULA REGINA KUSER FALCÃO
Researcher at Bioinformatics Multiuser Lab at Embrapa Agriculture Informatics, Campinas, SP, Brazil
paula.falcao@embrapa.br

FERNANDA STRINGASSI OLIVEIRA
Analyst at Bioinformatics Multiuser Lab at Embrapa Agriculture Informatics, Campinas, SP, Brazil
fernanda.oliveira@embrapa.br

ABSTRACT

Bioinformatics is an emergent biotechnological field of study marked by interdisciplinarity and complexity. It involves the application and development of computational tools to biological data in order to process, generate, and disseminate biological knowledge. Bioinformatics is characterized by an intense generation of data and information (configured as a context of big data and e-science), associated with the need for computational resources with high processing and storage capacities and highly qualified and interdisciplinary staff, often found only in academia. The objective of this paper is to describe the organizational model and collaborative innovation activities of the Bioinformatics Multi-user Laboratory (LMB, in the acronym in Portuguese). The LMB is a facility located at the Brazilian Agricultural Research Corporation (Embrapa), the main Brazilian agricultural research public institute, formed by 46 Research and Service Centers distributed throughout Brazil and by several laboratories and business offices abroad, in America, Africa, Asia and Europe. Its mission involves to contribute to the advance of the frontier of knowledge in bioinformatics by: incorporating new technologies and enabling efficient solutions to the demands related to this field; providing access to high performance computing infrastructure and developing human skills. Considering the importance of biotechnology in the context of agricultural research, Embrapa implemented the LMB in 2011, with the purpose of increasing the efficiency of the use of computational, human and technological resources of Embrapa by providing access to bioinformatics computational resources, offering research collaboration possibilities and consultation on project design and biological data analysis. A case-study was conducted based on documentary research and interviews. The main findings of this research are: the description of the organizational model of LMB, the management team and roles; the services it provides; its access policies and procedures of customer service.

Key-words: bioinformatics; genomics; agriculture; research laboratory; multiuser

INTRODUCTION

Contemporary agriculture is characterized by an intense incorporation of emergent technologies - such as biotechnology, nanotechnology, information technologies, precision agriculture, geographic positioning technologies - as well as sustainable and ecological concepts, all applied to solve agricultural challenges.

The agribusiness, as a whole, is one of the economic segments more influenced by modern biotechnology (Silveira *et al*., 2005). The combination of the discovery of Deoxyribonucleic Acid (DNA) and the deciphering of the genetic code, in the 1960s, led to a major scientific revolution, due to the use of different technological routes centered on recombinant DNA technology, related to genetic engineering, cell fusion and bioprocessing methods. This new scientific paradigm is characterized by a high ability to modify and control biological systems at cellular, sub-cellular and molecular levels.

The literature describes two waves of innovation in biotechnology: the first wave, occurred in the 1980s-1990s and refers to the use of recombinant DNA techniques (Swann & Prevezer, 1996); the second one occurred more recently and involves the development of monoclonal antibodies (mAb) (Fernald *et al.* , 2013). The authors point out that the "first wave of biotechnology" is now reaching a saturation stage, with radical innovations and new product developments at low levels. They stress out that subsequent technologies - such as combinatorial chemistry, bioinformatics, genomics, pharmacogenetics and gene therapy - can still be seen as a new wave of emerging and growing biotechnologies that have not yet reached the maximum of their innovative potential.

The emergent field of Bioinformatics involves the application of biological data and computational tools to generate, understand, process, organize and disseminate biological information (Spengler, 2000). It is a scientific area of study marked by high levels of complexity and interdisciplinary work, characterized by an intense generation of data and information (configured as a context of *Big Data*[1] and *e-Science*[2] leading to a need for computational resources with high processing and storage capacity. This field also demands a highly qualified and multidisciplinary staff, with competences in different fields of study, such as computing, biology and life sciences (a profile often found only in academia).

Considering the importance of biotechnology in the context of agricultural research, Brazilian Agricultural Research Corporation[3] (Embrapa, 2015) has implemented in 2011 the Bioinformatic Multi-user Laboratory (LMB), located in Campinas, SP, Brazil. The LMB objectives to contribute to the advance of the frontier of knowledge in Bioinformatics by: incorporating new technologies and enabling efficient solutions to the demands related to this field; providing access to high performance computing infrastructure and developing human skills. The laboratory aimed at increasing the efficiency of the use of Embrapa´s computational, human and technological resources by providing shared corporate

[1] Minelli *et al.* (2013) point out that the *Big Data* concept differs from the regular data analysis since it involves bigger data-sets whose size is greater than the capacity of common databases to storage, capture, manage and analyze data. The advances in hardware and software made viable the storage of a large volume of data and analytical tools allow the extraction of value from the data.

[2] e-Science can be considered a transformed scientific method which: is data intensive; involves the development of different methods to collect data (from various electronic sensors) or generate data (from simulators) and to store, analyze e transform them, generating new information and knowledge supported on high processing capacity of computers (Hey, 2009).

[3] Embrapa is the main Brazilian agricultural research public institute, formed by 46 Research and Service Centers, distributed throughout Brazil and by several international laboratories and business offices, in America, Africa, Asia and Europe.

Bioinformatic services. It can also provide Bioinformatic services to institutional partners such as universities, public research institutes and agricultural private companies.

This research aimed to identify and to understand the issues related to structuring scientific organizational models able to provide: efficient use of material resources and computing infrastructure; sharing of competences and skills; qualified technological services, in order to advance the frontier of knowledge in biotechnology. A qualitative research was conducted based on a case-study protocol (Yin, 2010) regarding the organizational practices and its results obtained by LMB. The next section presents a literature review related to the field of bioinformatics and to collaborative innovation concepts. It is followed by a description of the methods employed in this research and by its results. The conclusions are presented at the end of the article.

1. LITERATURE REVIEW

1.1 Bioinformatics applied to Agriculture

The Brazilian agricultural sector undergone a process of transformation from the 1990s on, with the incorporation of knowledge and technology to agricultural production processes that have led to increases in productivity and added value to the sector´s products (Castro, 2010). Contemporary agriculture is characterized by an intense incorporation of innovations and emerging technologies - such as biotechnology, nanotechnology, information technology, precision agriculture and geotechnologies - in its productive and management processes and the adoption of principles of agroecology and the promotion of sustainable farming practices.

In this context, the concept of agricultural biotechnology has emerged, which refers to the application of biotechnology techniques in the generation of products applied to the agricultural sector. In this new scenario, new technologies and processes (such as molecular markers of DNA and transgenic techniques) are used in the selection of plants and animals which have characteristics of interest to accelerate the results of breeding programs and reduce the investment needed to obtaining practical results and the transfer of genes between organisms in order to respectively give them new properties or generate byproducts (COSTA, 2011). In the second half of the 1990s, the emergence of automated DNA sequencers employed in the sequencing of the genome of various organisms led to an exponential growth in the number of sequences to be analyzed and stored, requiring computing resources each time more efficient and robust. So, in addition to data storage needs, it came into being a demand for computational processing power employed to store, manage, process, analyze and interpret genomic data with speed and accuracy.

The emergence of sequencing platforms called NGS (Next Generatio Sequencing), which occurred around 2005, triggered a greater demand for statistical methods and bioinformatics tools for managing and analyzing the massive amount of data generated by this new technology. In agriculture, the sequencing of the genome of animals and plants has the potential to bring enormous benefits to society at large. From the generated sequences, bioinformatics tools are used to identify, within these genomes, genes responsible for traits of economic interest. This knowledge can then be used to produce plants more tolerant to drought, pests and diseases, and to render the livestock production more healthy and productive improving the quality of their products (such as meat, milk, wool and eggs).

With the advancement of available technologies for testing, processing and generating biological information, other types of data and knowledge have been created in order to understand the organisms at a systemic level, studying the genes as parts of a complex network (and not separate entities) involving its expression in the organism and their interrelationship with the other genes. We live now at a Post-genomic Era (Espindola *et al*., 2010; Souza *et al*., 2014). In this context, the Omics Sciences were developed, referring to the overall assessment of biological systems through various sub-disciplines such as genomics, transcriptomics, proteomics and metabolomics[4].

Genomics was the first Omic to emerge, aiming to assign useful biological information for each gene and improve the understanding of how the different biological molecules contained within the cell combine to make viable organism. The need to understand an organism in a systemic way motivated the emergence of other omics (each with its own set of tools, techniques, software and databases), allowing, in addition to gene identification, the understanding its expression in organism and its interrelationship with other genes. Souza *et al*. (2014) consider that a key tool for the development of research projects in the Post-genomic Era is **bioinformatics.**

With the use of techniques, tools and expertise of bioinformatics it becomes possible to simulate the relationships between the different levels of Omics to generate economically useful knowledge and aplicable products *i.e.* innovations such as: the explanation of the molecular basis of some diseases; the identification of targets to improve strains; the understanding of how pathogens interact with living organisms; the generation of useful information for pharmacological studies; and the development of new industrial high value compounds destined to chemical, pharmaceutical and agronomic sectors, among others.

1.2 Bioinformatics: processes of knowledge generation and application areas

Prior to the emergence of bioinformatics, biological research was carried out from two main lines: using a living organism (***in vivo***) or in an artificial system (***in vitro***). Bioinformatics performs a biological study in a virtual system - ***in silico*** - using computers and computer tools to organize, analyze, integrate, process, model, simulate and store large volumes of data derived from *in vivo* and *in vitro* experiments. Bioinformatics arose from the need for automation of research processes in biology (in particular molecular biology) supported by the possibilities offered by increased data processing and storage capacity due to the revolution of Information Technology (IT) (Bongiolo, 2006).

Bioinformatics proposes new forms of scientific knowledge generation based on *in silico* experimentation that allow: the analysis of gene expression; assembly and analysis of genomes; the identification of molecular markers; the promotion of evolutionary studies; biological systems modeling; the prediction of protein structure and molecular interaction; performing interaction tests; inhibition or excitation of molecules; and creating inhibitors, and interference molecules, among other activities (Bongiolo, 2006; Espindola *et al*., 2010).

Bioinformatics handles heterogeneous data formats from structured and unstructured texts, to images, diagrams and drawings, raw genomic data (such as sequences and annotations), protein structures, gene expression profiles, diagrams, etc. (Romano *et al*. , 2011). In addition, the information available has grown exponentially along with the improvement of the means for storage and data analysis. Even só, there are still difficulties of intercommunication and interoperability.

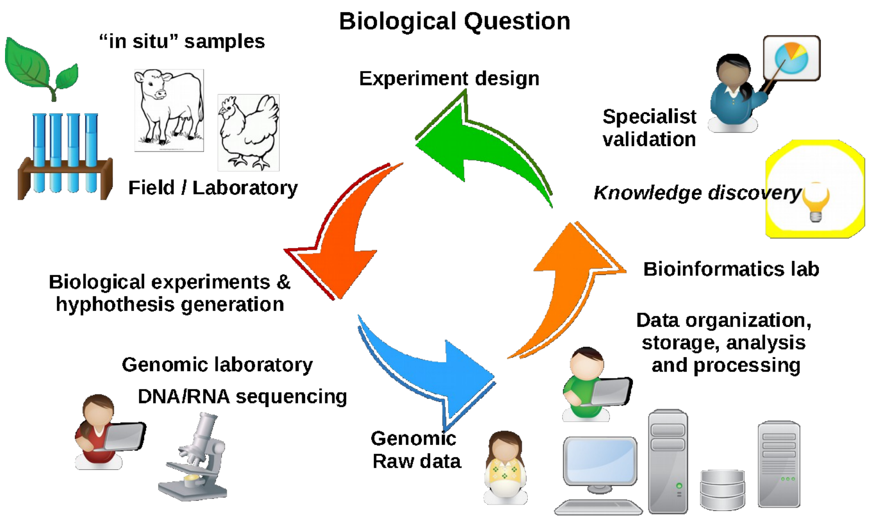
[4] According to Souza et al. (2014): (i) **genomics** corresponds to the acquisition of data relating to the genome, i.e. the full sequence of the genetic material called DNA (deoxyribonucleic acid) of an organism; (ii) **transcriptomics** relates to the knowledge of the transcriptome that is required by the cells (complete set of transcripts of a given organism, organ, tissue or cell line - messenger RNAs, ribosomal RNAs, transport RNA and microRNAs); (iii) **proteomics** refers to a systematic analysis of proteins by determining the identity of proteins of an organism and understanding their functions; (iv) and **metabolomics** involves an impartial quantitative and qualitative analysis of the complete set of metabolites present in cells, body fluids and tissues (called the metabolome).

The bioinformaticians employ IT tools to store, retrieve and process biological data, associated with statistical methods to analyze this data. Large part of the experiments and analysis performed *in silico* involve the sequential and/or parallel use of various software and the access and the query to various databases. Bongiolo (2006) indicates that projects to study genomes are initiated at the sequencing phase executed in biological laboratories, which generates sheets of raw data describing the DNA sequences, but still with no biological meaning. To analyze these sequences, the bioinformaticians employ various tools and computer programs. Usually these programs are executed manually by scientists or through the use of computer commands in the form of scripts. Although this approach allows a certain degree of automation, there are still deficiencies in regard to issues such as flexibility and interoperability.

The figure 1 describes the supplementary role of bioinformatics in studies and research of biological nature. The increasing volume and distribution of data sources and the implementation of new processes in bioinformatics facilitated the analysis phase. However, there is still significant demand for tools and semi-automatic to handle such volume and complexity systems.

Some important research fields in bioinformatics involve the research related to: databases; biological data analysis (applying data and text mining concepts) and the development of software and systems for data processing and for the implementation of pipelines[5] e scientific workflows[6]. Bongiolo (2006) points out that scientific workflows represent an interesting alternative to structure experiments in bioinformatics, providing the necessary support to the cycle of "execution and analysis" inherent in the processo of searching biological knowledge. Combined with Web services technology, they enable the creation of an environment of independence and interoperability between the various scientific applications and different databases to be used.

Figure 1: Bioinformatics applied in the generation of biological knowledge



Technical Note: Figure developed by the author

[5] A pipeline data processing is a partially ordered set of computational tasks, coordinated to process a large data set (Cingolani *et al.,* 2014).

[6] A scientific workflow is a set of sequential steps set to analyze a real process, in the form of a macro system that defines, executes and manages scientific applications by using different software tools. The workflow allows researchers with little programming skills to build complex applications for biological data analysis.(Emeakaroha *et al.*, 2013)

The use of bioinformatics tools for knowledge extraction from biological data is a complex activity. The study and analysis of more complex and comprehensive biological issues requires collaboration between different research groups located in different organizations, either to exchange knowledge and experiences or to share computing infrastructure for processing and storage.

The development of bioinformatics networks in Brazil was driven by the initiative of the São Paulo Research Foundation (FAPESP) in creating a virtual institute called Onsa network (Organization for Nucleotide Sequencing and Analysis) initially gathering 30 laboratories pf institutions located in São Paulo State. This institute supported the "FAPESP Genome Program" which ran from 1997 to 2008, providing resources to suppont research activities; to implement grants and to create the necessary infrastructure for sequencing and analysis of human genetic material and organisms of scientific interest. Were studied under the program the following genomes: sugar-cane, functional genome of the ox, human cancer, *Schistosoma*, *Xanthomonas*, *Xylella fastidiosa* and some agronomic and environmental genomes (Rede, 2015).

In early 1990s it was created the first bioinformatics lab in Brazil: the Bioinformatics Laboratory (LBI) of the Institute of Computing (IC) at the State University of Campinas (Unicamp). From the launch of the Onsa network until its consolidation, it was formed a critical mass of researchers and institutions working in genomics studies in the country, with the consolidation of several research centers in bioinformatics. The Ministry of Science Technology and Innovation (MCTI) also promoted incentives (through the National Council for Scientific and Technological Development - CNPq) to the development of laboratories and training of bioinformaticians by financing projects to conduct genomic studies of organisms of social, economic and regional interest.

The Brazilian Association of Bioinformatics and Computational Biology (AB3C) was established on June 12, 2004 as a scientific society dedicated to the encouragement of research and interaction of multiple related areas experts.

This section described the very wide field of research that can be developed using bioinformatics concept and tools. Considering the high diversification of applications and fields of research, cooperative and colaborative scientific and innovative strategies have to be employed - nationally and internationally – to integrate knowledge and to share resources, reducing time needed to data processing and increasing the efficiency and quality of generated results (Bongiolo, 2006).

1.3 Collaboration for Innovation

It is well established in the literature that the innovative process has become more and more structured as a collective initiative, rather than the result of isolated efforts of inventors and scientists. Some motivations for this trend are: a substantial increase of complexity in technological development processes; the asymmetry in access to various sources of knew knowledge and information; the limitation of resources (financial, human, infrastructure, and so forth), and the elevated risks associated to innovative endeavors.

Several authors - since Kline & Rosenberg (1986) up to Chesborough (2003) – have described the interactive, adaptive and multifaceted nature of innovative processes. In the 1990s, the literature enphasizes the collaborative aspect of innovation processes, that cross the organizational boundaries of the firm and involves the formation of alliances, cooperation and collective arrangements between several actors, such as research institutes, universities, private companies and government agencies.

Powell *et al*. (1996) indicate some motivations to establish a cooperative agreement for innovation: to **gain faster access to new markets or technologies**; to benefit from **economies of scale** resulting from joint research or joint production; to **reach sources of know-how and expertise** located outside the boundaries of the firm; and to **share the risks** of activities that are beyond the scope or capacity of a single organization.

Aruja (2000) points out some of the benefits of networking for innovation such as: **resources sharing**; **combination of knowledge**, **skills and physical assets** of several partners; direct access to **spillovers** that provide information on **new discoveries** or insights to solve problems. Each network is shaped and customized according to: the interests and needs of the parties; the types of skills of each actor involved, and the types of resources to be shared. Institutional and economic context as well as tax and regulatory issues can also influence the morphology of the arrangement to be formed.

Some factors encourage collaboration among organizations: repeated interactions between partners, the possibility of future partnerships and reputation (what becomes a reference of the reliability of the partner). In these cases, there is little need for hierarchical supervision, because the desire to continue the partnership discourages opportunism; in this case, the network shall be monitored by the partners themselves (Powell, 1990).

In the 2000s, the “Open Innovation” approach - coined by Chesbrough (2003) - has been adopted by several companies as a strategy to achieve business competitiveness. This paradigm assumes that firms can and should use external sources of ideas and internal as well as internal and external channels to market in order to advance their technologies.

The networks formed to decode the genome of several organisms - such as the Human Genome Project, the *Xylella fastidiosa* project, the Onsa network – exemplify some cooperative initiatives for the generation of new knowledge and innovation. These networks are collective arrangements marked by interaction, adaptation and negotiation among several actors, by shared resources and competences and by coordination strategies and organizational practices, established to promote collaborative innovation.

2. METHODS

The research question for this research was: considering the scientific context of Bioinformatics, which practices could be employed by a Public Research Institute to promote collaborative innovation through its Bioinformatics Multi-user Laboratory? (LMB, 2015). To answer this question, a qualitative research was conducted based on a case-study protocol (YIN, 2010).

3. RESULTS AND DISCUSSION

3.1 Antecedents: Bioinformatics at Embrapa

Since 2001, with the continued hiring of bioinformaticians, and from 2003 on, with the conduction of Bioinformatics Workshops to promote the interaction of Embrapa´s researchers working in Bioinformatics, the company institutionally has been recognizing the importance of this topic and articulating investments and scientific efforts in this domain. The development of corporate competencies in Bioinformatics is essencial for several Embrapa´s scientific initiatives, such as those related to the characterization and conservation of genetic resources and genetic improvement programs related to cattle and plants.

In 2003, the project "BIEM: Bioinformatics serving Embrapa Project and the Creation of Bioinformatics Network" was initiated funded by Embrapa Management System. This project had the main objective of establishing a technological base in bioinformatics and computational biology for the Embrapa and its partners, in the form of a bioinformatics network ensuring the generation of new products and process improvement related to the advancement of research in biology molecular.

The project had the following goals: (i) aggregate pre-existing skills at Embrapa in order to form a Bioinformatics network able to meet the growing demand of the projects conducted in within the company; (ii) make friendly tools available to network users, in order to enable the management of information provided by users and the information derived from the analysis; (iii) standardize, through the development of new software and promotion of user training, the application of these tools, making them available to all decentralized units of Embrapa; (iv) promote the basis for research on genome annotation area and structure-function relationship of macromolecules; (v) expansion of existing tools for bioinformatics; and (vi) training of different publics interested in applicatives in Bioinformatics and Computational Biology areas. This project created the basis for what in 2011 would become the Multi-User Laboratory of Bioinformatics (LMB).

In 2004, the first Embrapa´s Bioinformatics Workshop was organized with the following objectives: planning the management actions and implementing a Bioinformatics Network at Embrapa; establish evaluation criteria; disseminate basic bioinformatics knowledge among team members; promote the integration of groups and activities of the project and its action plans; and discuss the network dissemination mechanisms. In 2005 it was promoted a course of Bioinformatics Genomics emphasizing the applied aspect of the subject. In 2006, it was promoted the course of Structural Bioinformatics to present methods and *in silico* tools to study the relationship structure-function of proteins including analysis of structures, structure prediction, stability study, specificity, protein-protein interactions, among others.

By 2006, the genome sequencing process became fast and inexpensive promoting great cost savings. In this year it was conducted the 2nd Embrapa Bioinformatics Workshop which generated a document synthetizing the actions to be taken towards merging and integrating bioinfomatics activities at Embrapa, optimizing resources and creating a collaborative research platform. The project "PIBA - Research and Innovation in Bioinformatics for Agribusiness" was submitted to the Embrapa Management System (SEG). The project sought to facilitate bioinformatics solutions applied to agriculture and had the following goals: development of human resources by conducting training courses in bioinformatics; development of tools to assist the interpretation of data generated by genome and post-genome projects developed by Embrapa and its partners; work in data mining area from the standpoint of genomic and protein sequences; implement studies related to prediction, modeling, molecular dynamics of protein of interest; and establish a communication strategy and dissemination of Bioinformatics in within the company.

In 2007 it was created the Applied Bioinformatics Laboratory at Embrapa Informatics for Agriculture, and the Research, Develpment and Innovation (RD&I) Animal Genomics Network led by Embrapa Genetic Resources and Biotechnology (located in Brasília). In 2008, it was promoted the 3rd Embrapa´s Bioinformatics Workshop, which generated a proposal for RD&I Program of Bioinformatics to be created in the company, with the elaboration of a joint report on the Bioinformatics Platform.

In 2009 it was announced, by the President-Director of Embrapa, the creation of the LMB, followed by an investment of R$ 400,000.00. In 2010, there was another investment of R$ 1,000,000.00 for the assembly and infrastructure of the laboratory. In 2011, it was promoted the 4th Embrapa´s Workshop on Bioinformatics and set up a working group to define a deployment and operation plan for the multi-user laboratory bioinformatics, which was formally inaugurated in the same year.

3.2 Colaborative Innovation and Multi-User Labs at Embrapa

Silva Jr. (2014) addresses the relationship between sharing laboratory infrastructure and cooperative activities related to RD&I. The author highlights the significant increase in RD&I cooperative activities as a result of increasing complexity and the risks and costs associated with innovation process. The author specifically describes public-private partnerships (between public research institutions and firms) formed with the motivation of: having access to expertise that can not be generated in the firm; sharing and exchange of external resources; exchanges of knowledge and technologies; organizational learning; combination of complementary assets, among others.

Recently, Brazilian funding agencies have been opening calls for the creation of multi-user laboratory facilities like the call FAPESP/MCTI/FINEP/CT-INFRA2013 that aimed to develop the institutional research infrastructure, through the acquisition and support of multi-user research facilities, and the improvement of institutional research infrastructure. The Public Call MCTI/FINEP/CT-INFRA-PROINFRA-02/2014 aimed to support the acquisition of multi-user equipment.

Since 2008, Embrapa has been promoting some actions to map, assess, optimize and share its laboratory infrastructure. In that year a working group was created in order to elaborate a diagnosis of the current situation of the network of laboratories of the company and to propose actions to improve their efficiency and effectiveness, as well as to the modernization of processes, equipment and facilities, based on the goals and strategic directions of the company. This action was undertaken in association to the Growth Acceleration Program (PAC-Embrapa) implemented by the Federal Government. This survey selected the following priority issues were selected: recruitment of qualified human resources; adoption of quality management principles specific to the context of the company promoting greater reliability and traceability of results; promotion of efficiency and effectiveness of laboratory infrastructure; reduction of fragmentation of laboratories; multiplication of multi-user structures; integration among units and adequate infrastructure sharing; implementation of reference laboratories related to strategic issues.

In 2012, Embrapa initiated the development of a structuring project entitled "Strengthening the infrastructure of experimental fields and laboratories" which aimed to modernize, adapt and optimize the infrastructure of experimental fields and physical laboratories of Embrapa Units with emphasis on quality and relevance of the company's results, following the applicable rules and gearing up the guidelines of the environmental management and the new forest code. This project reinforces the importance of sharing concept and integration of Embrapa research structures, emphasizing multi-user laboratory facilities in order to promote synergy and reduce costs, waste and redundancies.

The LMB was not directly covered in the activities of this project, but it is part of Embrapa´s initiatives to share infrastructure, skills and optimize the use of company resources. In the context of Embrapa, the multi-user laboratories are those used for testing activities and analysis of high scientific complexity, involving multidisciplinary technical teams and highly specialized equipment (VAZ *et al.*, 2012). The company understands that these laboratories should work as shared research platforms available for use by multiple research units, and partner institutions.

3.3 Bioinformatics Multiuser Laboratory (LMB) at Embrapa

3.3.1 Motivation

Embrapa´s portfolio of RD&I projects encompass several actions related to: pre-breeding and genetic improvement of plants, animals and micro-organisms of economic interest; the molecular characterization of genetic resources and the discovery of products with biotechnological potential. A fundamental result obtained from all of these projects is the discovery of information related to the genome and/or transcriptome of the studied organisms, enabling the execution of subsequent research steps to obtain various products and applications such as: the development of species with enhanced economic characteristics; new biotech assets and improved technologies for the conservation and use of genetic resources.

With the advent of the "new generation of sequencing technologies", an unusually large volume of genomic data was produced, at a much lower cost (compared to the data obtained with the previous technologies). So genomic research activities generate now a huge amount of data which required multidisciplinary expertise and high-performance computing resources for its storage, management, processing and analysis. Thus, the challenge related to obtaining information of genetic origin moved from the generation of genomic data to the proper analysis and biological interpretation of large amounts of data. This paradigm shift in biological research strongly introduced and consolidated bioinformatics studies at Embrapa´s RD&I projects related to genetics studies.

This new scenario highlighted two major bottlenecks to Embrapa: the shortage of qualified human resources in the area and the resource limitations for the purchase of high-performance computational structure for storing data generated and performing bioinformatic analysis. Considering the high cost of acquisition and maintenance of computing infrastructure, it was unfeasible to create similar structures in each of 46 Embrapa´s research units. And neither existed - in the company - a compatible number of qualified bioinformaticians to act in só many research centers. Given these limitations, Embrapa has taken some strategic decisions to promote the optimization of existing research resources in order to maintain the competitiveness of the company. Thus, the Executive Board of Embrapa opted to build a multi-user bioinformatics laboratory structure, with the main core staying at Embrapa Agriculture Informatics, denominated the Multi-User Bioinformatics Laboratory (LMB).

The LMB was established by Deliberation Nº 55, 19/09/2011, published in the Bulletin of Administrative Communications (BCA) Nº 50, of 24/10/2011. Through the Normative Resolution No. 16, of 22/09/2011, the Implementation and Operation Plan of LMB was also approved (BCA, 2011).

3.3.2 Mission and Objectives of the Laboratory

The bioinformatics research activities involve mainly receiving, storing, processing and analysing raw data collected by laboratories operating in the sequencing of DNA and RNA. To ensure the efficiency of the field work, it is necessary to provide good planning and proper alignment of the experimental design so that it has an adequate amount of samples and that they contain the necessary characteristics for generating the raw data to be analyzed.

For the implementation of processing and analysis of this data, it is required the access to an infrastructure of high-performance computers with specific software for processing large amounts of genomic data, such as LMB.

LMB has the mission of provide bioinformatics solutions for research, development and innovation at Embrapa in a collaborative environment (LMB, 2015). The Laboratory provides consulting, training and collaborative services in bioinformatics area, and also provides shared access to its computer park. The LMB has a diverse team of researchers trained to work collaboratively on different areas of research such as genetic resources, pre-breeding, genetic improvement and biotechnology. The laboratory has three main lines of action:

- **Provide access to high-performance computing infrastructure** and software for genomic data analysis, with simplified usage policies;
- **Consulting and collaborative activities in biological data analysis** that require high-performance computing, either by the volume of data, or by the complexity of the analysis;
- **Training,** aiming to multiply skills through course promotion and other training activities.

In addition to these activities, the LMB has its own lines of research, developing software, workflows and providing training to the scientific community related to its tools and different types of analyzes.

3.3.3 Competences, Resources and Relationships at LMB

The LMB´s team has a diverse background including the following areas: biology; computer science; physics; applied mathematics; animal science; computer engineering; and systems analysis. All researchers have PhD degrees, and part of them specifically in the area of bioinformatics (a recent field of study). The laboratory staff today consists of 8 researchers and analysts Embrapa Agriculture Informatics and 9 interns and graduate students.

The Lab is structured to meet demands related to: assembly and annotation of genomes, transcriptomes and metagenomas; data analysis of gene expression (microarray, RNA-Seq); genotyping data analysis (SNP chips and sequencing), and the identification of molecular markers. Considering the complex nature of these analyzes and the need for qualified human resources for carrying out such processes, the LMB has acted as a fundamental asset in various research projects conducted by Embrapa and its partner institutions, evidencing its multidisciplinary, complementary, collaborative and innovative character.

Currently, the LMB team participates in more than 30 research projects, funded by Embrapa itself and by other public funding agencies such as CNPq and FAPESP involving more than 20 cultures and creations. The role and competences of LMB were communicated to the research units of Embrapa and to the scientific community. These publics demand several collaborative actions to LMB including collaborative research projects. Researchers at LMB perform various roles in these projects from: simple collaboration, to performing research activities, leading grups of actions and also entire projects, with activities directly or indirectly related to the area of bioinformatics.

Some of LMB users are research units of Embrapa such as: Embrapa Agroenergy, Embrapa Rice and Beans, Embrapa Coffee, Embrapa Dairy Cattle, Embrapa Cassava & Tropical Fruits, Embrapa Maize and Sorghum, Embrapa Southeast Livestock, Embrapa South Livestock, Embrapa Genetic Resources and Biotechnology, Embrapa Soybean, Embrapa Wheat, Embrapa Swine & Poultry. Other users are: universities (such as University of São Paulo -USP), the Genomic and Expression Laboratory (LGE) of Unicamp, Universidade Estadual Paulista - UNESP, the Fundação Universidade Federal de Mato Grosso do Sul - UFMS, Fedral University of Uberlândia - UFU, Federal University of Minas Gerais – UFMG and Federal University of Rio de Janeiro - UFRJ); Public Research Institutes (such as IAPAR, the Center of Excellence in Bioinformatics - CEBIO, National Laboratory of Biosciences - LNBio, the FAPESP Centralized Multi-User Laboratory); and private companies such as Sadia and cooperatives in beef cattle area.

3.3.4 LMB Organizational Model and Practices

3.3.4.1 Responsibilities of LMB´s Technical leader

The laboratory management activities are performed by a **technical leader** who works at the interface between the LMB and its users, forwarding their demands and defining the responsible teams. With regard to **requests to use of the computational resources** of the LMB, the technical leader makes an feasibility assessment of the request and, if feasible, demands the opening of the user account. In relation to the **requests for collaboration in research projects**, the leader first talks to the requester, who presents the project and the proposed collaborative activities. Then, the leader presents the to the LMB team, in order to determine how the activities would be distributed among members of the team.

Another important LMB action is related to **training**, that are organized and taught by LMB´s researchers aiming to meet the specific demands of the scientific community and disseminate new tools and analysis methodologies implemented in the LMB. The demand for training can be directly received from partners in project meetings. The LMB team also contribute to the post-graduate program in Genetics and Molecular Biology of the Institute of Biology (IB) of Unicamp. The cooperation involves the participation of Embrapa researchers in teaching activities and students supervision. Since 2011, the discipline "Special Topics in Genetics - Computational methods in Bioinformatics" has been offered to students of the program.

In addition to the processing of users demands and the management of partnerships, the technical leader has an interface with the Director od Embrapa Informatics for Agriculture and the Executive Board of Embrapa offering information on the operation of LMB and its results, as well as about the use of computing infrastruture, the users and the collaborations undertaken. Other responsibility of the technical leader refers to: prospection of financing sources to the purchase of computing infrastructure, staff training and participation in scientific conferences and as well as to hire employees, interns and grant fellows. The technical leader also acts as an interface between the LMB of the advisory committee, scheduling meetings and discussing with them several issues associated with the laboratory.

The advisory committee ir formed by seven researchers from Embrapa and partner institutions. The responsibilities of the advisory committee are: to contribute to the formulation of a strategic agenda for bioinformatics;to act in promoting the LMB; to analyze demands when solicited; to prospect scenarios and to internalize new technologies, and to suggest strategies of action, research and investment (BCA, 2011).

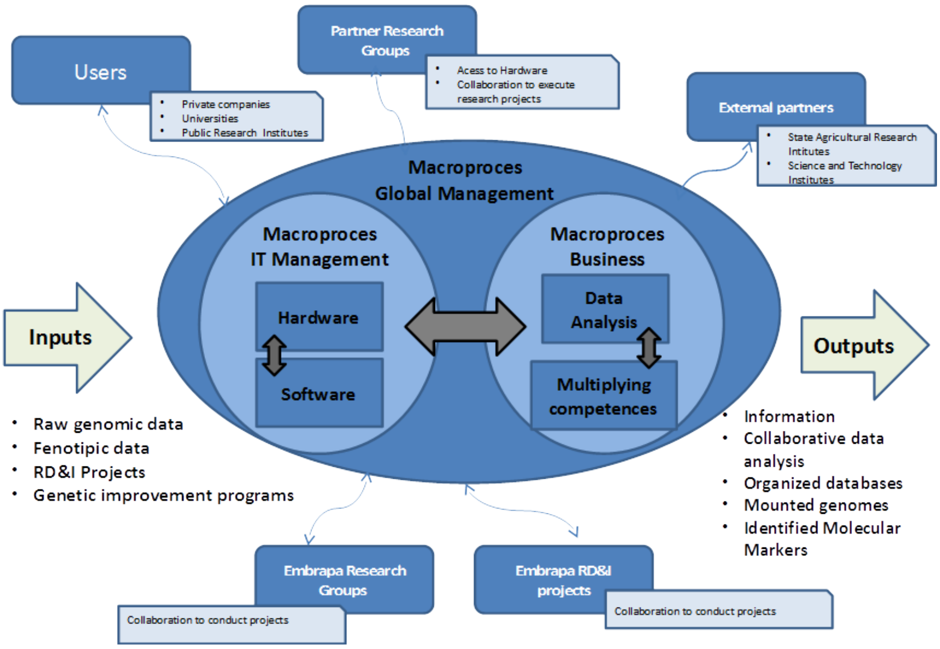
In 2013, Embrapa implemented a **corporate policy related to the management of its multi-user laboratories**. This regulation defines the conditions, rules and procedures for the use of Embrapa Multiuser Laboratories. These procedures consider, incisively, the interface of these laboratories with external institutions and the legal implications associated with collaborative innovation, such as the ownership jointly generated technologies, the definition of responsibilities and activities of each party under cooperative action and issues associated with the actual management of multi-user laboratory.

3.3.4.2 Management of user processes at LMB

The LMB management activities involve: receiving and controlling demands of projects and services; hiring staff (interns and fellows); acquiring inputs and equipment; and meeting with the advisory comitee and potential partners committee. Figure 2 describes the organizational model of the LMB and the ways to access the inputs and outputs as wll as the established relationships. The activities of providing services follow a service model, developed around the lab website which can be accessed at: www.lmb.cnptia.embrapa.br/

Through this website, potential customers can consult the LMB access policy, described in a very simplified way in order to facilitate the accessof users both to the infrastructure as well as to collaborative analysis in bioinformatics.

Figure 2: Organizacional Model and Macroprocesses – LMB (Vaz et al., 2012)


Partner Research Groups
Acess to Hardware
Collaboration to execute research projects
Users
Private companies
Universities
Public Research Institutes
External partners
State Agricultural Research Intitutes
Science and Technology Institutes
Macroproces Global Management
Macroproces IT Management
Hardware
Software
Macroproces Business
Data Analysis
Multiplying competences
Inputs
Raw genomic data
Fenotipic data
RD&I Projects
Genetic improvement programs
Outputs
Information
Collaborative data analysis
Organized databases
Mounted genomes
Identified Molecular Markers
Embrapa Research Groups
Collaboration to conduct projects
Embrapa RD&I projects
Collaboration to conduct projects

There are two main ways to access the LMB.

The first way refers to the **access to computing infrastructure**. Initially, the potential user should check the "Infrastructure" section of the website which describes the resources provided by the LMB (computational infrastructure and software). If it considers that the LMB has the needed conditions for the implementation of the selected analysis and that the equipment meets the user's need, the user must complete Form A, summing up information related to research design and the analyzes that would be made. The form will be reviewed by the technical leader, and based on the information described it will be defined the period of access authorization.

In order to proper manage and control de access and usage of LMB servers, there is a record with access information for each user account created, and a work area controlled by appropriate permissions. From the moment the access account is created, the user has remote access to a server where tasks can be submitted. Tasks that require high processing time must be submitted through a queue manager, which will set a priority of execution and manage the distribution of tasks among servers.

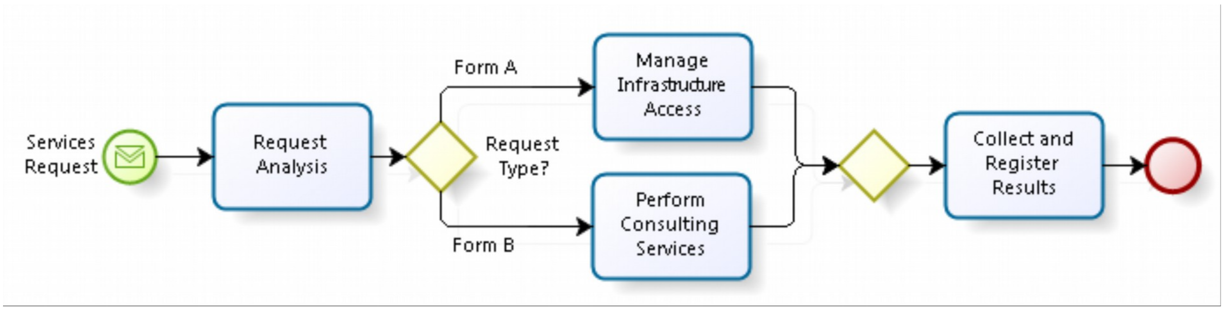
The use of computing infrastructure is periodically monitored and metrics are collected for analysis of the demand for processing and storage in relation to current server capacity for forecasting and planning upgrade of machinery, equipment and software. Systems management activities are of great relevance and impact on the operation of the LMB, in order to ensure the maintenance and availability of the computing infrastructure to its users and also to to be prepared to handle a large volume of data in various formats.

BML has currently a computacional infrastructure is composed by: 2 control nodes and 8 computer nodes totalizing 576 cores of processing power and 512 Gb of RAM by node, in general, but until 2Tb in some nodes; file storage capacity of 211Tb; and backup system composed by 48 LTO06 tapes with capacity of 300Tb of compressed data. All these resources are connected by a Gigabit Ethernet for management network, 1 NAS network 10 Git/s SFP+ and 1 SAN FC/8Gbs for data network.

The second way to access LMB refers to **collaborations in research projects and consultancy activities** involving research design, bioinformatics analysis and the processing of genomic data. In this case, the user must complete the Form B, describing the research design details. The consultancy may involve: from support activities to the experimental design for future data collection to execution of data analysis. The technical leader verify if the LMB is able to perform the job and, if so, a member of the team (or more) will be allocated for the task.

In both cases, the forms will be quickly evaluated, and the coordination of LMB comes into contact with the requesters. Figure 3 describes the managemtn of user processes at LMB.

Figura 3: Management of User processes at LMB


Services Request
Request Analysis
Request Type?
Form A
Form B
Manage Infrastructure Access
Perform Consulting Services
Collect and Register Results

Technical Note: Figure developed by the author

3.3.4.3 Executing biological data analysis

The execution of biological data analysis begins with the receipt of data to be analyzed. Generally the data is send in large files (with the size if gigabytes or more), which involves challenges regarding the security of transfering large volumes of data and information over the Internet. It demands the use of transfer tools and specific network settings for the proper functioning and ensuring the integrity of data received. It is also important to promote the control of the received data, which need to be stored in a standardized structure, with metadata for data categorization and access control.

Once the data set has been received and stored in the LMB, the researcher in charge need to prepare the computacional environment fot the analysis, providing the installation of bioinformatics software needed to run a specific pipeline (a set of computational tasks partially ordered and coordinated to perform to process the data set to be sent by the user , according to Cingolani *et al.,* 2014).

Most programs are free open-source sofwares and can be run via command line (scripts). Moreover there are also scientific workflow systems that define the sequential steps to analyze a real scientific process, selecting, executing and managing scientific applications. One example is the Galaxy software is a web-based platform that wraps several bioinformatics and allows experimentalists without informatics or programming expertise to perform complex large-scale analysis with just a web browser (Blankenberg et al, 2010; Cingolani et al, 2014).

After the pipeline is executed, it generates several result-files in various formats with data and metrics delivered via file upload. Given that these files contain lots of information, to facilitate the interpretation of results, the data is imported to web tools, for searching visualizing and integrating of genomic data, such as Blast (Johnson et al., 2008), Genome

Browser (Stein et al., 2002) and Intermine (Kalderimis et al., 2014).

In order to improve its scientific processes, the LMB team also develops software and workflows. The LMB development team generated customization of some free tools for internal use: (i) Galaxy; (ii) System Queue Manager (SGE) a tool executed by command line and used for monitoring and submitting jobs in the queue of LMB processes[7]; (iii) GBrowse is a web system and database for handling and genomes and annotations consultation; and (iv) Ganglia is a performance monitoring system computing environments.

Some of the developed software are: (i) KOMODO: for the detection of homologous genes significantly under- or over- represented among táxons; (ii) POTION: positive selection for and detection of genes involved in adaptive processes; (iv) JM-CNV: a precise algorithm and quick to combination of CNVs (Copy Number Variation) that overlap in a significant number of CNVRs (Copy Number Variable Regions) discrete. CNVs are regions of the genome that have a variable number of copies between individuals in a population, being used as molecular markers.

3.3.5 Results obtained at LMB

The LMB has today 31 active users (with ongoing activities) for access to its computing infrastructure and 13 new user accounts (and 30 closed user accounts). The accounts created correspond to users linked to five external institutions (USP, UNESP, Unicamp, UFRJ, UFMG) and 14 research units of Embrapa: more than 60 users have accessed the infrastructure and competences of LMB. In terms of scientific collaboration in data analysis, only in 2014, it were offered 32 analysis in the following categories: transcriptome and genome assembly; SNPs identification; quality analysis sequences; mapping reference-genome sequences; differential expression analysis; automatic annotation of genes; functional analysis of genes; quality analysis sequences; microbial diversity analysis; identifying repetitive sequences; CNVs identification; prediction of protein-coding genes.

Several genome were sequenced such as: nelore ox (*Bos taurus indicus*); caiaué (*Elaeis oleifera*); tambaqui fish (*Colossoma macropomum*). Some studies to identify genes of agricultural interest were conducted: eucalyptus; sugar cane; wheat; various cattle breed in order to analyze meat quality and tolerante to plagues such as the *Rhipicephalus microplus*; goats. Other studies were related to: the identification of molecular markers referring to the reproductive and performance of swine; racial characteristics of sheep and specie-specifics in fish; the search for improved efficiency of maize and sorghum plants to the use of phosphorus.

4. CONCLUSIONS

The implementation of the LMB was the result of an endeavour conducted by some researchers from the area of bioinformatics, hired from 2001 on, aiming to spread the implementation of this new scientific field at Embrapa with the promotion of interaction, training events and RD&I projects to structure a knowledge network around this theme in order to optimize resources and creating a research platform (this movement was described in section 3.1). As described in section 3.2, this process encountered a favorable environment at Embrapa, in termos of sharing laboratory facilities and optimizing research resources (human and physical) as consequence of PAC-Embrapa actions initiated in 2008. From 2011 on, various multi-user laboratories, including the LMB, were created at Embrapa as shared research platforms available for use by multiple research units, and several partner institutions.

[7] It can be acessed at the website of Son of Grid Engine community at: http://arc.liv.ac.uk/SGE/

An organizational model suited to collaborative innovation was implemented at LMB, providing a flexible framework to provide access to infrastructure, skills and knowledge. Aditionally, it contributed to the advancement of knowledge in bioinformatics with participation with the promotion of training and the offering of discipline and supervision at postgraduate courses. As highlighted by Powell *et al*. (1996) on the establishment of cooperation, this collaborative model allows LMB users to access shared infrastructure, know-how and expertise located outside their organizational boundaries. Similarly, the LMB team have access to new research problems that require specific combinations of knowledge, skills and computational tools. The resolution of these problems allow the Lab to maximize the use and occupation of their computing infrastructure and scientific personnel, allowing the development of know how and new competencies of LMB team.

The organizational model of LMB has to be adjusted to the established in the guidelines governing the operation of multi-user laboratories of Embrapa, including: the development of internal rules and establishing contractual instruments with external partners. It is necessary to counterbalance the use of legal instruments to ensure security to Embrapa´s activities and the provision of flexibility and agility in order to properly meet the demands of users and partners.

REFERENCES

Aruja, G. (2000) Collaboration Networks, Structural Holes and Innovation: A longitudinal study. Administrative Science Quarterly, 45, pp.425-455.

BCA - Boletim De Comunicações Administrativas – Empresa Brasileira de Pesquisa Agropecuária – Embrapa. Ano XXXVII - BCA Nº 50, de 24.10.2011.

Blankenberg, D. et al. (2010) Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology, 19-10.

Bongiolo, E. (2006) Análise Panorâmica da Bioinformática no Brasil: Propostas da Gestão de Pessoas para os Laboratórios de Pesquisa. 154 f. Monografia (Especialização) - Universidade do Extremo Sul Catarinense. Programa de Pós-Graduação em Gestão de Pessoas, Criciúma.

Castro, A.C. From catching-up to knowledge governance in the Brazilian Agribusiness. Desenvolvimento em Debate. v.1, n.2, 2010.

Chesborough, H. (2003) The Era of Open Innovation. MIT Sloan Management Review.

Cingolani, P., Sladek,R., Blanchette, M. (2015) BigDataScript: a scripting language for data pipelines. Bioinformatics 31.1 : 10-16.

Costa, A.M. (2011) Prospecção Gênica e Bioinformática In: Faleiro, F.G. , Andrade, S.R.M., Reis Jr, F..B dos (Eds). Biotecnologia: estado da arte e aplicações na agropecuária**.** pp. 121-141. Planaltina, DF: Embrapa Cerrados.

Embrapa - *Empresa Brasileira de Pesquisa Agropecuária.* (2015).[online]. www.embrapa.br/ [Accessed 27 Feb. 2015]

Emeakaroha, V. C. et al. (2013) Managing and Optimizing Bioinformatics Workflows for Data Analysis in Clouds. *Journal of Grid Computing,* Volume 11, Issue 3, 407-428.

Espindola, F. S. et al.. (2010) Recursos de Bioinformática Aplicados às Ciências Ômicas como Genômica, Transcriptômica, Proteômica, Interatômica e Metabolômica. *Biosci. J.*, Uberlândia, v. 26, n. 3, 463-477.

Fernald, K.D.S.et al. (2013) Limits to Biotechnological Innovation. *Technology and Investment*,Vol. 4 No. 3, 168-178.

Hey, T., Stewart,T., Tolle, K. M. (2009) Jim Gray on eScience: a transformed scientific method. In: Hey, T., Stewart,T., Tolle, K. M. (Eds) The fourth paradigm: data-intensive scientific discovery. Vol. 1. pp. Xvii-xxxi. Redmond, WA: Microsoft Research,

Johnson, M. et al. (2008) NCBI BLAST: a better web interface. Nucleic acids research 36, no. suppl 2, W5-W9.

Kalderimis, A. et al. (2014) InterMine: extensive web services for modern biology. Nucleic acids research, p.gku301.

Kline, S.J. Rosenberg, N. (1986) An Overview of Innovation. In: Landau, R. Rosenberg. N. The Positive Sum. pp. 275-305.Washington, National Academy Press.

LMB – Laboratório Multiusuário de Bioinformática. .[online]
https://www.agropediabrasilis.cnptia.embrapa.br/web/lmb/ [Accessed 31 May 2015]

Minelli, M., Chambers, M., Dhiraj, A. (2013) Big Data, Big Analytics: Emerging Business Inteligence and Analytic Trends for Today´s Business. pp. 1-18. Hoboken, NJ: John Wiley and Sons.

Powell, W. W. (1990) Neither market nor hierarchy: Network forms of organization. Research in organizational behavior, 12, 295-336.

Powell, W.W. Koput, K. W. Smith-Doerr, L. (1996) Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotech. Administrative Science Quarterly, 41,116-145.

Rede Onsa para a pesquisa genômica. Linha do Tempo.BV-CDI Fundação de Amparo à Pesquisa de São Paulo – FAPESP. [onlin] http://www.bv.fapesp.br/linha-do-tempo/1203/rede-pesquisa-genomica/ [Acessed 19 Jun 2015].

Romano, P., Giugno, R., Pulvirenti, A. (2011) Tools and collaborative environmentsfor bioinformatics research. Briefings in Bioinformatics. Vol 12, No. 6, 549-561.

Silva Jr, G.G. (2014) Infraestrutura Laboratorial e Cooperação Para P&D e Inovação. Radar: tecnologia, produção e comércio exterior**.** No 35. Brasília: Ipea, 31-40.

Silveira, J.M. F. J. Da , Borges, I. De C. , Buainain, A. M. (2005) Biotecnologia e Agricultura: da ciência e tecnologia aos impactos da inovação. São Paulo em Perspectiva, v. 19, n. 2,101-114.

Souza, L. De L., Rhoden, S.A. Pamphile, J.A. (2014) A Importância das Ômicas como Ferramentas para o Estudo da Prospecção de Microrganismos: Perspectivas e Desafios. Uningá Review V.18, n.2, 16-21.

Spengler, S. (2000) Computers and Biology: Bioinformatics in the Information Age. *T*ech.Sight. Science. Vol. 287 no. 5456, 1221-1223.

Stein, L. D. et al. (2002) The generic genome browser: a building block for a model organism system database. Genome research 12, No. 10,1599-1610.

Swann, P. Prevezer, M. (1996) A comparison of the dynamics of industrial clustering in computing and biotechnology. Research Policy 25, 1139-1157.

Vaz, G. J.; Giachetto, P. F.; Torres, T. Z.; Massruhá, S. M. F. S. (2012) Um modelo de estrutura organizacional em plataformas de e-science. In: Anais do Congresso da Sociedade Brasileira de Computação, 32., Curitiba. Computação e inovação: ampliando fronteiras para solução de desafios no Brasil: Sociedade Brasileira de Computação.

Yin, R. K. (2010) Estudo de Caso: planejamento e métodos. Porto Alegre: Bookman. 248p.