14

Improving Bio-Technology Processes Using Computational Techniques

Avinash Shankaranarayanan* and Christine Amaldas
Ritsumeikan Asia Pacific University
Beppushi, Oita, Japan

1. Introduction

Biology is the science of origin and evolution of life. Computational biology helps biologists to understand evolution using computer applications. Bioinformatics is a subset of computational biology centered at applying information and information processing tools to enable the development of biology for deciphering the human genome, biotechnologies, new legal and forensic techniques and futuristic medicines (Claverie et. al., 2003). More importantly, Bioinformatics can be defined as a science for solving complex biological problems using high performance computational tools. Sir Alfred Sanger won his Nobel Prize for sequencing insulin which triggered the modern era of molecular and dynamic biology that laid the foundation for molecular sequences (Claverie et. al., 2003). During the pre-computer era, time consuming and error prone methods of sequence analysis and storage were manually done. As the computing age took over, these sequences were input as computer data and stored in flat files or databases. Programs were written to enable error free comparison of existing sequences using preliminary pattern matching algorithms. Sequence comparison is one of the most commonly used computer applications in Bioinformatics research. The development of DNA sequencing tools and databases using information processing technologies has lead to the birth of Bioinformatics. The importance of biological sequences (DNA or proteins) is to provide a blueprint or map of the biological species (function). When the sequence is reduced to its respective letters (A, T, G, C) it becomes a unique identifier. This is a vital mechanism for computer scientists to store and retrieve data using a unique identifier (ID). A user can search and exactly pinpoint a particular gene in a database or flat file using the ID. The importance of applying identification or classification to sequences led to the annotation of genes. Biologists can retrieve meaningful information about the history of the genes using the unique ID (Shankaranarayanan, 2011). An organism's primary functional and hereditary information is stored as Deoxyribonucleic Acid (DNA), Ribonucleic Acid (RNA) and Proteins. All of them are linear chains composed of small molecules called macromolecules. These macromolecules are made up of a fixed set of alphabets varying in characteristics. The DNA is made up of four de-oxyribonucleotides namely, adenine (A), thymine (T), cytosine (C), and guanine (G). Similarly the RNA is made

*The Authors would like to take this opportunity to thank Professors Francisco P. Fellizar, Jr. and A Mani, Graduate School of Asia Pacific Studies, Ritsumeikan Asia Pacific University, for their valuable comments and support.

up of four ribonucleotides namely, adenine (A), uracil (U), cytosine(C) and guanine (G) and the proteins are made up of 20 amino acids (Gibas & Jambeck, 2001).

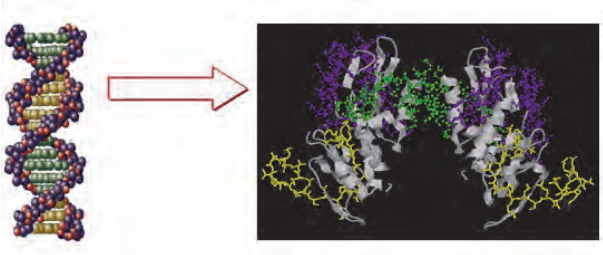

Fig. 1. DNA the fundamental building block of living Organisms.

Amino acids are the building blocks of proteins. An amino acid consists of a central carbon atom, linked to an amino acid group, a carboxylic acid group, a hydrogen atom, and a distinctive R group. The R group is often referred to as the side chain. With four different groups connected to the tetrahedral -carbon atom, -amino acids are chiral; the two mirror-image forms are called the l - isomer and the d - isomer. To build a simple tool for converting DNA to RNA, we can write a substitution algorithm to replace all the T's with the U's. This process of converting DNA to RNA or vice-versa is called transcribing. These macromolecules are defined as side-chains of defined components and are represented as a string of alphabets called sequences. The central dogma of molecular biology states that a DNA acts as a template to replicate itself; a DNA can transcribe itself into RNA and RNA into protein. Hence, the DNA contains the blue print of all living organisms where the DNA is used to replicate or reproduce similar organisms. The importance of applying identification or classification using sequence labels led to annotation of the genes where biologists could get meaningful information and history of the gene using the label information. Hence, an entire DNA sequence that codes the living organism is termed as a genome. Vast amounts of sequence data are stored in various remote database sites and are queried for sequence matching. This chapter is divided into two research phases. In Phase 1 we conduct an in situ experimentation of which the DNA of the input feed obtained from the Andaman and Nicobar Islands is analyzed and sequenced using Bioinformatics tools and computational techniques. In Phase 2, we setup a small scale bio-digester unit in a laboratory to test and verify the calorific content of the input feed (Biomass). We then go about improving the Biogas output using extraneous input feed materials and additives.

2. Understanding the importance of biotechnology

For centuries exploring and improving Biotechnology processes has been primarily a self-instigating approach by human beings simply because of something called the food-chain. Through the process of photosynthesis and transpiration plants convert sunlight and carbon dioxide (CO_2) into rich energy units to form the primary storage (namely food) devices of energy in the food chain. The primary, secondary and other types of food chains exist because of these mechanisms. Due to the decomposition of plants and animal dead matter, abundant sources of fossil fuels were discovered as the primary sources of energy. Economies around the world have come to depend upon fossil fuel resources so much that it is considered as a major bottleneck for many developing economies. Due to the vast exploitation and in-consideration towards preserving the Bio-diversity and Natural resources amongst

Nations, we have embarked upon a new journey where climate change and environmental de-gradation is threatening to put an end to our way of life. Energy is and will be the future blood line of modern economies. Food security and changes in life style has also been a major concern in the ever expanding human population. A part of the bio-technology industry's point of view: genetics will help alleviate world hunger. According to Nobel Prize winning economist Amartya Sen, people are hungry because they cannot afford to buy food owing to failing market economics. Hence, one part of the population is obese due to the substantial amount of food intake, while the other is under-nourished due to non-affordability of food as a commodity. International trade and economic policies have led to immense poverty, inequality and lack of access to food due to over population and poor governance. A majority of the islands in the Asia-Pacific region suffer from imported energy dependence. Island development problems are mostly related to imported fossil fuel dependence, fresh water availability and municipal solid waste management, associated with transportation and other issues. Most developed countries such as Japan and Singapore and developing countries such as India, China and the Philippines deal with hosts of Islands in the Asia-Pacific relying on rice and other Biomass as the primary sources of staple food for consumption. One such example is the Andaman and Nicobar islands of the Indian Sub-continent where Biomass samples from a digester unit have been analyzed using Bio-technology processes. Agricultural produce today is mainly dependent on a fossil based economy. Economic growth is conventionally measured as the percentage rate of increase in real Gross Domestic Product (GDP). Power production is the key to economic development in most countries. The power sector has been receiving inadequate priority in agricultural production in most developing countries. Energy production and supply in most islands depend mainly on expensive fossil fuels imported. These land masses are usually linked by a weak electricity grid connection like undersea submarine cables from the main-land. Prohibitive cost escalation due to distance and other geographic/demographic conditions have been met with increasing energy costs on the islands. The most promising resource that can abundantly store energy are that of available Biomass. Biomass based Bio-energy is capable of storing finite sources of energy that is replenishable in the form of aforestation and sustainable agriculture. The stored Biomass can later be used for electricity production and for transportation in the form of Bio-fuels such as compressed Biogas, hydrogen and other fuels. Bioenergy refers to the utilization of living plants as an energy source and Bioinformatics acts as a platform for utilizing computational tools and genetic blueprints to increase the yield of harvests. Bioenergy is extensively utilized worldwide; 13% of the worlds energy is utilized in the form of fuel firewood and woody biomass; 0.3% in the form of Bio-fuels (Keefe et. al., 2010). The United States for example, is utilizing 5% of its primary energy production from bio-energy (IEA, 2007). Bio-energy comprises of a range of material resources as distinguished below:

Woody Biomass - Wood and plant based materials utilized for heating or lighting purposes by rural population. For example, Small Islands of the Asia Pacific are affected by energy poverty.

Biomass - all forms of plant materials and organic waste sources that are also dried or fermented for gas (Biogas) or energy generation through co-firing.

Bio-fuels - also termed as Bio-diesel and ethanol / alcohols that are extracted from oil rich plants (energy crops) such as rapeseed and Jatropha used in transportation and liquid fuel appliances.

Renewable energy production have come more into public focus because of problems caused by the expected shortage of fossil fuels in the next few decades. Global warming due to CO_2 release from the burning of fossil fuels and woody biomass is causing climate change leading to a warmer planet and rising sea levels. These research challenges can be alleviated by the production of biogas from biomass sources (plant or organic waste materials) through biological processes (Angelidaki & Ellegaard, 2003). Anaerobic digestion (exposure to oxygen less environment) of plant biomass can be carried out in biogas plants though a series of metabolic processes (Daniels, 1992; Weiland, 2003; Yadvika et. al., 2004). Biogas is a high calorific mixture of gases having a high content of methane (CH_4). Anaerobic digestion is a biological process that produces gas mainly composed of CH_4 and CO_2 otherwise known as biogas in an oxygen free environment. It is produced from the decomposition of organic wastes (i.e. from biomass sources such as manure, food waste, wood waste, etc.).

Methane (CH_4): 40-70 vol. %.

Carbon dioxide (CO_2): 30-60 vol. %.

Other gases: 1-5 vol.% (including hydrogen (H_2): 0-1 vol% and hydrogen sulfide (H_2S): 0-3 vol.%).

Fig. 2. Biogas Compostion

Biogas producing microbial community (biogas microbes) consist of a large group of complex and differently acting microbe species, most notable the methane-producing bacteria (for example Eubacteria). The whole biogas production process can be divided into three stages: Hydrolysis, Acidification (Acidogenesis and Acetogenesis), and methane formation (Methanogenesis). The first stage is to take the organic ruminants of plant compounds including cell wall components such as cellulose and xylan which are hydrolyzed and converted into mono-, di- and oligosaccharides (Bayer et. al., 1992; Cirne et. al., 2007; Lynd et. al., 2002). During hydrolysis, the organic substances are divided into molecular components such as amino acids, glycerin, sugars and fatty acids. The hydrolysis process is conducted mainly by cellulolytic Bacilli and Clostridia, often utilized as a first step under anaerobic conditions. In the second stage, sugar intermediates are fermented to organic acids (acidogenesis) which in turn are converted to CO_2, Acetate, and H_2 by bacteria performing secondary fermentations (Drake et. al., 1997; Myint, 2007). In the acidification phase, microorganisms convert these intermediate products into H_2 and CO_2. The final stage is the methanogenesis stage that is conducted by Archaea which is constrained by a selection of few input substrates such as Acetate, CO_2 and H_2 and several C1 compounds like alcohols and formates that are transformed into methane and water (Deppenmeier, 1996) according to the equation:

$$CO_2 + 4H_2 -> CH_4 + 2H_2O \quad (1)$$

Numerous thermodynamic biochemical reactions are enabled based on closed interactions of two or more bacterial strands (for example H_2 feeding from algae). Methanogenic pathways have been analyzed by many models using enzymology (Schink, 1997; Schink et. al, 2006). The actual composition and interactions of biogas-producing microbial community is undefined as specific bacterium's could not be confirmed in the actual processes (Ferry, 1999; Reeve

et. al., 1997). It could be speculated that the origin of the input feed acquired from the digester unit might have been exposed to several dynamic environments and climates during acquisition stage. (Karakashev et. al., 2005; Shigematsu, 2004) state that the influence of physio-chemical parameters on population structure and efficiency of biogas formation still needs to be extensively investigated (Schlüter et. al., 2008). The calorific value is the most important factor that determines how much energy content is available in the biomass. It is used to estimate the energy potential of the input feed (Biomass fed into a fermenter unit). The calorific value of biogas is estimated to be about 6 kWh/m^3 of gas produced which roughly corresponds to around half a liter of diesel oil. The net calorific value depends on the efficiency of the burners or appliances. Methane gas is the valuable component under the aspect of using biogas as a fuel. Although the calorific value here is a general value assumed, the actual value varies based on the biomass used as feed. The energy potential is determined by the calorific value which has the capacity to produce energy (electricity or gas). An experimental setup was created to understand and verify the calorific values of the biomass taken from a small scale digester unit in the Andaman Islands. In this section we have discussed how Bioinformatics as a science could help us identify and improve our energy dependent economies using bio-energy sources. In the next section we will explore the various tools used by biologists and computational scientists.

3. Bioinformatics tools

Bioinformatics heavily relies upon statistical and analytical methods of processing biological data. Some of the important biological research aims at studying the evolutionary effects of gene mutation and similarities between gene sequences using computer technology. This aids biologists to find and cure disease causing viruses by applying new and faster methods of drug discovery in the laboratory. When looking for ways to improve agricultural yield either for food production or for the production of Bio-fuels(bio-energy production), biologists need to explore the realms of genetic blueprints of the plant biota to improve growing conditions and provide for favorable yield. Biologists often require sequence comparison and alignment applications such as Basic Local Alignment Search Tools (BLAST)(Altschul et. al., 1992), ClustalW (Higgins et. al., 1996) and Tandem Repeats(Benson, 1999) which are effectively utilized for processing large sets of gene sequences (plant and animal) for similarity matching. BLAST and Tandem Repeats are used for finding the similarities and mutational history between sequences, while ClustalW (Higgins et. al., 1996) is used for studying evolutionary relationships. Biologists often need to detect similarities between different genomic sequences. Therefore, BLAST was introduced in 1990. It was used for searching databases such as the NCBI databases that stores sequences of different species used for optimal local alignment (similarity) for a given input query of sequences. The BLAST algorithm has similarities with the approximation of Smith-Waterman Algorithm which uses a heuristic approach. (Waterman, 1981) algorithm is slow but guarantees to achieve the best possible alignment based on optimal input parameters. It sacrifices some accuracy to substantially increase the speed of the search. BLAST uses a heuristic search method to make assumptions about the data based on previous experiences. It does not guarantee to find the best alignment in all possible circumstances. Global alignments need to use gaps (representing insertions/deletions) while local alignments can avoid them by aligning regions between gaps. ClustalW (Thompson et. al., 1994) can be classified as a bioinformatics application having semi-regular computational patterns, which means the algorithms are composed of both synchronous and asynchronous steps. The basic idea behind the ClustalW

algorithm for building multiple alignments is centred on aligning the most related sequences first. The exponential rise in the size of datasets increases the problems related to the scalability of existing bioinformatics programs and tools. One approach to solving this problem is to break the problems into a number of sub-problems which could be done either in the algorithmic level (re-programming for parallel processing such as MPI) or dataset parallelism where the data is broken down depending on the number of available processors. Bioinformatics applications along with well defined outputs heavily rely on various methods of pattern recognition and statistical methods of information processing (on sequences). Bioinformatics tools had previously been extensively investigated (Shankaranarayanan, 2011), most of the existing tools were either too low level (complicated); too expensive for most laboratories to afford; and inflexible towards customizing requirements for heterogeneous networks and computational environments without significant technical expertise. Hence, high performance computational resources, such as Cluster and Grids enable faster results when parallelized as discussed in the next section.

4. High performance computational challenges in bioinformatics

Substantial discoveries of new life forms and drugs takes place on a daily basis leading to biological data being stored into remote databases (resources). The exponential increase in the size of datasets makes it mandatory for biologists to opt for better methods of crunching genomic data. Throughput is a form of measure by which the performance of an application is quantified. There are many methods by which an application or algorithm can be optimized as enumerated below (Shankaranarayanan, 2011):

I Algorithmic Optimization

II Heuristics Approach

III Statistical approach

IV High performance approaches

Optimization is the process of modifying a system to improve its efficiency (Optimization definition, 2010). Algorithmic optimization generally focuses on the Quality of Service (QoS) characteristics of a system namely execution time, memory usage, disk space, bandwidth or other resources. This will usually require a tradeoff among the different available resources. For example, increasing the size of cache improves runtime performance, but also increases the consumption of memory. Another mechanism is to optimize the loops in the program code which is termed as loop optimization. Heuristics is a technique designed to solve a problem that ignores whether the solution can be proven to be correct (as used by BLAST), but which usually produces a good solution or solves a simpler problem that contains or intersects with the solution of the more complex problems. Heuristics are intended to gain computational performance or conceptual simplicity potentially at the cost of accuracy or precision (Heuristics definition, 2010). BLAST gene sequencing application is a good example of applying heuristics to improve the performance of its search algorithms. The statistical approach deals with utilizing probabilistic methods of approaching a problem algorithmically. Mechanisms such as Marcov models of computing are applied to obtain better results at the expense of processing time. Benson's method of finding approximate tandem repeats applies both heuristic and statistical methods for computing tandem repeats (recently termed as micro satellites). Another approach to improving the throughput was to apply distributed computing techniques to biology problems and applications. Recently,

biologists have started to conduct in silico experiments. The term 'in silico' refers to conducting biological experiments using computational techniques. Biologists have started to realize and utilize high performance distributed mechanisms to improve the running time of bioinformatics applications by dividing the datasets into subunits to be processed individually in parallel (simultaneous execution of processes on number of machines) by available processors (Compute Nodes). The World Wide Web (WWW) technology enables biologists and computer scientists to collaborate and conduct remote scientific experiments using computational tools over the Internet. The exponential rise in the size of the datasets acts as a bottleneck to scalability of existing bioinformatics applications and tools causing system crashes, malfunctions (Data loss, etc) and sometimes even closure of research projects due to the level of complexity involved. This sudden increase in data, leads to resource allocation problems that cause poor performance and failure of applications. Therefore, high-end computational tools and optimal resource allocation strategies becomes vital to coping with such problems. Typical Bioinformatics applications face problems of application and resource scalability when exposed to exponential increases in the input side (data set sizes). This type of computational problems (usually NP-complete) requiring high end compute resources are usually solved using high performance distributed systems such as Clusters / Grids or supercomputers. Cluster computing is a 'task farming' system of computing that breaks the given problem into numerous sub-problems; and individual nodes work on each of the sub-problems in parallel. The cluster utilizes a centralized head node that 'farms out' sub-tasks to a static set of loosely coupled nodes. Grid computing is a more complicated infrastructure that provides better efficiency than cluster computing as it tries to improve upon scalability through the dynamic addition of new nodes and efficient allocation of resources. Although distributed approaches tend to inhibit application performance due to external factors such as latency, bandwidth and scalability issues, when properly applied can boost application performance manifolds.

Examples of a computationally expensive biological study

Dr Stanley Burt's group (Shankaranarayanan, 2011) studied the enzyme mechanism of many enzymes involved in cancer. For an enzyme named Ras, which is mutated in over 30% of known cancers, they modeled 1,622 atoms of the protein by molecular mechanics and only 43 atoms by quantum chemistry. These studies took several years and were bound by limited computational power. To calculate reaction surfaces normally takes several months of time on High Performance Computers (HPC's). Luthey Schulten's group at Illinois(Shankaranarayanan, 2011) did molecular dynamics simulations of Imidazole Glycerol Phosphate Synthase, an enzyme involved in making DNA and RNA. It took 10 hours, 12 hours, and 40 hours respectively to animate one nanosecond on three cluster machines (with different processor speeds). It took many nanoseconds of simulation to just relax the systems to prepare for further simulations. It has been estimated that to go from nanoseconds to milliseconds will require an increase in computer capacity of approximately 1,000,000. This can only be achieved by applying optimal high performance hardware and software techniques to improve the overall throughput of these tools.

Experiences by David Baker's group at Illinois

Drug design using computational tools has become the defacto standard in applying

high performance tools using distributed techniques for drug discovery. A great recent example is the discovery of Gleevec, an inhibitor of protein kinase activity, which brings about complete and sustained remission in nearly all patients in the early stages of chronic myeloid leukemia. If the structure of the protein is known, docking calculations can be performed. This usually involves docking thousands of molecules into an active site and scoring the resultant interaction. If the docking is done with rigid molecules, the calculations are fairly trivial. If, however, flexibility is allowed, and most proteins and ligands do flex, then the problem becomes enormously computationally expensive (Shankaranarayanan, 2011). If the protein structure is not known, and the protein is not similar to another one, then one must perform ab initio structure determination. David Baker's group at Illinois took approximately 150 CPU days to determine the structure of the CASP6 target T0281. Also to do a docking interaction between two proteins took 15 CPU days. He makes particular note that his group is limited by computational power. A computational Grid test bed would have ideally satisfied the computational requirements of the research group. In order to process large sets of genomic data, high performance computational tools are available on distributed computing platforms such as Clusters and Grids.

5. Methodology

The sources of evidence for this study were practical documentation, interviews, direct observation, participatory-observation, field trips to the Islands of Andaman and Nicobar in India, Marinduque, the Philippines and a post survey of the events on a one-one basis with various Island authorities. The first purpose was to collect data using these sources and second, to convey three essential data collection principles namely: using multiple sources of evidence instead of a single source; creating a study database; and maintaining the chain of evidence. The chapter is divided into two research phases. In Phase 1 we conduct an in situ experimentation of which the DNA of the input feed is analyzed and sequenced using Bioinformatics tools and computational techniques. In Phase 2, we setup a small scale bio-digester unit in a laboratory to test and verify the calorific content of the input feed (Biomass). We then go about improving the Biogas output using extraneous input feed materials and additives. The data will then be analyzed at three levels. These will be at the Biogas generation (Creation mechanism) level, Capture or detection (Monitoring) level, and Gas Cleaning (Membrane and other technologies utilized).

6. Phase 1: Experimentation (in situ) using our input feed

The enzymology of methanogenic pathways has been evaluated in detail using biological modeling and systems approaches (Blaut, 1994; Deppenmeier, 2002; Ferry, 1992; 1999; Reeve, 1992). But, the composition and interactions within a biogas-producing microbial community namely the breakdown of specific bacterium to the overall process, is mainly unknown during mesophilic or thermophilic stages of Biogas production. Moreover, the influence of physio-chemical parameters on population structure and efficiency of biogas formation is still under investigation (Karakashev et. al., 2005; Shigematsu, 2004). Hence, improvements to the production of biogas are next to impossible without knowing the calorific value of the input feed. The composition of biogas-producing microbial communities is determined through the construction of 16S-rDNA clone libraries and subsequent sequencing of 16S-rDNA amplicons as done by (Huang et. al., 2002; Klocke, 2007; Mladenovska, 2003). Moreover, Polymerase Chain Reaction Single Strand Conformation Polymorphism (PCR-SSCP) followed

by sequencing of obtained DNA-molecules will help understand the community structures in a biogas reactor (Chachkhiani et. al., 2004). Many methanogenic communities were analysed by using themcrA gene as a phylogeneticmarker (Lueders, 2001). Development of third-generation ultrafast sequencing technologies such as pyrosequencing and related computational tools have led to the realization of cost-effective large-scale environmental shotgun sequencing projects (Schluter et al., 2008). Bioinformatics for the interpretation of metagenomic data has constantly evolved and improved (Raes, 2007) wherein recently, a novel gene finding algorithm allows the exploitation of the limited information contained in the 250 nucleotides reads generated by 454-pyrosequencing for the prediction of coding sequences has been developed (Krause, 2007). Here, the insight is into the metagenome of a biogas-producing microbial community residing in the main fermenter of a small production-scale bio-digester unit (Batch fed rural Andaman Islands) where in the obtained nucleotide sequence data has been analyzed at the single read and contig level for their genetic information content by applying different bioinformatics mechanisms.

6.1 DNA preparation from the fermenter unit

A fermentation sample was taken from a small scale bio-digester unit at an Agricultural site in the Andaman and Nicobar Islands in October 2010. The sample was stored in entirely filled, screw capped bottles and transferred to a laboratory. The analyzed sample was then fed into a custom build bio-digester unit consisting of a fermenter and a storage reservoir that was continuously fed with a mix of Corn/Maize silage (33%), Rice Husk (30%) and low amounts of chicken manure (approx. 2%). The substrate was fermented at approximately 41 °C at a pH-value of 7.3. The retention period of the substrate was 30 - 55 days. First microscopic analysis of the fermentation sample was carried out within 2 hours upon sampling. Samples were diluted with two parts of sterile tap-water. The diluted fermentation sludge was strained for 30 min by the addition of 2 mg/ml 4, 6-diamidino- 2-phenylindole hydrochloride (DAPI). Bacteria were visualized through the use of a fluorescence microscope. A 20 g aliquot of the fermentation sample was used for total community DNA preparation by applying a CTAB (cetyltrimethy - ammonium bromide) containing DNA extraction buffer as described by (Entcheva, 2001; Henne, 1999). The obtained DNA pellet was re-suspended in 8ml TE buffer. The final purification step included ten DNA-eluates which were pooled and subjected to precipitation using 40ml NaCl (5M) and 2ml ethanol −20 °C. After centrifugation (15,500 rpm, 6 min) the DNA-pellet was re-suspended in 100 ml TE buffer. DNA concentration was analyzed by gel-electrophoresis. The applied method yielded a highly pure genomic DNA. Sequencing of the genomic DNA derived from the biogas reactor sample was done by applying the whole-genome-shotgun sequencing approach using a third party vendor. Approximately 7 µg of DNA-preparation were used to generate a whole-genome-shotgun library according to the protocol supplied by the manufacturer. After titration, 4.5 DNA-copies per bead were used for the main sequencing run. After emulsion PCR and subsequent bead recovery, 1,100,000 DNA-beads were subjected to sequencing.

6.2 Identifying cellulosome genes on assembled contigs

To search for contigs encoded cellulosome proteins, all protein sequences associated with the annotation term 'Cellulosome' were collected from the NCBI sequence database and imported into a BLAST database. The obtained gene sequences were subjected to BLAST searches which proved to be computationally expensive. The exponential rise in the size of the datasets acted as a bottleneck to scalability of our existing bioinformatics applications and tools

```
Obtained Sequence Assembly
No of reads 525,042
No of bases 121,554,183 bases
Avg read length 239.9 bases
No of large contigs 6,752
No of bases in large contigs 11,797,906 bases
Avg large contig size 1,348 bases
Largest contig 31,533 bases
No of all contigs 57,108
No of bases in all contigs 22,724,756 bases
Percentage of assembled bases (%) 16.04%
```

Fig. 3. Initial Results from the Andaman Islands DNA Sample prepared.

causing system crashes, malfunctions (for example data loss) due to the level of complexity involved. This sudden increase in data, led to resource allocation problems that caused poor performance and failure of applications. Typical Bioinformatics applications face such problems of application and resource scalability when exposed to exponential increases in the input side. To tackle this kind of computational complexities (usually NP-complete problems), we required high end compute resources that utilized high performance distributed systems such as computational Grids. Hence a previously researched computational platform namely $A^{3p}viGrid$ (Shankaranarayanan, 2011) was utilized for the sequencing process using Multi-Agents.

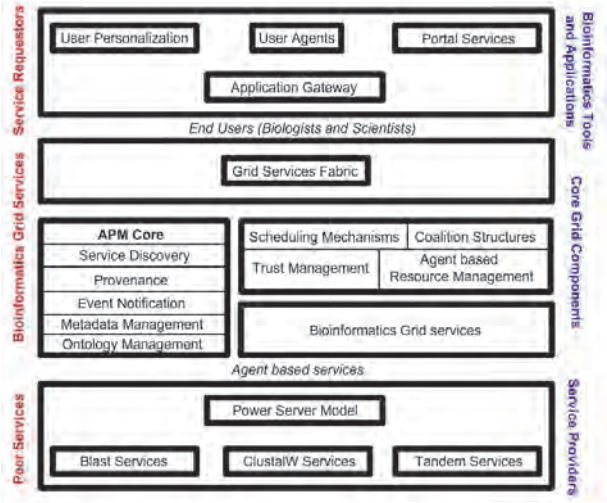

Fig. 4. The Block Diagram of the $A^{3p}viGrid$ System

$A^{3p}viGrid$ works on the principle of the power server model of computing. Each of the clients ran the $A^{3p}viGrid$ server, a simplistic http web server running services in the form of CGI/Perl wrapper Scripts. The client side coding model enables the developer to develop services using the common gateway interface (CGI) which can use any of the languages that support CGI scripting. For the sake of simplicity and rapid development of services we have used Perl as

the language of choice due to its availability and portability for most platforms. The $A^{3p}viGrid$ uses a decentralized directory structure (APM) to enable peers to register and de-register peers and their respective services (Shankaranarayanan, 2011). A set of 128 nodes were used for job processing. All the nodes ran $A^{3p}viGrid$ web servers. The Blast.apm file, a directory structure file that is local to all nodes was downloaded by all the peers as part of the initialization phase. This file contains information such as location information of nearby agents, domain and IP address and other important data. Each of the nodes compute the ideal set of nodes using a basic ping test based on the Blast grid service list. As all the nodes are capable of receiving jobs, one of them was randomly chosen for job execution (Originator). Our Fasta formatted Sequence database (DNA sequence Cellulosome from Biogas Digester Unit) was used to evaluate the Blast searches. The input query file was obtained, and a set of jobs for job processing was prepared using the optimal coalition list. Based on QoS characteristics namely Latency, Load and CPU time, the Originator of the job computed the most optimal coalition. Once the coalition list was computed the data files were migrated using the POST method to all the members of the coalition. Each of the coalition members started to search using the input query files and outputted the results to an output file. The output of the Search Phase was then appended to a file using POST back to Originator where the results were formatted using the Blast format perl script and stored as a file at the originator. Each of the agents ran on a virtual machine (VM) test bed having their own execution environments. For the sake of true heterogeneous functionality and testing, four operating environments were deployed namely: Fedora Linux Core, Windows Vista Ultimate, Mac OS Leopard and Sun's Open Solaris 10. Each of the agents were given a resource limit which shared the following specifications: 10 GB disk space; 4 GB RAM and Dual 2 GHZ CPU Cores. All VM's were equally created as disk images and were run on 10 networked computers each hosting the four agents (on four core operating environments). Gigabyte iRAM modules were installed towards testing the improvements in I/O access to the data file where all VM's were equally loaded using the Virtual Box open source virtualization software.

6.3 Initial results and discussion

The turnaround and compute time were recorded as follows: we assume N data distributed over P = 2_d tasks, with N an integer multiple of the computation costs which comprise of the initial comparisons performed during the communication phase where d = log P. The former involves a total of P = 2_d comparisons, while the latter requires at most (N_d (d+1) / 2) comparisons. Because the algorithm is perfectly balanced, we assume that idle time is negligible. Our results were obtained by running Gridblast code on Linux Clusters (Fedora Core) with 2.0 GHz Duo core CPU's and 4GB RAM. A heterogeneous set of peers having different configurations were used for running the algorithm as a Grid service using the $A^{3p}viGrid$ agents running on their VM's or individual user space. In this project, our custom created DNA sequence Cellulosome obtained from the Biogas Digester Unit has been used as the database. The size of this sequence is 121,554,183 base pairs (bp). A BLAST search for all contigs of the metagenome data set from the biogas digester unit were searched against the cellulosome protein database, which was carried out and the best matching contigs were identified.

To improve application and agent specific performance, customized Virtual execution environments (Virtual Machines) were created for each of the agents running the $A^{3p}viGrid$ service. An increase in performance after initialization and execution of agents on the VM's was observed. The initial data set was stored and written to scratch disks created in RAM

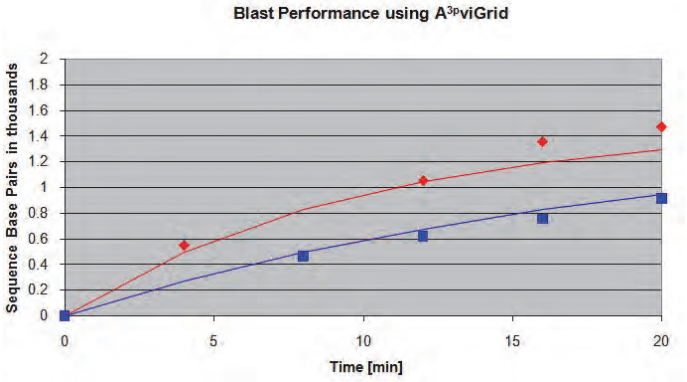

Fig. 5. Graph Showing the Performance of the Sequencing using $A^{3p}viGrid$ platform

along with accessing and storing results on the iRAM installed on the head node (where the initial job was submitted). From the data recorded, we estimated that the initial turnaround time was affected due to an increase in latency posed by the VM's during initialization and data retrieval. The overall turnaround time almost increases two fold during initial execution as resources are allocated dynamically by the agents during execution. The researchers observed that once the data was made available, the execution time was decreased more than half after the agent and its environment were initialized. A two-fold speedup can be observed based on running agents in virtual machines as the input/output data access time is cut by half as resources and data were made available locally to the agents using virtual machines. A coalition based approach to solving a known problem in bioinformatics was undertaken. The use of RAM based scratch disks proved useful in improving the execution times of the BLAST searches on the small scale Grid test bed.

7. Phase 2: Experimentally evaluating the calorific value and energy content of input feed

The calorific value is the most important factor that determines how much energy content is available in the biomass. This in turn is used to estimate the energy potential of the input feed (Biomass fed into a fermenter unit). The calorific value of biogas is ideally estimated to be about 6 kWh/m^3 of gas produced which roughly corresponds to around half a liter of diesel oil. The net calorific value depends on the efficiency of the burners or appliances. Methane gas is the valuable component under the aspect of using biogas as a fuel. Although the calorific value here is a general value assumed, the actual value varies based on the biomass used as feed. The energy potential is determined by the calorific value which has the capacity to produce energy (electricity or gas). An experimental setup (Bio-digester Unit) was created to understand and verify the calorific values of the biomass taken from a small scale bio-digester unit in the Andaman Islands. Gas-liquid chromatography (GLC), or simply gas chromatography (GC), is a common type of chromatography used in analytic chemistry for separating and analyzing compounds that can be vaporized without decomposition. In the initial phase the organic fractions were obtained from a small scale Bio-digester unit in the Andaman Islands. Typical uses of GC include testing the purity of a particular substance, or separating the different components of a mixture (the relative amounts of such

components can also be determined). In some situations, GC may help in identifying a compound. In preparative chromatography, GC can be used to prepare pure compounds from a mixture. In our experiments as discussed below, there is a need for finding out the percentage of pure gas components from our biogas mixture obtained in the laboratory. In gas chromatography, the moving phase (or "mobile phase") is a carrier gas, usually an inert gas such as helium or an un-reactive gas such as nitrogen. The stationary phase is a microscopic layer of liquid or polymer on an inert solid support, inside a piece of glass or metal tubing called a column. The instrument used to perform gas chromatography is called a gas chromatograph (or "aerograph", "gas separator") . The gaseous compounds being analyzed interact with the walls of the column, which is coated with different stationary phases. This causes each compound to elute at a different time, known as the retention time of the compound. The comparison of retention times is what gives GC its analytical usefulness. A gas sample was also obtained to do a comparative study. An initial laboratory setup of the a mini scale Bio-digester unit was setup for the purpose of obtaining sample gas and comparing the measurements with the gas sample obtained from the commercial setup. The Thermal conductivity Detector (TCD) and The Pulsed Flame Photometric Detector (PFPD) were used to detect the gas mixtures with varying temperature ranges in the column. The signal samples were monitored in a work station and graphs were generated accordingly. The phase 2 of our research does a comparative analysis of the differences of creating biogas from commercial bio-digester setup like that of our test case plant at Andaman Islands with that of the experimental verification done using laboratory equipment with similar conditions. The results and inferences are discussed in this chapter later on. In the following sections we will discuss the Laboratory preparation of Biogas creation, capture, monitoring and cleaning; inference of our experimental analysis; and the graphs obtained from the GC detectors that detected the different gas mixtures obtained from the experimental setup.

7.1 Biogas generation

The Initial Setup:

In the initial phase the organic fractions were obtained from a small scale rural Biogas digester unit in the Andaman Islands. A gas sample was also obtained for further comparative study. As shown in Figure 4 an initial Laboratory setup of the a mini scale Bio-digester unit was setup for the purpose of obtaining sample gas and comparing the measurements with the gas sample obtained from a commercial setup. The Thermal conductivity Detector (TCD) and The Pulsed Flame Photometric Detector (PFPD) were used to detect the gas mixtures with varying temperature ranges in the column. A minimum sample of 1 ml gas is constantly required for running the gas chromatography equipment. Flow controllers control the percentage of gas mixtures needed for measuring the signals correctly. Extremely efficient filters remove moisture and oxygen content to extend column life and improve system performance.

The steel vessel shown in Figure 4 holds the Organic fractions obtained from the small scale digester unit. The mixture is sealed in a closed flash with a rubber cork to have anaerobic conditions similar to the bio-digester unit. After all the connections and necessary equipments are installed the water in the bath (steel vessel) is heated to 40 degrees centigrade and run for several weeks. Initially the experiment failed due to too much input fed in. In order to increase the calorific content and give the bacteria a better environment, a 10% quantity of high grain fibre was introduced into the mixture. The experiment was rerun and successful biogas was obtained in about a week's time. The gas bubbles can be seen as shown Figure 6.

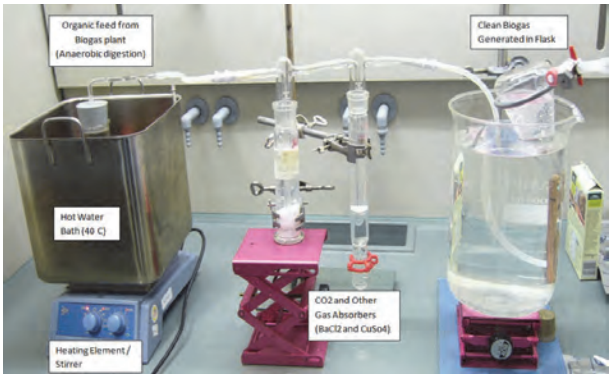

Fig. 6. Our Laboratory Assembled Bio-Digester Unit

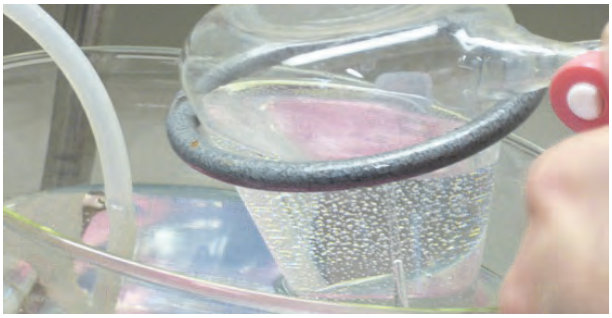

Fig. 7. Biogas Generated

7.2 Capture, detection and monitoring

TCD is most commonly used for the detection of inorganic gases wherein the dual channel TCD can automatically shut off filament current to prevent detector damage in the event of an air leak or loss of carrier gas. Individual control of detector and filament temperature allows optimization of detector performance and decreases maintenance costs.

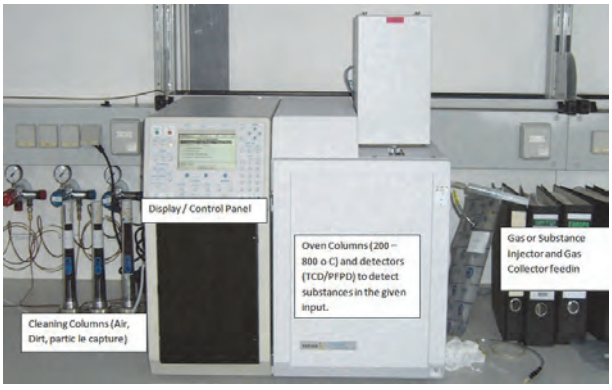

Fig. 8. Capture, Detection and Monitoring

The Pulsed Flame Photometric Detector (PFPD)

The Pulsed Flame Photometric Detector (PFPD) was developed in the early 1990's by Dr. Aviv Amirav (Varian, 2011). Unlike the traditional flame photometric detector which has a continuous flame, the PFPD is based on a pulsed flame for the generation of flame chemiluminescence's. The detector operates with a fuel rich mixture of hydrogen and air. This mixture is ignited and then propagates into a combustion chamber three to four times per second where the flame front extinguishes. Carbon light emissions and the emissions from the hydrogen/oxygen combustion flame are completed in two to three milliseconds, after which a number of heteroatomic species give delayed emissions which can last from 4 to 20 milliseconds. These delayed emissions are filtered with a wide band pass filter, detected by an appropriate photomultiplier tube, and electronically gated to eliminate background carbon emission. In a conventional flame photometric detector (FPD), a sample containing heteroatoms of interest is burned in a hydrogen-rich flame to produce molecular products that emit light (i.e., chemiluminescent chemical reactions). The emitted light is isolated from background emissions by narrow bandpass wavelength-selective filters and is detected by a photomultiplier and then amplified. The detectivity of the FPD is limited by light emissions of the continuous flame combustion products including CH^*, CO_2^*, and OH^*. Narrow bandpass filters limit the fraction of the element-specific light which reaches the PMT and are not completely effective in eliminating flame background and hydrocarbon interferences. The solution to this problem, conceived by Professor Amirav of Tel Aviv University was to set the fuel gas (H_2) flow into the FPD so low that a continuous flame could not be sustained. But by inserting a constant ignition source into the gas flow, the fuel gas would ignite, propagate back through a quartz combustor tube to a constriction in the flow path, extinguish, then refill the detector, ignite and repeat the cycle (Varian 2010). The result was a pulsed flame photometric detector (PFPD). The background emissions from the hydrogen-rich air: hydrogen flame (approximately 10 mL/min H_2 and 40 mL/min Air) is a broad band chemiluminescence. The combustion of hydrocarbons is highly exothermic, rapid and irreversible, producing a light emission by the hydrocarbon products equal to the time for the flame to propagate through the combustor lasting 2 to 3 milliseconds. Many of the chemiluminescent reactions of the heteroatoms such as S, P, N , etc., are less energetic and more reversible, and proceed after the temperature behind the propagating flame has dropped. These heteroatom emissions are therefore delayed from the background emissions (Varian, 2011). By using the leading edge of the flame, the background emission triggers a gated amplifier with an adjustable delay time. Heteroatomic emissions can be amplified to the virtual exclusion of the hydrocarbon background emission through selective amplification of the element-specific emissions which are the basis of the PFPD's unique sensitivity and selectivity.

7.3 Gas cleaning (membrane)

The biogas obtained contains other gases such as H_2S and CO_2 which requires cleaning. Hence a cleaning apparatus was setup. It consists of a trickle water bath that is used in reusing the water used for gas cleaning. A couple of air compressors is used to generate 4 - 10 Bars of pressure over the gas used for compression. A set of two membranes were used for separating the CO_2 and CH_4 from the compressed gas column.

Tiny droplets of water are used for washing down the mixture which is then taken into the absorption chamber at 10 Bar pressure.

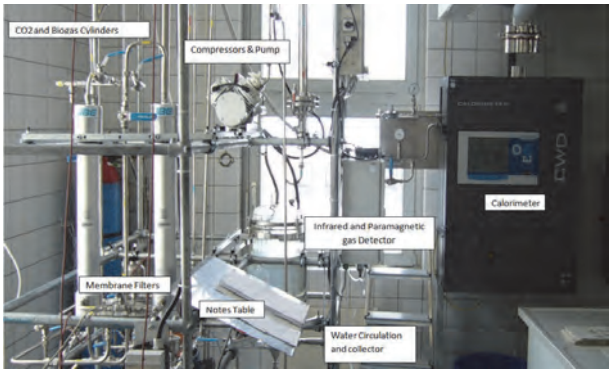

Fig. 9. Gas Detection and Cleaning under Laboratory Conditions

$$CO_2 + H_2O - > CO_2Gas \tag{2}$$

The desorption column on the other hand uses the gas compressor to reduce the pressure to < 1 Bar. The glass cylinder is used to see the water level and how much fresh water might be needed to be added on. Membranes utilize steel wool or other materials to create water droplets into a mild flow. The output from the desorption device is the then sent to the two membrane columns and the gas stream is again raised from 2 Bar to 10 Bar of pressure and is sent back to the absorption column. The membranes have no holes and gas diffusion takes place to separate the CO_2 from the CH_4. The calorimeter also called as the Wobber Index is used to measure the CH_4 content at 10 Bar pressure. The Wobber index is an indicator of the calorific value of the gas mixture. The detector on the left of the calorimeter as shown is the paramagnetic and infrared radiation detector used in detecting gas mixtures of upto 10 PPM. Gas from the scrubbing column has a level of moisture content (H_2O) in it with very low specific heat value. The detector is also used to find out the moisture content in the gas obtained after cleaning. The process could be repeated a number of times to obtain high quality CH_4 Gas. In the next section we will discuss the results obtained from the Gas Chromatography.

7.4 Results and Conclusion

As shown in the figures below, we can note major differences in the signal measurements due to the following:

- The experiments have slight variations to the conditions in the bio-digester unit. Eg. Temperature variations, Human handling and gas capture.
- The initial results are for different gas mixtures having slight feed variations.
- Channel 1 indicates signals for CO_2, H_2S, CH_4 and Nitrogen.
- The TPD used is a universal detector of all gas mixtures.
- The PFPD measures the sulphur content alone as observed in the graphs generated.
- Significant changes in signal levels can be observed.
- We understand that very similar conditions of a commercial bio-digester unit are impossible to mimic given the quantity of the organic fractions and the conditions

simulated. Hence, we went in for DNA extraction, sequencing and calorific content testing and verification.

- It can be observed from the graphs that the signals obtained for P, CH_4, CO_2, Sulphur, etc are comparable to the commercial setup as the input organic fractions were obtained from the Andaman Bio-digester unit.

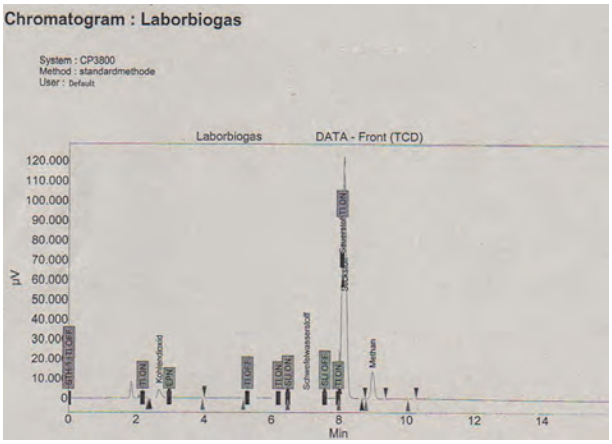

Fig. 10. TCD results from Laboratory Gas Sample

Index	Name	Time$[Min]$	Qty $[Vol-\%]$	Height [µV]	Area [µV Min]	Area %
2	CO_2	2.66	3.56	4261.3	518.8	2778
5	H_2S	7.02	0.01	6.1	2.3	0.012
8	O_2	8.09	17.67	70192.8	3841.9	20.568
8	N	8.15	66.31	122241.7	12344.1	66.085
10	CH_4	8.99	3.31	13595.6	1600.1	8.566
Total			90.85	234571.5	18679.1	100.00

Table 1. TCD results from Laboratory Gas Sample

Index	Name	Time$[Min]$	Qty $[Vol-\%]$	Height [µV]	Area [µV Min]	Area %
2	CO_2	2.36	42.12	25278.1	6140.0	18.551
5	H_2S	7.02	0.02	16.1	5.6	0.017
8	O_2	8.07	0.56	2312.4	121.7	0.368
8	N	8.20	3.94	10676.1	732.6	2.214
10	CH_4	8.73	53.22	140050.2	25707.3	77.670
Total			99.85	202506.2	33098.2	100.000

Table 2. TCD results from Andaman Islands Gas Sample

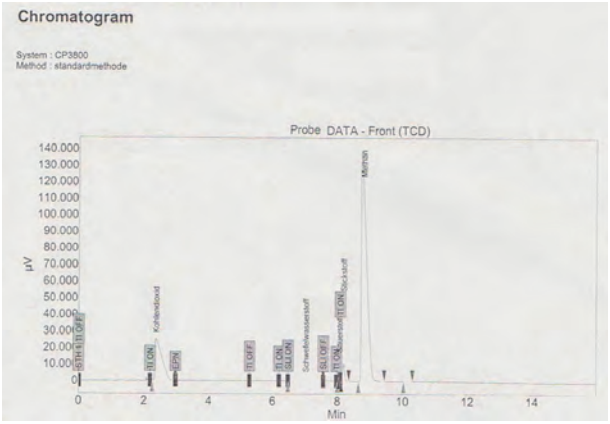

Fig. 11. TCD results from Laboratory Gas Sample

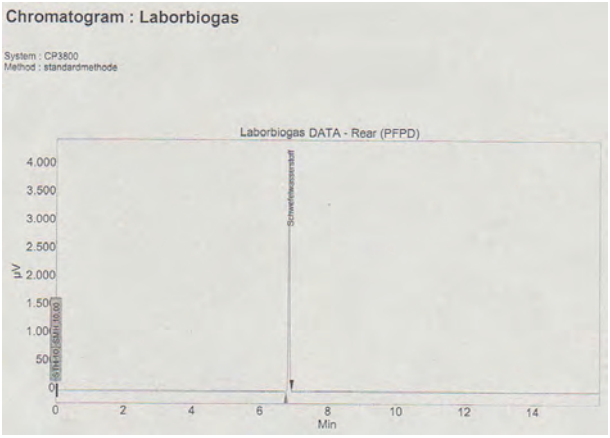

Fig. 12. TCD results from Laboratory Gas Sample

Index	Name	Time$[Min]$	Qty $[Vol-\%]$	Height [µV]	Area [µV Min]	Area %
1	H_2S	6.85	49.97	4250.2	190.8	100.000
Total			49.97	4250.2	190.8	100.000

Table 3. PFPD results from Andaman Islands Gas Sample

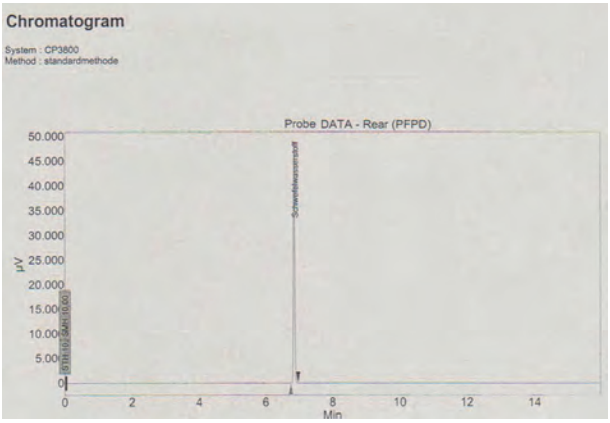

Fig. 13. PFPD results from Laboratory Gas Sample

Index	Name	Time$[Min]$	Qty $[Vol - \%]$	Height [µV]	Area [µV Min]	Area %
1	H_2S	6.85	210.90	48439.5	2117.4	100.000
Total			210.90	48439.5	2117.4	100.000

Table 4. PFPD results from Andaman Islands Gas Sample

In this chapter we discussed the importance of Bioenergy; the role played by Bioinformatics and associated Computational tools. More importantly we identified the various tools used by biologists in everyday drug discovery and genetic engineering; the various bottlenecks incurring due to the exponential increase in datasets due to new discoveries and publication of data regularly using the world wide web; finally we can understand and utilize high performance computational approaches, tools and platforms (such as Grids and Supercomputers) to solve these problems on a day to day basis.

8. References

Angelidaki, I., Ellegaard, L., (2003). Codigestion of manure and organic wastes in centralized biogas plants: status and future trends. *Appl. Biochem. Biotechnol*. 109, pp. 95–105.

Altschul, S.F., Gish W, Miller W, Myers EW, Lipman DJ., (1992). Basic local alignment search tool. *Journal of Molecular Biology* 215, pp. 403–410.

Bayer, E.A., Belaich, J.P., Shoham, Y., Lamed, R., (2004). The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol*. 58, pp. 521–554.

Benson, G., (1999). Tandem repeats finder - a program to analyze DNA sequences. *Nucleic Acids Research*. 27, pp. 573–580.

Benson, G., (2001). Tandem Cyclic Alignment. *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching (CPM)*, LNCS, 2089, pp. 118–130.

Blaut, M., 1994.Metabolism of methanogens. Antonie Van Leeuw. 66, pp. 187–208.

Burt, (2006). *Written Statement of Dr. Stanley Burt for the Senate Committee on Commerce, Science, and transportation Subcommittee on Technology, Innovation, and Competitiveness.*, www.nscee.edu, Last accessed on: 22 Nov 2010.

Claverie, J.M. and Notredame, C., (2003). *Bioinformatics for Dummies, Wiley Publishing Inc., U.S.A.*

Chachkhiani, M., Dabert, P., Abzianidze, T., Partskhaladze, G., Tsiklauri, L., Dudauri, T., Godon, J.J., 2004. 16S rDNA characterisation of bacterial and archaeal communities during start-up of anaerobic thermophilic digestion of cattle manure. Bioresour. Technol. 93, pp. 227–232.

Cirne, D.G., Lehtomaki, A., Bjornsson, L., Blackall, L.L., (2007). Hydrolysis andmicrobial community analyses in two-stage anaerobic digestion of energy crops. *J. Appl.Microbiol*. 103, pp. 516–527.

Daniels, L., (1992). Biotechnological potential of methanogens. *Biochem. Soc. Symp*. 58, pp. 181–193.

Deppenmeier, U., 2002. The unique biochemistry of methanogenesis. Prog. Nucleic Acid Res. Mol. Biol. 71, pp. 223–283.

Deppenmeier, U., Muller, V., Gottschalk, G., 1996. Pathways of energy conservation in methanogenic archaea. Arch. Microbiol. 165, pp. 149–163.

Drake, H.L., Kusel, K.,Matthies, C., (2002). Ecological consequences of the phylogenetic and physiological diversities of acetogens. *Antonie Van Leeuw*. 81, pp. 203–213.

Drake, H.L., Daniel, S.L., Kusel, K., Matthies, C., Kuhner, C., Braus-Stromeyer, S., (1997). Acetogenic bacteria:what are the in situ consequences of their diversemetabolic versatilities? *Biofactors* 6, pp. 13–24.

Entcheva, P., Liebl,W., Johann, A., Hartsch, T., Streit,W.R., 2001. Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Appl. Environ. Microbiol. 67, pp. 89–99.

Ferry, J.G., (1992). Biochemistry of methanogenesis. *Crit. Rev. Biochem. Mol. Biol*. 27, pp. 473–503.

Ferry, J.G., (1999). Enzymology of one-carbon metabolism in methanogenic pathways. *FEMS Microbiol. Rev*. 23, pp. 13–38.

Gibas, C., and Jambeck, P., (2001). Developing Bioinformatics Computer Resources, *Ø'Reilly & Associates, Inc., U.S.A*.

Henne, A., Daniel, R., Schmitz, R.A., Gottschalk, G., 1999. Construction of environmental DNA libraries in Escherichia coli and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl. Environ. Microbiol. 65, pp. 3901–3907.

Heuristics definition, (2010). http://en.wikipedia.org/wiki/Heuristics, Last accessed on 04 Oct 2010.

Higgins, D.G., Thompson, J.D., and Gibson, T.J., (1996), Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, 266, pp. 383–402.

Huang, L.N., Zhou, H., Chen, Y.Q., Luo, S., Lan, C.Y., Qu, L.H., 2002. Diversity and structure of the archaeal community in the leachate of a full-scale recirculating landfill as examined by direct 16S rRNAgene sequence retrieval. FEMSMicrobiol.Lett. 214, pp. 235–240.

IEA (2007), Bioenergy, Potential Contribution of Bioenergy to the World's Future Energy Demand, *IEA Bioenergy Programme, Paris:IEA*

Karakashev, D., Batstone, D.J., Angelidaki, I., (2005). Influence of environmental conditions onmethanogenic compositions in anaerobic biogas reactors. *Appl. Environ. Microbiol*. 71, pp. 331-338.

Phil O'Keefe, Geoff O'Brien, and Nicola Pearsall, (2010). The Future of Energy Use, *Earthscan*, pp. 276, 2nd Ed, ISBN 9781844075041

Klocke, M., MÍ ahnert, P., Mundt, K., Souidi, K., Linke, B., 2007. Microbial community analysis of a biogas-producing completely stirred tank reactor fed continuously with fodder beet silage as mono-substrate. Syst. Appl. Microbiol. 30, pp. 139–151.

Krause, L., McHardy, A.C., Nattkemper, T.W., Puhler, A., Stoye, J., Meyer, F., 2007. GISMOŮgene identification using a support vector machine for ORF classification. Nucleic Acids Res. 35, pp. 540–549.

Lueders, T., Chin, K.J., Conrad, R., Friedrich, M., 2001. Molecular analyses of methyl-coenzyme M reductase alpha-subunit (mcrA) genes in rice field soil and enrichment cultures reveal the methanogenic phenotype of a novel archaeal lineage. Environ. Microbiol. 3, pp. 194–204.

R.S. Meyers, R. Amaro, Z. Luthey-Schulten, and V.J. Davisson, (2005). Reaction Coupling through Interdomain Contacts in Imidazole Glycerol Phosphate Synthase. *Biochemistry*, 13; 44(36):119, pp. 74–85.

Lynd, L.R., Weimer, P.J., van Zyl, W.H., Pretorius, I.S., (2002). Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol. Mol. Biol*. Rev. 66, pp. 506-577.

Mladenovska, Z., Dabrowski, S., Ahring, B.K., 2003. Anaerobic digestion of manure and mixture of manure with lipids: biogas reactor performance and microbial community analysis.Water Sci. Technol. 48, pp. 271–278.

Myint, M., Nirmalakhandan, N., Speece, R.E., (2007). Anaerobic fermentation of cattle manure: modeling of hydrolysis and acidogenesis.*Water Res*. 41, pp. 323-332.

Optimization definition, (2010). http://en.wikipedia.org/wiki/Optimization, Last accessed on 4 November, 2010.

Raes, J., Foerstner, K.U., Bork, P., 2007. Get the most out of your metagenome: computational analysis of environmental sequence data. Curr. Opin. Microbiol. 10, pp. 490–498.

Reeve, J.N., (1992). Molecular biology of methanogens. *Annu. Rev. Microbiol*. 46, pp. 165-191.

Reeve, J.N., NÍ olling, J., Morgan, R.M., Smith, D.R., (1997). Methanogenesis: genes, genomes, and who's on first? *J. Bacteriol*. 179, pp. 5975-5986.

Schink, B., (1997). Energetics of syntrophic cooperation inmethanogenic degradation. *Microbiol. Mol. Biol. Rev*. 61, pp. 262-280.

Schink, B., (2006). Syntrophic associations in methanogenic degradation. *Prog. Mol.Subcell Biol*. 41, pp. 1-19.

Schlüter et. al., (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology, *J Biotechnol*. 2008 Aug 31;136(1-2), pp. 77-90 PMID: 18597880.

Shankaranarayanan, (2011). Evaluating Biotechnology Processes Using Distributed Systems, Master of Philosophy Thesis. SASTRA University.

Shigematsu, T., Tang, Y., Kobayashi, T., Kawaguchi, H., Morimura, S., Kida, K., (2004). Effect of dilution rate on metabolic pathway shift between aceticlastic and nonaceticlasticmethanogenesis in chemostat cultivation. *Appl. Environ.Micro-biol*. 70, pp. 4048-4052.

A. Schluter et al., 2008, The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technologyJournal of Biotechnology 136, pp. 77–90

Stanley K. Burt, Phillip Yen, Oscar N. Ventura and Raul E. Cachau, (2007). Fast pattern recognition of protein three dimensional features using a bit-pattern approach as a prescreen, *Biophysical Journal* (S), pp. 567A–567A.

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

Varian (2011). Varian, Inc, www.varianinc.com, Last accessed:10 Feb 2011.

Waterman, Smith, Temple F.; and Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences". *Journal of Molecular Biology*. 147, pp. 195–197. doi:10.1016/0022-2836(81)90087-5

Weiland, P., (2003). Production and energetic use of biogas from energy crops and wastes in Germany. *Appl. Biochem. Biotechnol*. 109, pp. 263–274.

Yadvika, Santosh, Sreekrishnan, T.R., Kohli, S., Rana, V., (2004). Enhancement of biogas production from solid substrates using different techniques-a review. *Bioresour. Technol*. 95, pp. 1–10.