SMGroup

Genomics

Aburjaile FF[1],Santana MP[1], Viana MVC[1], Silva WM[1], Folador EL[1], Silva A[2] and Azevedo V[1]

[1]Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brazil

[2]Universidade Federal do Pará, Bélem, Brazil

***Corresponding author:** Vasco Azevedo, Universidade Federal de Minas Gerais/Instituto de Ciências Biológicas, Minas Gerais, Brazil, Email: vasco@icb.ufmg.br

Published Date: February 15, 2015

ABSTRACT

In the last few years, the technologies have revolutionized new areas in science especially those involving genomics, proteomics and transcriptomics. This chapter approach each of these areas showing their concepts, importance and applications.

Keywords: Genomics; Structural genomics; Functional genomics; Transcriptomics; Proteomics

STRUCTURAL GENOMICS

Introduction

Structural genomics is the analysis of sequence and structure of genome elements such as genes, regulators and mobile elements. Sequencing is necessary to obtain this information. The identification of variants among genomes has applications in evolution studies, health, biotechnology and the comprehension of relation between genotype and phenotype. Knowledge of genome sequence is fundamental to obtain this kind of information. Sequencing is the process of characterization of nucleotide sequence at a region (Targeted Genome Sequencing, TGS) or in the entire genome (Whole Genome Sequencing, WGS).

The first complete genome sequence of an organism was achieved in 1995, from the Gram-negative *Haemofilus influenzae*. Since then, the number of sequenced genomes is increasing, mainly because of new technologies that reduce the required cost and time, and new bioinformatics tools able to process the volume of data generated. Initiated formally in 1990, the Human Genome Project (HGP) was an international collaborative project that sequenced the human genome in 13 years at a cost of approximately US $ 2.7 billion [1].

Genome Sequencing

The procedure for sequencing complete genomes involves fragmentation of DNA molecule, creation of genomic libraries, reading nucleotide sequence of each fragment and reassembles the genome by alignment of the *reads*. Since the 1970s, different methods have been developed to determine the nucleotide sequence of a DNA molecule.

Sequencing generations and technologies

In 1977, Sanger and Nicklen published a DNA sequencing method by chain termination. This method required the creation of libraries by insertion of DNA fragments to be sequenced into cloning vectors (plasmids or artificial chromosomes) and amplification *in vivo*. Four sequencing reactions are performed for each fragment, each one containing a proportion of one nucleotide without 3' OH group (dideoxy nucleotides), which causes termination of synthesis of a new DNA strands in random positions. The generated fragments are separated by size on gel electrophoresis and the nucleotide sequence is determined by the last nucleotide of each fragment [2]. Over the years, improvements have been incorporated into the method, such as *in vitro* amplification and automation by thermal cyclers, labeling the amplified fragments with fluorescent dideoxynucleotides (ddNTP's), as well as automatic sequencing based on reading fragments by capillary electrophoresis. The sequenced fragments generated by this method have a maximum size between 800 and 1000 bases. According to the National Human Genome Research Institute (NHGRI), in 2001, the cost to sequence 1 Mb (1,000,000 bases) and the complete human genome were respectively U$ 5,292 and U$ 95,293,072. In 2007, the cost was reduced to U$ 397 and U$ 7,147 [3]. The amount of data generated by chain termination technology was inadequate for sequencing projects and to obtains complete genomes, even with the reduction of required costs and time.

In 2005, Next Generation Sequencing (NGS) platforms were launched, belonging to the second generation. These platforms are characterized by (i) the use of genomic libraries independent of cloning and (ii) a higher throughput compared to the first generation, by means of parallel sequencing of thousands or millions of fragments. However, the necessity of amplification in the preparation of the libraries can bias the quantification of fragments, insert sequence errors during the replication and increase the complexity of sample preparation. In addition, the smaller size of the *reads* and the large amount of generated data make the assembly of a more complex genome and creates a demand for computational structures with greater processing and storage capacity.

in 2014, the utilization of these platforms, made the cost for sequencing 1Mb and a human genome decreased to, respectively, U$ 102 and U$ 3,063 million in 2007 to $ 0.05 to U$ 4905 [3]. The first released NGS technology was pyrosequencing technology. The sequencing reaction occurs in emulsion oil droplets distributed in wells of a fiber optic slide. Each oil droplet contains one bead linked to a single library fragment and PCR reagents. The reaction receives a flow of each nucleotide at a time followed by a wash step. The incorporation of a nucleotide is detected by the releasing of pyrophosphate. This technology is used in 454 Genome Sequencer platform (Roche), launched in 2005, being able to generate 200.00 *reads* of 110 base pairs (bp), and generating approximately 20MB of data. In 2007, Illumina/Solexa launched 1G Genetic Analyzer platform. In this technology, the fragments connected to adapters are denatured and its ends hybridized to complementary sequence adapters present on the surface of a solid blade. Complementary strand are synthesized by extension of the adapter attached to a solid plate. Also in 2007, Applied Biosystems launched SOLiD platform, based on ligation sequencing. A new complementary DNA strand is synthesized by ligation of an oligonucleotide to a primer and cleavage of oligonucleotide 3' end, releasing a fluorescent color, which represents the first two nucleotides of the incorporated oligonucleotide. The new DNA strand is removed by cleavage of the last nucleotide primer and a new cycle begins, initiated one position behind in relation to previous cycle. A color code and the sequence of each cycle colors are used to determine the nucleotide sequence. In 2010, Ion Torrent (Applied Biosystems) released the Personal Genome Machine, which sequencing chemistry is similar to pyrosequencing, however the incorporation of nucleotides is detected by pH variation. This technology uses a semiconductor in place of fluorescence and a scanning camera, resulting in higher speed, lower cost and smaller equipment. Since then, updates of each platform were launched, in order to increase the size of fragment sequences and reduce the required time, cost and labor.

The third generation is characterized by (i) non-interruption of the sequencing reaction to detect each nucleotide incorporated (flow type of each nucleotide), (ii) larger *reads* and (iii) sequencing of a single molecule. This has reduced the sequencing time and simplified genome assembly in relation to the second generation, as *reads* sizes exceeds 1000pb. This generation technologies are based on imaging of singles molecules of DNA polymerase as they synthesize a single molecule of DNA; threading DNA polymerase to or next a nanopore and detection of nucleotide passage; or imaging of individual DNA strands by advanced microscopy. One example is the Platform II PacBio RS (Pacific Biosciences), launched in 2013, that generate *reads* longer than 20kb.

Genomic libraries

The DNA library preparation is one of the fundamental steps for the success of a genome project. Despite the variation in protocols, the similarities are fragmentation of the sample to sizes between 50 and 500 nucleotides, selection of fragments by size, ligation to adapters. The adapters

often contain elements for ligation on a solid surface, primer annealing sites for amplification and sequencing, and a barcode sequence that makes possible to sequence samples of different sources.

In single-end or fragments libraries, adapters are ligated to one fragment, generating one *read*. In paired-end libraries, the ends of a DNA fragment are *read*, generating two *reads*, separated by a gap of known size. In mate-pair libraries, DNA fragments are connected to an adapter, circularized, and their ends are *read*, generating two *reads*, separated by a gap of known size. The difference between paired-end and mate-pair library is circularization and larger fragments of the second one. The latter two libraries provide an approximate distance between two *reads*, and this information facilitates the genome assembly and makes it possible to detect genome deletions, inversions, duplications and nucleotide insertions.

Applications

In studies with the genomic DNA, NGS technologies have been used in sequencing of complete genomes or fragments, analysis of genetic variation, chromatin immunoprecipitation followed by sequencing (ChIP-Seq), epigenetic markers, identification of genes associated with disease, genetic screening, monitoring of infectious agents, forensic research, metagenomics and agrigenomics. Thousands of sequencing projects and millions of genomes were released. Among the RNA sequencing applications are analysis of gene expression and single cell transcriptome. In gene expression studies, NGS made the prior knowledge of genes unnecessary to the identification and quantification of transcripts, alternative splicing and sequence variations. The various applications of NGS led to the launch of benchtop sequencers, cheaper, faster, easier to use and accessible to small laboratories. These include GS Junior (Roche), Mi Seq (Illumina) and Personal Genome Machine (Life Technologies).

Genome Assembly

After sequencing the genome is reconstructed from fragments *reads*, in a process called genome assembly. The assembly is performed by means of alignment and overlap of *reads* to generate *contigs (Figure 1), and contigs* overlap. It is important that all nucleotides of a genome are represented among the sequence *reads*. The *theoretical (or expected) coverage* of a genome refers to the average expected times a nucleotide is sequenced. The calculation is performed by multiplying the size of the *reads* by the number of *reads,* and dividing the result by the size of the sequenced genome. The *depth of coverage* refers to the number of times a particular nucleotide of the genome is represented in the *reads.* This number varies because the fragment library preparation, sequencing and the process of assembling the genome generate deviations in the distribution of *reads.* The *breadth of coverage* is the percentage of a given genome sequenced by a number of *reads.*

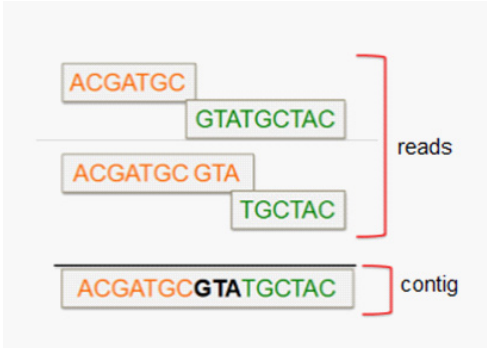

Figure 1: Assembly of *contigs* sequences by *reads* alignment.

Data analysis

The first assembly step is data analysis. Raw data from sequencer is converted to file formats used by assembly software. The main formats are FASTQ and FASTA. FASTA (Fast Alignment) format stores nucleotide sequences (Figure 2), while FASTQ stores nucleotide sequence and quality scores. Removal of poor quality sequences, and barcode *reads* is required, usually performed automatically by the assembly software. *Reads* quality is measured by the Phred index, a measure of error probability in identification of each nucleotide in a *read*.

```
>Sequencia_1
ATCGATCGATGCTAGCATCGATCGATCGATCAGCTAGCTAGCTACTACGATCGATCA
GTCAGCTAGCATGCATCGATCGATCGATCGATCGATCAGCTAGCTAGCTAGCATCGA
GCTAGCTAGCATCGATCGATCAGTCAGCATGCATGCATCGATGCACACACACACA
CACACCACACACACGTGTGTCAGCTAGGCTCGCGCGCGCGCCCGTACGATCGGCCAC
ATCGATCGATGCTAGCATCGATCGATCGATCAGCTAGCTAGCTACTACGATCGATCA
>Sequencia_2
ATCGATCGATGCTAGCATCGATCGATCGATCAGCTAGCTAGCTACTACGATCGATCA
GTCAGCTAGCATGCATCGATCGATCGATCGATCGATCAGCTAGCTAGCTAGCATCGA
GCTAGCTAGCATCGATCGATCAGTCAGCATGCATGCATCGATGCACACACACACA
ATCGATCGATGCTAGCATCGATCGATCGATCAGCTAGCTAGCTACTACGATCGATCA
GTCAGCTAGCATGCATCGATCGATCGATCGATCGATCAGCTAGCTAGCTAGCATCGA
```

Figure 2: Fasta file containing two sequences (multifasta file).

Assembly

To assembly a genome is necessary to find overlaps between *reads* and generate a consensus, or contiguous, sequence. The assembly software uses algorithms that create substrings of each *read*, called k-mers, and test alignments with k-mers of other *reads* (Figure 3) to generate contiguous sequences. The use of paired libraries facilitates the assembly process because the known distance among the pairs of readings. For example, if each paired *reading* belongs to a different *contig*, it is possible to determine the distance and orientation between these *two contigs*. A genome can be assembled using *de novo* or *ab initio* methods and by reference method.

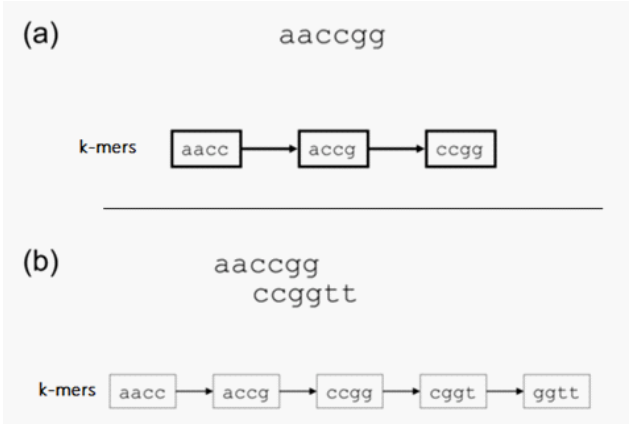

Figure 3: **(A)** k-mers of 4 nucleotides based on the *read* sequence "aaccgg".

(B) Alignment of the *reads* "aaccgg" and "ccggtt" based on k-mers.

The *de novo* assembly is performed only by *reads* overlapping, not requiring the use of a reference genome mapping (Figure 1). The most commonly used algorithms are the Greedy, OLC (Overlap-consensus) and De Bruijn graph. This method allows identification of absent regions in a reference genome. The difficulty of this method is repetitive regions where *reads* can be aligned in wrong places.

In the assembly by reference, the *reads* are mapped in a phylogenetically close genome. This method has lower computational cost and is useful in identifying SNPs, insertions and deletions between the reference and the new genomes (Figure 4). Another advantage is the resolution of repetitive regions, because the *reads* are distributed at the reference regions. However, absent sequences in the reference genome will not be mapped.

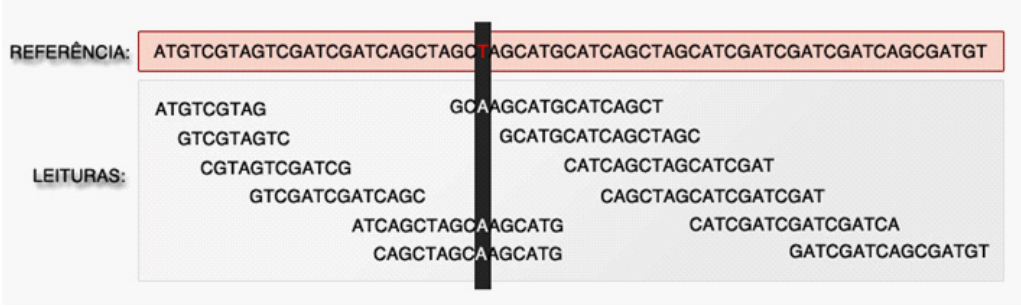

Figure 4: Mapping *reads* on a reference genome.

Scaffolding and closure of gaps

The *contigs* have to be ordered and oriented by overlaps between their ends (*de novo* assembly), mapping on reference genome or optical mapping. The latter is done by mapping

restriction sites *in vitro* and comparison to the *contigs*. A contig is called *scaffold* after its position and orientation is defined. The not identified overlaps between the *contigs* and, mainly, uncovered regions cause gaps in the assembly. Gaps can be closed *in silico* by overlapping *scaffold* ends, extension of *scaffold* by paired *reads* libraries, mapping *reads* in another reference genome in the region equivalent to the gap. The closing is accomplished by *in vitro* primer design which is long to the edge of two adjacent *scaffold*s and subsequent amplification of the region of the gap, usually by the Sanger method by generating sequences which can cover or extend the gap ends of the *scaffold*s. The third-generation sequencing platforms which generate long readings are becoming obsolete closing gaps *in vitro*, mainly due to the cost required. If a genome is still contains gaps after assembly, it is called a *draft*. The most commonly used parameters to measure the quality of an assembly is the number of *contigs*, minimum and maximum contig size and N50. In a good assembly, the number of *contigs* and its sizes is, respectively, the minimum and the maximum possible. The N50 is the size of the contig which the sum with the others, in descending order of size, is greater than or equal to 50% of the assembled genome.

Genome Annotation (Automatic and Manual Steps)

After assembly, the structure and function of genome elements as genes, regulators and other non-coding sequences have to be identified to posterior analysis. This process is called genome annotation. Gene prediction can be made empirically, by similarity with known sequences in databases and proteins, or ab initio, by computational techniques, using algorithms to find stop codons, coding sequences (CDS), start codons.

Normally, the algorithms to predict coding sequences are based on Markov chain models. In Prokaryotes, gene prediction is based on Open Reading Frames (ORFs) in the six possible frames. An ORF is a sequence that initiates at a start codon and finish at the nearest 3' stop codon. The prediction of false positives, incorrect start codons and overlapping genes are the main problems of this method. This problem can be minimized by searching for ribosome binding sites (RBS) to indicate a true start site and homologies in closely-related organisms.

The occurrence of introns and alternative splicing makes Eukaryotic gene prediction more complex. The prediction is based on recognizing signal functions on DNA strand and posterior accurate prediction of gene structure and organization. Sequences are classified as coding and non-coding and functions are assigned by alignments with functionally-related documented sequences, involved in transcription, translation and splicing.

After prediction, annotation can be done automatically, by align similarity with sequences at databases and transferring the information about their function to the new sequence. Databases can hold information of nucleic acids (e.g. NCBI, EMBL), protein sequences (e.g. UniProtKB), domains (e.g. Pfam) and structures (e.g. Protein

Data Bank), metabolic pathways (e.g. Kyoto Encyclopedia of Genes and Genomes), gene

expression (e.g. dbEST), gene ontology and evolutionary relationships. Different databases are integrated by automatically annotation services, as provided by RAST (Rapid Annotation using Subsystem Technology). Automatic annotation can provide a great amount of information about a new genome from its sequence. However, not all databases are fully curated, updated or have standard terminology. Manual curation is done by checking the results for different databases to identify protein domains and functions, correct start codon positions, gene names, gene products and to identification of frameshift.

After annotation, the genome can be submitted to a database, as submission is a pre-requisite to publication in scientific journals. Once this data is available, it can be used researches about genetic diversity, evolution, biotechnology and health.

Comparative Genomics

After sequencing, assembly and annotation of a genome, different analyzes will try to find out to remove as much information as possible from a genome of an organism. In this context, there is the comparative genomics in order to seek from genomic data, which are the between intra and inter-specific relationships. This can be studied using the homology the genes, loci, functions, processes and categories.

Currently, there are a number of tools available for the comparison of genomes at the level of their genes and proteins, among them: BLAST, but their visualization is somewhat limited, hindering an overview of the genomes studied. However, emerged over the last decades other tools, which integrate different databases and integrate visualization and their processing, widely used by bioinformatics are they: CGView [5] and BLAST Ring Image Generator (BRIG). With these tools you can see the structure and organization of genes in one or more genomes through a graphical interface. The other programs, such as the PIPS, are already able to report data pathogenicity, improving the entire information and providing the user various relevant information to do exploration of the body [6].

All of these tools require a computing power to process a large amount of data, since genomes can be processed genomes from of simple organisms (prokaryotes) to more complex organisms (eukaryotes). Another important detail to highlight is the fact that numerous incomplete genomes are being deposited in the National Center for Biotechnology Information (NCBI). Incomplete genomes are those assigned "*draft*", which will have a percentage of genetic information lost one time analyzed using these tools,this is because part of the "gap" cannot be correlated to anything since it is said to be unknown. Therefore, we emphasize the importance of finalizing a genome in all steps, assembly and annotation, thus avoiding the loss of biological information.

FUNCTIONAL GENOMICS

Transcriptomics

Concepts and evolution

Transcriptomics is a global study of gene expression levels of all the genes in an organism (genome-wide expression profiling), in order words, a set of detectably transcripts in a given cell. Transcripts of the profiles are modulated by factors such as environment, time, diseases and response to adverse processes.

The first idea of transcriptome emerged in the late 70's,described as the Dot-Blot technique in order to use nucleic acids for the simultaneous analysis of some genes in arrays. In 1995, with the evolution of different technologies, this method has been enhanced by improving their level of sensitivity in optical detection. In 2002, the Affymetrix launch the expression of identification of unknown genes and detection to splice variants. Finally, in 2008, the RNA-seq technique revolutionized history with sequencing from new generation technologies.

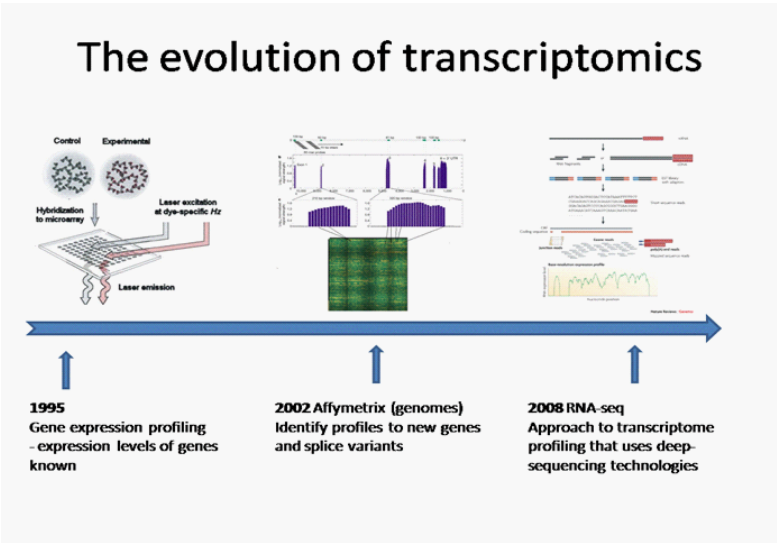

Figure 5: The evolution of transcriptomics.

Different RNAs

Total RNA is the combination of coding and noncoding; being the first comprises the messenger RNA (mRNA), transcript that carries the code information for proteins synthesis. In prokaryotes, the mRNA can be transcribed in operons or single genes and contain the exact transcript, and in eukaryotes, the mRNA has fragments that are translated in proteins (exons) and other regulatory regions (introns) that are removed by a process call splicing.

The noncoding RNAs are a combination of different classes: ribosomal RNA (rRNA), transfer

RNA (tRNA), small RNA (sRNA) and long noncoding RNA (lncRNA). This topic will focuses in the non-coding ones, since the coding ones represents less than 2% of the total transcripts of human cells, and to be more precisely the emphases will be given to the regulatory RNAs, which are need to maintain genetic integrity and gene expression in a fine tune. There is still a lot to find out about those types of RNA and as a result of the use of new technologies there set are continually rising.

The tRNAs were known to be responsible for transport specific amino acids for protein synthesis and also recognize their specific codons, but recently another function were attribute to this RNA, when they are exposed to environmental stress in eukaryotes. After cleaved by ribonucleases, tRNAs derived fragments are able to inhibits translation and consequently forms stress granules where the non-translated mRNAs are storages [7]. Therewith, now the tRNAs are also recognized as a new class of regulatory RNAs in response to stress.

The sRNA are responsible for mediate post-transcriptional gene silence (PTGS) and transcriptional gene silence (TGS). In PTGS, the target is cleaved or transcriptionally repress after is bound with the small RNA. TGS is a protective mechanism of genome integrity that keeps a heterochromatic state through DNA methylation or histone modifications. One of the most known type of sRNA, in eukaryotes, are the miRNA, noncoding RNAs responsible for regulates the expression of genes in a sequencing specific matter. Others types are siRNA, piRNA and sRNAs (prokaryotes). siRNAs are double-stranded RNA (dsRNA), like miRNA, which directs transcriptional silence of targets, participates in histone modification, heterochromatin formation. Piwi-interacting RNA (piRNA) is a defense system against transposons that acts through deposition of modifications in histones, inducing a heterochromatic state. The piRNAs are present in the genome in clusters and are a germ line specific class from both male and female organisms, and different from the others classes of sRNAs, piRNA requires the Piwi protein to be generated.

In prokaryotes, regulatory noncoding RNAs are transcripts of 50 to 250 nucleotides known as sRNAs that can either bound to mRNA and regulates gene expression or bind to proteins and alter their function. sRNAs can act on a target in *cis or trans*, and are involved in stress response, virulence, quorum sensing, DNA elimination, among others. A recently discover mechanism is the clustered regularly interspaced short palindromic repeat (CRISPR) pathway which are responsible for direct sequence-specific cleavage of foreign DNA.

Another type of regulatory RNA are the lncRNAs, transcripts with proximally 200 nucleotides or more that is not translate into proteins. They include different classes of RNA like enhancer RNAs (eRNA), small nucleolar RNA (snoRNA), intergenic and overlapping transcripts. Independent of the class, the lncRNA are cell specific with high selective pressure, considering the genetic sequences, the promoters and there function structure, and also can function in *cis* and in *trans*, affecting neighbor genes or more distant ones.

An important aspect of the lncRNAs is that they interact with diverse proteins, forms nucleus domains, interacting two distinct chromosomal regions and modulates the chromatin. The lncRNA can act as a regulatory element, having an influence in histone modification, activating and repressing, even silence alleles of several genes (genomic imprinting). One example is XIST, a lncRNA involved in the inactivation of the X chromosome in mammals, and aberrant XIST showed to be involved in human cancer and an alteration on a XIST regulator (TSIX – antisense noncoding regulator) result in alteration X inactivation and embryonic lethality. lncRNA are also involved in the development of central nervous system, heart among other organs and environment responses.

The field of study of the ncRNA is just getting started and there is still a lot to learn about then and what they influence in the eukaryote and prokaryotes life, and because of that the scientific world is focusing all efforts on the study of those transcripts, the real responsible for the complexly of organisms.

METHODOLOGY OF STUDY: ADVANTAGES AND DISADVANTAGES

Microarray X RNA-seq

Microarray and RNA-seq are experimental techniques to measure the expression levels of the transcripts. These technologies have driven the functional genomics research in different organisms, prokaryotes and eukaryotes, for studies of physiological or pathological conditions. Microarrays allow a quantification of nucleic acids (mRNA form of genomic DNA or cDNA) by hybridization to complementary array. However, the RNA-seq technique, by sequencing of RNAs, can apply the same protocol for various interests, such as total RNA, mRNA, small RNA, such as miRNA, tRNA and rRNA.

There are a number of advantages and disadvantages about these two techniques. Microarray techniques are limited to detect transcripts, whose genomic sequences are their known; however, the microarrays are excellent for specific identification of known variants, for example in diagnostic testing, such as: cystic fibrosis. Nevertheless, RNA-seq is used for detection of known and unknown transcripts sequences and detects splicing and variants, especially you used for further depth studies in any organism, prokaryotes and eukaryotes.

Real Time PCR

Real time PCR is a golden standard technique of gene expression that detects the product as soon it amplifies. The sample (RNA of study) is first reverse transcribed and then a mix similar with the conversional PCR plus a fluorescent agent is added in the sample for amplification. Real time PCR are applied in a variety of biological fields such as microbiology, veterinary science, pharmacology, biotechnology among others and can be useful in detecting SNP variation, allelic discrimination, forensic studies, quantification of pathogens and genetically modified organisms (GMOs), food quality, besides gene expression and ncRNA analysis.

This methodology has some advantages over the conversional PCR considering specially the measure of the DNA concentration, and another relevant fact is that the results obtained can be either qualitative or quantitative (qPCR), different from de conventional. Real time PCR is widely used technique and so became necessary an experimental protocol with detail information, being a simple and efficient instrument of research is require a maximal quality control so the results can be more reliable and possible to compare. Real-time PCR is very good for the few genes already known, however when do you want to see all a comprehensive analysis already exist other techniques are more suitable and effective.

Applied Biotechnology: Looking in to the Future

In the past, the focuses of genetics studies were or in the genome or in proteins, but the discovery of new classes of RNA and what they can do opened up a whole world of different approaches to old problems. Besides the knowledge of the biology of the specie at a given moment and situation, obtained especially by the mRNA, now the RNA is also being used in medical treatment, agriculture improvement, and in others fields involving biotechnology.

The development of global transcriptome profile using next generation sequencing as a functional genomics tool allows a better selection of possible target for further studies. With that in mind plus the fact that constant improvement and discoveries are being made in the field of science, the techniques applied now that focused on RNA made incredible breakthrough and are a promise in various fields of research.

Ageing

The use of noncoding RNA like miRNA was first tested in the improvement of lifespan by in *Caenorhabditis elegans* and from that different authors pursued this topic. miRNAs modulates protein homeostasis, mitochondrial metabolism, stress response, including DNA damage [7]. Using new sequencing technologies among other, researchers discovered that many *C. elegans* miRNA including *lin-4* changed their expression during life and in the absence of *lin-4* the animals decreased their lifespan [8]. This and other breakthroughs demonstrated that miRNAs as capable of predicting the longevity and remaining lifespan, and that possibly others ncRNAs are also involved in ageing [7].

Gene function and vaccine production

Next generation sequencing became an extremely important tool for the understanding of gene function, since the costs of sequencing for the whole genome and transcriptome is more cheap and fast than previously methodologies. The study of the transcriptome allows us to understand what happens with the microorganism and the host during the infection and select potential virulent factors with more precision. This selection is possible thanks to studies of different expression in environment (e.g. culture medium) that simulates the conditions faced by the pathogen or by a double transcriptome study where both the microorganism and the host are analyzed *in vitro* or *in vivo*.

Insect management

Insects pests are a major problem in agriculture and the use of pesticides are slowly decreasing since it is toxic for humans, fact that brought the necessity for nontoxic approaches. RNA interference (RNAi), a process of specific gene silence based on dsRNA, was susceptibly tested in deleting essential genes in insects and by doing so caused lethality [9]. Both miRNA and siRNA are pathways endogenous of insects and plants, and the siRNA are recognizing as a major antivirus defense mechanism of insects. Even so this strategy proved to have potential not only in killing pests but also in the control of diseases transmitted by insects [10]. The problem encounter for this method was the way of delivery, since dsRNA apply do not replicate in the host and because of that need to be constantly supplied. Two popular approaches are the use of RNAi pesticides that are safe for humans and beneficial insects, and use of transgenic plants that already produce dsRNAs [10] but in this case brings further problems that will not be discussed here.

Disease therapy

The RNAi therapy shown to be effective in treating a variety of disease from viral disease to single nucleotide polymorphisms (SNP) and oncogenes, but as mentioned in the topic above the delivery and possible toxicity of the RNAi are still a concern [11].

Human immunodeficiency virus (HIV) was one of the first target by RNAi in the attempt to silence an infectiousness targets, but since it is a highly mutate virus there is still a lot to figure it out until this approach be successful. A second method were then tested to improve the effectiveness of the treatment which instead of focusing on the virus the target were a host receptor require for the infection [12]. The use of RNAi to treat disease are in progress and apart from HIV others illness are also in experimental testing and among then are hepatitis C virus (HCV), amyotrophic lateral sclerosis (ALS), Huntington, different types of cancer and more [11,13].

Approaches that manipulates splicing also function as disease therapy by correcting aberrant splicing or inducing then. Mutation that disrupt or create new splice sites are very common, in both intronic and exonic location, and are known to cause genetic diseases (e.g. Duchenne muscular dystrophy, Miyoshi myopathy, Cancer) [13].

PROTEOMICS

The development of new technologies, involved in next-generation sequencing technology and access the complete genomes has promoted an increase of information about genomic data for several organisms. Thus, the functional genomics come with aim of complementing genomic data, where through of transcriptomic and proteomic studies, has promoted the description of the gene and protein functions as well their interactions. However, unlike of genomic and transcriptomic studies, the proteomic try to evaluate the protein expression at a given condition. Thus, a proteomic study provides valuable information about: protein synthesis, qualitative and quantitative information of protein product, post-translational modifications, protein-protein interaction and subcellular localization.

Currently, due the great advances both in classical and high-throughput proteomic technologies, new advances has been observed in the identification of targets for infectious diseases, inflammatory diseases, chronic conditions, cancer, drug discovery. Furthermore, through of comparative proteomic screening, between two conditions, i.e normal and diseased cells, allow the identification of differentially expressed proteins, that can promoted the identification of novel targets and biomarkers.

The proteomic allows the analysis of the gene expression on a large scale; this study is basically divided into two steps: (i) separation of the protein extract through of the intrinsic fundamental properties of each protein and (ii) identification of proteins through of the sequencing of peptides predigested enzymatically. However, the success of a proteomic study is due the complimentarily by bioinformatics analysis; where several softwares are developed to auxiliary in the data analysis, i.e., correlation with genomic data, furthermore, several public data repository for *proteomics* data were created. Thus, proteomics is presented as a multidisciplinary discipline in the investigation medical, biologic and biochemistry.

Gel-Based Proteomic

Gel-based methods like gel electrophoresis are routinely used in proteomic analysis to separation of molecules or peptides. Two-dimensional gel electrophoresis (2-DE) is a classical proteome technique, applied to resolution of protein from complex protein mixture. 2-DE is used to separate proteins in two dimensions: (i) during the process of isoelectric focalization (IEF) the proteins are separated according to the isoelectric point (pI) and (ii) using a SDS-PAGE the proteins are separate by molecular weight. After the electrophoretic resolution the spot/protein are submitted to tryptic digestion and identified by mass spectrometry (MS). Thus, the combination of the 2-DE with the technique MALDI-TOF-MS (matrix-assisted laser desorption/ionization – time of flight), has promoted great insight in the study proteomics.

The 2-DE analysis allows the generation of proteins maps, where spots/proteins differentially expressed are visualized, as well the presence of post-translational modifications (PTM) and isoform of proteins. However, a disadvantage of 2-DE classical is gel-to-gel variation, thus to suppress this variation the two-dimensional fluorescence difference gel electrophoresis (2-D DIGE) was developed. This technique allows the simultaneous separation of two samples different pre-labeled with cyanine dye (CyDye) fluorophores in the same 2-DE. Furthermore, the great advance of this technique is the realization of a multiplex assay, where an internal standard is used for normalization of data that minimize the experimental variation between gels and favor the quantitative analysis of the proteins [14].

Gel-Free Proteomic

Due the great advances in the biotechnological, new technologies also has been developed in proteomic area, where several advances have been observed in recent years, in order to address the weaknesses observed in 2-DE. Thus, the development of gel-free methods, such as liquid

chromatography (LC) MS/MS promoted a great advance in the proteomic field, where are obtained higher sensitivity, accuracy and resolution in proteomic study of large-scale. In this approach normally utilize a microcapillary reverse phase (RP) μLC system to separate the protein extract previously cleaved into peptides and *in tandem* analyzed by MS/MS, using electrospray ionization (ESI) coupled the mass analyzer (time-of-flight (TOF), quadrupole and ion trap). Thus, LC-MS/MS analysis associated the database searching has promoted a high-throughput proteomic analysis.

The abundance changes of proteins present in complex biological mixtures is a major challenge in the proteomic study, thus with the introduction of gel-free methods, several workflows to quantification have been developed such as: (1) Chemical labeling: utilize thiol-specific tags to covalently label in cysteine residues through of the isotope-coded affinity tag (ICAT) reagent approach, in addition other approach used to quantification using the chemical labeling is the technique iTRAQ, where is based in the labeling of an amine reactive, isobaric, isotope tag. (2) Metabolic labeling: metabolic precursors (amino acids, carbohydrates, organic salts) are introduces in the culture media and changed the molecular weight of the proteins, these metabolic precursors are labeling, through of the technique of stable isotope labeling of amino acids in cell culture (SILAC). (3) Label-free quantification: no label type is used to quantify the samples. The quantification is based on (i) spectral counting; where the quantification is achieved by total number of MS/MS spectra for a protein, and (ii) intensity of the chromatographic peak area under the curve or intensity of the precursor ion MS spectra [Otto A, Becher D, Schmidt F. Quantitative proteomics in the field of microbiology. Proteomics. 2014; 14:547-565].

Proteomic in Applied Microbiology and Biotechnology

The proteomic study offer a global survey of cellular changes in different physiological states at protein level, thus several studies using distinct or combined proteomic approach has been applied both in the applied microbiology and biotechnology.

In the microbiology field several proteomic studies have been performed with lactic bacteria, due the importance of this group of bacteria in the aliment industry, act as components of probiotic products and mainly in the fermentation process of yoghurt and cheese as well in the long-term storage of these products. Thus, proteomic studies related to knowledge of the process adaptive of *Lactobacillus*, *Propionibacterium* and *Bifidobacterium*, to acid condition and bile salts, showed a set of proteins that favor the survival and adapt theses bacteria in theses environmental that are present in the gastro-intestinal tract [15-17]. In addition, the combination of bioinformatics tools and proteomic techniques has been applied to explore the host–bacteria interaction process, through of the characterization of surface-exposed proteins main protein fraction used by bacteria to promote the interaction with receptors of the host-cell [18,19]. However, these studies favor the selection of strain bacterial that can be used in the alimentary industry.

The genomic data from cancer patients has promoted information about variations, mutations and molecular pathways this disease. However, these alterations reflect in the gene expression

of the cell cancer and consequently affected the function and stability of the proteins coding by these genes. Thus, proteomic study allow unveil this alteration and identify proteins involved in this pathway. In addition, currently the proteomic has been applied in the research of potential biomarker to have utilized in diagnostic of various type of cancer. In this type of study the biological samples evaluated from blood specimens like serum or plasma, thus quantitative proteomic both label-based and label-free has promoted the identification of several regulated differentially proteins, when compared the normal cell and tumor cell [18-20].

Currently with the technology advances, great efforts have been directed to the study of sustainable energy using mainly the bioenergetics process, which aims at energy production from biological materials and photosynthetic organism. A of the applied bioenergetics process is at production of biofuels. Thus, several proteomic studies are used to characterize the main raw material used to produce biofuels. Proteomic study performed with the plants: *Sorghum* sugarcane and maize, and algae and cyanobacteria that also are used in biofuels fabrication, have promoted the characterization of proteins present in different subcellular compartments, involved in several biological process as well the identification of proteins associated the resistance and adaptation to several stress conditions. These studies combined different proteomic approach and the results obtained in these studies theses several source used in the bio-ethanol production, promote information about of the physiology of this organism as well the identification of targets that can be used in its improvement main during the cultivation in different climatic conditions, favoring the biofuels production [21].

Application to a Bacterial Protein-Protein Interaction

With the advent of second and third generation of sequencing technologies of there has a significant reduction of cost and time, and simultaneously happend an amount of sequenced genomes. With that, the genome sequencing became more accessible therefore increasing the amount of biological data such as DNA, and proteins available in public databases about of various organisms [22-24].

DNA, RNA and proteins do not act individually in a living organism they work together interacting with each other other, physicaly or regulatory, in such a way they can execute their biological processes. Knowing whereby these elements are grouped and how they interact is a relevant area of study and research, allowing the understanding of a given organism, disease or biological process at systemic level. Therefore, the data generated from DNA, RNA, proteins and other molecules, enabled researchers to broaden the fields of research to better understand organisms and/or diseases, search for new pharmaceuticals, veterinary, agricultural and biotechnology compounds.

An important tool for understanding how these molecules act in complex biological systems are protein-protein interactions (PPI) that, pair-to-pair, form a complex set of biological

interactions called interactome [25, 26]. A PPI is composed of nodes that represent proteins (A and B), and edges that connect these proteins (nodes). The edge that connects both proteins do not have source and destination, thus forming a non-directional interaction, being equivalent the interpretation that the protein A interacts with protein B or that protein B interacts with the protein A.

Each interaction has two elements, called binary, and several binary interactions form a complex network of PPI that may represent all interactions of an organism or pathogen-host (Figure 6). The literature describes methods for the identification of PPI, being classified as: (i) experimental- yeast-two-hybrid (Y2H), protein chip, tandem affinity purification followed by mass spectrometry (TAP-MS); - biophysical- atomic force microscopy (AFM), analytical ultracentrifugation (UC), nuclear magnetic resonance (NMR)or (ii) computational - three-dimensional structure, text mining, phylogenetic profile, phylogenetic tree, gene co-localization, interolog mapping-based and machine learning.

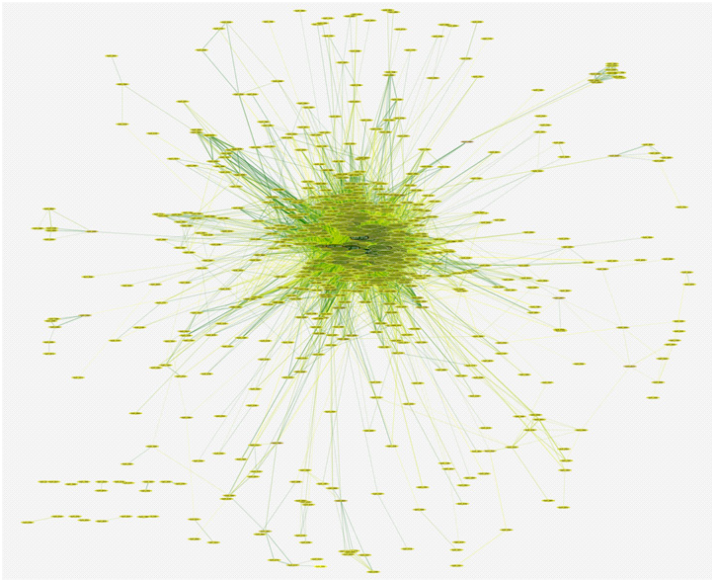

Figure 6: Example of a protein-protein interaction network, made up of several pairs of interaction.

In time, the methods can be classified by data generated aslow-throughput or high-throughput. High-throughput methods are able to detect greater amount of interactions, but have low specificity, while the low-throughput methods, which test specific interactions, have high specificity. More specific is the method, most sure has that the interaction occurs within the organism and, the more sensitive is the method, the more interactions are known. Experimental and computational methods complement each other, both being necessary for the scientific community to better explore the data [27,28].

Regardless of the method used, a network interaction can provide researchers with more knowledge about an organism, making possible to think about new biological hypotheses and direct new experiments.

In this sense, though indirectly, network interaction can be used for biotechnological purposes. By better knowing the organism, this can be genetically modified and have his behavior changed, aiming at greater efficiency or quality in the production of a product of our interest, as well as, can be genetically modified to produce or metabolize a product that is harmful to human, animal health or the environment, such as bio-remediation of soils. Aided by a network of PPI this is possible.

Assuming that some gene, óperon or protein involved in a biotechnological process of interest be known, with the help of a PPI network, it is possible to identify the proteins that interact with (partner) the protein of interest. By identifying proteins an interaction, may also be known their functions allowing to better understand the biological process or even the biochemical pathway involved in biotechnological process of interest. This information, aggregated to the knowledge about the organism, can help a researcher to understand the organism at systemic level, allowing new hypotheses to be tested in the laboratory. Still, with a network of interaction topological analysis can performed to identify biologically important proteins as proteins Hubs that have a high degree of interaction with other proteins, and proteins Betweenness that make the connection between the functional modules in a network of interaction. These both protein class, Hub and Betweenness, are considered essential proteins [28].

Information of interactions, protein Hubs and between allies proteomic experiments or information to RNA-Seq, which test a condition of biotechnological interest and measure the expression of transcripts or proteins, extend the prospects of identifying targets or metabolic pathways relevant. Besides allowing know which existing proteins and the expression of these, it is possible to meet with who these proteins interact and work together for the realization of biotechnological condition tested.

References

1. National human genome research institute. 2012.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977; 74: 5463-5467.
3. National human genome research institute. 2014.
4. Otto A, Becher D, Schmidt F. Quantitative proteomics in the field of microbiology. Proteomics. 2014; 14: 547-565.
5. Grant JR, Stothard P. The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res. 2008; 36: W181-184.
6. Soares SC, Abreu VA, Ramos RT, Cerdeira L, Silva A. PIPS: pathogenicity island prediction software. PLoS One. 2012; 7: e30848.
7. Kato M, Slack FJ. Ageing and the small, non-coding RNA world. Ageing Res Rev. 2013; 12: 429-435.
8. Kato M, Chen X, Inukai S, Zhao H, Slack FJ. Age-Associated Changes in Expression of Small, Noncoding RNAs, Including microRNAs, in C. Elegans. Rna. 2011; 17: 1804–1820.
9. Baum James A, Thierry Bogaert, William Clinton, Gregory R Heck, Pascale Feldmann, et al. Control of Coleopteran Insect Pests through RNA Interference. Nature Biotechnology. 2007; 25: 1322–1326.
10. Burand John P, Wayne B Hunter. RNAi: Future in Insect Management. J Invertebr Pathol. 2013; 112: S68–74.
11. Hannon GJ, Rossi JJ. Unlocking the potential of the human genome with RNA interference. Nature. 2004; 431: 371-378.

12. Boden D, Pusch O, Lee F, Tucker L, Ramratnam B. Human immunodeficiency virus type 1 escape from RNA interference. J Virol. 2003; 77: 11531-11535.

13. Havens MA, Duelli DM, Hastings ML. Targeting RNA splicing for disease therapy. Wiley Interdiscip Rev RNA. 2013; 4: 247-266.

14. Arentz G, Weiland F, Oehler MK, Hoffmann P. State of the art of 2D DIGE. Proteomics Clin Appl. 2014.

15. Hamon E, Horvatovich P, Bisch M, Bringel F, Marchioni E, et al. Investigation of Biomarkers of Bile Tolerance in Lactobacillus Casei Using Comparative Proteomics. J Proteome Res. 2012; 11: 109–118.

16. Leverrier P, Dimova D, Pichereau V, Auffray Y, Boyaval P, et al. Susceptibility and Adaptive Response to Bile Salts in Propionibacterium Freudenreichii: Physiological and Proteomic Analysis. Appl Environ Microbiol. 2003; 69: 3809–3818.

17. Borja Sánchez,Marie-Christine Champomier-Vergès,Birgitte Stuer-Lauridsen, Patricia Ruas-Madiedo,Patricia Anglade, et al. Adaptation and Response of Bifidobacterium Animalis Subsp. Lactis to Bile: A Proteomic and Physiological Approach. Applied and Environmental Microbiology. 2007; 73: 6757–6767.

18. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, et al. Prediction of Surface Exposed Proteins in Streptococcus Pyogenes, with a Potential Application to Other Gram-Positive Bacteria. Proteomics. 2009; 9: 61–73.

19. Meyrand M, Guillot A, Goin M, Furlan S, Armalyte J, et al. Surface Proteome Analysis of a Natural Isolate of Lactococcus Lactis Reveals the Presence of Pili Able to Bind Human Intestinal Epithelial Cells. Mol Cell Proteomics. 2013; 12: 3935–3947.

20. KoÄ□evar N, Hudler P, Komel R. The progress of proteomic approaches in searching for cancer biomarkers. N Biotechnol. 2013; 30: 319-326.

21. Ndimba BK, Ndimba RJ, Johnson TS, Waditee-Sirisattha R, Baba M, et al. Biofuels as a Sustainable Energy Source: An Update of the Applications of Proteomics in Bioenergy Crops and Algae. J Proteomics. 2013; 93: 234–244.

22. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013; 155: 27-38.

23. Casals F, Idaghdour Y, Hussin J, Awadalla P. Next-Generation Sequencing Approaches for Genetic Mapping of Complex Diseases. J Neuroimmunol. 2012; 248: 10–22.

24. van Dijk EL, Auger H2, Jaszczyszyn Y3, Thermes C2. Ten years of next-generation sequencing technology. Trends Genet. 2014; 30: 418-426.

25. Ngounou Wetie AG, Sokolowska I, Woods AG, Roy U, Deinhardt K, et al. Protein–protein Interactions: Switch from Classical Methods to Proteomics and Bioinformatics-Based Approaches. Cell Mol Life Sci. 71: 205–228.

26. Rao VS, Srinivas K, Sujini GN2, Kumar GN1. Protein-protein interaction detection: methods and analysis. Int J Proteomics. 2014; 2014: 147648.

27. Harrington ED, Jensen LJ, Bork P. Predicting biological networks from genomic data. FEBS Lett. 2008; 582: 1251-1258.

28. Nicola J Mulder, Richard O Akinola , Gaston K Mazandu , Holifidy Rapanoel. Using Biological Networks to Improve Our Understanding of Infectious Diseases. Computational and Structural Biotechnology Journal. 2014; 11: 1–10.