Computational and Comparative Investigations of Syntrophic Acetate-oxidising Bacteria (SAOB)

Genome-guided analysis of metabolic capacities and energy conserving systems

Shahid Manzoor
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2014

Acta Universitatis agriculturae Sueciae
2014:56

Cover: Bioinformatics helping the constructed biogas reactors to run efficiently.
(photo: (*Shahid Manzoor*)

ISSN 1652-6880
ISBN (print version) 978-91-576-8060-0
ISBN (electronic version) 978-91-576-8061-7

Print: SLU Service/Repro, Uppsala 2014

Computational and Comparative Investigations of Syntrophic Acetate-oxidising Bacteria (SAOB) – Genome-guided analysis of metabolic capacities and energy conserving systems.

Abstract

Today's main energy sources are the fossil fuels petroleum, coal and natural gas, which are depleting rapidly and are major contributors to global warming. Methane is produced during anaerobic biodegradation of wastes and residues and can serve as an alternative energy source with reduced greenhouse gas emissions. In the anaerobic biodegradation process acetate is a major precursor and degradation can occur through two different pathways: aceticlastic methanogenesis and syntrophic acetate oxidation combined with hydrogenotrophic methanogenesis. Bioinformatics is critical for modern biological research, because different bioinformatics approaches, such as genome sequencing, *de novo* assembly sequencing and transcriptomics sequencing are providing a distinctly better understanding at the genomic level by predicting genes and pathways and by deciphering the relationships between genotype and phenotype.

This thesis describes the genomic analysis of three syntrophic acetate-oxidising bacteria (SAOB), namely *Tepidanaerobacter acetatoxydans*, *Clostridium ultunense* and *Syntrophaceticus schinkii*. These isolates have the ability to perform syntrophic acetate oxidation in the presence of a partner methanogen, which ultimately produces methane in the final step of anaerobic digestion. The genomes were assembled using NGS data and the genomic behaviour was determined through genome annotation. Metabolic pathway analysis revealed the physiological attributes of the SAOB regarding substrate utilisation, intermediate metabolism, energy conservation and genes of the Wood-Ljungdahl pathway, which are known to be involved in acetate oxidation.

The results showed that the three SAOB use contrasting strategies for syntrophic acetate oxidation (SAO): *T. acetatoxydans* possesses all genes involved in the W-L pathway except formate dehydrogenase and thus requires a syntrophic formate-utilising methanogenic partner; *S. schinkii* possesses the complete set of genes required for the W-L pathway to oxidise acetate in the presence of a hydrogen-utilising methanogenic partner; and *C. ultunense* uses different ways to oxidise acetate because it does not contain the complete set of W-L pathway genes. Moreover, the three SAOB differ from each other as regards organisation of the W-L pathway genes operon.

Keywords: hydrogenotrophic methanogenesis, biogas, Wood-Ljungdahl pathway, next generation sequencing, *de novo* assembly, acetogenesis, metabolism.

Author's address: Shahid Manzoor, SLU, Department of Animal Breeding and Genetics, P.O. Box 7023, 750 07 Uppsala, Sweden
E-mail: Shahid.Manzoor@slu.se

Dedication

To my beloved parents, wife, daughters and lovely son.

Det är skönare lyss till en sträng, som brast, än att aldrig spänna en båge.
Verner von Heidenstam

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Manzoor S., Bongcam-Rudloff E., Schnürer A., Müller B. (2013) First genome sequence of a Syntrophic Acetate-Oxidizing Bacterium, *Tepidanaerobacter acetatoxydans* strain Re1. Genome Announc. 1 *:e00213-12; doi:10.1128/genomeA.00213-12.*

II Manzoor S., Müller B., Bongcam-Rudloff E., Schnürer A. (2013) Draft genome sequence of *Clostridium ultunense* strain Esp, a Syntrophic Acetate-Oxidizing bacterium. Genome Announc. 1: *e00107-13; doi:10.1128/genomeA.00107-13.*

III Manzoor S., Müller B., Niazi A., Schnürer A., Bongcam-Rudloff E. (2014) Working draft genome sequence of the mesophilic acetate oxidizing bacterium *Syntrophaceticus schinkii* strain Sp3. Standards In Genomic Sciences. (Submitted)

IV Manzoor S., Müller B., Niazi A., Schnürer A., Bongcam-Rudloff E. (2014) Genome guided analysis of the metabolic capacities and energy conserving systems of the SAOB *Tepidanaerobacter acetatoxydans*. PLOS ONE (Submitted)

V Manzoor S., Müller B., Niazi A., Schnürer A., Bongcam-Rudloff E. The syntrophic acetate oxidizing bacterium *Syntrophaceticus schinkii* strain Sp3-A genome scale analysis. (Manuscript)

The contribution of Shahid Manzoor to the papers included in this thesis was as follows:

I **Papers I-III** Assembled the three SAOB genomes using NGS data, performed annotations and planned and wrote the articles, with suggestions and comments from the supervisors.

II **Paper IV-V** Performed the genome scale and metabolic pathways analysis and planned and wrote the articles with comments and suggestions from the supervisors.

Additional publications

- Manzoor S., Niazi A., Bejai S., Meijer J., & Bongcam-Rudloff E. (2013) Genome Sequence of a Plant-Associated Bacterium, *Bacillus amyloliquefaciens* Strain UCMB5036 Genome Announc.1: e00111-13.

- Niazi A., Manzoor S., Bejai S., Meijer J., & Bongcam-Rudloff E. (2014) Complete genome sequence of a plant associated bacterium *Bacillus amyloliquefaciens* strain UCMB5033. Standards In Genomic Sciences, 9(3). Doi: 10.4056/doisigs.4758653.

- Niazi A., Manzoor S., Asari S., Bejai S., Meijer J., & Bongcam-Rudloff E. (2014). Analysis of the genome of *Bacillus amyloliquefaciens* subsp. plantarum UCMB5113: a rhizobacterium that improves plant growth and stress management. PLOS ONE, (Accepted).

- Mushtaq M., Manzoor S., Pringle M., & Rosander A. (2014) Draft genome sequence of *Treponema phagedenis* strain V1, isolated from bovine digital dermatitis. Standards In Genomic Sciences. (Submitted)

- Niazi A., Bejai S., Maqbool K., Manzoor S., Meijer J., & Bongcam-Rudloff E. Comparative genomics reveals novel insights into the divergence between *Bacillus amyloliquefaciens* strains. (Manuscript)

Abbreviations

SAO	Syntrophic acetate oxidation
SAOB	Syntrophic acetate oxidising bacteria
NGS	Next generation sequencing
RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
PCR	Polymerase chain reaction
bp	Base pair
Kb	Kilo base pair
Mb	Mega base pairs
Gb	Giga base pairs
BLAST	Basic local alignment search tool
DM	Data mining
ML	Machine learning
RefSeq	Reference sequence
W-L	Wood-Ljungdahl pathway
GO	Gene ontology
COG	Cluster of orthologous group
ORF	Open reading frame
CDS	Coding sequence
TF	Transcription factor

"*New problems demand new solutions. New solutions create new problems*."

(Solomon Short)

1 Introduction

The emerging field of sciences known as bioinformatics started its journey fairly modestly, with the almost manually assembled *Caenorhabditis elegans* genome (Consortium, 1998). It gathered some momentum on its way to the next milestone, with the first human draft genome (Consortium, 2001), and since then it has developed to offer full-blown services and to become a common nexus in the field of life sciences.

Bioinformatics is a highly interdisciplinary field that is progressing at an exponential pace due to recent advances in new technologies, software and production of large volumes of data. Now, this relatively new science of bioinformatics is at the heart of many aspects of modern biological research, particularly whole genome studies and next generation sequencing (NGS) data analysis. The NGS technologies have changed the path of genomics and proteomics research by providing massive parallel sequencing with high accuracy and greater throughput. This is now driving biological research to several applications at the DNA or RNA level, including whole genome sequencing, *de novo* assembly, resequencing and transcriptomics sequencing, which may lead to a better understanding of cells by predicting genes and cellular pathways (Li *et al.*, 2010), as well as improving understanding of the relationships between genotype and phenotype (Vallender, 2011; Voelkerding *et al.*, 2010).

Today's industrial society is heavily dependent on energy and energy consumption is predicted to increase by up to 57% by 2030 (Administration, 2007). At the same time, the main sources of energy fossil fuels (oil, coal, gas), are depleting at a rapid rate are becoming more difficult to extract and are increasing the risk of environmental pollution. In addition, combustion of these fossil fuels is a main contributing factor to global warming by increasing the greenhouse gas effects. In recent years all these factors have resulted in a demand for alternative sustainable energy sources such as methane-containing

biogas (bio-methane) that can be produced by anaerobic degradation of organic material and can replace fossil fuels. Anaerobic digestion can also be a suitable solution for waste and wastewater treatments, where the residues can be used as organic-mineral fertiliser to avoid the negative effects of chemical fertiliser manufacture.

A number of microorganisms are involved in the multi-step process of anaerobic degradation, where they work in close interaction with each other. During this process, acetate is the most important precursor of methane (Kaspar & Wuhrmann, 1978; Hobson & Shaw, 1974) and two different pathways are responsible for the production of methane from acetate: aceticlastic methanogenesis, which is carried out by acetate-cleaving methanogens, or hydrogenotrophic methanogenesis, which involves two reactions. First syntrophic acetate-oxidising bacteria (SAOB) convert acetate to hydrogen (H_2) and carbon dioxide ($CO_{2)}$ and then hydrogen-consuming methanogens reduce CO_2 to methane (Karakashev *et al.*, 2006; Zinder & Koch, 1984). The development of syntrophic acetate oxidation (SAO) depends on many factors, including ammonia level, acetate concentration, temperature, dilution rate and methanogenic population structure (Hao *et al.*, 2011; Schnurer & Nordberg, 2008; Karakashev *et al.*, 2006; Shigematsu *et al.*, 2004; Ahring *et al.*, 1993; Petersen & Ahring, 1991). Several recent studies of methanogenic systems have emphasised the significance of SAO for biogas reactors (Westerholm *et al.*, 2012; Hao *et al.*, 2011; Sasaki *et al.*, 2011; Shimada *et al.*, 2011), whereas previously the main focus was on the activity of the aceticlastic methanogens (Karakashev *et al.*, 2006). Consequently, further research within this field, especially with the support of bioinformatics is important to improve our understanding of molecular mechanisms and metabolic pathways of SAOB. This knowledge can be applied to improve the efficiency of these organisms in biogas reactors.

This thesis describes the results of a SAOB project in which informatics technology was applied in order to assemble and annotate some SAOB genomes, using NGS data analysis. This was followed by metabolic pathway analysis using bioinformatics to identify the associated molecular mechanisms such as substrate utilisation, energy conservation and acetogenesis.

"*Understanding the laws of nature does not mean that we are immune to their operations*"

(Solomon Short)

2 Background

2.1 Importance of bacteria

People generally associate bacteria with disease, but the reality is that only a minority of bacteria are pathogenic (disease-causing), while most others are non-pathogenic (beneficial for human and plant life and environment or neutral; not interacting at all). There are billions of microbes (bacteria) that have symbiotic-relationships with other organisms (human, plant, animals), as was first discovered in 1869 (Bary, 1954). Recent novel discoveries demonstrates the new ways in which microbes (bacteria) can be beneficial for the whole planet (Stephanie, 2011), such as: some specific bacterial strains, *e.g. Alcanivorax,* can be used to clean up environmental pollutants, some deep sea bacteria *e.g. Shewanella,* digest toxic waste; *Geobactor* bacteria with nanowires have the ability to immobilise harmful material such as uranium, avoiding contamination of groundwater; some specific types of bacteria are able to consume non-biodegradable plastic, and some types of bacteria have the ability to use copper from the environment to metabolise methane and thus eliminate both greenhouse gases and toxic heavy metals.

2.2 Anaerobic environment

Anaerobic bacteria may be obligate anaerobes, which can only grow in anoxic environments, since oxygen is toxic for them, and they depend on other substrates as electron acceptors. However, some aerotolerant anaerobes can survive in the presence of oxygen. Anaerobic microorganisms are considered to be one among the oldest life forms on the earth. They have the most diverse metabolic pathways and lifestyles ever developed through evolution. For example, they have various anaerobic respiration and energy gaining reactions, fermentation pathways, carbon fixation mechanisms and syntrophic carbon

pathways, enabling life on the thermodynamic edge. Anaerobic microbial communities play a substantial role in eliminating natural contamination and, more importantly, in various biotechnological processes, including wastewater treatment, anaerobic digestion of biomass and biowaste and biofuel production (de Souza, 2013; Koukkou, 2011; Diaz, 2008; Gallert & Winter, 2008; Ohmiya *et al.*, 2005; Scow & Hicks, 2005).

2.3 Importance of Biogas

A large number of organic waste types can be used as substrate for the production of biogas containing methane (CH_4) during anaerobic degradation including animal manure, agricultural residues and by-products, sewage sludge, source-separated household wastes and organic industrial waste (Angelidaki *et al.*, 2011; Ahring, 2003). Bio-methane is an energy-rich component of the biogas and can be a possible replacement for fossil fuels in the production of electricity, heat & power, chemicals and materials and vehicle fuel (Weiland, 2010). The use of biogas as an alternative energy source can also offer a great benefit in the form of a more healthy environment by reducing the emissions of ammonia and methane that otherwise occur from composting or storage of untreated animal manure (Borjesson & Mattiasson, 2008; Borjesson & Berglund, 2007), and ultimately contribute to global warming.

2.4 Syntrophic relationship

Syntrophy can be defined as any type of crossfeeding of molecules between microbial species, but a restricted definition is applied for anaerobic syntrophic metabolism. In that case, syntrophy is a close mutualistic interaction in a very specific nutritional situation where the level of intermediates exchanged between the partners must be kept low for efficient cooperation and syntrophic partners combine their metabolic capabilities to catabolise a substrate that neither one of them can catabolize alone. Syntrophy operates close to thermodynamic equilibrium, whereby both partners must share the limited energy released by their overall reactions (Sieber *et al.*, 2012). Syntrophic metabolism occurs ubiquitously in the microbial community without restriction to any particular phylogenetic group of microorganisms.

The first syntrophic interaction identified was between phototrophic green sulphur bacteria and chemolithotrophic sulphate-reducing bacteria, with exchange of sulphur compounds between the partners (Biebl & Pfennig, 1978). Subsequently, a thermodynamically interdependent syntrophic lifestyle was

discovered between fermentative bacteria and methanogenic archaea that involved the exchange of hydrogen/formate between the partners (McInerney *et al.*, 1979; Bryant *et al.*, 1967). Syntrophic metabolism is found in three groups within the phylum *Firmicutes;* i) Members of the family *Syntrophomonadaceae,* syntrophically metabolise fatty acids in association with hydrogen/formate using microorganisms (Sobieraj & Boone, 2006); ii) syntrophic species in the *Desulfotomaculum* lineage (*D. thermocisternum* and *D. thermobenzoicum*) syntrophically oxidise propionate (Nilsen *et al.*, 1996); and iii) members of the family *Thermoanaerobacteraceae* syntrophically oxidise acetate, *e.g. Thermoacetogenium phaeum (Hattori et al., 2000)* and *Syntrophaceticus schinkii* (Westerholm *et al.*, 2010).

2.5 Syntrophic acetate oxidation (SAO)

Barker (1936) first hypothesised that methanogenesis from acetate could be possible in a two-step reaction whereby SAOB oxidise acetate into hydrogen/formate (H_2/CO_2) and then hydrogenotrophic methanogens convert CO_2 into methane. This reaction was confirmed around half a century later by Zinder & Koch (1984).

During syntrophic acetate oxidation (SAO), the SAOB and the methanogen are mutually dependent on each other for performing their metabolic activities, because acetate oxidation can only proceed if the hydrogen/formate level is kept low by the syntrophic methanogenic partner (Schink, 2002; Schink, 1997; Stams, 1994). Under standard conditions, the oxidation of acetate to CO_2 and H_2 is a thermodynamically unfavourable reaction that can only proceed when hydrogenotrophic methanogens consume the hydrogen (Hattori, 2008; Stams, 1994). SAOB works more efficiently if the H_2/formate level is kept low, but the level must also be sufficient to favour the hydrogenotrophic methanogens. These contradictory conditions require the H_2 and formate level to be maintained within a low narrow range (Stams, 1994).

A number of studies have shown that SAO contribute at a relatively high level in conversion of acetate to methane in biogas reactors operating under diverse conditions (Westerholm *et al.*, 2012; Hao *et al.*, 2011; Sasaki *et al.*, 2011; Shimada *et al.*, 2011; Laukenmann *et al.*, 2010; Schnurer & Nordberg, 2008; Karakashev *et al.*, 2006; Shigematsu *et al.*, 2004; Schnurer *et al.*, 1999). In addition, SAO have been observed to occur in a diverse range of natural environments, including lake sediments (Nusslein *et al.*, 2001), oil reservoirs (Gray *et al.*, 2011; Jones *et al.*, 2008), soil (Chauhan & Ogram, 2006) and rice paddy soil (Rui *et al.*, 2011; Liu & Conrad, 2010).

2.6 Syntrophic acetate oxidising bacteria (SAOB)

The SAOB belong to the physiological group of acetogens which are obligate anaerobes and as a group differ widely in their morphological, nutritional and physiological properties. The term 'acetogen' has been previously defined by Drake (1994) as:

> Acetogen: an anaerobe that can use the acetyl-CoA pathway 1) as a mechanism for the reductive synthesis of acetyl-CoA from CO_2, 2) in a terminal electron-accepting, energy-conserving process, and 3) as a mechanism for the fixation (assimilation) of CO_2 in the synthesis of cell carbon.

SAO has been found to occur in biogas processes (Schnurer & Nordberg, 2008; Karakashev *et al.*, 2006; Karakashev *et al.*, 2005; Schnurer *et al.*, 1999; Zinder & Koch, 1984), where methane is generally considered to be produced as a result of aceticlastic methanogenesis (Zinder, 1984). Interestingly, with the increased ammonia levels released during the degradation of protein-rich materials, SAOB take over and oxidise acetate to H_2/formate and CO_2. This is enabled thermodynamically by the consumption of hydrogen through hydrogenotrophic methanogenesis, resulting in the production of methane (Westerholm *et al.*, 2012; Schnurer & Nordberg, 2008). Other factors such as operating parameters, acetate concentration and microbial community structures may also be responsible for this shift (Karakashev *et al.*, 2006; Karakashev *et al.*, 2005). The process of biogas production through anaerobic digestion can be categorised into four different stages: hydrolysis, fermentation, anaerobic oxidation and methanogenesis (Figure 1).

During the hydrolysis phase the original complex organic compounds (carbohydrates, proteins, fats) are broken down into simpler soluble monomers compounds (sugars, amino acids, and fatty acids) due to the action of exoenzymes secreted by hydrolytic bacteria (Ramsay & Pullammanappallil, 2001; Mackie *et al.*, 1991).

In the next step, several anaerobic oxidation reactions or fermentative processes are carried out by a number of fermentative bacteria and hydrogen-producing acetogens, resulting in the degradation of soluble compounds and the production of acetate, formate, carbon dioxide, hydrogen, alcohols, short-chain fatty acids and ammonia (Angelidaki *et al.*, 2011). At this stage, syntrophic association is required for the removal of products by methanogens to make this step thermodynamically favourable (Schink, 1997).

In the final stage, methanogenic *archaea* produce methane as the major product of their energy metabolism (Whitman *et al.*, 2006) and there can be two possibilities regarding these methanogens: 1) Acetate-cleaving aceticlastic methanogens (which perform acetotrophic methanogenesis) or 2) hydrogen-

consuming hydrogenotrophic methanogens (which perform hydrogenotrophic methanogenesis) (Demirel & Scherer, 2008).

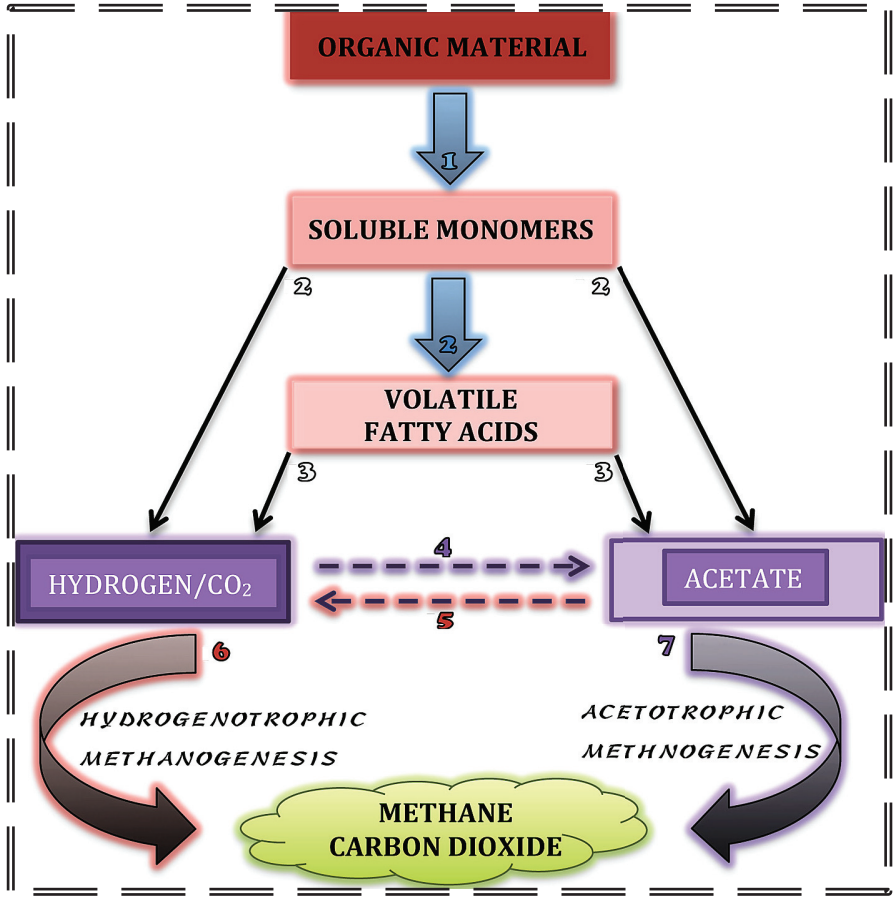

Figure 1. The process of anaerobic degradation of organic materials to produce methane and carbon dioxide may comprise the different stages: 1) Hydrolysis, 2) fermentation, 3) anaerobic oxidation, 4) hydrogen oxidation 5) syntrophic acetate oxidation (SAO), 6) hydrogenotrophic methanogenesis, and 7) aceticlastic methanogenesis. Modified from Zinder (1984).

A thermophilic SAOB *Reveribactor* was described for the first time in 1988 (Lee & Zinder, 1988), but unfortunately that bacterium was lost before any phylogenetic position could be established. Since then, a restricted number of SAOB have been isolated and characterised, including the thermophiles *Thermacetogenium phaeum* (Hattori *et al.*, 2000) and *Thermotoga lettingae* (Balk et al., 2002), the thermotolerant *Tepidanaerobacter acetatoxydans* (Westerholm *et al.*, 2011b) and the mesophiles *Clostridium ultunense* (Schnurer *et al.*, 1996), and *Syntrophaceticus schinkii* (Westerholm *et al.*,

2010). Of all these SAOB, only *T. lettingae* belongs to the phylum Thermotogae, while the other four belong to the phylum Firmicutes. The SAO phenotype has also been found in some other bacterial phyla such as members of *Geobactor*, and *Anaeromyxobacter* in anoxic rice paddy soil (Hori *et al.*, 2007), *Smithella* in methanogenic crude oil-degrading enrichment cultures (Gray *et al.*, 2011), *Betaproteobacteria* and *Nitrospira* in lake sediments (Schwarz *et al.*, 2007). Some syntrophic bacteria have also been shown to possess the ability to oxidise acetate in syntrophic association with bacteria other than hydrogen-consuming methanogen partners. *e.g.* the alkali bacterium *Contubernalis alkalaceticum* can oxidise acetate in co-culture with an hydrogenotrophic alkaliphilic sulphate reducer to produce sulphide (Zhilina *et al.*, 2005) and *Geobacter sulfurreducens,* can oxidise acetate syntrophically with nitrate or sulphate-reducing hydrogenotrophic bacteria (Cord-Ruwisch *et al.*, 1998), which may also indicate that syntrophy is probably not a common component in biogas-producing consortia.

To date, only a limited number of SAOB have been sequenced and published. These include *Th. phaeum* (Oehler *et al.*, 2012), *T. acetatoxydans sp. Re1* (**I**), *C. ultunense strain* Esp (**II**), *S. schinkii* strain Sp3 (**III**), and *C. ultunense* strain BS (Wei *et al.*, 2014).

2.7 Bioinformatics

The 'bio-' part of the term bioinformatics refers to biology, more particularly conceptualisation of biology in terms of macromolecule, while '-informatics' refers to the disciplines (applied maths, computer science and statistics) involved in managing and revealing information related to these molecules. The term was first coined by Paulien Hogeweg in 1979 for the study of informatics processes in biotic systems (Hogeweg & Hesper, 1978). However, Dr. Margaret Oakley Dayhoff (1925-1983) is considered the main founder of the field of bioinformatics. Her main interest was in using informatics to deduce the evolutionary history of biological kingdoms, phyla, and taxa from protein sequence alignments. Since her ground-breaking work, the field of bioinformatics has broadened to encompass many new fields of expertise and the production of biological data at a phenomenal rate makes the use of computers indispensable for biological research. Major research efforts nowadays include sequence alignment, gene finding, genome assembly, genome annotation, protein structure prediction, protein structure alignment, prediction of gene expression, comparative genomics, protein-protein interactions, genome-wide association studies, modelling of evolution, and proteomics studies (Visel *et al.*, 2009; Vastrik *et al.*, 2007; Snel *et al.*, 2005;

Ouzounis & Valencia, 2003; Stein, 2001a; Henikoff *et al.*, 1997; Jacob, 1977). The interrelationships between different fields of science and bioinformatics are illustrated in Figure 2.

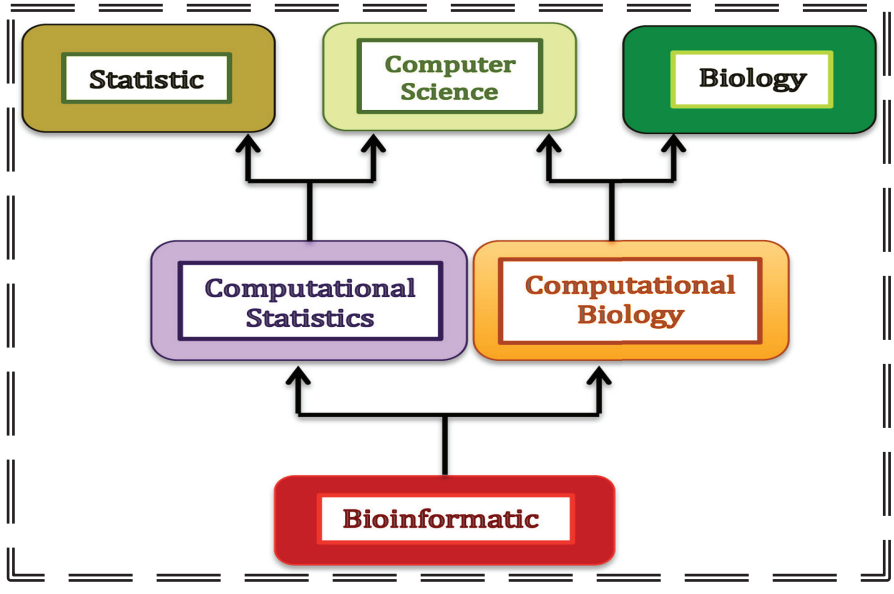

Figure 2. Interrelationships between different fields of science within the bioinformatics field.

2.8 Data mining

The revolutionary changes in biomedical research and biotechnology and the explosive growth of high-throughput data in biological sciences, together with advances in digital storage, computing and information and communication technologies in the 1990s, have begun to transform biology from a data-poor into a data-rich science. These advances are responsible for the gradual transformation of biology research from classic hypothesis approaches (in which a single answer is provided for a single question) to data-driven research (in which more than one possible answer is given at a time and we have to seek the hypothesis which best explains the answers).

In response to this exponential growth in biological data, the statistics, computer science and artificial intelligence began to be applied to the extraction of knowledge from these huge datasets, giving birth to the new discipline of data mining (DM). The DM field attained notable maturity in the late 1990s and its effectiveness in dealing with very large datasets has been proven in a range of areas of applications such as marketing, banking, weather forecasting, medical informatics, biology *etc*. The term DM can have a number

of different meanings within a wide range of contexts, but with reference to bioinformatics it can be defined as the set of techniques and working trends used for discovering meaningful and novel relationships and patterns in biological data that were previously unknown. Due to the diverse nature of biological data, several preparatory steps such as selection, cleaning/trimming, preprocessing, and transformation may be required prior to analysis to uncover information, which is the ultimate goal of DM. All these steps are illustrated in Figure 3.

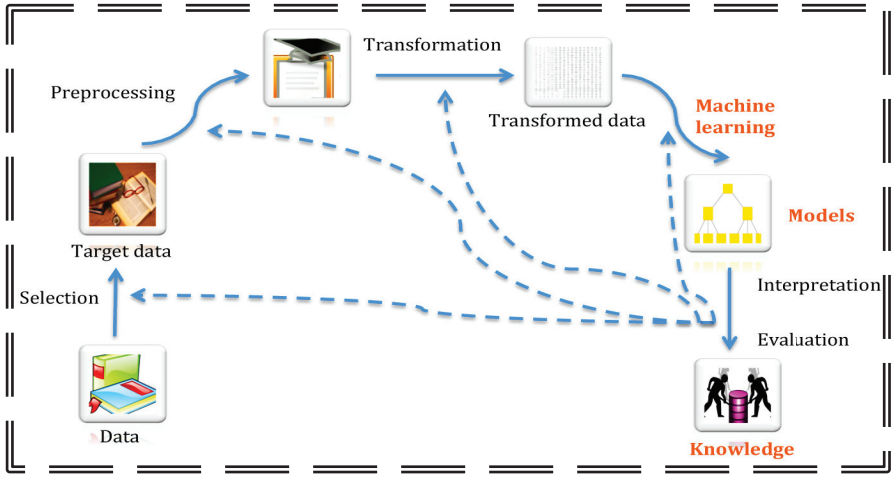

Figure 3. The general steps involved in performing a common data mining (DM) task.

2.9 Machine learning

Machine learning (ML) is one of the most representative tasks of many DM applications that can be used to solve problems by employing the sample data or past experiences. In biological jargon, the goal of ML is to understand the relationship between the observations collected and the experimental results obtained. There are a number of examples showing the wide use of ML in a range of applications, especially in bioinformatics (Zhaoli, 2012):

- Molecular biology research produces dynamic data and novel concepts as a result of different experiments, so it is essential to apply techniques such as ML that can be adapted efficiently to these fast-evolving environments.
- ML has the ability to handle effectively the huge amount of data produced through the novel high-throughput devices, in order to predict underlying relationships that are not immediately noticeable, even to experts.
- In most biological experiments, researcher can only specify input-output data pairs, but are not able to describe the relationships between different

interrelating attributes. ML is an approach with the ability to find internal structures in the existing data and produce approximate models and results.

- Due to the complex nature of biological organisms, the most difficult task is to extract short, understandable and relevant information without non-desirable results. This can be done with ML dynamic approaches by describing the underlying hidden characteristics.
- The constant improvements in ML allow it able to handle the abundance of missing and noisy data from many biological experiments.

The development of ML techniques means that researchers in the bioinformatics community are now capable of dealing with the two most frequently occurring situations in bioinformatics applications: i) Limited number of samples (the curse of dataset sparseness), and ii) several thousand of features characterising each sample (the curse of dimensionability) (Derman, 1961). The general scheme of some applications of ML techniques in bioinformatics is illustrated in Figure 4 (Larranaga *et al.*, 2006).

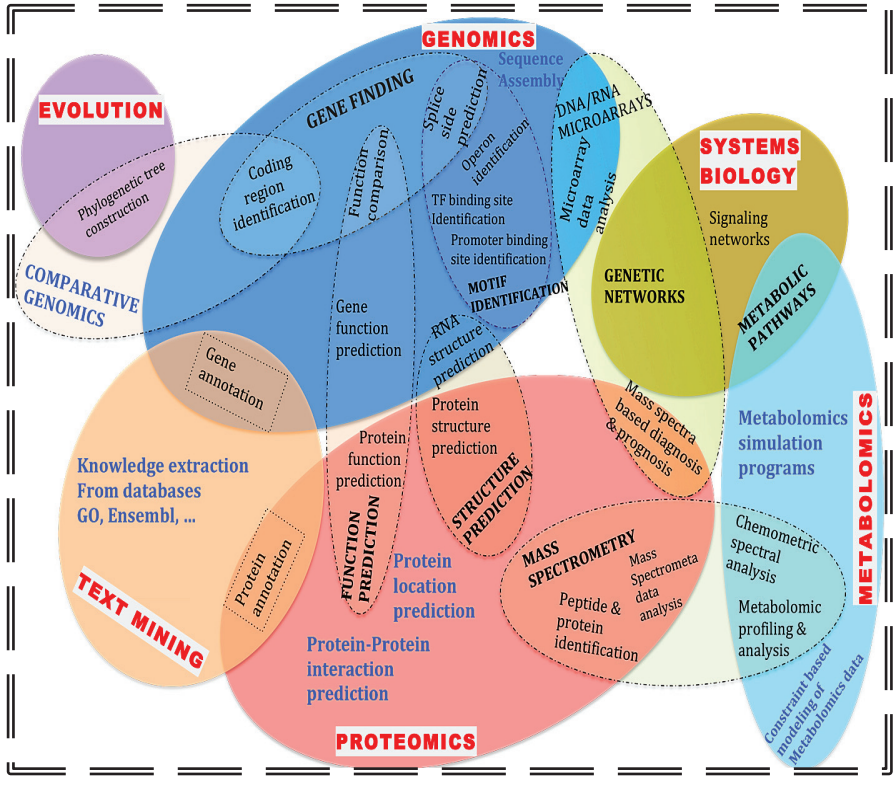

Figure 4. General scheme of applications of machine learning techniques in bioinformatics.

2.10 Next-generation sequencing

In biological sciences, the deciphering of DNA sequences has become instrumental for further studies. Sanger sequencing (Sanger *et al.*, 1977) (first-generation technology) allowed scientists to extract information from biological systems and led to a number of accomplishments, including the completion of the only finished-grade human genome sequence (International Human Genome Sequencing, 2004). A number of limitations of Sanger technology, such as scalability, throughput and speed, catalysed the development of a new technology known as next-generation sequencing (NGS) which ignited a revolution in genomic science with many ground breaking discoveries. In a broader prospective, the NGS technologies have enabled comparative genomics by resequencing related organisms/species/strains to understand how genetic differences affect phenotypic features. Some advances leading to the DNA revolution are listed in Table 1.

Table 1. *Description of the stepwise evolution of DNA-based technologies.*

Year	Discoverer(s) Inventor(s)	Discovery(ies) Invention(s)
1953	James Watson and Francis Crick	Deduce DNA's conformation from experimental clues and model building.
1958	Matthew Meselson and Franklin Stahl	Demonstrate how DNA replicates.
1961-1963	-	Researchers crack the genetic code linking gene and protein.
1964	Robert Holley	Complete the first nucleotide sequence of the gene encoding yeast alanine tRNA.
1972	Paul Berg and colleagues	Create first recombinant DNA molecule.
1977	Frederick Sanger, Allan Maxam, and Walter Gilbert	Pioneer for DNA sequencing(Sanger *et al.*, 1977).
1985	Kary Mullis	Invents Polymerase chain reaction (PCR)(Mullis *et al.*, 1986)
1986	Leroy Hood and Lloyd Smith	Automate DNA sequencing to sequence human genome.
1986-87	James Watson	United States DOE officially begins human genome project. US NIH takes over the genome project.
1990	-	Sequencing of human and model organism genomes begins. BLAST algorithm developed to align DNA sequences.
1994	-	Detailed genetic map of the human genome was published.
1995	J. Craig Venter and his team	The Institute of Genomic Research (TIGR) published first genome sequence of organism *Haemophilus influenzae(Fleischmann et al., 1995).*

1996	-	The international human genome project consortium established “Bermuda rules” for public data release.
1999	-	First human chromosome sequence was published.
2000	-	Fruit fly genome was sequenced, First assembly of the human genome completed by the UCSC group.
2001	-	Science and Nature published the first draft of human genome sequence.
2003	-	The human genome sequence completed, 2 years earlier than planned.
2004	-	Introduction of Massively parallel sequencing platforms giving rise to the “Next Generation Sequencing” (NGS).

The ability of NGS technologies to produce voluminous data at low cost proved attractive for many applications, including variant discovery by resequencing targeted regions of interest or whole genomes, *de novo* assembly of bacterial and eukaryotic genomes, transcriptomics of cells, tissues and organisms (RNA-seq) (Wang *et al.*, 2009), genome-wide profiling by using other sequence-based methods (ChIP-seq, DNase-seq) (Wold & Myers, 2008), and species classification and/or gene discovery by metagenomics studies (Petrosino *et al.*, 2009).

The ongoing journey of DNA sequencing that started in the late 1980s (capacity: 10 Kb per 4-hours run) has continued with improvements and step-wise upgrades *e.g.* in late 1990s with capillary sequencers (capacity: 50 Kb per 1-hour run), in 2005 with massive parallel pyrosequencing (capacity: 20 Mb per 5-hours run), in 2007 with sequencing by synthesis (capacity: 1 Gb per 5-days run), and in 2010 with single molecule sequencing (capacity: 100 Gb per 5-days run). At present (2014), the level of efficiency has reached at a point where a whole human genome can be sequenced in just 15 minutes.

Now these NGS technologies have the ability to generate short reads, typically 100-150 bp for Illumina (Loman *et al.*, 2012), 75 bp for SOLiD (Miller *et al.*, 2012), 400–600 bp for 454 (Loman *et al.*, 2012), and ~200 bp for Ion Torrent (Loman *et al.*, 2012), and long reads of up to 20 Kb for Pacific Biosciences, but also with a higher error rate (Nagarajan & Pop, 2013; Koren *et al.*, 2012; Niedringhaus *et al.*, 2011).

2.10.1 Shotgun library sequencing

DNA sequencing with the chain termination method was initially restricted to fairly short fragments (100-1000 bp). To overcome this problem, a new sequencing method was designed to analyse sequences larger than 1000 bp, up to entire chromosomes, and whole genomes, which is known as *shotgun sequencing*. The *shotgun sequencing* approach works very well for small (unicellular) to medium size genomes with less repetitive regions, where the

shotgun-sequencing library is prepared by randomly shearing the amplified genomes into many small fragments (50-8500 bp). The library of subfragments is sampled at random, and a number of sequence reads generated. Overlapping ends are used for construction of *contigs* and eventually the full genome is assembled based on a high coverage of reads. A schematic graphical illustration of a shotgun sequencing approach is presented in Figure 5.

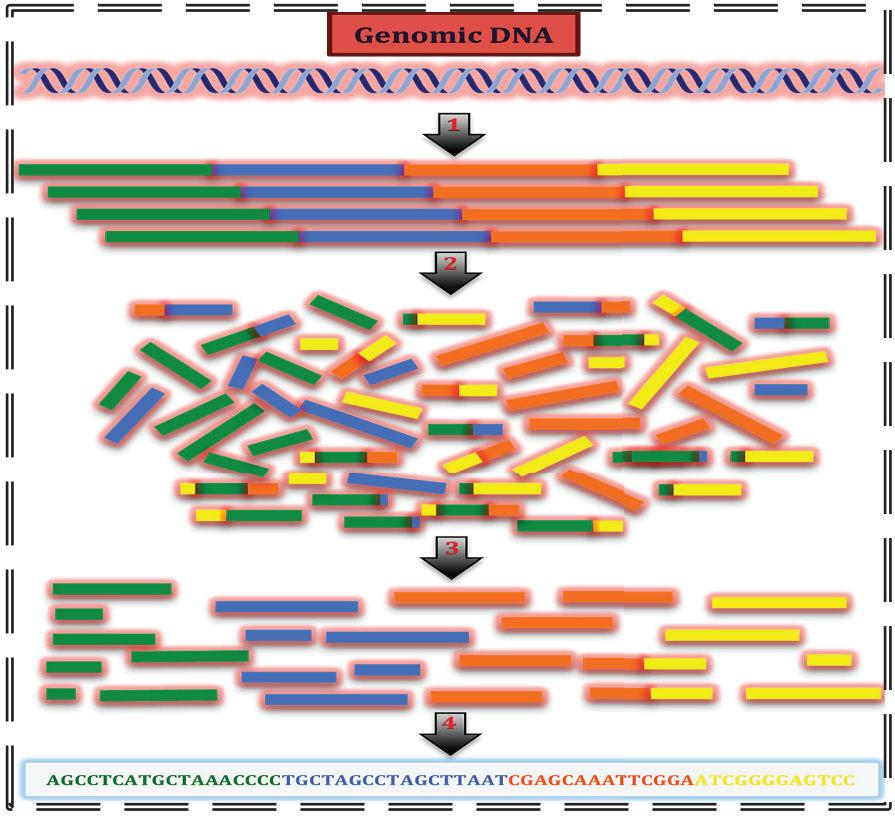

Figure 5. Schematic overview with different steps labelled with numbers showing the stepwise process for shotgun library preparation and assembly: 1) Genome cloning, 2) multiple genomes are sheared into variable sized unordered fragments, 3) overlapping fragments are aligned through computational assembly, and 4) consensus sequence is produced on the basis of overlapping segments. Higher coverage results in better quality of the consensus sequence.

On the other hand, large genomes add complexity by the high percentage of repetitive DNA (greater than 50% for the human genome) making the shotgun sequencing approach less reliable for large genomes (Venter, 2006). For these reasons, it was necessary to lower the computational load during the assembly process, which was done by the *hierarchical shotgun sequencing* approach. In

hierarchical approach the amplified genome is first sheared randomly into larger pieces (50-200 Kb) with different ends, which are subsequently used with enough coverage to find a scaffold that covers the entire genome. This scaffold is called a *tilling path*. Once a tilling path has been found, the large fragments are sheared at random into smaller fragments and can be sequenced using the shotgun sequencing method. A schematic graphical illustration of a hierarchical shotgun sequencing approach is presented in Figure 6.

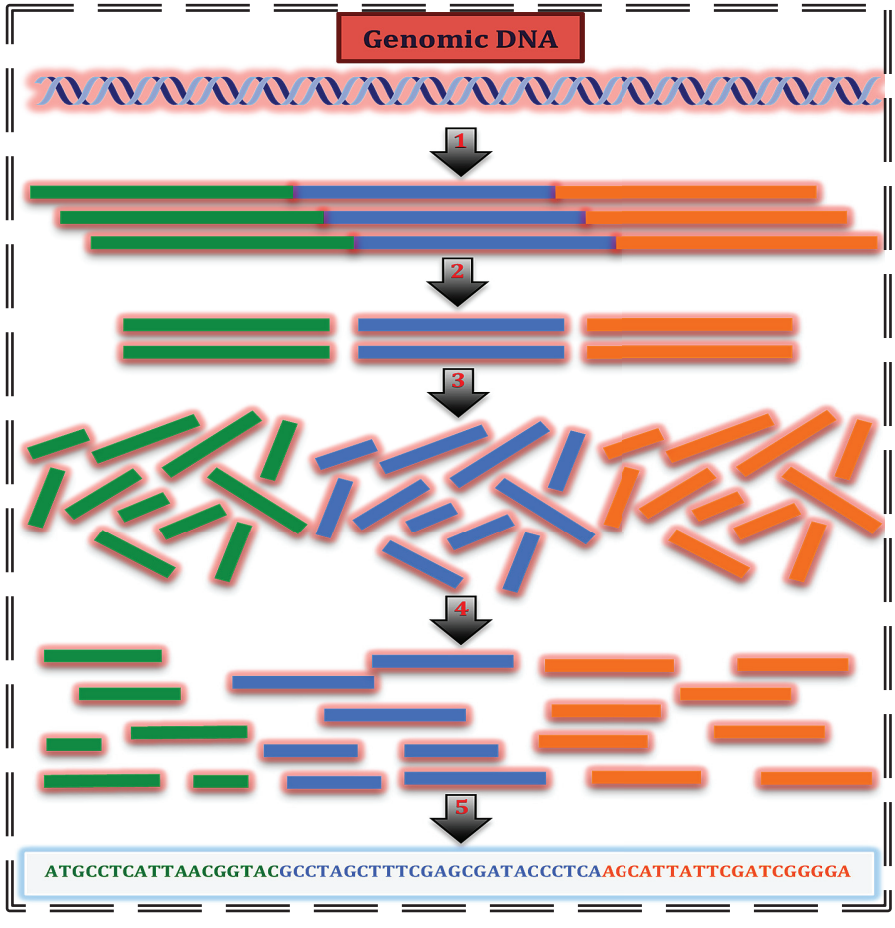

Figure 6. Schematic overview with different steps labelled with numbers showing the stepwise process for hierarchical shotgun library preparation and assembly: 1) Genome cloning, 2) genomes divided into large fragments of known order, 3) multiple large fragments are sheared into variable sized unordered segments, 4) overlapping segments are aligned through computational assembly, and 5) a consensus sequence is produced on the basis of segments overlap. Higher coverage can be the sign of better quality of the consensus sequence.

The SAOB genomes sequencing data used in this thesis (**I-III**) was first produced by shotgun sequencing with ion semiconductor sequencing based on H^+ detection, where reads with average sizes of 206, 191, 205 bp were produced by Ion Torrent™ technology using chip type 316-D in SciLifeLab Uppsala, Sweden.

2.10.2 Pair-end library sequencing

The NGS technologies are transforming the field of genomic science (Schuster, 2008) by providing the ability to read DNA fragments in a highly parallel manner to generate massive amounts of sequencing data with higher coverage and greater throughput. This has made it possible to sequence a whole genome in a relatively short time with unprecedented accuracy (Metzker, 2010). However, the short read length limits the use of this enormous sequencing power for many biological applications.

Paired-end sequencing is a strategy for improving DNA sequencing efficiency and enabling biological applications. This sequencing method produces two sequencing reads from each end of a DNA fragment. The size of the DNA fragment is usually known (referred to as *insert size*) and can range between short inserts of 100-200 bp or long inserts of 6,000-20,000 bp. Paired-end reads have three prominent properties: i) Both reads are from the same strand, ii) the approximate distance separating the two reads is known, and iii) the forward and reverse reads are identified as such, and as a pair. This makes the paired-end libraries useful for establishing *contig* order in a scaffold, mapping large rearrangements or insertion/deletions (Indels) to a reference genome or establishing a sequence within flanking repeats.

The sequencing data for plant growth-promoting rhizobacteria (PGPR), which is presented as novel information in this thesis essay were generated by paired-end sequencing library from Illumina Genome Analyzer with insert size ~200 bp at Uppsala Hospital, Sweden. A schematic illustration of paired-end library preparation and analysis is presented in Figure 7.

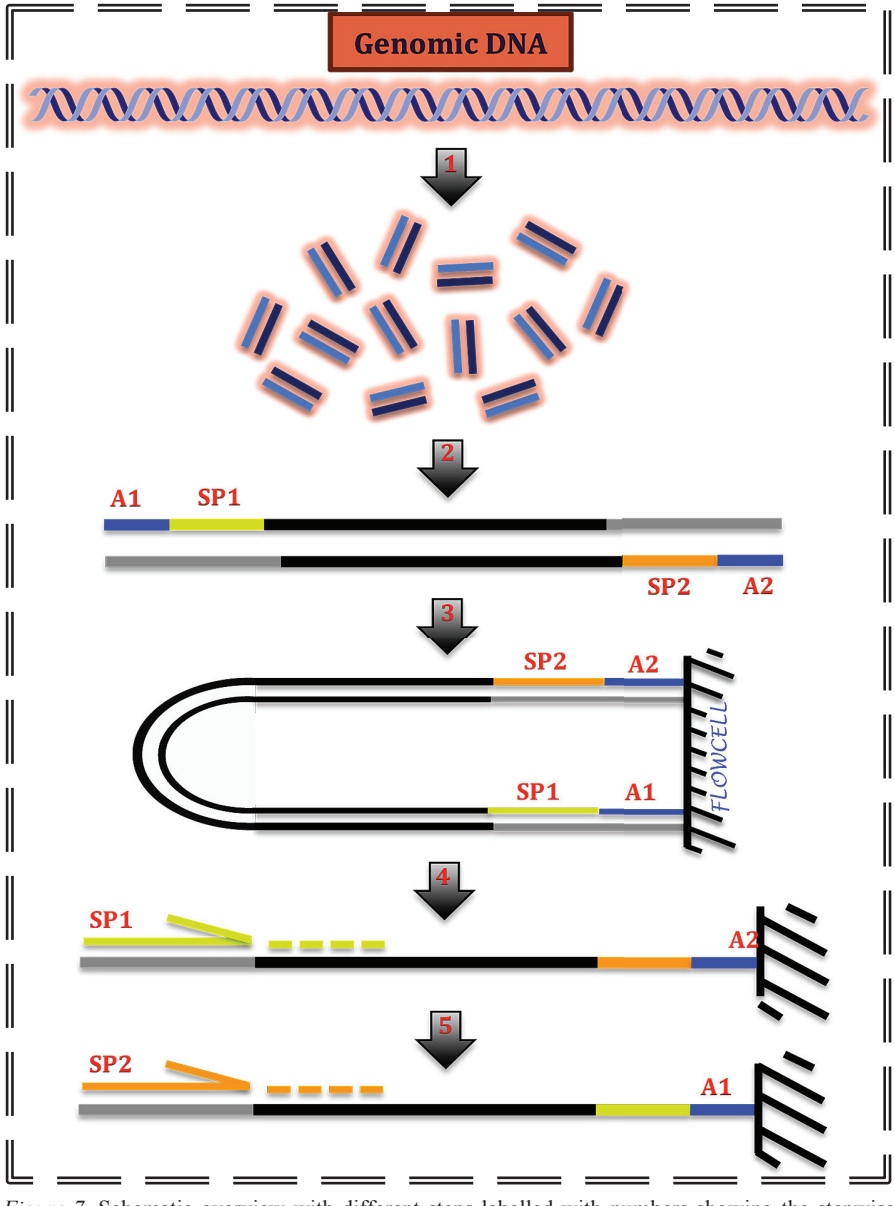

Figure 7. Schematic overview with different steps labelled with numbers showing the stepwise process for paired-end library preparation: 1) Genomic DNA, 2) fragments (2-10 Kb), 3) ligate adaptors, 4) generating clusters, 5) sequencing first end, and 6) regenerating clusters and sequencing paired end.

2.10.3 Mate-pair library sequencing

Mate-pair library sequencing differs from paired-end library sequencing only in the way of library preparation. Mate-pair library preparation generates long insert paired-end DNA libraries with size ranging from 2-10 Kb fragments, which are sequenced from both ends for providing information on how nucleotides far apart are linked together. Mate-pair library sequencing is more suitable for studying different applications such as *de novo* sequencing, genome finishing, structural variant detection and identification of complex genomic rearrangements. A combination of *mate-pair library sequencing* and short insert *paired-end* reads can be used as a powerful tool to generate read length for maximal sequencing coverage across whole genomes.

The SAOB genomes sequencing data generated by shotgun sequencing from Ion Torrent^{TM} technology were also sequenced in this thesis using mate-pair library sequencing from Illumina HiSeq-2500 with library size 3 Kb in SciLifeLab Stockholm, Sweden, to finish the draft SAOB genomes (**II, III**) (data not presented). A schematic graphical illustration of mate-pair library preparation is presented in Figure 8.

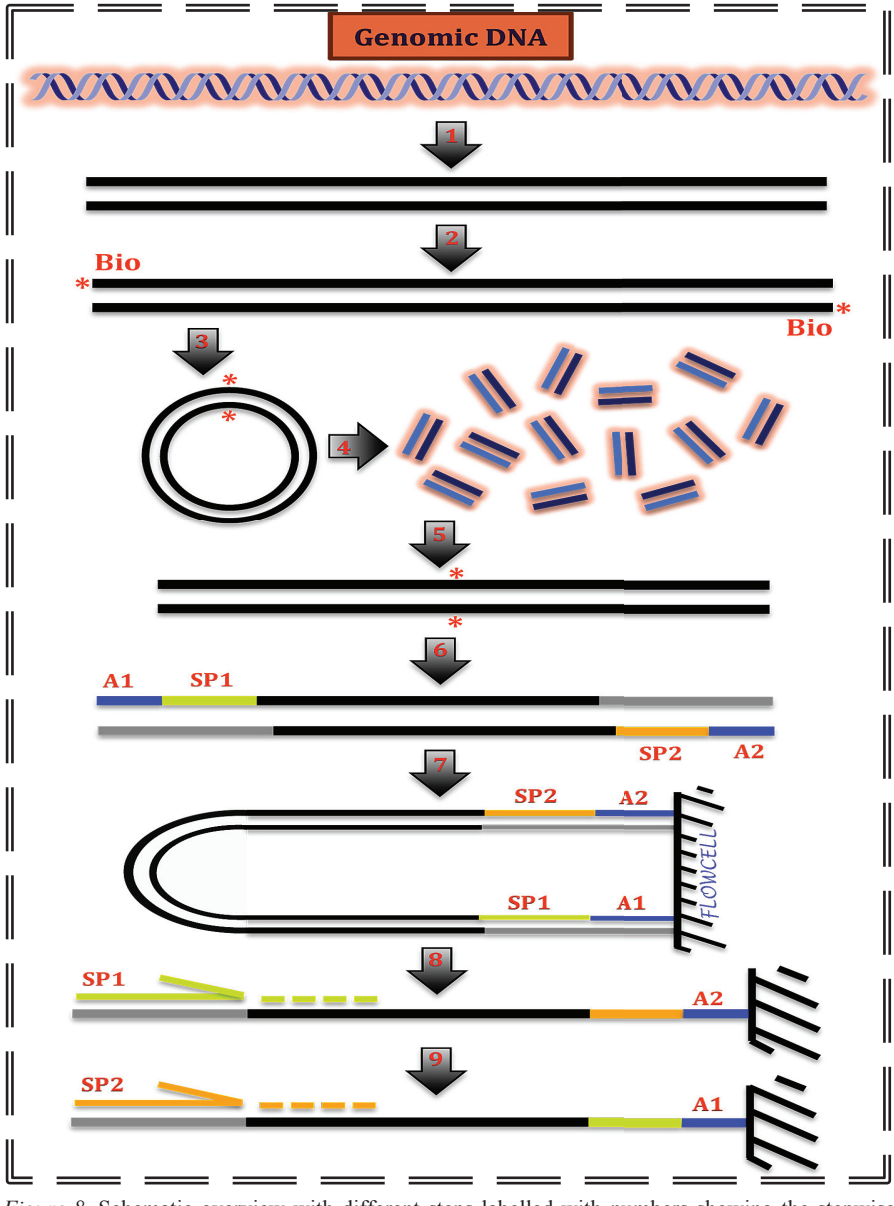

Figure 8. Schematic overview with different steps labelled with numbers showing the stepwise process for mate-pair library preparation: 1) Genomic DNA, 2) fragments (2-10 Kb), 3) biotinylate ends, 4) self-circularisation, 5) small fragments (400-600 bp), 6) enriched biotinylated fragments, 7) ligate adaptors, 8) generating clusters, 9) sequencing first end, and 10) regenerating clusters and sequencing paired end.

2.11 The *-omics* age

The completion of the genome sequence for *C. elegans* (Consortium, 1998) provided inspiration for the Human genome project, which had started in 1990 (Watson & Cookdeegan, 1991) and was successfully completed with the announcement of a first human draft genome sequence (Consortium, 2001), followed by a finished sequence (Collins *et al.*, 2004). Over the past decade, overlapping developments in molecular research technologies along with the rapid developments in information technology have collectively generated a huge flood of data that can be considered a revolutionary advancement in this field. However, this might not have been possible without a number of small advances (Table 2). This whole very fast journey of molecular biology to today's era of '*omics'* including genomics, transcriptomics, proteomics *etc.,* could not have been conceptualised without the support of bioinformatics.

Table 2. *Bioinformatics contributions that have played a key role for the advancement of life sciences.*

Year	Discoverer(s) Developer(s) Inventor(s)	Discovery(ies) Technology(ies) Invention(s)
1965	Margaret Dayhoff	Atlas of Protein Sequences
1966	National Library of Medicine	Development of Medline
1968	U.S. Advanced Research Projects Agency	Development of Internet
1972	R. Tomlinson	Development of Email
1980	P.J. Stoehr and G.N. Cameron	The EMBL Nucleotide Sequence Database, EMBL, Germany (Stoesser *et al.*, 1997)
1982	W.Goad	Development of GenBank
1986	Amos Bairoch	Swiss-prot launched
1988	National Library of Medicine	Founding of the National Center for Biotechnology Information (NCBI) and creation of public databases systems for their use (Altschul *et al.*, 1990)
1988	Chris Sander at EMBL	EMBnet was created to be able to distribute the EMBL database
1990	S.F. Altschul and D.J. Lipman	Development of the BLAST algorithm …
1991	T. Berners-Lee and R. Cailliau	Development of the World Wide Web (WWW), CERN
1995	Michael Ashburner	EMBL-EBI, keep ENA, ENSEMBL, UniProt
1997	National Center for Biotechnology Information	Development of PubMed
1998	L. Page, S. Brin	Development of Google
1999	Ewan Birney	ENSEMBL Genome Browser
2000	Jim Kent	UCSC Genome Browser

3 Aims of this thesis

Anaerobic digestion technology is an interesting and promising method for the production of biofuel. Through this processes various organic waste steams can be degraded by microorganisms and converted to biogas, a renewable energy source. The main aim of the studies described in this thesis was to enhance knowledge of the SAOB key organisms in anaerobic digestion, using bioinformatics methods. The SAOB studied are all acetogens and are active in the biogas production step with the ability to reverse their metabolic flow from acetate production to acetate oxidation to produce hydrogen/formate, which is later used by methanogens for the production of methane. Acetate oxidation is a thermodynamically very difficult reaction that only proceeds at low hydrogen levels, which is achieved by syntrophic growth with partner methanogen. At present little is known about these organisms and how they manage to be active with energy levels scarcely enough for growth. Thus to learn more about these microorganisms, the specific aims of the work presented in Papers I-V were to:

- Sequence the genomes of three syntrophic acetate-oxidising bacteria (SAOB), in order to provide a platform for further investigation of the syntrophic lifestyle
- Perform structural and functional annotations for the sequenced SAOB genomes in order to allow investigations of their genomic behaviour
- Start analysing metabolic pathways, such as substrate utilisation, intermediate metabolism, energy conservation and acetogenesis, in order to improve understanding of these SAOB
- Understand the processes of acetate reduction versus oxidation through the Wood-Ljungdahl (W-L) pathway.

4 Methods

Different bioinformatics methods were used in papers I-V. In this section some of these methods are described in order to provide an understanding of bioinformatics resources, genome assembly algorithms, annotation processes, and metabolic pathway analysis.

"If it were easy, it would have been done already"
(Solomon short)

4.1 DNA isolation

Table 3. *Statistics on temperature and growth time to isolate DNA of SAOB genomes.*

Organism	Isolation source	Temp	Time	Kit used
T. acetatoxydans strain Re1	Continuously stirred laboratory-scale reactor designated Recirc (Westerholm *et al.*, 2011b).	37 °C	1 weeks	Blood & Tissue Kit from Qiagen
C. ultunense strain Esp	Sludge of an upflow anaerobic filter treating wastewater from a fishmeal-producing factory (Westerholm *et al.*, 2010).	37 °C	4 weeks	Blood & Tissue Kit from Qiagen
S. schinkii strain Sp3	Sludge of an upflow anaerobic filter treating wastewater from a fishmeal-producing factory (Westerholm *et al.*, 2010).	37 °C	4 weeks	Blood & Tissue Kit from Qiagen

4.2 Genome assembly

Next generation DNA sequencing platforms are characterised by high parallel operation, high throughput and low cost, but unfortunately produce short length reads, with the exception of Pacific Biosciences machines (Korlach *et al.*, 2010; Eid *et al.*, 2009). All these sequencing machines detect the target DNA molecule and produce an output in the form of reads consisting of single letter base calls, plus a numerical quality value for each base call (Ewing & Green, 1998). These NGS platforms can also have different error profiles including enrichment of base call error towards the 3′ ends of reads, compositional bias against high GC sequence and inaccurate determination of sequence repeats (Harismendy *et al.*, 2009; Dohm *et al.*, 2008; Huse *et al.*, 2007). With these different error profiles, the genome assembly procedure becomes more difficult and complex, especially when dealing with short and high-throughput reads (Miller *et al.*, 2010).

The genome assembly process can be linked to a jigsaw puzzle, where different tasks or challenges have to be complete in order to solve the puzzle. Each read (piece) has to be placed in the correct position in the puzzle. This task is directly linked to the ‘quality’ of the solution of the puzzle, *i.e.* the only available information relating to the correct position of a read (piece) comes from its neighbouring reads. The larger the number of reads (pieces in the puzzle), the greater the complexity in determining the correct position. Moreover, there can be ambiguity due to the positioning of similar reads which share similar suitable locations in the puzzle. However, some reads may have unique features and these serve as unique indicators to aid further read

positioning (Pevzner *et al.*, 2001). A schematic graphical illustration of a typical genome assembly process is presented in Figure 9.

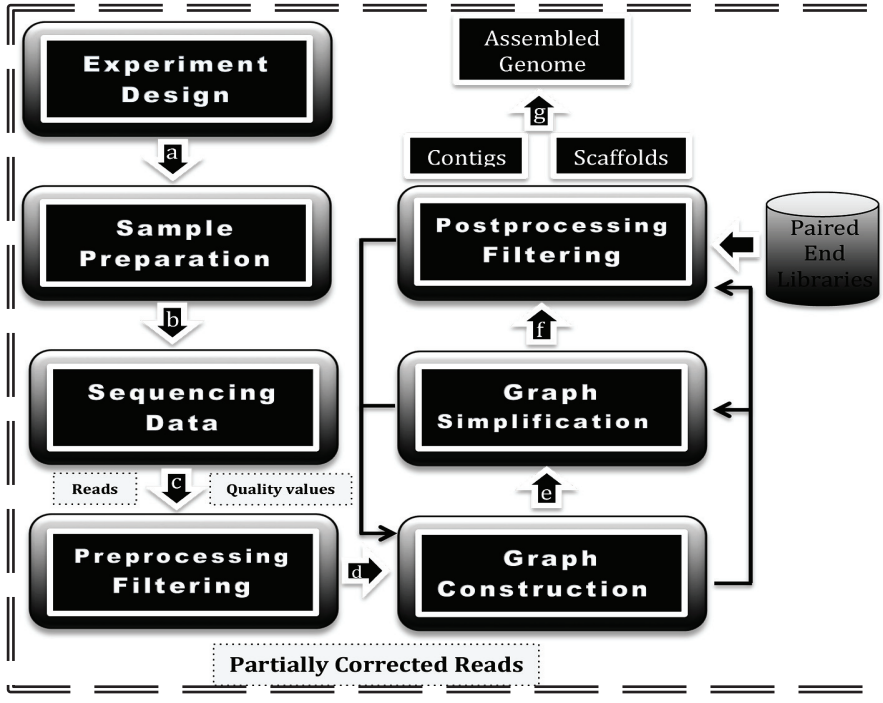

Figure 9. Schematic view of the different stages in next generation genome assembly. a) Genome assembly starts with experimental design where experimental, technical and computational issues have to be considered, b) DNA from a single cell, a clonal population or a heterogeneous collection of cells is extracted for sequencing, c) prepared DNA samples are sequenced by selected technology, with maximum possible read length and appropriate coverage depth, d) erroneous short reads are detected and corrected by a preprocessing filtering step, e) a graph model is constructed for organising short read sequences into a compact form and for creating longer sequences during assembly, f) the graph is simplified by reducing the number of graph nodes and edges and removing erroneous values and g) contigs are built in a postprocessing filtering step. The contigs are then extended into scaffolds and attempts are made to detect misassembled contigs. Modified from Sara (2013).

4.2.1 Mapping assembly

During *mapping assembly,* also known as *reference-guided assembly,* shotgun reads are assembled against an existing backbone sequence, resulting in a target-assembled genome that is similar, but not necessarily identical, to the backbone sequence. The idea behind reference-guided assembly is to use a reference genome from the same organism or a closely related species as a map to guide the assembly process. This approach is useful for resequencing applications (Pop *et al.*, 2004). However, even if the target and reference

genomes are closely related, the species-specific attributes must be addressed separately to obtain the maximum information on the target genome. These different forms of polymorphisms between the reference and the target genomes can be:

- *DNA divergence*: Some low similarity regions can be present in the target genome where the reference and the target genomes have diverged significantly from each other during evolution (Figure 10A)
- *Genomic rearrangements*: A number of DNA segments have a different order or orientation in the target genome compared with the reference genome (Figure 10B)
- *Insertion in the target genome*: Some DNA segments are part of the target genome, but are absent in the reference genome (Figure 10C)
- *Deletion from the target genome*: Some DNA segments are only present in the reference genome (Figure 10D).

4.2.2 *De novo* assembly

De novo assembly is a process of assembling the short reads to reconstruct the full-length sequence of a target genome that may be a novel organism for which no map or guidance is available. Compared with mapping assembly, assembly by this approach commonly poses more computational challenges (Martin & Wang, 2011). The generation of high-throughput short reads from NGS technologies makes this process a highly memory-intensive computational task. To deal with this problem, most assemblers format their input reads data using graph data structures. Even though the principle is the same, different assemblers may vary concerning initial graph construction, configuration, traversing, and simplification processes (Zhang *et al.*, 2011).

In this thesis, the *de novo* assembly approach was used for constructing the genomes of two SAOB (**II, III**) and their working draft genome sequences were produced. The genomes were sequenced using the Ion Torrent PGM™ systems, resulting in a total of 2,631,078 bp (**II**) and 2,985,963 bp (**III**) single end reads with a mean length of 206 bp.

The sequencing data produced by sequencing machines are often provided as raw reads and are not always correct and precise in their entire length, which may introduce artefacts into the genome assembly. For example, for an *M. tuberculosis* genome, the assembly N50 substantially increased, from 1 Kb to 30 Kb, after filtering of the reads. Moreover, due to the massively parallel nature of sequencing technologies, it is not worthwhile inputting too many reads to the assembler, as this can dramatically increase the memory requirements and can also result in a lower quality assembly due to the excess of error reads. To avoid this, the preassembly quality was checked using the

FastQC tool. There are two different approaches for dealing with these identified low quality nucleotides in the raw reads data. The first approach is to correct the reads by superimposing them on each other, which usually works at the level of k-mers and is implemented by several tools, such as Quake (Kelley *et al.*, 2010) and ALLPATHS-LG (Gnerre *et al.*, 2011). The second approach tries to surgically eliminate only low quality regions and has been used in several studies, *e.g* prior to genome assembly, transcriptome assembly, metagenome reconstruction and RNA-Seq.

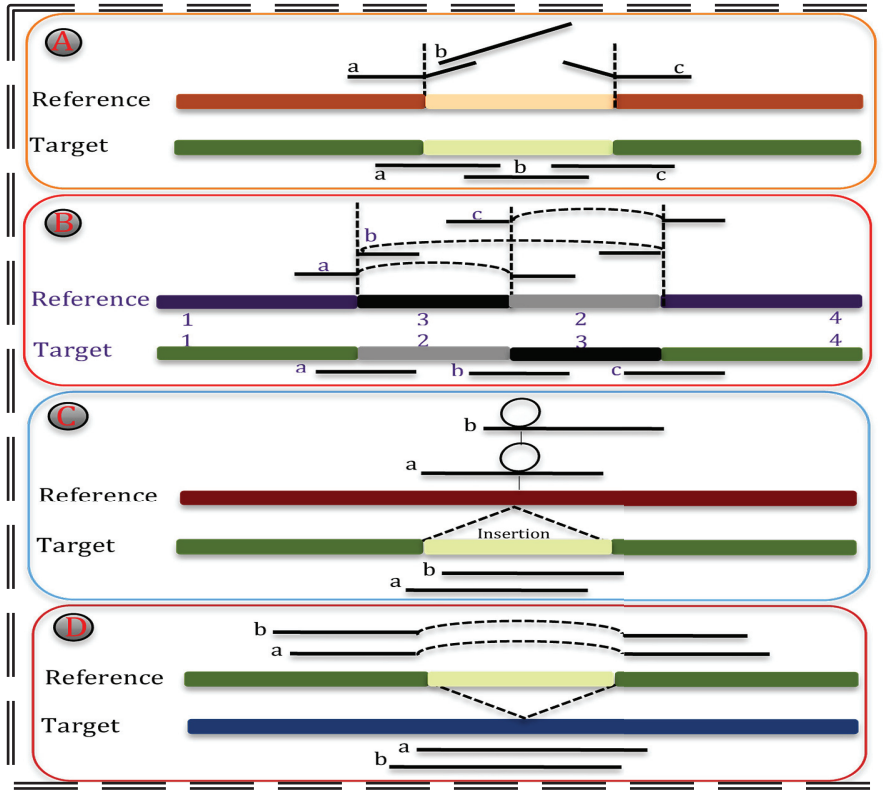

Figure 10. Illustration of different types of polymorphisms between the reference and the target genomes. A) Region of divergence between the reference and the target genomes, marked with light colours. The angled line shows the parts of the reads not matching the reference genome, B) the genomic rearrangement between the reference and target genomes. Note that Parts 2 and 3 from the target genome appear in a different order in the reference genome, while a, b, and c reads match the reference genome at disjoint locations. The dotted lines connect sections of a read that are adjacent in the target genome, C) insertion in the target genome, with circles indicating the portions of two reads that do not align to the reference genome and d) insertion in the reference genome. The upper alignment of reads in the reference indicates the presence of that insertion and dotted lines indicate the stretch of the reads needed to align to the reference. Modified from Mihai (2004).

In this thesis, prior to the genome assembly process, a preprocessing reads trimming procedure was performed using the *Sickle* tool (Joshi & Fass, 2011), which is a windowed adaptive trimming tool for FASTQ files. It uses sliding windows (size 10% of the whole read length) along with quality and length thresholds to determine when quality is sufficiently low to trim the 3′-end of reads and also determines when the quality is sufficiently high to trim the 5′-end of reads. It also discards reads based upon the length threshold.

All filtered reads were then input to *MIRA* 4.0 (**M**imicking **I**ntelligent **R**ead **A**ssembly) and *Newbler* 2.8 assemblers separately to perform the *de novo* assembly. Misassemblies were corrected by visualising the assemblies and read mappings with a graphical visualisation tool called *Tablet* (Milne *et al.*, 2010). A whole genome comparison tool called *Mauve (Darling et al., 2004)* was used to fill some gaps between *contigs* by aligning the assemblies produced by these two assemblers.

For another SAOB genome presented in paper (**I**), a PGPR genome included in this thesis as additional work, two PGPR genomes (unpublished), and a methanogen *Achaea* genome (syntrophic partner of SAOB)(unpublished), a comparative assembly approach was used. In this *de novo* assembly was combined with reference-guided assembly to fill the gaps among the *de novo* assembled scaffolds and finish all these genomes in a single chromosome.

This comparative assembly approach comprised a multi-steps process, where the sorted *de novo* assembled scaffolds were first concatenated by aligning them against the referenced-guided consensus sequence, which resulted in large scaffolds. This was done in two steps: i) two overlapping adjacent scaffolds were merged into one larger scaffold, and ii) if a subsequence was inserted in the reference-guided consensus sequence between two adjacent scaffolds, such a subsequence was inserted and scaffolds were concatenated into one scaffold, which filled the gap between these scaffolds. The species-specific features in these genomes (insertions, deletions) were also identified by aligning the reference-guided consensus sequence against the *de novo* assembled scaffolds (Figure 11). The *polymerase chain reaction* (PCR) amplification method was used to confirm all the insertion and deletion regions in the target genome and to fill the remaining gaps between scaffolds and end up with a compete genome sequence in a single chromosome.

For this comparative assembly approach the whole genome shotgun sequences were obtained using the Ion Torrent PGM™ systems with chip type 316-D, which produced single chip data with a total number of 3,333,516 reads with an average length of 191 bp (**I**). All steps such as preassembly quality checking, read filtering, *de novo and* mapping assembly, correction of

misassemblies and genome alignment were performed as described in the previous section.

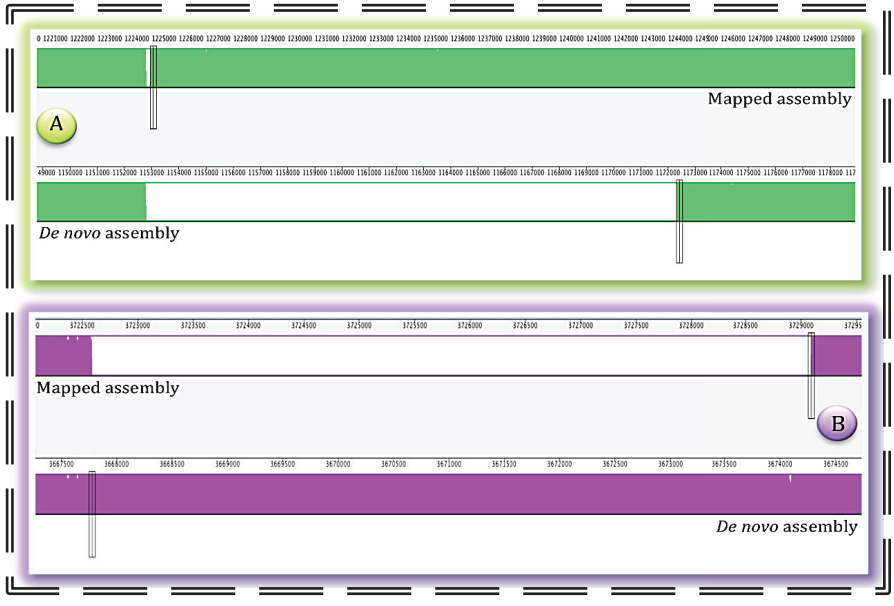

Figure 11. Illustration of identified insertion and deletion regions by aligning the consensus sequence of reference-guided assembly against scaffolds of *de novo* assembly. A) putative insertion region identified in a consensus sequence of the reference-guided assembly and B) putative deletion region identified in a consensus sequence of the reference-guided assembly.

In this thesis, the *Velvet* assembler for short reads was also used to reconstruct a PGPR genome using Illumina paired-end reads with a library size of 180 bp. It was observed that *Velvet* assembly of Illumina data produce similar statistics to *Newbler* assembly of 454 and Ion Torrent data (the sff files produced by the Ion Torrent platform are compatible with *Newbler* assembler). However, getting the best results from *Velvet* is slightly more tricky than using *Newbler*. Whether using *Velvet* or *Newbler,* the first important step for *de novo* assembly with short reads can be an aggressive pre-filtering of the raw reads data. *Velvet* does not consider quality scores, perhaps indicating the utility of filtering steps. In summary, to obtain good results from *Velvet,* several steps can play a vital role and should be included, such as count per base quality scores and trimming the reads, elimination of singletons, if possible eliminating reads with Ns and discarding reads with low average quality.

4.3 Sequence alignment

Due to the high availability of molecular sequences in DNA and protein sequence databases, *sequence alignment* has become one of the most useful and powerful tool. Determining the arrangement of DNA, RNA, or protein sequences in order to identify conserve and similar regions can provide insights into the functional, structural or evolutionary role of the sequences. The basic idea behind sequence alignment is to find *conservation* between sequences, as it is believed that evolutionarily conserved sequences can be an indication of active functionality (Keane *et al.*, 1998).

A number of computational approaches have been researched and as a result of those studies, methods including *dynamic programming*, which involves an exhaustive search of solutions, *heuristic algorithms* and *probabilistic methods* have been designed for large-scale database searches. The last step of a sequence alignment procedure is a scoring system to evaluate all possible alignments and finalise an optimal alignment based on the highest score. In the scoring system there is a scoring function, usually in the form of *substitution matrices* between residues. For example, the two most frequently used scoring matrices for protein sequences are the *Point Accepted Mutation* (PAM) (Dayhoff, 1978) and the BLOcks of amino acid *Substitution Matrix* (BLOSUM) (Henikoff & Henikoff, 1992). In addition two parameters, the gap opening penalty and the gap extension penalty, are used to discourage gaps (insertions and deletions) in the alignment.

Synteny analysis was performed for the genomes of the three newly sequenced SAOB isolates in this thesis (**III**), because it is important in genome comparison to reveal genomic evolution of related species (Tamames, 2001). The synteny regions identified consist of genomic fragments from different species, which originated from a certain common ancestor. The genes located in these syntenic fragments called *syntenic genes,* are orthologs and often share similar functions. A built-in tool in the MaGe annotation pipeline was used to perform synteny analysis. All proteins of bacterial genomes present in the NCBI RefSeq database were compared to identify the *syntenic genes*. The putative orthologous relations between two genomes were defined as; gene couples satisfying the *bi-directional best hit* (BBH) criterion or a *blastP* alignment threshold, with a minimum of 35% sequence identity on 80% of the length of the smallest protein. These criteria were subsequently used to search for conserved gene clusters such as synteny groups (syntons) by allowing all possible kinds of chromosomal rearrangements such as inversion and insertion/deletion. A gap parameter representing the maximum number of consecutive genes which were not involved in a synteny group was set to five genes. For graphical visualisation of these computed synteny groups, a tool

called *lineplot* was used. It draws a global comparison, based on synteny results between two bacterial genomes, and produces a diagram to give an overview of the conservation of synteny groups between these two genomes.

Local alignment

Local alignment identifies regions of similarity within long sequences that are usually widely divergent as a whole sequence. It can be a powerful tool for some specific tasks, such as comparing protein sequences that share a common *motif* (conserved pattern) or domain (independent folded unit) but differ elsewhere; comparing DNA sequences that share a common *motif* but differ elsewhere; and comparing protein sequences against genomic DNA sequences (long stretches of uncharacterised sequence). It is more sensitive when comparing highly divergent sequences.

Global alignment

Global alignment finds best matches of sequences in their entirety and tries to optimise alignment along the entire length of both sequences. So-called *semi-global alignment* finds best matches of both sequences without penalising gaps on the ends of the alignment.

Pairwise alignment

Pairwise alignment is used to find the best matching for a subsequence (*local*) or for an entire sequence (*global*), between two sequences at a time. The primary methods to perform pairwise alignments include the *dot-matrix method,* which is most likely the oldest visual representation when comparing two sequences (Maizel & Lenk, 1981). *Dynamic programming* algorithms were invented for *global alignment* by Needleman and Wunsch (1970) and for *local alignment* by Smith and Waterman (1981). In these methods, a substitution matrix is used for protein alignment and it assigns scores to amino acid matches or mismatches and a gap penalty for matching an amino acid in one sequence to a gap in the other. *Word methods* are heuristic methods that are used for searching against a large number of sequences to find local similarities for a specific sequence. This method implies that an optimal alignment is not guaranteed to be found, but on the other hand it is computationally more efficient than dynamic programming for alignments.

For the genomes of the three-sequenced SAOB identified here (**I-III**), the genomic regions containing tandem duplications of protein coding genes were identified. The genes were considered to be tandem duplicate genes if they showed more than 35% identity on 80% of the length of the smallest protein and were separated by a maximum of 5 consecutive genes.

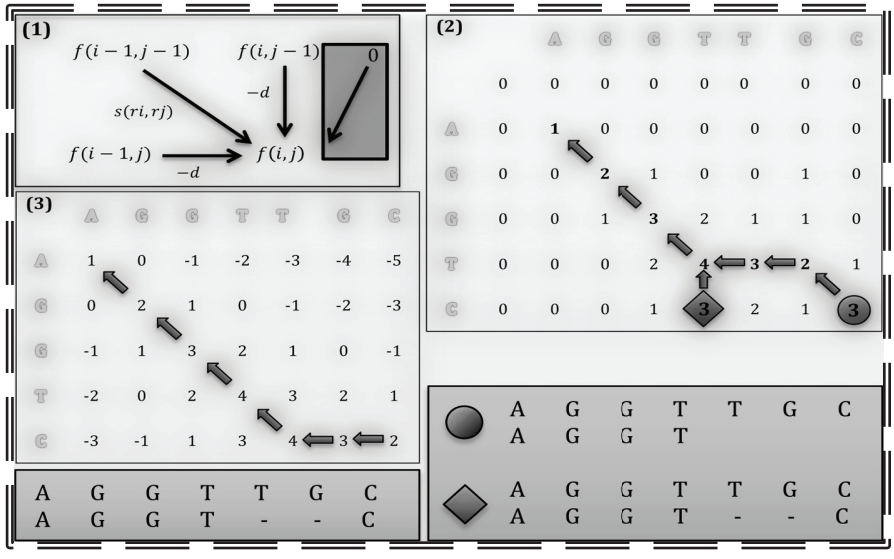

Figure 12. 1) The Needleman and Wunsch (1970) forward algorithm to recursively compute the entries of the alignment matrix. The grey rectangle indicates the additional parcel of the Smith and Waterman algorithm (1981). 2) Example of an alignment matrix and the back-tracking algorithm used to find the best–scoring *global alignment*. In the Needleman-Wunsch algorithm, the back-tracking reconstruction of the alignment goes from the low right corner to the top left of the matrix. 3) Example of an alignment matrix and the back-tracking algorithm used to find the best-scoring *local alignment*. In the Smith-Waterman algorithm, the back-tracking starts at the highest scoring value in the lowest row. The back-tracking reconstruction stops when a 0 value is found. Adopted from Balding *et al*. (2007).

Alignment can be performed for protein sequences and nucleotide sequences, but protein sequence alignment appears more advantageous for a number of reasons: i) redundancy of amino acid codons is not considered in nucleotide alignment, ii) due to large alphabet of characters for protein sequences it is easy to obtain statistically significant alignment, iii) nucleotide sequence databases contain much more sequences apart from protein coding sequences (only ~1.5% of the human genome codes for proteins), so it is more efficient to search against proteins rather than scanning the entire genome, and iv) nucleotide sequence alignment does not consider the more similar structures of some amino acids to others and the similar functional role they have in the protein. Since not all important functional regions in the nucleotide sequence code for proteins and there are some particular situations when nucleotide alignment is required, *i.e.* to compare regions in the nucleotide sequence that encode for functional RNA molecules rather than proteins or regions that serve as binding sites for transcription factors, there is no protein sequence to work with. Nevertheless, it is better to narrow down the possible

homologs with a protein alignment, and after that nucleotide alignment can be performed to compare the remaining sequences.

To allow the maximum annotation of unknown genes identified in the newly sequenced genomes of the three SAOB isolates (**I-III**), we also opted to perform the alignment for protein sequences. As an example pairwise alignment between two proteins *fhs1* and *fhs2* (formyltetrahydrofolate synthetase a key enzyme of the W-L pathway identified in the sequenced genome of *T. acetatoxydans*), is presented in Figure 13.

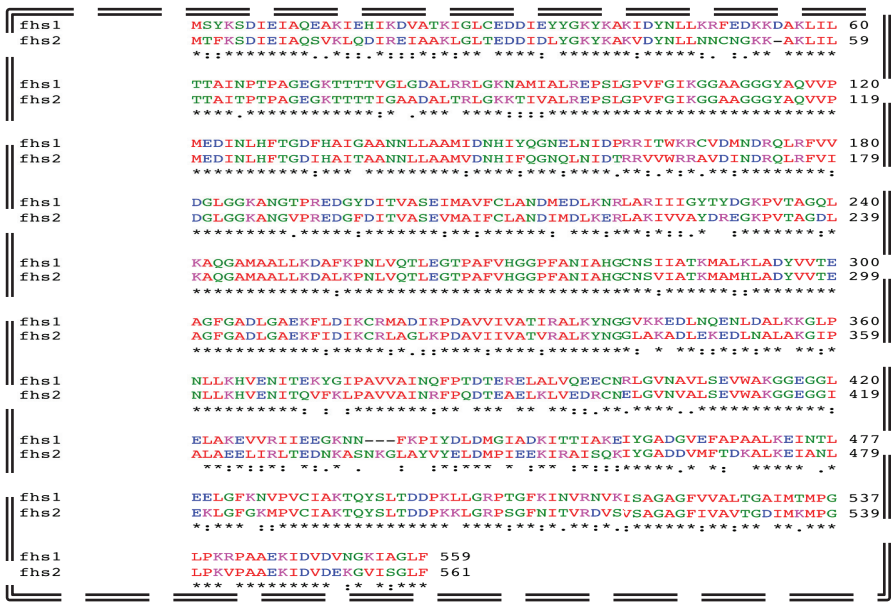

Figure 13. A sequence alignment produced by *ClustalW*, of two *T. acetatoxydans* proteins (formyltetrahydrofolate synthetase). Amino acids are colour coded by their physicochemical properties. *Red*: small, hydrophobic, aromatic, not Y. *Blue*: acidic. *Magenta*: basic, not H. Green: hydroxyl, amine, amide, and basic. *Grey*: others. Alignment keys are "*": identical residues, ":": conserved substitutions (same colour group), ".": semi-conserved substitution (similar shapes), "-": Inserted gap in a sequence.

Multiple sequence alignment

Multiple sequence alignment is an extended version of pairwise alignment, and is used to identify conserved sequence regions among a group of sequences. The most popular multiple sequence alignment tool *ClustalW* (Thompson *et al.*, 1994) is based on the *progressive algorithm* by Feng and Doolittle (1987). The algorithm first finds the identities by using the most similar sequences to minimise the alignment errors and gaps are discouraged by penalties to reduce their presence and benefit established identities. In addition, initial sequences

are selected on the basis of weights, which are calculated from the branch length of the guiding tree. This selection is performed in *ClustalW* by using the neighbour-joining method (Saitou & Nei, 1987).

Iterative methods improve the dependence which the progressive algorithm shows in the pairwise alignments calculated for the guided tree matrix. An iterative approach is to optimise the initial global alignment by realigning randomly selected sequence subsets. This sequence subsets selection procedure is reviewed by Hirosawa *et al.* (1995). In this method the newly aligned sequence subsets are realigned back into the multiple sequence alignment and used in the next iteration.

On the basis of 16S rRNA gene sequence, the most closely related species to the newly sequenced genome of SAOB (**III**) was identified from a group of acetogenic bacteria, which produce acetate trough the W-L pathway. The multiple sequence alignment tool *MUSCLE* (Edgar, 2004b; Edgar, 2004a) with a *progressive* and *iterative* approach was used due to its reasonable speed performance and iterative search for a better local optimum. A benchmark published for different protein alignment algorithms (Nuin *et al.*, 2006) is summarised in Table 4.

Table 4. *Evaluation of different multiple sequence alignment algorithms on the basis of average accuracy values obtained from Simprot's simulated sequences and all BAliBASE's references.The execution times are normalized to Mafft FFT-NS-2, the fastest of the algorithms. Updated from Nuin et al. (2006).*

Program	Publishing Year	Simprot	BAliBASE	Time (s)
ClustalW	(Thompson *et al.*, 1994)	0.789	0.702	22
Dialign 2.2	(Morgenstern, 1999)	0.755	0.667	53
T-Coffee	(Notredame *et al.*, 2000)	0.836	0.735	1274
POA	(Lee *et al.*, 2002)	0.752	0.608	9
Mafft FFT-NS-2	(Katoh *et al.*, 2002)	0.839	0.701	1
MUSCLE	(Edgar, 2004b)	0.830	0.731	4
Mafft L-NS-i	(Katoh *et al.*, 2005)	0.865	0.758	16
ProbCons	(Do *et al.*, 2005)	0.867	0.762	354
Dialign-T	(Subramanian *et al.*, 2005)	0.775	0.670	41
Kalign	(Lassmann & Sonnhammer, 2005)	0.803	0.708	3

4.4 Databases

A *database* is basically a systematically organised or structured repository of data that allow easy retrieval, update, analysis and output of a piece of information. Databases can be classified into different types on the basis of

their underlying data models, which are used to describe how the data are stored and retrieved in the database.

Flat file database model

Flat files were the first form of database implementation in the computing machines, the indexed sequential access method (ISAM) and the virtual storage access method (VSAM) being two of the more common file formats of the time. Files have no clear advantages but rather have several limitations, such as only random or sequential access to information, data duplication, lack of security, data isolation and high maintenance cost.

Hierarchical database model

To overcome the limitations of a flat file database, another type of database has been developed, a so-called *hierarchical database* that uses a one-to-many relationship for data elements. This model uses a tree structure to link a number of disparate elements to one owner in a parent/child manner. The advantages of this model include efficient searching, less data redundancy, more data independence, security, and integrity, but at the same time it has several limitations, such as complex implementation, inability to handle many-to-many relationships and lack of structural independence. The popularity of the hierarchical model coincided with the popularity of the *Network database model,* which modelled data more naturally by providing the possibility of many-to-many relationship.

Relational database model

Of all existing database models, the most successful is the *relational database model*, which was developed in 1970 by a British computer scientist (Edgar Frank Codd). He proposed thirteen rules (numbered from zero to 12), which define the characteristics of relational databases (Codd, October 21, 1985; Codd, October 14, 1985). A relational database permits the definition of data structures, storage and retrieval operations and integrity constraints. Moreover, it organises the data and relationships between them only in the form of tables. Relational databases have particular importance for the field of bioinformatics for the reason that most of the annotation systems are based on underlying relational databases. This makes it possible to look at biological phenomena on a scale that was previously not possible, *e.g.* all genes in a genome, all transcripts in a cell or all metabolic processes in a tissue. According to Roos (2001)*"we are swimming in a rapidly growing sea of data ... how do we keep from drowning?"*. Relational database systems also have certain shortcomings

as regards storing images, large text items and complex objects (Elsmari & Navathe, 2000).

Object oriented database model

In the early 1990s, *object oriented database systems* came onto the market to fulfil the requirements of complex applications *e.g.* (Atkinson *et al.*, 1993). The object-oriented system is based mainly on the concepts of *encapsulation*, *polymorphism* and *inheritance* and stores data in the form of objects and values. Object-oriented databases have a number of advantages over their predecessors, such as providing persistent storage of complex data types, ease of update and increased data security. However, in recent years both the research and business communities have moved toward use of *object-relational databases*, which are a hybrid of object and relational databases.

Other database systems

The life sciences represent a data intensive world that requires superior data management and advanced analytics and molecular biology research. For all these factors, over the past decade the database system that has attracted most attention for storing and managing life sciences data is the *data warehouse system*. The concept behind the data warehouse system is a database solution designed to collect and maintain cleansed and consolidated data from various sources in a secure environment to facilitate performance management, decision-making, strategic planning and execution for an organisation.

Data warehouse systems have been successfully implemented in life sciences (Nazari *et al.*, 2013; Haider *et al.*, 2009; Smedley *et al.*, 2009; Topel *et al.*, 2008; Prather *et al.*, 1997; BioMart Central Portal; Oracle) thanks to their friendliness, usefulness, different levels of user expertise, seamless integration of data and remote data access by special web services protocols.

4.5 Computing systems

The requirements for high speed computing have increased with the prodigious output of NGS data. *High performance computing* (HPC) storage and retrieval systems have become the only practical way to sift through the data to discover useful insights. It is also worth noting that computational processing of the NGS data is only half the computational problem associated with *genomics*. Managing and storing the huge volume of data is also an immense challenge. In this regard new database systems *e.g.* NoSQL, are becoming significant and experiencing growing use in big data and real-time web applications (Andlinger, 2013). *Cloud computing* is distributed computing over a network

and allows abstraction by the end users over the underlying applications and computer infrastructure. It could be the future solution for computing services and data storage.

UPPNEX is a national resource for the NGS community in Sweden and provides large-scale storage and computational resources for NGS projects. It is part of *Uppsala Multidisciplinary Center for Advanced Computational Science* (UPPMAX), which is Uppsala University's resource of high performance computers, large-scale storage and high performance computing (HPC). For this thesis, both UPPNEX and UPPMAX facilities were used for data storage and to perform bioinformatics analysis for the NGS data, as described in the appended papers (**I-III**).

4.6 Biological databases

Biological databases serve to store, organise, and analyse vast amounts of *data,* which are available in the form of sequences and structures of proteins and nucleotides. In addition, they serve as repositories for life sciences *information* that is gathered by means of scientific experiments, published literature, high-throughput sequencing and computational analysis (Attwood *et al.*, 2011).

4.6.1 Primary databases

Primary nucleotide sequence databases

GenBank (the genetic sequence database) is physically located in Bethesda, USA (Benson *et al.*, 2013) and is accessible through the NCBI (National Centre of Biotechnology Information) portal over the Internet (http://www.ncbi.nlm.nih.gov/).

EMBL (European Molecular Laboratory) is physically located at Hinxton, UK. It is part of the European Nucleotide Archive (ENA) (Cochrane *et al.*, 2013) and now is administered by EBI (European Bioinformatics Institute) (https://www.ebi.ac.uk/).

DDBJ (DNA Databank of Japan) (Ogasawara *et al.*, 2013) is physically located in Japan (http://www.ddbj.nig.ac.jp).

The GenBank, EMBL and DDBJ databases have collaborated since 1982 and are now linked with each other and synchronise their entries once every 24 hours according to the International Nucleotide Sequence Database Collaboration (INSDC) (Figure 14).

In 1982 GenBank had 606 sequences and by 1984 the number had reached 3,424 sequences, which seemed a lot at that time. However, since then the database has continued to grow exponentially and by the August 2013 release

GenBank contained publically available nucleotide sequences for over 2,80,000 formally described species (Benson et al., 2014).

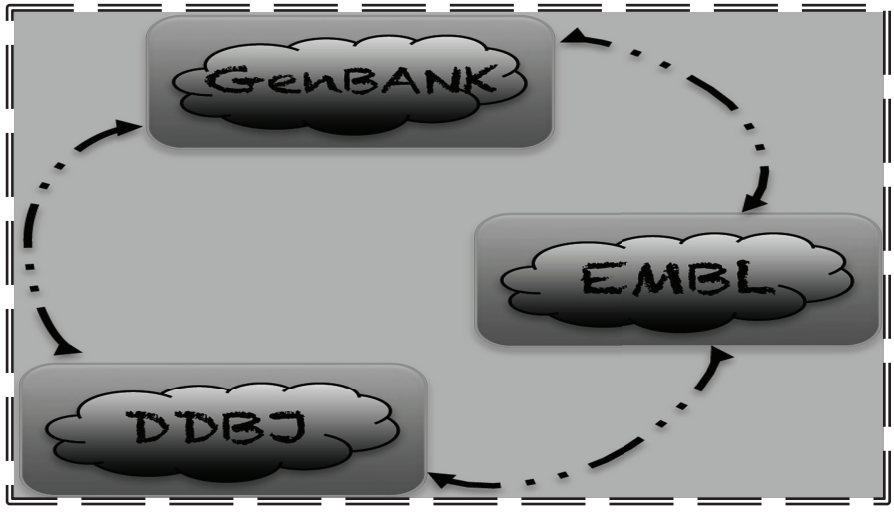

Figure 14. Links between the GenBank, EMBL and DDBJ databases, which synchronise their entries, once every 24 within the International Nucleotide Sequence Database Collaboration (INSDC).

Primary protein sequence databases

PIR-PSD (Protein Information Resource – Protein Sequence Database) was developed in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to help in the identification and interpretation of protein sequence information. It was the world's first database of classified and functionally annotated protein sequences and was extended from the *Atlas* of protein sequence and structure, which was developed by Margaret Dayhoff (1965-1978). PIR-PSD is a comprehensive, non-redundant and expertly annotated protein sequences database, organised by an underlying object relational database management system.

SWISS-PROT is a curated protein sequence database, which provides a high level of annotation. It contains some distinct features such as: i) the data in each entry are considered separately as core data and annotation. The core data consist of sequences entered in common single letter amino-acid code and the related references. The annotation contains information on the function or functions of the protein; ii) minimal redundancies due to merging the protein sequences submitted by separate entries; and iii) it is cross-referenced with 24 different databases.

TrEMBL (Translated EMBL) is a computer-annotated protein sequence database which was developed as a supplement of SWISS-PROT. It contains the translation of all coding sequences present in the EMBL nucleotide database that have not been fully annotated. For this reason it may contain the sequence of proteins that are never expressed and identified in the organisms.

UniProt is a comprehensive, fully classified and accurately annotated protein sequence knowledge base (Consortium, 2010). The UniProt Consortium is a result of joint efforts by the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). It is comprised of four components: 1) the UniProt archive, 2) the UniProt knowledge base, 3) the UniProt reference clusters and 4) the UniProt metagenomic and environmental sequence database.

In this thesis, the newly sequenced SAOB genomes were annotated for identified genes (**I-III**) using nucleotide sequence and protein sequence databases. Running an automated pipeline, the fraction of genes to which a specific function could be assigned was 50-60% of the total number of genes, which is the normal range for most of the sequenced bacterial genomes (Loewenstein *et al.*, 2009; Nimrod *et al.*, 2008). The remaining 30-40% of genes could be either homologous to genes of unknown function and are typically referred to as *conserved hypothetical* genes, or they do not have any known homologs and are called *hypothetical, non-characterised* or *unknown*. These can be real genes with no known function or they can be artefacts of the gene prediction process.

To annotate these hypothetical genes as much as possible, a PSI-BLAST search was performed against the primary protein sequence databases to predict some function for those hypothetical/unknown genes, particularly from regions, clusters, and operons of interest. PSI-BLAST (Altschul *et al.*, 1997) is a BLAST-like tool to find distant relatives based on iterative protein profiles and can find much more remote matches. It is normally used to identify the possible evolutionary relationship between the sequences. These conserved sequence motifs are commonly used along with structural and functional information to reveal many important features such as the catalytically active sites of enzymes, nuclear localisation signals in TFs or any other functional motifs in protein domains.

Primary structure databases

PDB (Protein DataBank) stores three-dimensional structures of proteins along with other biologically important molecules. The data in PDB are organised in flat files.

CDS (Cambridge Structural Database) was developed by the University of Cambridge to store the published three-dimensional structure of small organic molecules. It only stores small peptides, including neuropeptides and monomers and dimers of nucleic acids.

4.6.2 Secondary databases

Secondary databases contain information derived from primary databases and most of the secondary nucleotide sequence databases are simply sub-collection of sequences culled from one or more of the primary databases.

Secondary nucleotide sequence databases

FlyBase is a Berkeley Drosophila genome project (St. Pierre et al., 2014), while *AceDB* is a database management system for the *Caenorhabditis elegans* genome project. *GOBASE* is an organelle genome database that organises and integrates diverse data related to mitochondria and chloroplasts (O'Brien *et al.*, 2009) and the *Omniome database* is a comprehensive microbial resource (CMR) which makes data accessible from all of the completely sequenced bacterial genomes.

Secondary protein sequence databases

The *CluSTr* (Clusters of SWISS-PROT and TrEMBL proteins) database provides an automatic classification of entries of primary protein sequence databases (SWISS-PROT and TrEMBL) into groups of related proteins. In the database the clustering is done on the basis of analysis of all pairwise comparisons between protein sequences (Kriventseva *et al.*, 2001). It also links to InterPro (integrated documentation resource) to integrate information on protein families, domains and functional sites from member databases (PROSITE, Pfam, PRINTS and ProDom).

The *COGs* (Clusters of Orthologous Groups) database contains orthologous groups of proteins, which are generated by comparing the protein sequences of complete genomes (Tatusov *et al.*, 1997). Being a database of phylogenetic relationships, it can help to assign function to new protein sequences without going through laborious biochemical discovery processes.

The *PRINTS* database contains the patterns of protein sequences in the form of '*fingerprints*' that represent a set of motifs rather than a single motif (Attwood *et al.*, 1994). Along with the name, accession number and number of motifs information in the database, each entry also contains some additional information such as: i) cross-links to other databases for more detailed information about the characterised family, ii) number of motifs that make up

the fingerprint and how many sequences of that family contain that fingerprint, and iii) information on actual fingerprints.

PROSITE is a domain database for functional characterisation and annotation. It contains information on protein families, domains and function sites, along with amino acid patterns and profiles in these (Sigrist *et al.*, 2013). Each entry contains information about the patterns and related text description, plus references and links to all the protein sequences which contain that pattern.

The *Pfam* database contains protein families, which are sets of protein regions that share a significant degree of sequence similarity (Finn *et al.*, 2014). Each family contains information on the source used to make the entry and the method used for it, seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family, the hidden Markov model (HMM) profile, and the complete alignment of all the sequences identified in that family.

The *Protein Domain* (*ProDom)* database contains homologous domains that have been automatically identified with the sequence comparison and clustering from the UniProt knowledge database (Bru *et al.*, 2005). Rather than patterns identification, the database mainly focuses on identifying complete and self-contained domains, which can be used to analyse the structural and functional homology relationships between the protein sequences.

For some unknown proteins, sequence alignment or searches against protein databases with previously annotated proteins is not an effective way to identity their possible function and structure. The reason is that these unknown proteins are too distantly related to any protein of known structure. Another option to annotate these types of proteins is to search for a particular cluster of residue types, which are known as protein pattern, motif, signature or fingerprint. The reason is that during evolution, the folding patterns of proteins are often preserved, a feature which can be explored through structure-based comparisons to identify homologs when sequence-based comparisons become futile. These motifs developed in proteins because of particular requirements on the structure of specific regions of a protein and can be important *e.g.* for their binding properties or for their enzymatic activities. For this reason InterproScan (Mulder & Apweiler, 2007) was used in this thesis to perform searches against a range of domain or motif databases to annotate these *unknown*, *hypothetical* proteins in the newly sequenced SAOB genomes (**I-III**). This approach allowed the prediction of putative functions for several *hypothetical genes* in our regions and operons of interest. To extract the maximum information from the newly sequenced SAOB genomes, all identified genes were classified according to their homologous relationships

using the COGs database. Because orthologs typically have the same function, these relations yielded a number of functional predictions for poorly characterised genes.

Secondary structure databases

NDB (Nucleic acid Database) contains three-dimensional structural information about experimentally determined nucleic acids and assemblies. The *SCOP* (Structural Classification of Proteins) database is used to store manual classification of protein structures in a hierarchical way with many levels. The *CATH* (Class, Architecture, Topology, Homologous super family) database is used to store classification hierarchies and is a subset of PDB.

4.6.3 Metabolic pathway databases

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a knowledge base developed by Minoru Kanehisa's group at Kyoto University (Kanehisa & Goto, 2000), which consists of four databases: 1) Pathways (molecular interactions and reactions), 2) genes, 3) metabolites, and 4) functional interpretation.

BioCyc serves as a collection of organism-specific databases (Caspi *et al.*, 2012), which rely on the underlying methodology and data model developed by Peter Karp's group at SRI International (http://www.sri.com/). It includes EcoCyc (pathway/genome database of *Escherichia coli* K-12), MetaCyc (reference database of metabolic pathways), an open chemical database (database of metabolites) and organism-specific databases.

In this thesis (**IV, V**), to examine the physiological characteristics such substrate utilisation, acetate transportation, energy conservation, intermediate metabolism and acetogenesis for the newly sequenced SAOB genomes, several metabolic pathways were identified and analysed using different pathway databases such as KEGG and MicroCyc. *MicroCyc* serves as a collection of microbial pathway/genome databases (PGDBs) and was developed for the MicroScope projects (Vallenet *et al.*, 2006).

4.7 Information sources and annotations

When a certain molecular entity such as a nucleic acid, protein sequence, protein structure or a cell is known, *annotation* is the next process of adding the layers of analysis and interpretation for these entities, which can be used to extract their biological significance and to identify their role in some biological processes. It is a multi-step process, *e.g.* nucleotide-level annotation, protein-level annotation, and process-level annotation. A schematic data flow for a

genome annotation starting with the finished sequence is illustrated in Figure 15.

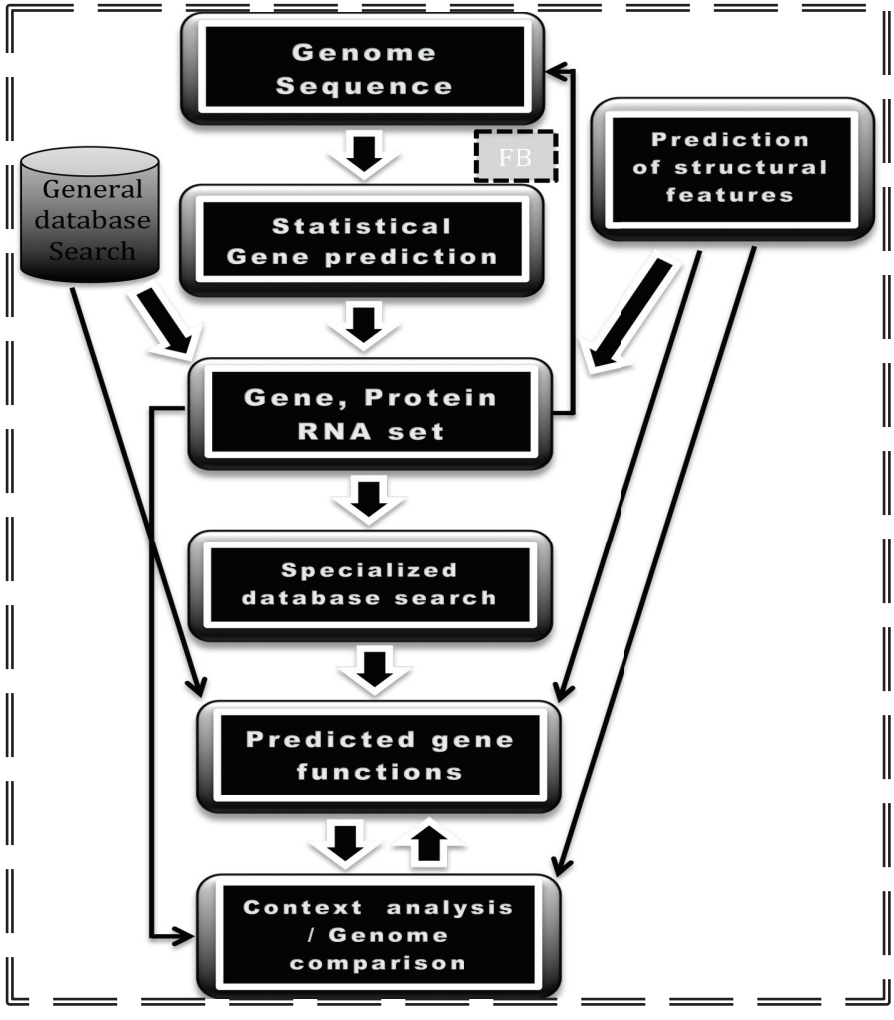

Figure 15. Illustration of a generalised data flow for a genome annotation. *FB*: feedback from gene identification for correction of sequencing errors, primary frameshifts. *Statistical gene prediction*: identification of open-reading frames (ORFs) (nucleotide-level annotation), *Prediction of structural features*: prediction of putative protein functions, signal peptide and transmembrane segments *etc*. (protein-level annotation). *General database search*: searching sequence databases for sequence similarity using BLAST-like tools. *Specialised database search*: searching domain databases such as Pfam and Prosite and conserved domain genome-oriented databases such as COGs for identification of orthologous relationship and refined functional prediction, and metabolic databases such as KEGG for metabolic pathway reconstruction (process-level annotation).

4.7.1 Nucleotide-level annotation

Nucleotide-level annotation or *structural annotation* is the first step in a genome annotation where punctuation marks are identified in the newly sequenced genomes, such as prediction of protein-coding genes, structural RNAs, tRNAs, small RNAs, pseudogenes, direct and inverted repeats, insertion sequences, transposons and other mobile genetic elements. In contrast to eukaryotic genomes, this step has the marked advantage of lacking intron in prokaryotic genomes (*e.g.* Bacteria, Archaea). Despite this, there can be some complications with prokaryotic genomes due to the overlapping ORFs, use of alternative start codon, use of codons other than the traditional ATG and possible dual function of stop codon (stop codons TGA or TAG may encode selenocysteine and pyrrolysine). Several approaches have been developed for accurate prediction of translation initiation sites in prokaryotes (Hu et al., 2009; Yada et al., 2001).

In the early 1980s the first generation of gene prediction algorithms used a local Bayesian approach to analyse one ORF at a time (Fickett, 1996; Gribskov *et al.*, 1984; Staden & Mclachlan, 1982). Subsequently, the second generation of gene prediction algorithms analysed the global properties of the genomic sequence (*e.g. di-, tri-, tetra-nucleotide composition*) to exploit discrimination between coding and non-coding regions using statistical and computational methods. Several programmes such as *Glimmer* (Salzberg *et al.*, 1998) and *GeneMark* (Lukashin & Borodovsky, 1998) use this approach. Gene prediction methods then entered into another generation by performing similarity searches (Guigo *et al.*, 2000) with known protein sequences, like the BLASTx and FASTA gene predication programmes. These third generation programmes combine execution of multiple gene-calling programmes with similarity-based methods.

4.7.2 Protein-level annotation

Functional annotation, the second step in the process of genome annotation, involves assigning a biological function to the predicted genes or proteins. This is similar to asking "*what*" after answering the question of "*where*" in the genome (Stein, 2001b). Protein alignments and searches against protein sequence databases are some common approaches to perform this level of annotation. The reason is that many genes that encode similar proteins also share varying amounts of sequence homology, but this can be difficult for some genes because of the intrinsic nature of evolutionary processes (Stein, 2001b). These diverse situations can be addressed by searching against experimentally verified or highly conserved protein domain databases.

4.7.3 Process-level annotation

In *process-level annotation* higher-level classification is performed for annotated genes and proteins by placing them into their respective biological pathways, which helps to understand their functional roles in the organism. A common method for this annotation is through the creation of the Gene Ontology (GO) vocabulary, which assigns a functional role to a gene in a hierarchical way (Ashburner *et al.*, 2000). For example, a term such as *enzyme* can lead to a more specific enzyme such as *alcohol dehydrogenase*, which can be further associated with a number of other descriptions. Due to this level of flexibility, many pathway databases, *e.g.* KEGG, use GO terms as a framework for describing metabolic and component interactions.

In order to annotate the newly sequenced SAOB genomes in this thesis (**I - III**), structural annotation was performed to identify the genomic objects and their properties, such as number of protein-coding genes, GC%, ribosomal RNAs, tRNA, average CDS length, repeated regions, average intergenic length and protein coding density. Different algorithms were used to predict the putative CDSs within these sequenced genomes such as: i) *Glimmer* (Gene Locator and Interpolated Markov ModelER) (Salzberg *et al.*, 1998), which primarily searches for long open reading frames (ORF) and uses a fixed order Markov model or interpolated Markov model, ii) *AMIGene* (Annotation of MIcrobial Genes) application (Bocs *et al.*, 2003), which first constructs a Markov model for input genomic sequence (the gene model), followed by the combination of well-known gene-finding methods and an heuristic approach for the selection of the most likely CDSs, and iii) *Prodigal* (PROkaryotic DYnamic programming Gene-finding Algorithm) (Hyatt *et al.*, 2010), which has salient features such as speed, accuracy and specificity. tRNAs were predicted using the *tRNAScanSE* tool (Lowe & Eddy, 1997), which combines several algorithms to identify tRNAs with high accuracy and also has the ability to distinguish between active tRNAs and tRNA pseudogenes. Repetitive sequences (a sequence present twice or more with a high degree of similarity within a large sequence) were detected using the Repseek programme (Achaz *et al.*, 2007). PHAST (PHAge Search Tool) (Zhou *et al.*, 2011) tool was used to identify the phage regions in the sequenced SAOB genomes. The transporter database (Saier *et al.*, 2014) was used to identify the transporter genes in the sequenced genomes (**V**).

As a following step, functional annotation for all predicted CDSs was carried out using different approaches as described in the above sections. SignalIP neural network software (Vonheijne, 1986) was used for signal peptide prediction and transmembrane regions were predicted through the TMHMM server, which analyses the physical constraints of both soluble and

membrane-based sequences with up to 90% accuracy (Krogh et al., 2001). Predicted genes were also subjected to manual analysis using the MaGe web-based platform, which provides functional information on proteins that was used to assess and correct genes predicted through the automated pipeline.

Web-based annotation systems

The first automated systems, *MAGPIE* (Gaasterland & Sensen, 1996) and *GeneQuiz* (Scharf *et al.*, 1994), were developed to provide biological function assignments to the genes. Subsequently, several *web-based tools* came into the field offering a number of services to annotate and analyse the genomic data. In contrast to web servers, several *semi-standalone* systems allow more control of the annotation process and unrestricted access to the data with enhanced security, but also require computational hardware and necessary expertise to integrate the tools within the system. Some of these web-based and stand-alone annotation pipelines are listed in Table 5.

Table 5. *List of web-based and stand-alone annotation systems.*

Program Name	URL
BASYs	http:/basys.ca/basys/cgi/submit.pl
RAST	http:/rast.nmpdr.org/
MaGe	http://www.genoscope.cns.fr/agc/microscope
MAGPIE	http:/magpie.ucalgar y.ca/
GenDB	http:/www.cebitec.uni-bielefeld.de/ groups/brf/software/gendb_info/
IGS	http:/ae.igs.umar yland.edu/cgi/index. cgi
JCVI	http:/www.jcvi.org/cms/research/ projects/annotation-ser vice
CGP	http:/nbase.biology.gatech.edu/
GenePRIMP	http:/geneprimp.jgi-psf.org/login
Integrated Microbial Resource Expert Review (IMG-ER)	http:/merced.jgi-psf.org/cgi-bin/er/main.cgi
xBASE	http:/xbase.ac.uk.annotation
Stand-alone annotation	*systems*
NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP)	http:/www.ncbi.nlm.nih.gov/genomes/frameshifts/frameshifts.cgi
DIYA	http:/sourceforge.net/projects/ diyg/
Ergatis	http:/ergatis.sourceforge.net
Computational Genomics Pipeline	http:/jordan.biology.gatech.edu/ jordan/software/cg-pipeline

Genome-annotation browsers

Bacterial genome annotations produced using computational pipelines need to be reviewed for accuracy and error correction through visualising the sequence information. Using a sequence viewer that is characterised to show the individual genes based on their sequence often does this work. Some of these *genome-annotation browsers* are listed in Table 6.

Table 6. *List of some genome annotation browsers.*

Browser Name	URL
Artemis	http:/www.sanger.ac.uk/resources/software/ artemis/
GBrowse	http:/gmod.org/wiki/Gbrowse
Apollo	http:/apollo.berkeleybop.org/current/index.html
CGView	http:/wishart.biology.ualberta.ca/cgview/
Manatee	http:/manatee.sourceforge.net/
WebGBrowse	http:/webgbrowse.cgb.indiana.edu/cgi-bin/webgbrowse/uploadData
NCBI Genome-Workbench	http:/www.ncbi.nlm.nih.gov/projects/gbench/
UCSC Microbial-Genome Viewer	http:/microbes.ucsc.edu/
IMG	http:/imgweb.jgi-psf.org/archaeal_qa/doc/ findGenomes.html

MaGe (**Ma**gnifying **Ge**nomes) system was used in this thesis to perform the structural and functional annotation and metabolic pathways analysis for the newly sequenced SAOB genomes (**I-V**). (Vallenet *et al.*, 2013; Vallenet *et al.*, 2006). The *RAST* (**R**apid **A**nnotation using **S**ubsystem **T**echnology) annotation system was also used to annotate the sequenced genomes, which provides the additional facility to correct the annotations which come from automated pipeline by '*walking the genome*' to look for genes that need to be deleted, inserted or re-annotated (Aziz *et al.*, 2008). The *BASys* (**B**acterial **A**nnotation **S**ystem) annotation system was also used, which provides extensive textual annotation by using a collection of more than 30 programmes and results in an annotation output that contains ~60 subfields for each gene (Van Domselaar *et al.*, 2005). The *Artemis* tool was used to visualise genome annotations (**I-III**) to identify the alternative start codon for doubtful CDSs that could be predicted by automated annotation pipelines. (Carver *et al.*, 2012; Rutherford *et al.*, 2000). Microbial genome annotation with these automatic pipelines can also possibly present poor annotation and errors. For this reason manual curation was performed to identify and remove these types of annotation errors from our newly sequenced and annotated SAOB genomes.

4.8 Limitations of biological databases

There is no doubt that the existing biological databases serve as a great source of life sciences information from search areas including genomics, proteomics and metabolomics. While acknowledging all these pros of biological databases, it should be noted that there are a number of possible cons for reasons such as:

- The new genomes submitted may have misannotations and errors that ultimately propagate into secondary and domain-specific databases
- The entries in the databases may contain spelling mistakes and can be omitted from the search results for example there are 128 proteins in the UniProt database that contain the misspelled word '*syntase*' instead of the correct word '*synthase*'
- A particular gene in the databases may have more than one product name, for example of the current set of 2,696 microbial genome sequences in RefSeq, 23,843 are identified by at least two different product names. The worse example of this is gene "*tnp*", which has 151 different product names in the database
- There can be inconsistencies with the names of the proteins in the database, for example there are currently 53,035 proteins in UniProt with a name containing two words at the same time 'name:*hypothetical'* AND 'name:*protein*'. Moreover, there are 5,178,212 proteins that contain the words '*uncharacterized*' and '*protein'*, which may be real genes with no known function or can be artefacts of gene prediction
- Several bacterial genomes can have multiple species and strains, which can be sequenced and annotated separately by different groups. This may also introduce inconsistencies in the biological databases.

5 Summary of papers

This section presents the background, results and discussion of papers 1-V, following with conclusions and ends up with future perspectives for the SAOB project.

"*Never trust a tall dwarf. He's lying about something*"
(Solomon Short)

5.1 Background

Today's industrial society is greatly dependent on energy. There is a growing demand for new alternative energy sources other than fossil oil, coal and gas because of the depletion of reserves of these fossil fuels at a high rate. In this regard, biogas containing energy-rich methane, produced during anaerobic degradation of biomass has promising potential to secure future energy supply with reduced greenhouse gas emissions (Bauer *et al.*, 2012). Methane is formed as the end product during anaerobic degradation of organic material. Acetate is the main precursor for methane production during this process and can be converted to methane through two distinct pathways: aceticlastic methanogenesis and hydrogenotrophic methanogenesis (SAO). The second pathway encompasses a two-step reaction in which acetate is first oxidised to H_2/formate and CO_2 by acetate-oxidising bacteria and subsequently CO_2 is converted to methane in a syntrophic association with hydrogenotrophic methanogens. For a long time methane was considered to result only from the action of aceticlastic methanogens. In fact, acetate oxidation is a thermodynamically very difficult reaction that only proceeds at low hydrogen levels. As a result, studies on the operation and optimisation of biogas reactors have tended to focus on maintaining the activity of the aceticlastic methanogens (Karakashev *et al.*, 2006), with the role of SAO neglected. However, the significance of SAO has recently been emphasised in a number of studies of methanogenic systems (Sun *et al.*, 2014; Westerholm *et al.*, 2012; Hao *et al.*, 2011; Sasaki *et al.*, 2011; Shimada *et al.*, 2011).

Only a few SAOB have been isolated and characterised and little is known about their physiology and biochemistry. Therefore in the work described in this thesis, we were interested in research within this area to increase understanding of the nature and lifestyle of SAOB. To do this, we sequenced three SAOB genomes and performed structural and functional annotations to reveal the genomic behaviour of these SAOB genomes (**I-III**). Subsequently we started performing metabolic pathways analysis (**IV, V**) in order to understand physiological attributes such as substrate utilisation, acetate transportation, energy conservation, intermediate metabolism and acetogenesis.

The three SAOB isolates chosen for genome sequencing (*Tepidanaerobacter acetatoxydans* sp. Re1, *Clostridium ultunense* strain Esp *and Syntrophaceticus schinkii* strain Sp3) have a proven ability to oxidise acetate in co-culture in the presence of hydrogen-consuming *Methanoculleus* sp. and have been identified as key organisms for efficient biogas production from protein-rich materials (Moestedt *et al.*, 2014; Sun *et al.*, 2013; Westerholm *et al.*, 2012; Westerholm *et al.*, 2011a). Determination of the whole genome sequence for these SAOB was intended to provide insights into

the genetic and biochemical diversity of acetate-oxidising microorganisms with the ability to thrive in thermodynamically difficult situations and different environmental conditions.

5.2 Genome assembly

The comparative assembly approach employed, which combined reference-guided assembly with *de novo* assembly, worked well and was used to assemble the complete *Tepidanaerobacter acetatoxydans* genome sequence as a single circular chromosome. The resulting genome had a size of 2,761,252 bp (**IV**), which is ~1.3 Kb greater than the other available genome of *T. acetatoxydans* (NC_015519) in the GenBank database. In the case of *Clostridium ultunense* and *Syntrophaceticus schinkii,* no reference genomes were available at the time of study. Therefore using the *de novo* assembly approach, 281 scaffolds with a total size of 6,159,766 (**II**) and 215 scaffolds with a total size of 3,196,921 bp (**III**) were produced in the form of working draft genome sequences.

The genomes of all *Thermoanaerobacterales* sequenced to date vary in size with the range 1.4 - 3.3 Mb, where *Coprothermobacter proteolyticus* DSM 5265 is the smallest and *Thermanaeromonas toyohensis* ToBE DSM 14490 is the largest member of the class. *Tepidanaerobacter acetatoxydans* and *Syntrophaceticus schinkii* match this range very well, with a size of 2.7Mb and 3.1 Mb, respectively.

In contrast to the *T. acetatoxydans, S. schinkii* and a recently published *Clostridium ultunense* strain BS (3.2 Mb) (Wei *et al.*, 2014) genomes, *Clostridium ultunense* Esp was almost double the size. Some clostridial species also have a genome size in this range (*C. beijerinckii* NCIMB 8052: 6 Mb, *C. saccharoperbutylacetonicum* N1-4(HMT): 6.6 Mb). Using a local BLAST tool against the updated 16S rRNA genes database, only one 16S rRNA gene could be identified in the sequenced *C. ultunense* genome. This is in accordance with our findings in the wet laboratory, where only one 16S rRNA species was identified. On the basis of synteny analysis, the most closely related organism to *C. ultunense* is *Alkaliphilus metalliredigens,* which also possesses more or less the same genome size, number of tRNA genes and protein-coding density as *C. ultunense* Esp (see Table 11 for comparison). However, to confirm the assembly one more time, the genome is also being sequenced using Illumina HiSeq-2500 mate-pair sequencing with a library size of 3 Kb (data not yet analysed).

5.3 Genomic features

Among the total number of predicted ORFs of the three SAOB isolates ~61–72% have been assigned tentative functions. A comparison of genomic statistics for three SAOB genomes with their closely related organisms is summarised in Table 11.

5.3.1 Tandem duplication

In prokaryotic genomes, operons mostly become into existence through clustering of related genes, which also reflect their functional context. Tandem duplication of related protein coding genes acts as a driving evolutionary force in the origin and maintenance of gene clusters (Reams & Neidle, 2004). It occurs frequently and facilitates the organisms in adaptation to a large number of diverse conditions in their niche (Romero & Palacios, 1997; Anderson & Roth, 1977). Moreover, tandem gene duplication can augment gene expression and protein dosage, which can allow cells to proliferate under growth-limiting conditions. Tandem duplications of protein coding genes were identified in the genomes of the SAOB isolates. The results showed that *C. ultunense* contains a higher gene duplication rate than the other sequenced SAOB genomes and the model acetogenic organism. This provides justification for the comparably larger size of the organism (Table 7). It also permits adaptation of this organism to a wide range of conditions in its competitive environment.

Table 7. *Comparison of tandem duplicated protein-coding genes identified in the sequenced SAOB and model acetogenic genomes.*

Organism	Genomic regions	No. of duplicated genes
T. acetatoxydans Re1	45	126
C. ultunense Esp	**91**	**221**
S. schinkii Sp3	56	163
Th. phaeum DSM 12270	59	176
Moorella thermoacetica ATCC 39073	55	177

5.3.2 CRISPRs defence system

CRISPRs (**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats) are widespread in many bacterial and almost all archaeal genomes sequenced (Horvath & Barrangou, 2010). They can be used to understand the bacterial defence mechanism (Barrangou *et al.*, 2007).

In the genome of *T. acetatoxydans,* two identified operons are responsible for encoding *cas* proteins (operon 1:TepiRe1_0109-0115, and operon 2:

TepiRe1_0129-0134) (**IV**). The *C. ultunense* genome harbours three operons for encoding *cas* proteins (operon 1: CULT_210003-210006, operon 2: CULT_1190006-11900012, and operon 3: 1170013-1170017). The *S. schinkii* (**V**) genome contains one operon for encoding *cas* proteins (SSCH_830002-830009). This *cas*/CRISPR system is a prokaryotic defence mechanism, which provides immunity against invading mobile genetic elements such as phages and plasmids in an RNA interference-like manner. These CRISPR-associated sequence (*cas*) genes are often directly adjacent to the CRISPR loci. These loci typically consist of different numbers of non-contiguous repeats with lengths ranging from 20 to 47 bp (Haft *et al.*, 2005) and unique spacers of different length and sequence between the repeats. Of the total of 2,091 completely sequenced prokaryotic genomes deposited in the GenBank database (2012-07-04), CRISPRs are present in ~55% of bacterial genomes and ~87% of archaeal genomes. In all CRISPR-containing organisms the occurrence of CRISPR loci ranges from one to 21 and only a small number of organisms (28), including *T. acetatoxydans, C. ultunense* and thermophilic SAOB *Th. phaeum* harbour 10 or more CRISPRs loci, whereas *S. schinkii* contains 8 CRISPR loci in the genome. This may be an indication that these organisms have to cope with mobile genetic elements in their particular environment, which results in it acquiring a higher number of CRISPR loci in their genomes. Based on this, it can also be speculated that the increased size of *C. ultunense* might be the result of the activity of mobile genetic elements, because a large number of CRISP loci can be considered to be a sign of a niche where the organism is more prone to attachment by mobile genetic elements.

5.3.3 Phage identification

The mostly widely accepted paradigm for viral-bacterial interactions was first suggested by Thingstad and Lignell (1997), and is termed "*phage kill the winners*". Based on the ubiquitous presence of these bacterial viruses, it can be concluded that they exit wherever bacteria exit. Bacteriophages are found commonly in two categories, *i.e.* virulent or lytic, which kill the host immediately, and temperate or lysogenic, which recombine with their host DNA and remain dormant, but can cause cell lysis under certain environmental conditions. These viruses may have single or double stranded RNA or DNA genomes ranging in size from a few thousand to half a million base pairs in length (Casjens, 2005). Among the tailed, filamentous, pleomorphic and polyhedral bacteriophages, tailed phages are the largest and most widespread group (Ackermann, 2001).

As mentioned in the previous section, our CRISPR analysis showed that the genomes of the sequenced SAOB isolates had acquired more CRISPR loci, perhaps to cope with mobile genetic elements due to their microbe-rich environment. Related to this, in the complete genome sequence of *T. acetatoxydans*, two tailed phages (temperate) were identified by *in silico* analysis (**IV**) (Table 8). In our first wet-laboratory experiment, we also obtained an indication for their presence, but this needs to be confirmed. These phages might impair growth of *T. acetatoxydans* in pure and syntrophic culture.

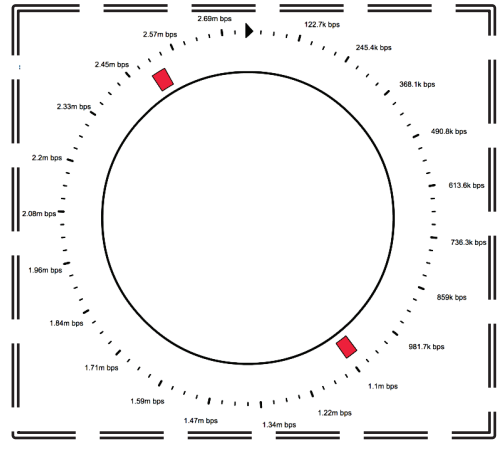

Figure 16. Graphical illustration of two phage regions identified in the complete genome sequence of *Tepidanaerobacter acetatoxydans* Re1.

One complete set of genes for a phage was also found in the *S. schinkii* genome (Figure 17) (**V**). The most closely related SAOB *Th. phaeum* to *S. schinkii* also contains one complete phage genome, which also falls in the small group of organisms that acquired higher number of CRISPR loci into their genomes. No phage region was identified in the genome of *Clostridium ultunense* Esp.

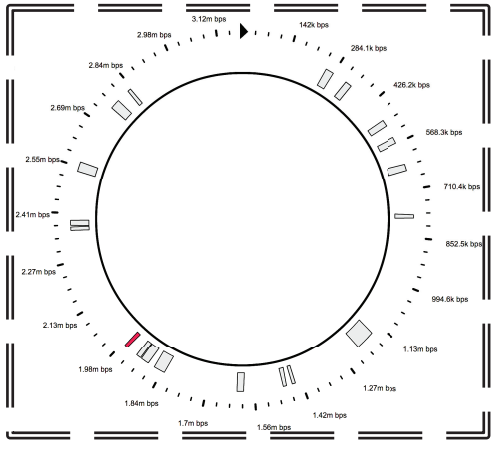

Figure 17. Graphical illustration of a phage region identified in the genome sequence of *Syntrophaceticus schinkii* Sp3.

Table 8. *Statistics on two phage regions identified in T. acetatoxydans strain Re1 genome.*

Start	End	Length Kb	Total proteins	Phage proteins	Hypothetical proteins	Bacterial proteins	tRNA	Status
1,075,607	1,111,233	35.6	51	28	22	1	1	Complete
2,500,103	2,538,024	37.9	49	34	13	2	0	Complete

However, there is little or no information regarding the impact of viral activity on the syntrophic oxidising bacterial community in biogas reactors, which might have deleterious effects under certain factors such as pH, ammonia concentration or temperature. A thorough investigation is needed to determine the role of phages for the biogas reactor SAOB community.

5.4 Comparative analysis

5.4.1 COG analysis

Of the total number of predicted CDSs of each of three SAOB genomes identified, it was possible to allocate ~75-81 % to the 21-23 functional COGs (**I-III**), which is in the same range as described for other acetogenic bacteria such as *Acetobacterium woodii* and *Moorella thermoacetica*. The number of genes in four major COG categories, including amino acid transport and metabolism, energy metabolism, carbohydrate transport and metabolism and inorganic ion transport and metabolism, were found to be high in *T. acetatoxydans and C. ultunense* (Table 9), revealing their efficient carbon, amino acid and energy metabolism. This is also an indication of their efficiency in utilising nutrients that enable them to survive in different difficult situations. In contrast, *S. schinkii* has comparatively fewer genes for carbohydrate transport and metabolism (Table 9), which explains the observed inability of this organism to grow apart from on a restricted number of substrates, such as ethanol, betaine and lactate (Westerholm *et al.*, 2010).

The COG analysis conclusively reflected the habitat adaptation of the SAOB genomes on account of their competitive environments and also confirmed their growth habits and ability for substrate utilisation as previously observed by Westerholm *et al.* (2010) (2011b).

Table 9. *Comparison of COG analysis for the three newly sequenced SAOB genomes.*

Organism	Total COGs (%age)	Total No. genes	Amino acid transport & metabolism	Carbohydrate transport & metabolism	Energy production & conservation	Inorganic ion transport & metabolism
T. acetatoxydans Re1	81.25	2,158	10.80	9.11	5.91	3.87
C. ultunense Esp	81.41	5,248	10.53	7.08	5.60	6.18
S. schinkii Sp3	75.07	2,586	9.84	4.00	5.92	5.97

5.4.2 Synteny analysis

Phylogenetic analysis revealed that the closest relatives of *T. acetatoxydans* are members of the genera *Thermovenabulum, Tepidanaerobacter, and Thermosediminibacter* (Westerholm *et al.*, 2011b). Synteny analysis again confirmed this phylogenetic relationship inferred by 16S rRNA sequence analysis. *Tepidanaerobacter acetatoxydans* has the highest number of orthologues (1,294 or 48.72%) relative to *Thermosediminibacter oceani* (Figure 18A) (**IV**), an anaerobic thermophilic bacterium isolated from marine sediment (Pitluck *et al.*, 2010).

In contrast, synteny analysis of *Clostridium ultunense* Esp showed no high similarity to acetogens and or to sulphate reducers, which share phenotypic features, at first glance. Instead, *Alkaliphilus metalliredigens* appears to be the closest relative of *C. ultunense,* which has the highest number of orthologues (2,271 or 35.23%) relative to *A. metalliredigens* (Figure 18B). *Alkaliphilus metalliredigens* is a strict anaerobic metal-reducing bacterium, which belongs to the genus *Alkaliphilus,* but is only distantly related to other commonly studied iron-reducing microorganisms (Ye *et al.*, 2004). All the genes involved in the W-L pathway were also identified in the genome. This led us to speculate that this organism might also belong to the physiological group of acetogens, which might be the reason for the distant relationship to other iron-reducing microorganisms.

Synteny analysis confirmed that *S. schinkii* is the closest relative of *Th. phaeum* by having the maximum number of orthologues (1,788 or 51.90%) relative to *Th. phaeum* (Figure 18C). Both *S. schinkii* and *Th. phaeum*, are known as SAOB, but differ clearly in their substrate utilisation patterns. Moreover, in contrast to the thermophilic *Th. phaeum*, *S. schinkii* possesses mesophilic characteristics and cannot switch to a chemolithoautotrophic lifestyle and use sulphate as a terminal electron acceptor.

Moreover synteny analysis was performed for these newly sequenced SAOB isolates with SAOB (*Th. phaeum*), sulphate-reducing bacteria (*Desulfosporosinus orientis*), acetogenic bacteria (*M. thermoacetica, C. ljundanlii* and *T. oceani*), and *A. metalliredigens* (Table 10). All of these organisms possess diverse abilities, such as SAO, CO_2 utilisation through W-L pathway and sulphate reduction, but still share 25-50% similarity for both genomic organisation and functionality. This may indicate that all these organisms have evolved these specific diverse abilities by interacting with niche-associated microbes to survive in their specific competitive environments. All three-sequenced SAOB showed more similarity to acetogens than to each other except for *S. schinkii.* However, this organism showed the second highest synteny to sulphate reducers after its closest relative SAOB genome, which was also observed by Müller *et al.* (2013) on the basis of a single *fhs* gene sequence.

It has been shown that all the SAOB belong to the physiological group of acetogens that are widely distributed in a range of phylogenetic classes where acetogenesis appears as a metabolic feature, rather a phylogenetic trait. Besides this, SAO is not a common physiological feature of acetogens. This provides reason to suggest that the physiological feature of SAO is acquired by these SAOB for their habitat adaptation to survive in a specific anoxic and very competitive environment.

Table 10. *Comparison of synteny analysis for the genome of the three SAOB isolates determined here with thermophlic SAOB (Th. phaeum), sulphate-reducing bacteria (Desulfosporosinus orientis), acetogenic bacteria (M. thermoacetica, C. ljundanlii, and T. oceani), and A. metalliredigens.*

Organism	*T. acetatoxydans*	*C. ultunense*	*S. schinkii*
T. acetatoxydans	-	27%	30%
C. ultunense	43%	-	33%
S. schinkii	35%	25%	-
M. thermoacetica	36%	25%	36%
T. oceani	**49%**	20%	29%
A. metalliredigens	44%	**35%**	31%
Th. phaeum	37%	25%	**51%**
C. ljundanlii	36%	27%	29%
D. orientis	38%	31%	37%

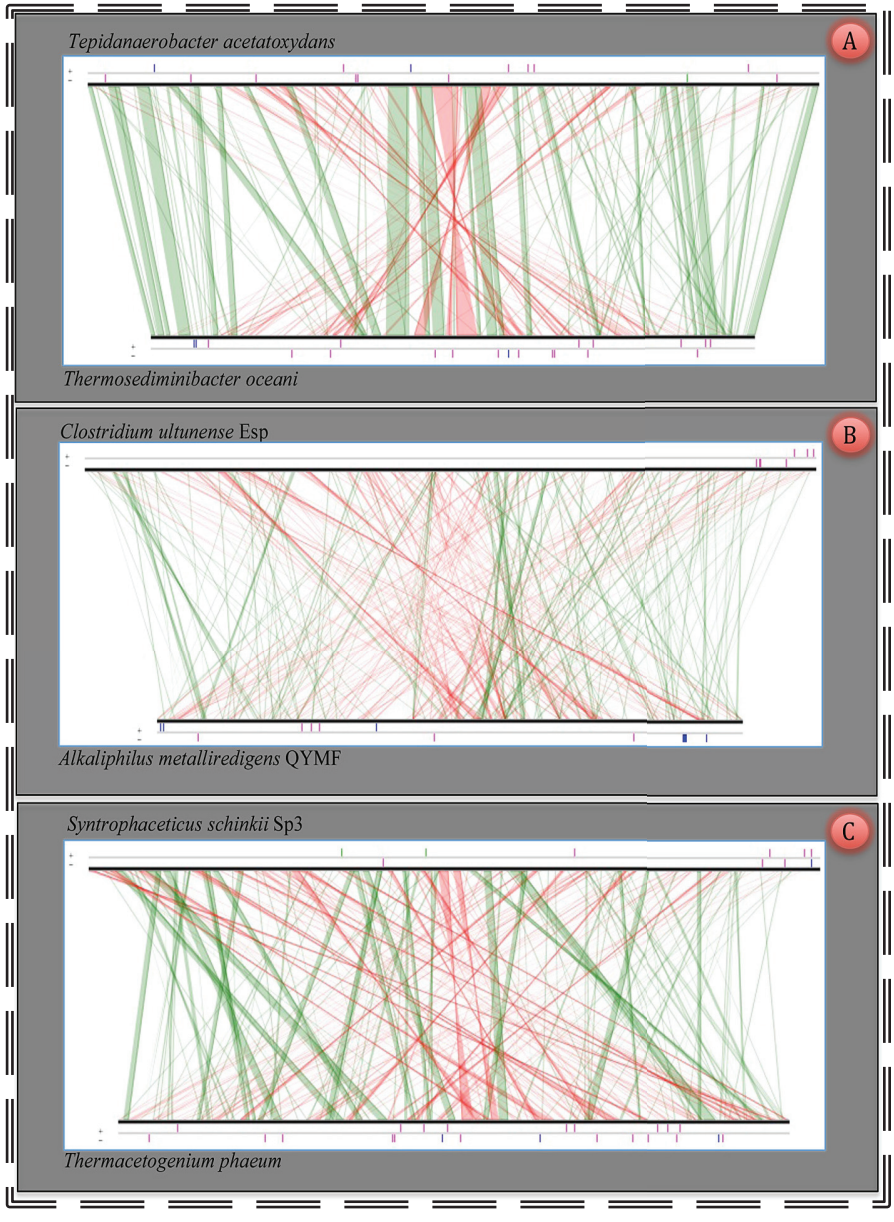

Figure 18. Synteny-based comparison of A) T. acetatoxydans Re1 genome with that of the closely related T. oceani*, B) C. ultunense strain Esp genome with that of closely related A.* metalliredigens strain QYMF and C) *S. schinkii Sp3 genome with that of the closely related Th. Phaeum. Linear comparison of all predicted gene loci was performed using built-in tool in MaGe platform with the synton size of 3 genes. The lines indicate syntons between two genomes. Red lines show inversions around the origin of replication. Vertical bars on the boarder line indicate different elements in genomes such as pink: transposases or insertion sequences: blue: rRNA and green: tRNA.*

5.5 Phenotypic features

5.5.1 Sporulation

All SAOB were shown to be able to form endospores. The master regulator Spo0A needed for sporulation (Paredes-Sabja *et al.*, 2014) was identified as a single copy in *T. acetatoxydans* and *S. schinkii* (TepiRe1_1496, SSCH_630004), while *C. ultunense* possesses two copies (CULT_1250001-290061), but all three organisms lack genes encoding the phosphorylase (Spo0F, Spo0B), as has been observed in other clostridia.

All the sporulation-specific sigma factors were identified as a single occurrence in *T. acetatoxydans* (SigE: TepiRe1_1251, SigG: TepiRe1_1252, SigF: TepiRe1_1488, SigK: TepiRe1_1533), and *S. schinkii* (SigE: SSCH_460001, SigG:SSCH_1070017, SigK: SSCH_700030), which lacks one SigF. In contrast, *C. ultunense* possesses two occurrences for each sigma factor in the genome (SigE: CULT_1130013-310020, SigG: CULT_1130012-310019, SigF: CULT_1000022-290013 and SigK: CULT_130076-310020) .

The master regulator plays the central regulatory role in sporulation. It is an element of the effector pathway responsible for the activation of sporulation genes in response to nutritional stress. *Clostridium ultunense* Esp uniquely contains two copies of this master regulator and all sigma factors, which is not common in other SAOB and acetogens sequenced.

Even though biogas reactors are fed with nutrients and microbes face less risk of starvation, these organisms seem to have adapted to their habitat to deal with any unfavourable conditions, in which they can survive through forming endospores.

5.5.2 Oxygen tolerance

The presence of two putative manganese containing catalases and two putative rubrerythrin encoding genes in *T. acetatoxydans* (TepiRe1_0143, 2025, 0311, 1181), *S. schinkii* (SSCH_1760003, 2560004, 590006, 180042), and *C. ultunense* (CULT_140064, 10174, 110003, 140065) can be an indication that all these three organisms possess the ability to be tolerant to small amounts of oxygen, as contained by all sequenced SAOB and certain acetogens (Karnholz *et al.*, 2002).

In biogas reactors, some oxygen tolerance might be advantageous due to the brief exposure to oxygen for feeding purposes. The presence of these genes is in good agreement with the non-disrupted growth of these SAOB isolates in biogas reactor environment. Thus, micro-aerotolerance and adaptation to unstable redox conditions is of advantage and gives increased competiveness.

5.5.3 Selenocysteine-containing proteins

Selenium is an essential trace element for many organisms and is present in proteins in the form of selenocysteine (Sec) residue (Stadtman, 1996; Bock *et al.*, 1991). Sec is known as the 21st naturally occurring amino acid and it is co-translationally inserted into proteins by recoding opal (UGA) codons. Interestingly, most of the organisms that are able to decode Sec use this amino acid only in a small set of proteins or even in a single protein. Genes encoding key components of Sec-decoding (SelA, SelB, SelC, and SelD) and selenouridine-utilising (SelD) machinery are widely distributed in bacterial genomes (Zhang *et al.*, 2006).

Selenocysteine containing proteins were identified in *C. ultunense* (Selenoprotein B: CULT_240034, 490013, 970008) in three occurrences and in *S. schinkii* (Selenoprotein B: SSCH_440005, 960008) in two occurrences but seem not to be expressed in the *T. acetatoxydans* genome. Although all three genomes contain a single copy of the L-selenocysteinyl-tRNA^{Sec} (selA: TepiRe1_1824, selA: CULT_160013, selA: SSCH_110006), *T. acetatoxydans* and *S. schinkii* each also contain a single copy of the selenophosphate synthase (selD: TepiRe1_0473, selD: SSCH_970007), and *C. ultunense* possesses two copies (selD: CULT_1030023, 1300012). Furthermore, *S. schinkii* contains a single copy of the selenocysteinyl-tRNA specific elongation factor (selB: SSCH_110004) and *C. ultunense* possesses two copies (selA: CULT_1030025, 210029), but it was not identified in the *T. acetatoxydans* genome. It seems that only *S. schinkii and C. ultunense* have the ability to express selenocysteine proteins, while *T. acetatoxydans* is incapable of expressing these proteins.

5.5.4 Secretion pathways

In the *Tepidanaerobacter acetatoxydans* genome, a total of 628 CDSs were predicted to encode proteins having at least one transmembrane helix, including 28 putative ATP-binding cassette (ABC) transport systems (Table S2 in **IV**), 6 triparite ATP-independent transporters (TRAP) (Table 5 in **IV**), 9 secondary sodium/solute transporters ((Table 4 in **IV**) and 15 putative phosphotransferase systems (PTS) (Table S1 in **IV**) to transport phosphoenolpyruvate-dependent sugar and sugar derivatives such as lactose, cellobiose, mannose, fructose, sorbose, galacticol, glucitol, sorbitol and N-acetyl glucosamine (**IV**). The TRAP transporters might be an adaptation to the syntrophic life style of this organism. In contrast, the physiological relatives of *T. acetatoxydans* (*M. thermoacetica* and *A. woodii),* harbors only one or none TRAP, respectively.

Syntrophaceticus schinkii contains 123 transporter genes, which are distributed into 65 transport systems (Table S1 in **V**). This is a smaller number

than in the model organism for acetogens, *M. thermoacetica*, which explains the heterotrophic growth ability of the organism with a limited number of substrates (Westerholm *et al.*, 2010). We were able to identify at least 27 ABC-type transport systems in *S. schinkii* (Table S1 in **V**), but both *S. schinkii* and its most closely related thermophilic SAOB *Th. phaeum* do not harbour any PTS and TRAP transport systems in their genomes. This explains the frailty for growth of *S. schinkii* on sugar and sugar derivatives, as previously reported elsewhere (Westerholm *et al.*, 2010).

It was not possible to identify genes coding for the TAT translocation pathway or genes harbouring the N–terminal twin-arginine in the *T. acetatoxydans*. However, *S. schinkii* and *C. ultunense* encode TatA and TatC proteins (TatA:SSCH_510014, TatC:SSCH_360036) (TatA:CULT_950006, TatC:CULT_360037), but both do not have a TatB gene, which is similar case to SAOB *Th. phaeum*, model acetogen *M. themoacetica*, and *Bacillus subtilis*, where the Tat machinery for protein transport across the inner membrane is composed of A and C subunits. Thus TatB-like component is apparently not required (Jongbloed *et al.*, 2004).

5.5.5 Motility

Previous studies have reported that *T. acetatoxydans* shows a spinning movement, and *C. ultunense* demonstrates slightly tumbling motility, which was confirmed here by the identification of a complete set of flagellum genes in their genomes (**IV**). In contrast, *S. schinkii* (**III**) does not contain any flagellum-related genes, confirming its observed immobility.

Motility might be an important factor in explaining how SAOB find their methanogenic partner and get into close contact in order to establish an efficient syntrophic cooperation. In addition, it plays a role in the organisms locating their nutrients in their environment.

Table 11. *Genomic comparison on the genomes of three-sequenced SAOB isolates (**I, II, III**) with their synteny-based closely related genomes.*

Organism	Genome size (bp)	G+C (%)	Number of scaffolds	Total genes	tRNA genes	rRNA genes	Protein-coding genes	Pseudo gene (Partial genes)	Genes with function prediction	Genes assigned to COGs	Average CDS length	Protein coding density
**T. acetatoxydans* Re1	2,761,252	37.53	01	2,852	52	06	2,656	05	2,053	2,158	912.22	87.07
***C. ultunense* Esp	6,159,766	40.87	281	6,446	103	07	5,711	149	4,277	5,248	843.25	83.81
****S. schinkii* Sp3	3,196,921	46.59	215	3,688	50	05	3,445	90	2,099	2,586	754.59	75.05
**T. oceani* strain DSM16646	2,280,035	46.82	01	2,453	51	3	2,327	126	-	1,948	866.16	86.84
***A. metalliredigens* QYMF	4,929,566	36.82	01	5,667	106	20	4,939	138	3,357	3,602	776.34	83.08
****Th. phaeum* DSM12270	2,939,057	53.87	01	3,090	50	09	3,079	19	2,427	2,165	821.68	84.77

5.5.6 Substrate utilisation

Tepidanaerobacter acetatoxydans has the ability to ferment sugar derivatives (glucose, fructose, mannose, lactose, cellobiose, and salicin) (Westerholm *et al.*, 2011b). This was confirmed here with the identification of all the enzymes required for the Embden-Meyerhof-Parnas (EMP) pathway, which are organised in three clusters in the genome (**IV**). All the genes needed for the EMP pathway were also found in the *S. schinkii* genome (**V**), but this organism showed ineptness to grow on sugar and sugar derivates. A possible reason for this is the absence of a glucose uptake system in the genome, as confirmed by the non-appearance of PTS system genes in the *S. schinkii* genome (**V**).

Tepidanaerobacter acetatoxydans can grow heterotrophically on puruvate, malate, citrate, 1,2 propandiol, glycerol and dimethylamin (Westerholm *et al.*, 2011b). Likewise three potential malate dehydrogenases (TepRe1_0566, 0498, 1804) and a citratelyase complex (TepRe1_2273-2275) were identified in the genome. *Tepidanaerobacter acetatoxydans* does not contain the classical gluconeolytic enzyme fructose-1, 6 bisphosphatase, but instead has a pyrophosphate-dependent fructose-6-phosphate-1-transferase (TepRe1_0478) in its genome. *Tepidanaerobacter acetatoxydans* has been shown to grow without any supplemental amino acids, which was confirmed here by the presence of all genes necessary for biosynthesis of amino acids expect for arginine (**IV**). It was also supported by the COGs analysis results, which showed a high number of genes falling into the category of amino acid transport and metabolism, explaining very well its slow growth even without yeast extract. Thus this organism might need to invest more proteins for substrate uptake or protein biosynthesis.

Syntrophaceticus schinkii can only grow heterotrophically on a restricted number of substrates (Westerholm *et al.*, 2010), as was supported by the COG analysis (**III),** which revealed a lower number of genes for carbohydrate transport and metabolism (Table 9).

5.5.7 Intermediate metabolites

The important precursor (tetrahydrofolate) of W-L pathway could be synthesised as a result of two pathways: *de novo* synthesis or salvage pathway. Dealing with this, all the genes required for *de novo* synthesis pathway were identified in *Syntrophaceticus schinkii* except for one (folA). Though on the basis of some similarity of the gene (SSCH_1160025) with dihydropteridine reductase [EC 1.5.1.34] from *Thermus thermophiles,* we speculate that the product of this gene can be the alternative of the missing substrate in the

genome. In addition, *S. schinkii* also harbours genes responsible to perform the salvage pathway (SSCH_1250009, SSCH_630016-17).

The two pathways for cobalt insertion are, *i.e.* the "early " and the "late" cobalt insertion pathways and in *S. schinkii* one gene is absent for "early" and two genes are missing for "late" pathways. Similarly the most closely related SAOB *Th. phaeum* also lacks one gene for "early" and two genes for "late" pathways (**V**). It seems that only "early" cobalt insertion pathway is working in *S. schinkii*, because "late" cobalt insertion pathway needs oxygen to be functional. Moreover, due to the absence of heme biosynthesis genes in the *S. schinkii*, apparently the organism is unable to synthesize cytochromes and its closely related thermophilic SAOB *Th. phaeum* also does not contain genes that could be involved in cytochrome biosynthesis.

The genome of *T. acetatoxydans* does not contain the complete set of genes required for syntheses of cobalamin, naphtoate, dihydroopterin, folate, pyridoxal-5-phosphate and phosphopantothenate derivatives, which can be the explanation of the higher requirements of supplements for this organism during laboratory cultivation (**IV**).

5.5.8 Energy conservation

The genomic analysis of sequenced homoacetogenic bacteria has revealed different types of possible homoacetogenic metabolisms for energy conservation: i) A sodium pumping Rnf complex, ii) cytochromes and menaquinone iii) a proton pumping Rnf complex, and iv) hydrogenases and an F_1F_0-type ATPase using the generated proton/sodium gradient, leading to ATP formation. *Syntrophaceticus schinkii* contains genes encoding six subunits of the Rnf complex (rnfCDGEAB: SSCH_420047-420053), which is also present in many clostridial species, including *C. tetani* and *C. kluyveri,* but missing in the genome of *M. thermoacetica,* which produces a proton gradient via cytochromes and quinones. Interestingly the closest relative of *S. schinkii,* a thermophilic SAOB *Th. phaeum* does not harbour an Rnf complex (Table 12). *Syntrophaceticus schinkii* (**V**) does not harbour cytochromes and quinones, but instead it creates a metabolic mechanism for energy conservation by generating a proton gradient via the Rnf system, similar to the homoacetogen *C. ljungdahlii. S. schinkii* also contains F_1F_0-type ATP synthase, which in turn could used the membrane potential produced by Rnf complex to generate ATPs. In contrast, closely related SAOB *Th. phaeum* contains F_1F_0-type ATP synthase and, additionally menaquinone-7, which could be involved in electron transport and proton translocation (Table 12).

Clostridium ultunense also possess an energy conservation system similar to *S. schinkii,* by generating a proton gradient via the Rnf system and for use F_1F_0-type ATP synthase was also found in the genome (Table 12).

In syntrophic environments, the partner methanogenic archaea also conserve energy as part of a membrane-bound electron transport system using hydrogenases, which are distinct from the Ech hydrogenase, along with heterodisulphide reductase. Although *S. schinkii* is not an archaeon, it always works syntrophically together with this methanogenic archaea to oxidise acetate, but it also has two heterodisulphide reductase gene clusters (Table 1 in **V**) that are homologous to the three-subunit heterodisulphide reductase from *Methanothermobacter marburgensis*.

The *T. acetatoxydans* genome is differs in term of energy conservation from the other sequenced SAOB genomes. An electron transport complex Rnf was identified in the *T. acetatoxydans* genome (**IV**). The organization of the genes (*rnfCDGEAB)* is similar to that found for a number of Clostridia (Biegel et al. 2011). This Rnf complex can generate a membrane potential however, for use no F_1F_0 ATP synthase could be identified in the *T. acetatoxydans* genome (Table 12). Instead, two V-ATPases were found in the genome, which in prokaryotic organisms consist of nine different subunits. V-ATPases build up sodium or proton gradients at the expense of ATP (**IV**). The closely related sequenced *T. oceanii* also harbours two V-ATPase operons, but the second one seems completely absent in the other sequenced homoacetogenic genomes.

During heterotrophic growth, *T. acetatoxydans* has the ability to generate energy through substrate level phosphorylation, but for acetate oxidation it is still an open question how energy is obtained by this organism. In this regard, our growth experiment showed that during methane production with a syntrophic partner, this organism remains active without showing any further growth (Figure 3 in **IV**).

A comparison of energy conservation systems in SAOB (*T. acetatoxydans, C. ultunense, S. schinkii, Th. phaeum*) and acetogens (*T. oceani, M. thermoacetica, C. ljungdahlii, A. woodii*) is presented in Table 12.

Table 12. *Comparison of energy conservation systems in SAOB and acetogens genomes.*

	T. acetatoxydans	*C. ultunense*	*S. schinkii*	*T. oceani*	*Th. phaeum*	*M. thermoacetica*	*C. ljungdahlii*	*A. woodii*
(Fe) hydrogenase (hydABCDE)	TepiRe1_2033-2037	CULT_140050, 390014, 390015	SSCH_90017-90019	+	+	+	+	+
(Fe) hydrogenase (hydABC)	TepiRe1_2699-2701)	-	SSCH_210008, 600010-600011, 1120014-1120015)	+	+	+	+	+
Ech hydrogenase (EchABCDEF)	-	-	SSCH_170021 - 170026	-	+	-	-	-
Rnf complex (RnfABCDGE)	TepiRe1_2026-2031	CULT_860012-860017	SSCH_420047-420053	+	-	-	+	+
formate dehydrogenase (alpha, beta)	-	CULT_390009-390010 (fdhA, fdhB)	SSCH_1520003 – 1520002 (fdhA, fdhB)	-	+ (fdhA, fdhB)	+ (fdhA, fdhB)	+ (fdhA)	+ (fdhF1, fdhF2)
hyc formate dehydrogenase	-	CULT_1180002, 1180008, 2160004	-	-	+	+	-	-
F_1F_0 ATP synthase	-	CULT_20134 - 20141 CULT_220003 - 2200010	SSCH_240003 – 2400010	+	+	+	+	+
V-type ATP synthase operon 1	TepiRe1-0557 - 0565	-	-	+	-	-	-	+
V-type ATP synthase operon 2	TepiRe1_2235-2244	-	-	+	-	-	-	-
(NiFe)-hydrogenase	-	CULT_60081 - 60083	SSCH_30031 – 30033	-	+	-	+	-
Sulfhydrogenase hyd(alpha, beta, delta)	-	-	SSCH_370001-370003 (hydABD)	-	+	-	-	-

5.5.9 Acetogenesis

Tepidanaerobacter acetatoxydans is a homoacetogen that has the ability to produce acetate as the only end product through the W-L pathway when it grows heterotrophically (Westerholm *et al.*, 2011b). This pathway has been suggested to be used in reverse for oxidation of acetate in syntrophy with a methanogen (Müller *et al.*, 2013; Schnurer *et al.*, 1996). All the genes involved in the W-L pathway were found in the genome, and were also previously partially identified by Müller *et al.* (2013) with one exception: no formate dehydrogenase was found, confirming the inability of the organism to establish an autotrophic lifestyle when grown on CO_2 and H_2 (Westerholm *et al.*, 2011b) (**IV**). As a consequence of the absence of formate dehydrogenase (Figure 19) *T. acetatoxydans* can oxidise acetate through the W-L pathway in reverse direction only to formate and CO_2. In this regard the isolated methanogenic partner MAB2 has been shown to have the ability to utilise both hydrogen and formate (Schnurer *et al.*, 1999).

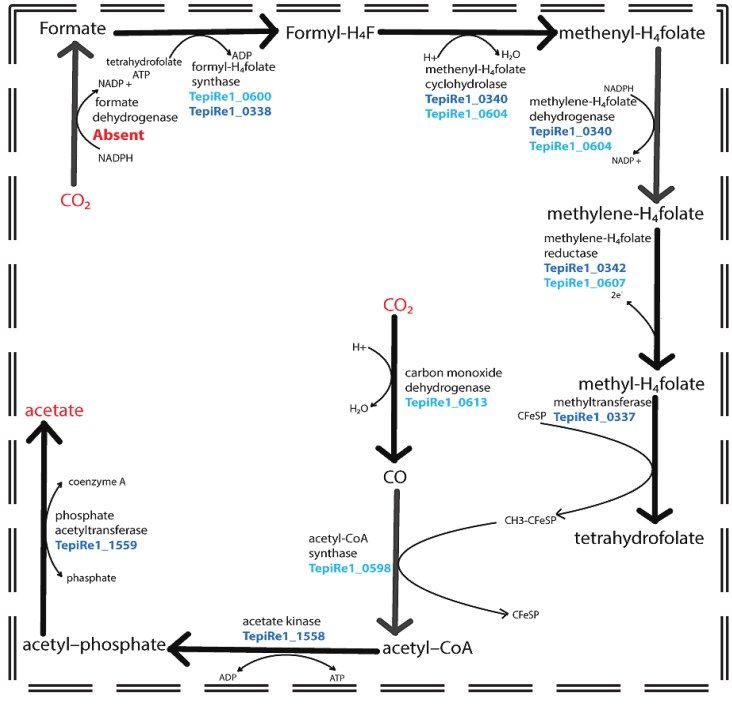

Figure 19. The Wood-Ljungdahl pathway with specific *Tepidanaerobacter acetatoxydans Re1* genes predicted to encode each step in the pathway.

Syntrophaceticus schinkii is also incapable of growing autotrophically on H_2 and CO_2 (Westerholm *et al.*, 2010), although all the enzymes needed for this growth through W-L pathway could be identified in the genome (**V**) (Figure 20). We can thus conclude that this SAOB theoretically has the ability to grow autotrophically, but we did not observe this ability in the laboratory culture. *Syntrophaceticus schinkii* uses the W-L pathway in reverse to oxidise acetate to H_2 and CO_2.

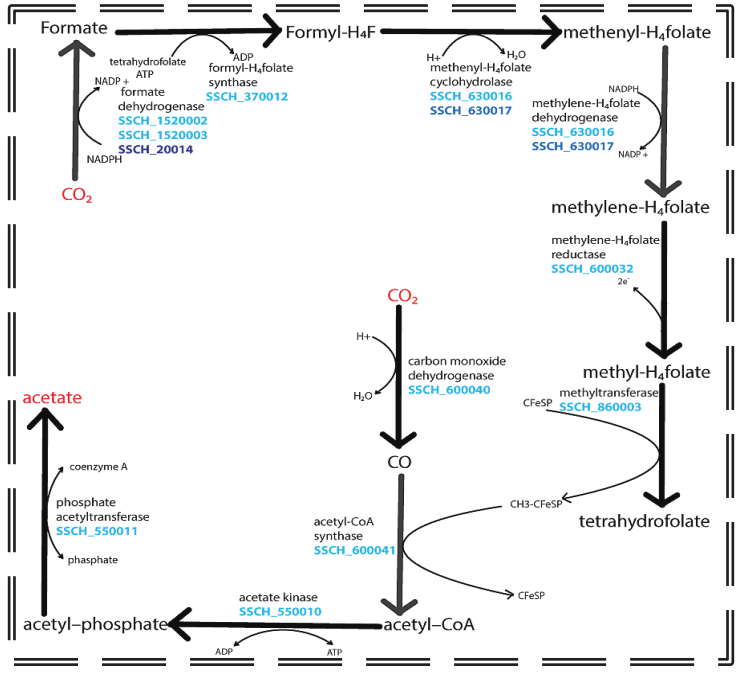

Figure 20. The Wood-Ljungdahl pathway with specific *Syntrophaceticus schinkii Sp3* genes predicted to encode each step in the pathway.

Interspecies H_2 transfer from the fermentative community of bacteria to methanogens is of great significance. Accordingly, hydrogenases were identified in the genomes of SAOB sequenced here (*T. acetatoxydans* and *S. schinkii*). Since all the genes have been encoded except for formate dehydrogenase, that is not a problem for the methanogen archaea MAB2. It was found that the pathway could be used reversely. *T. acetatoxydans* and *S. schinkii* degrade acetate syntrophically and form H_2 and CO_2 (and possibly formate), which can be used further by the methanogenic partner to form methane. The important question here is whether the same enzymes are responsible for the reactions in both directions (acetate formation and acetate oxidation). On the basis of our analysis, it can be speculated that enzymes such

as formate dehydrogenase, CO dehydrogenase and methylene-THF reductase can be more interesting due to their significant role concerning energy conservation in both directions.

Clostridium ultunense revealed an ability to oxidise acetate syntrophically in the presence of a methanogenic partner, but not all the genes of the W-L pathway could be identified in the genome (Figure 21). Thus it is possible that this pathway is not used as was assumed for the other SAOB. It appears that the sequenced SAOB use different ways to oxidise acetate. It has been shown that *Thermotoga lettingae,* a known SAOB, also does not use the W-L pathway. However, comparison to *T. lettingae* on genome scale was not possible because its genome has not yet been sequenced.

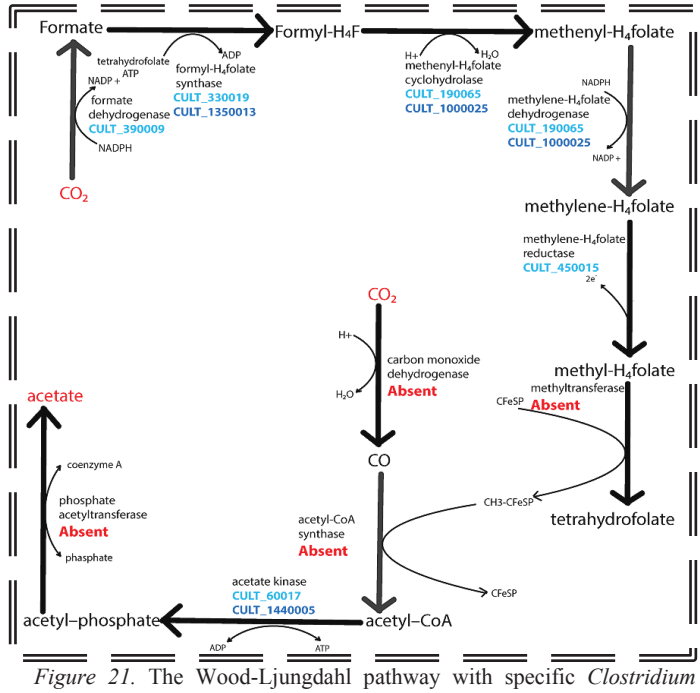

Figure 21. The Wood-Ljungdahl pathway with specific *Clostridium ultunense Esp* genes predicted to encode each step in the pathway.

5.5.10 Wood-Ljungdahl pathway

All the genes involved in the methyl branch of the W-L pathway were found as duplicates in the *T. acetatoxydans* genome, where they are organised in a separate cluster (**IV**), as previously reported by Müller *et al.* (2013). According to preliminary mRNA studies, this cluster may work as an alternative set of genes, which are required for the intermediate C1 carbon metabolism, when the

W-L operon goes to downregulation. Most of the W-L pathway genes were encoded only once in the *S. schinkii* genome (**V**).

Comparison of W-L pathway genes represented in the operon of the SAOB genomes (*S. schinkii, T. acetatoxydans*, *Th. phaeum*) and acetgogens (*M. thermoacetica, A. woodii)* showed clearly that the organisation of the *S. schinkii* operon is strictly identical to that of its most closely related SAOB, *Th. phaeum* (Figure 22). Moreover, more genes of the W-L pathway are organized in the three SAOB genomes, whereas in acetogens these genes are dispersed over the genome.

The formation of the W-L pathway operon showed clearly that the genomes of the three-compared SAOB are more similar in their organisation of the W-L pathway genes as a single operon, in contrast to acetogens. This more compact organisation of genes in SAOB might play an important role in syntrophic acetate oxidation by using the W-L pathway in reverse, which has not been observed for acetogens.

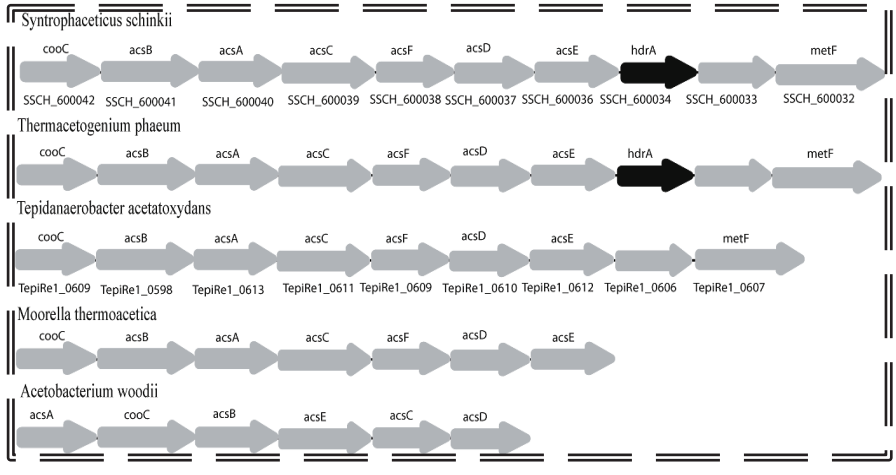

Figure 22. Comparison between the W-L pathway gene clusters from SAOB (*S. schinkii, T. acetatoxydans, Th. phaeum)* and those from the acetogens (*M. thermoacetica, A. woodii)*. cooC, CODH accessory protein; acsA, CODH/ACS complex, CODH subunit; acsB, CODH/ACS complex, ACS subunit; acsC, corrinoid iron-sulphur protein large subunit; acsF, CODH accessory protein similar to CooC; acsD, corrinoid iron-sulfur protein small subunit; acsE, CODH/ACS complex, methyltransferase subunit; hdrA, heterodisulphide reductase; metF, methylene-tetrahydrofolate reductase.

"*There is no such thing as absolute truth. That is absolutely true*"
(Solomon Short)

6 Conclusions

Three genomes of SAOB isolates were sequenced in this thesis using Ion Torrent technology. One complete genome as a single circular chromosome was produced using the combined approach of *de novo* and mapping assembly, while two working draft genomes were generated using the *de novo* assembly approach. All three genomes of SAOB isolates differ in size, but contain genes for all essential functions that are required for a free-living life style and also acquired some specific genes related to their habitat adaptation, *e.g.* genomic analysis of the three sequenced SAOB revealed their habitat adaptation of acquiring more CRISPR loci to strengthen their defence mechanism against possible attack by mobile genetic elements in their specific microbe-rich environment.

The presence of all the enzymes required for the Embden-Meyerhof-Parnas (EMP) pathway in the *T. acetatoxydans* genome confirmed its ability to grow with sugar and sugar derivatives. In contrast, all the genes needed for EMP pathway were also found in the *S. schinkii* genome but the absence of PTS system in the organism justify its inability to grow with sugar and sugar derivatives.

Genome scale analysis indicated that the three sequenced SAOB use contrasting strategies for SAO: *T. acetatoxydans* possesses all genes involved in the W-L pathway with one exception (formate dehydrogenase) and therefore requires a syntrophic formate-utilising methanogenic partner; *S. schinkii* possesses the complete set of genes required for the W-L pathway to oxidise acetate in the presence of a hydrogen-utilising methanogenic partner; and *C. ultunense* uses different ways to oxidise acetate, because it does not contains the complete set of genes responsible for the W-L pathway. Moreover, the three SAOB are different from each other as regards the organisation of the W-L pathway genes operon.

Genomic scale comparison revealed that *S. schinkii* and *C. ultunense* possess a similar energy conservation system by generating a proton gradient via the Rnf system and for use both also possess F_1F_0-type ATP synthase. In contrast *T. acetatoxydans* surprisingly does not contain F_1F_0-type ATP synthase and instead possess two V-ATPase operons, which build up sodium or proton gradients at the expense of ATP.

7 Future perspectives

The results presented in this thesis provide a genomic picture of syntrophic acetate-oxidising bacteria. The genomic behaviour and features underlying syntrophic growth, substrate utilisation, energy conservation, defence and habitat adaptation described in this thesis can be used as milestones for further development to understand the syntrophic lifestyle more deeply. They also provide a platform within which to investigate the use of sequenced SAOB in biogas reactors for bio-fuel (methane) production.

The successful use of microorganisms in biogas reactors requires a thorough understanding of the processes and modes of action underlying their ability to directly and indirectly perform their role in syntrophic biomass degradation. The use of RNA sequencing (RNA-seq) analysis for identifying these essential building blocks of biogas-producing bioreactors will provide a valuable snapshot of RNA presence and quantity in a genome at a given moment in time. However, because the transcriptome of a cell is dynamic, it continually changes, as opposed to a static genome.

Moreover, sequenced SAOB isolates grow in different environments and in different ecological settings for example, how their metabolism and gene expression patterns change when as a acetogen play a role in a microbe-community or as a syntrophical role with their methanogenic partner that can act as a H_2/formate sink. In turn, proteomic and microarray studies should help us to improve the annotation of the genomes, by allowing us to identify specific biochemical functions and metabolic role of genes of unknown or partially known function.

In addition, to expose the co-related genes that might be involved in acetate-oxidation metabolism by using W-L pathway in reverse direction, co-relation network analysis will be crucial. This analysis will allow us to identify

the co-regulated genes of W-L pathway to understand their functional roles for syntrophic acetate oxidation.

The results presented in this thesis provide information about the possible threat to biogas reactors arising from the presence of viral genomes, which can become active and disturb the biomass degradation process under specific conditions. A comprehensive study is required to investigate the possible role of these '*killers of the winners'*.

"*I am all in favour of keeping dangerous weapons out of the hand of fools. Let's start with typewriters.*"

(Solomon Short)

8 Acknowledgements

I must start with the name of the most Merciful, Omnipotent and Omnipresent ALLAH who has made me energetic enough to accomplish this task with concentration and devotion and finally blessed me with success. I have no more words to thank Him enough. All of my devotions and tributes to His HOLY PROPHET MUHAMMAD (Peace be Upon Him) Who's teaching enabled us to recognize our creator.

The studies in this thesis were carried out at the Department of Animal Breeding and Genetics, Swedish University of Agricultural Science, Uppsala, Sweden. The SLU fund for internationalisation of postgraduate studies provided travel grants for my visit at Saclay, France and Omaha, USA.

I would like to express my sincere gratitude to all of you at the HGEN who have helped and support me during my time as a PhD student. There are some of you that I like to give special credit:

Erik Bongcam-Rudloff, main supervisor: Thanks for believing in me and for encouragement and most of all for your great support. Thanks for your sincere care for all the time from day first to the end of my stay here in Uppsala. I am grateful for the trust and freedom you placed in me.

Bettina Müllar, co-supervisor: You are friendly and hardworking person and thanks for always talking time to discuss my project and work with me in the lab. I really enjoyed traveling with you to Austria and discussing a lot of things other than work as well. You are my dear friend and I really appreciate your honesty and professionalism.

Johan Meijer, co-supervisor: Thank you for always being kind and supportive, and patiently tolerating and answering my questions. I will never forget a mushroom picking trip with you and after that your lovely instructions for cooking and preserving those mushrooms.

Anna Schnürer: It was great to have such an experienced scientist as you involved in Biogas project. Thanks for your support and discussion for my project and your dedication and great knowledge about SAOB, I always considered you as one of my supervisor.

I would also like to say many thanks to all my family members, friends and colleagues in the Punjab University, Lahore, Pakistan for their prayers for me and support in all matters of life. There are three persons that I would like to express my special gratitude:

M. Ehsan Malik, one of the great professors in the Punjab University and for me just likes a mentor for my academic professional life. I really appreciate your sincere support for all professional matters during my study leave absence from the University.

Harris Mubeen & Naseer bin Zaheer, My dear brothers and always my trusty advisors for all good and hard times in my life. I really know that both of you are the persons who made all this possible for me to come over here Uppsala for such a long time for my PhD studies, because without your unconditional support and sincere care for my parents in Pakistan how it can be possible to happen. I have no more words to say thanks to both of you. Thanks to naseer bhai for letting me drive your car whenever I visited Pakistan ☺.

I really say so many thank all of my Pakistani friends and theirs families for providing us a very friendly social environment to survive here in Uppsala for such a long time. I really enjoyed a lot and also get energy for work again after discussing with all of you about politics, games and all others matters on weekend's gatherings.

I would like to acknowledge all the people who have worked with me in the bioinformatics group.

I would like to thank the **Higher Education Commission** (HEC) of Pakistan and **University of the Punjab**, Lahore, Pakistan for their support of my PhD studies.

I have no words to express my heartiest gratitude to my parents and two sisters for their great love, continuous supports, encouragement and uncountable prays for throughout of my life.

I would also like to thank to my beloved wife, Mona, from the core of my heart for her efforts that she made for my well being, for her patience, allowing me to concentrate on my work and for her unfathomable love and off course my daughters and son for their innocent smiles and especially evening fights with Ibrahim which always relax me.

I guess its done now ☺

References

Achaz, G., Boyer, F., Rocha, E.P.C., Viari, A. & Coissac, E. (2007). Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics,* 23(1), pp. 119-121.

Ackermann, H.W. (2001). Frequency of morphological phage descriptions in the year 2000. *Archives of Virology,* 146(5), pp. 843-857.

Administration, E.I. (2007). Energy Information Administration (2007) International Energy Outlook 2007 (US Government Printing Office, Washington, DC). *US Government Printing Office.*

Ahring, B.K. (2003). Perspectives for anaerobic digestion. *Adv Biochem Eng Biotechnol,* 81, pp. 1-30.

Ahring, B.K., Schmidt, J.E., Winthernielsen, M., Macario, A.J.L. & Demacario, E.C. (1993). Effect of Medium Composition and Sludge Removal on the Production, Composition, and Architecture of Thermophilic (55-Degrees-C) Acetate-Utilizing Granules from an Upflow Anaerobic Sludge Blanket Reactor. *Applied and Environmental Microbiology,* 59(8), pp. 2538-2545.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology,* 215(3), pp. 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25(17), pp. 3389-3402.

Anderson, R.P. & Roth, J.R. (1977). Tandem Genetic Duplications in Phage and Bacteria. *Annual Review of Microbiology,* 31, pp. 473-505.

Andlinger, P. *RDBMS dominate the database market, but NoSQL systems are catching up*. URL: http://db-engines.com/en/blog_post/23.

Angelidaki, I., Karakashev, D., Batstone, D.J., Plugge, C.M. & Stams, A.J.M. (2011). Biomethanation and Its Potential. *Methods in Enzymology: Methods in Methane Metabolism, Pt A,* 494, pp. 327-351.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. & Consortium, G.O.

(2000). Gene Ontology: tool for the unification of biology. *Nature genetics,* 25(1), pp. 25-29.

Atkinson, M., Bancilhon, F., DeWitt, D., Dittrich, K., Maier, D. & Zdonik, S. (1993). *The Object-Oriented Database System Manifesto*.

Attwood, T.K., Beck, M.E., Bleasby, A.J. & Parrysmith, D.J. (1994). Prints - a Database of Protein Motif Fingerprints. *Nucleic Acids Research,* 22(17), pp. 3590-3596.

Attwood, T.K., Gisel, A., Eriksson, N.E. & Bongcam-Rudloff, E. (2011). Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective. In: A. Mahdavi, M. (ed. *Bioinformatics - Trends and Methodologies* InTech[2014/05/12/09:28:19].

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. (2008). The RAST server: Rapid annotations using subsystems technology. *BMC genomics,* 9.

Balding, D.J., Bishop, M.J. & Cannings, C. (2007). *Handbook of statistical genetics*. 3rd edition. ed. Chichester, England: John Wiley & Sons.

Balk, M., Weijma, J. & Stams, A.J.M. (2002). Thermotoga lettingae sp nov., a novel thermophilic, methanol-degrading bacterium isolated from a thermophilic anaerobic reactor. *International Journal of Systematic and Evolutionary Microbiology,* 52, pp. 1361-1368.

Barker, H.A. (1936). On the biochemistry of the methane fermentation. *Archiv für Mikrobiologie,* 7(1-5), pp. 404-419.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science,* 315(5819), pp. 1709-12.

Bary, A.d. (1954). *Die Erscheinung der Symbiose Vortag, gehalten auf der Versammlung Deutscher Naturforscher und Aerzte zu Cassel*. Strassburg: K.J. Tru bner.

Bauer, W., Bauer, S. & Bauer, T. (2012). Energy and Greenhouse Gas Analysis for Biogas Power Plants. *International Conference on Renewable Energ ies and Power Quality (ICREPQ'12)* Santiago de Compostela (Spain): European Association for the Development of Renewable Energies, Environment and Power Quality (EA4EPQ).

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2013). GenBank. *Nucleic Acids Research,* 41(D1), pp. D36-D42.

Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2014). GenBank. *Nucleic Acids Research,* 42(D1), pp. D32-D37.

Biebl, H. & Pfennig, N. (1978). Growth Yields of Green Sulfur Bacteria in Mixed Cultures with Sulfur and Sulfate Reducing Bacteria. *Archives of Microbiology,* 117(1), pp. 9-16.

BioMart Central Portal.

Bock, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B. & Zinoni, F. (1991). Selenocysteine - the 21st Amino-Acid. *Molecular Microbiology,* 5(3), pp. 515-520.

Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G. & Medigue, C. (2003). AMIGene: Annotation of MIcrobial genes. *Nucleic Acids Research,* 31(13), pp. 3723-3726.

Borjesson, P. & Berglund, M. (2007). Environmental systems analysis of biogas systems - Part II: The environmental impact of replacing various reference systems. *Biomass & Bioenergy,* 31(5), pp. 326-344.

Borjesson, P. & Mattiasson, B. (2008). Biogas as a resource-efficient vehicle fuel. *Trends Biotechnol,* 26(1), pp. 7-13.

Bru, C., Courcelle, E., Carrre, S., Beausse, Y., Dalmar, S. & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Research,* 33, pp. D212-D215.

Bryant, M.P., Wolin, E.A., Wolin, M.J. & Wolfe, R.S. (1967). Methanobacillus omelianskii, a symbiotic association of two species of bacteria. *Arch Mikrobiol,* 59(1), pp. 20-31.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J. & McQuillan, J.A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics,* 28(4), pp. 464-469.

Casjens, S.R. (2005). Comparative genomics and evolution of the tailed-bacteriophages. *Current Opinion in Microbiology,* 8(4), pp. 451-458.

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Pujar, A., Shearer, A.G., Travers, M., Weerasinghe, D., Zhang, P. & Karp, P.D. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research,* 40(Database issue), pp. D742-53.

Chauhan, A. & Ogram, A. (2006). Phylogeny of acetate-utilizing microorganisms in soils along a nutrient gradient in the Florida everglades. *Applied and Environmental Microbiology,* 72(10), pp. 6837-6840.

Cochrane, G., Alako, B., Amid, C., Bower, L., Cerdeno-Tarraga, A., Cleland, I., Gibson, R., Goodgame, N., Jang, M., Kay, S., Leinonen, R., Lin, X., Lopez, R., McWilliam, H., Oisel, A., Pakseresht, N., Pallreddy, S., Park, Y., Plaister, S., Radhakrishnan, R., Riviere, S., Rossello, M., Senf, A., Silvester, N., Smirnov, D., ten Hoopen, P., Toribio, A., Vaughan, D. & Zalunin, V. (2013). Facing growth in the European Nucleotide Archive. *Nucleic Acids Research,* 41(D1), pp. D30-D35.

Codd, E.F. (October 14, 1985). Is Your DBMS Really Relational? ComputerWorld.

Codd, E.F. (October 21, 1985). Does Your DBMS Run By the Rules? ComputerWorld.

Collins, F.S., Lander, E.S., Rogers, J., Waterston, R.H. & Conso, I.H.G.S. (2004). Finishing the euchromatic sequence of the human genome. *Nature,* 431(7011), pp. 931-945.

Consortium, C.e.S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science,* 282(5396), pp. 2012-8.
Consortium, I.H.G.S. (2001). Initial sequencing and analysis of the human genome. *Nature,* 409(6822), pp. 860-921.
Consortium, U. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research,* 38, pp. D142-D148.
Cord-Ruwisch, R., Lovley, D.R. & Schink, B. (1998). Growth of Geobacter sulfurreducens with acetate in syntrophic cooperation with hydrogen-oxidizing anaerobic partners. *Applied and Environmental Microbiology,* 64(6), pp. 2232-2236.
Darling, A.C.E., Mau, B., Blattner, F.R. & Perna, N.T. (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research,* 14(7), pp. 1394-1403.
Dayhoff, M.O. (1978). Protein segment dictionary 78: from the Atlas of protein sequence and structure. In: Foundation, N.B.R. (ed. Silver Spring, Md.,.
de Souza, W.R. (2013). Microbial Degradation of Lignocellulosic Biomass. In: Chandel, A. (ed. *Sustainable Degradation of Lignocellulosic Biomass - Techniques, Applications and Commercialization* InTech[2014/04/17/10:12:57].
Demirel, B. & Scherer, P. (2008). The roles of acetotrophic and hydrogenotrophic methanogens during anaerobic conversion of biomass to methane: a review. *Reviews in Environmental Science and Bio/Technology,* 7(2), pp. 173-190.
Derman, C. (1961). Adaptive-Control Processes, a Guided Tour - Bellman,R. *Management Science,* 7(4), pp. 450-450.
Diaz, E. (2008). *Microbial Biodegradation: Genomics and Molecular Biology*. 1 edition. ed. Norfolk, U.K: Caister Academic Press.
Do, C.B., Mahabhashyam, M.S.P., Brudno, M. & Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research,* 15(2), pp. 330-340.
Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research,* 36(16).
Drake, H. (1994). Acetogenesis, Acetogenic Bacteria, and the Acetyl-CoA "Wood/Ljungdahl" Pathway: Past and Current Perspectives. In: *Acetogenesis* Springer US, pp. 3-60.
Edgar, R.C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5, pp. 1-19.
Edgar, R.C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32(5), pp. 1792-1797.
Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X.X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C.C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A.,

Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science,* 323(5910), pp. 133-138.

El-Metwally, S., Hamza, T., Zakaria, M. & Helmy, M. (2013). Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. *PLoS Comput Biol,* 9(12).

Elsmari, R. & Navathe, S.B. (2000). *Fundamentals of Database Systems*: Addison-Wesley.

Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research,* 8(3), pp. 186-194.

Feng, D.F. & Doolittle, R.F. (1987). Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees. *Journal of Molecular Evolution,* 25(4), pp. 351-360.

Fickett, J.W. (1996). Finding genes by computer: The state of the art. *Trends in Genetics,* 12(8), pp. 316-320.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J. & Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research,* 42(D1), pp. D222-D230.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. & Al, E. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* 269(5223), pp. 496-512.

Gaasterland, T. & Sensen, C.W. (1996). Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie,* 78(5), pp. 302-310.

Gallert, C. & Winter, J. (2008). Propionic acid accumulation and degradation during restart of a full-scale anaerobic biowaste digester. *Bioresource Technology,* 99(1), pp. 170-178.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S. & Jaffe, D.B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A,* 108(4), pp. 1513-1518.

Gray, N.D., Sherry, A., Grant, R.J., Rowan, A.K., Hubert, C.R.J., Callbeck, C.M., Aitken, C.M., Jones, D.M., Adams, J.J., Larter, S.R. & Head, I.M. (2011). The quantitative significance of Syntrophaceae and syntrophic partnerships in methanogenic degradation of crude oil alkanes. *Environmental microbiology,* 13(11), pp. 2957-2975.

Gribskov, M., Devereux, J. & Burgess, R.R. (1984). The Codon Preference Plot - Graphic Analysis of Protein Coding Sequences and Prediction of Gene-Expression. *Nucleic Acids Research,* 12(1), pp. 539-549.

Guigo, R., Burset, M., Agarwal, P., Abril, J.F., Smith, R.F. & Fickett, J.W. (2000). Sequence similarity based gene prediction. *Genomics and Proteomics*, pp. 95-105.

Haft, D.H., Selengut, J., Mongodin, E.F. & Nelson, K.E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol,* 1(6), pp. 474-483.

Haider, S., Ballester, B., Smedley, D., Zhang, J.J., Rice, P. & Kasprzyk, A. (2009). BioMart Central Portal-unified access to biological data. *Nucleic Acids Research,* 37, pp. W23-W27.

Hao, L.P., Lu, F., He, P.J., Li, L. & Shao, L.M. (2011). Predominant contribution of syntrophic acetate oxidation to thermophilic methane formation at high acetate concentrations. *Environmental Science & Technology,* 45(2), pp. 508-13.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X.Y., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S. & Frazer, K.A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology,* 10(3).

Hattori, S. (2008). Syntrophic acetate-oxidizing microbes in methanogenic environments. *Microbes and Environments,* 23(2), pp. 118-127.

Hattori, S., Kamagata, Y., Hanada, S. & Shoun, H. (2000). Thermacetogenium phaeum gen. nov., sp nov., a strictly anaerobic, thermophilic, syntrophic acetate-oxidizing bacterium. *International Journal of Systematic and Evolutionary Microbiology,* 50, pp. 1601-1609.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K. & Hood, L. (1997). Gene families: The taxonomy of protein paralogs and chimeras. *Science,* 278(5338), pp. 609-614.

Henikoff, S. & Henikoff, J.G. (1992). Amino-Acid Substitution Matrices from Protein Blocks. *Proc Natl Acad Sci U S A,* 89(22), pp. 10915-10919.

Hirosawa, M., Totoki, Y., Hoshida, M. & Ishikawa, M. (1995). Comprehensive Study on Iterative Algorithms of Multiple Sequence Alignment. *Computer Applications in the Biosciences,* 11(1), pp. 13-18.

Hobson, P.N. & Shaw, B.G. (1974). Bacterial Population of Piggery-Waste Anaerobic Digesters. *Water Research,* 8(8), pp. 507-516.

Hogeweg, P. & Hesper, B. (1978). Interactive Instruction on Population Interactions. *Computers in Biology and Medicine,* 8(4), pp. 319-327.

Hori, T., Noll, M., Igarashi, Y., Friedrich, M.W. & Conrad, R. (2007). Identification of acetate-assimilating microorganisms under methanogenic conditions in anoxic rice field soil by comparative stable isotope probing of RNA. *Appl Environ Microbiol,* 73(1), pp. 101-9.

Horvath, P. & Barrangou, R. (2010). CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science,* 327(5962), pp. 167-170.

Hu, G.Q., Zheng, X.B., Zhu, H.Q. & She, Z.S. (2009). Prediction of translation initiation site for microbial genomes with TriTISA. *Bioinformatics,* 25(1), pp. 123-125.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L. & Mark Welch, D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology,* 8(7).

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics,* 11.

International Human Genome Sequencing, C. (2004). Finishing the euchromatic sequence of the human genome. *Nature,* 431(7011), pp. 931-45.

Jacob, F. (1977). Evolution and Tinkering. *Science,* 196(4295), pp. 1161-1166.

Jones, D.M., Head, I.M., Gray, N.D., Adams, J.J., Rowan, A.K., Aitken, C.M., Bennett, B., Huang, H., Brown, A., Bowler, B.F.J., Oldenburg, T., Erdmann, M. & Larter, S.R. (2008). Crude-oil biodegradation via methanogenesis in subsurface petroleum reservoirs. *Nature,* 451(7175), pp. 176-U6.

Jongbloed, J.D.H., Grieger, U., Antelmann, H., Hecker, M., Nijland, R., Bron, S. & van Dijl, J.M. (2004). Two minimal Tat translocases in Bacillus. *Molecular Microbiology,* 54(5), pp. 1319-1325.

Joshi, N. & Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.21).

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* 28(1), pp. 27-30.

Karakashev, D., Batstone, D.J. & Angelidaki, I. (2005). Influence of environmental conditions on methanogenic compositions in anaerobic biogas reactors. *Applied and Environmental Microbiology,* 71(1), pp. 331-338.

Karakashev, D., Batstone, D.J., Trably, E. & Angelidaki, I. (2006). Acetate oxidation is the dominant methanogenic pathway from acetate in the absence of Methanosaetaceae. *Applied and Environmental Microbiology,* 72(7), pp. 5138-5141.

Karnholz, A., Kusel, K., Gossner, A., Schramm, A. & Drake, H.L. (2002). Tolerance and metabolic response of acetogenic bacteria toward oxygen. *Applied and Environmental Microbiology,* 68(2), pp. 1005-1009.

Kaspar, H.F. & Wuhrmann, K. (1978). Kinetic-Parameters and Relative Turnovers of Some Important Catabolic Reactions in Digesting Sludge. *Applied and Environmental Microbiology,* 36(1), pp. 1-7.

Katoh, K., Kuma, K., Toh, H. & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research,* 33(2), pp. 511-518.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research,* 30(14), pp. 3059-66.

Keane, J., Nicoll, J., Kim, S., Wu, D.M.H., Cruikshank, W.W., Brazer, W., Natke, B., Zhang, Y.J., Center, D.M. & Kornfeld, H. (1998). Conservation of structure and function between human and murine IL-16. *Journal of Immunology,* 160(12), pp. 5945-5954.

Kelley, D.R., Schatz, M.C. & Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biology,* 11(11).

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D. & Phillippy, A.M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology,* 30(7), pp. 692-+.

Korlach, J., Bjornson, K.P., Chaudhuri, B.P., Cicero, R.L., Flusberg, B.A., Gray, J.J., Holden, D., Saxena, R., Wegener, J. & Turner, S.W. (2010). Real-Time DNA Sequencing from Single Polymerase Molecules. *Methods in Enzymology, Vol 472: Single Molecule Tools, Pt A: Fluorescence Based Approaches,* 472, pp. 431-455.

Koukkou, A.-I. (2011). *Microbial Bioremediation of Non-metals: Current Research*: Horizon Scientific Press.

Kriventseva, E.V., Fleischmann, W., Zdobnov, E.M. & Apweiler, R. (2001). CluSTr: a database of clusters of SWISS-PROT plus TrEMBL proteins. *Nucleic Acids Research,* 29(1), pp. 33-36.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology,* 305(3), pp. 567-580.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafe, G., Perez, A. & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics,* 7(1), pp. 86-112.

Lassmann, T. & Sonnhammer, E.L.L. (2005). Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics,* 6.

Laukenmann, S., Polag, D., Heuwinkel, H., Greule, M., Gronauer, A., Lelieveld, J. & Keppler, F. (2010). Identification of methanogenic pathways in anaerobic digesters using stable carbon isotopes. *Engineering in Life Sciences,* 10(6), pp. 509-514.

Lee, C., Grasso, C. & Sharlow, M.F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics,* 18(3), pp. 452-464.

Lee, M.J. & Zinder, S.H. (1988). Isolation and Characterization of a Thermophilic Bacterium Which Oxidizes Acetate in Syntrophic Association with a Methanogen and Which Grows Acetogenically on H2-CO2. *Applied and Environmental Microbiology,* 54(1), pp. 124-129.

Li, R.Q., Fan, W., Tian, G., Zhu, H.M., He, L., Cai, J., Huang, Q.F., Cai, Q.L., Li, B., Bai, Y.Q., Zhang, Z.H., Zhang, Y.P., Wang, W., Li, J., Wei, F.W., Li, H., Jian, M., Li, J.W., Zhang, Z.L., Nielsen, R., Li, D.W., Gu, W.J., Yang, Z.T., Xuan, Z.L., Ryder, O.A., Leung, F.C.C., Zhou, Y., Cao, J.J., Sun, X., Fu, Y.G., Fang, X.D., Guo, X.S., Wang, B., Hou, R., Shen, F.J., Mu, B., Ni, P.X., Lin, R.M., Qian, W.B., Wang, G.D., Yu, C., Nie, W.H., Wang, J.H., Wu, Z.G., Liang, H.Q., Min, J.M., Wu, Q., Cheng, S.F., Ruan, J., Wang, M.W., Shi, Z.B., Wen, M., Liu, B.H., Ren, X.L., Zheng, H.S., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X.Y., Lu, Z.H., Zheng, H.C., Li, Y.R., Steiner, C.C., Lam, T.T.Y., Lin, S.Y., Zhang, Q.H., Li, G.Q., Tian, J., Gong, T.M., Liu, H.D., Zhang, D.J., Fang, L., Ye, C., Zhang, J.B., Hu, W.B., Xu, A.L., Ren, Y.Y., Zhang, G.J., Bruford, M.W., Li, Q.B., Ma, L.J., Guo, Y.R., An, N., Hu, Y.J., Zheng, Y., Shi, Y.Y., Li, Z.Q., Liu, Q., Chen, Y.L., Zhao, J., Qu, N., Zhao, S.C., Tian, F., Wang, X.L., Wang, H.Y., Xu, L.Z., Liu, X., Vinar, T., Wang, Y.J., Lam, T.W., Yiu, S.M., Liu, S.P., Zhang, H.M., Li, D.S., Huang, Y., Wang, X., Yang, G.H., Jiang, Z., Wang, J.Y., Qin, N., Li, L., Li, J.X., Bolund, L.,

Kristiansen, K., Wong, G.K.S., Olson, M., Zhang, X.Q., Li, S.G., Yang, H.M., Wang, J. & Wang, J. (2010). The sequence and de novo assembly of the giant panda genome. *Nature,* 463(7279), pp. 311-317.

Liu, F.H. & Conrad, R. (2010). Thermoanaerobacteriaceae oxidize acetate in methanogenic rice field soil at 50 degrees C. *Environmental microbiology,* 12(8), pp. 2341-2354.

Loewenstein, Y., Raimondo, D., Redfern, O.C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J. & Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome Biology,* 10(2).

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J. & Pallen, M.J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology,* 30(5), pp. 434-+.

Lowe, T.M. & Eddy, S.R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research,* 25(5), pp. 955-964.

Lukashin, A.V. & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research,* 26(4), pp. 1107-1115.

Mackie, R.I., White, B.A. & Bryant, M.P. (1991). Lipid-Metabolism in Anaerobic Ecosystems. *Critical Reviews in Microbiology,* 17(6), pp. 449-479.

Maizel, J.V. & Lenk, R.P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc Natl Acad Sci U S A,* 78(12), pp. 7665-9.

Martin, J.A. & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics,* 12(10), pp. 671-682.

McInerney, M.J., Bryant, M.P. & Pfennig, N. (1979). Anaerobic bacterium that degrades fatty acids in syntrophic association with methanogens. *Archives of Microbiology,* 122(2), pp. 129-135.

Metzker, M.L. (2010). Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation. *Nature Reviews Genetics,* 11(1), pp. 31-46.

Miller, J.M., Malenfant, R.M., Moore, S.S. & Coltman, D.W. (2012). Short Reads, Circular Genome: Skimming SOLiD Sequence to Construct the Bighorn Sheep Mitochondrial Genome. *Journal of Heredity,* 103(1), pp. 140-146.

Miller, J.R., Koren, S. & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics,* 95(6), pp. 315-327.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. & Marshall, D. (2010). Tablet-next generation sequence assembly visualization. *Bioinformatics,* 26(3), pp. 401-402.

Moestedt, J., Nordell, E. & Schnurer, A. (2014). Comparison of operating strategies for increased biogas production from thin stillage. *Journal of Biotechnology,* 175, pp. 22-30.

Morgenstern, B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics,* 15(3), pp. 211-8.

Mulder, N. & Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol,* 396, pp. 59-70.

Müller, B., Sun, L. & Schnürer, A. (2013). First insights into the syntrophic acetate-oxidizing bacteria – a genetic study. *Microbiologyopen,* 2(1), pp. 35-53.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. & Erlich, H. (1986). Specific Enzymatic Amplification of DNA Invitro - the Polymerase Chain-Reaction. *Cold Spring Harbor Symposia on Quantitative Biology,* 51, pp. 263-273.

Nagarajan, N. & Pop, M. (2013). Sequence assembly demystified. *Nat Rev Genet,* 14(3), pp. 157-67.

Nazari, A., Karami, M., Safdari, R. & Ashrafi, M.Y. (2013). Optimizing Disease Management with Data Warehousing. *Life Science Journal,* LifeSciJ2013;10(10s), pp. 356-359.

Needlema.Sb & Wunsch, C.D. (1970). A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins. *Journal of Molecular Biology,* 48(3), pp. 443-&.

Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P. & Barron, A.E. (2011). Landscape of Next-Generation Sequencing Technologies. *Analytical Chemistry,* 83(12), pp. 4327-4341.

Nilsen, R.K., Torsvik, T. & Lien, T. (1996). Desulfotomaculum thermocisternum sp nov, a sulfate reducer isolated from a hot North Sea oil reservoir. *International Journal of Systematic Bacteriology,* 46(2), pp. 397-402.

Nimrod, G., Schushan, M., Steinberg, D.M. & Ben-Tal, N. (2008). Detection of Functionally Important Regions in "Hypothetical Proteins" of Known Structure. *Structure,* 16(12), pp. 1755-1763.

Notredame, C., Higgins, D.G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology,* 302(1), pp. 205-217.

Nuin, P.A.S., Wang, Z.Z. & Elisabeth, R.M. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics,* 7.

Nusslein, B., Chin, K.J., Eckert, W. & Conrad, R. (2001). Evidence for anaerobic syntrophic acetate oxidation during methane production in the profundal sediment of subtropical Lake Kinneret (Israel). *Environmental microbiology,* 3(7), pp. 460-470.

O'Brien, E.A., Zhang, Y., Wang, E., Marie, V., Badejoko, W., Lang, B.F. & Burger, G. (2009). GOBASE: an organelle genome database. *Nucleic Acids Research,* 37(suppl 1), pp. D946-D950.

Oehler, D., Poehlein, A., Leimbach, A., Muller, N., Daniel, R., Gottschalk, G. & Schink, B. (2012). Genome-guided analysis of physiological and morphological traits of the fermentative acetate oxidizer Thermacetogenium phaeum. *BMC genomics,* 13, p. 723.

Ogasawara, O., Mashima, J., Kodama, Y., Kaminuma, E., Nakamura, Y., Okubo, K. & Takagi, T. (2013). DDBJ new system and service refactoring. *Nucleic Acids Research,* 41(D1), pp. D25-D29.

Ohmiya, K., Sakka, K. & Kimura, T. (2005). Anaerobic bacterial degradation for the effective utilization of biomass. *Biotechnology and Bioprocess Engineering,* 10(6), pp. 482-493.

Oracle *Oracle Life Sciences Data Hub*. URL: http://www.oracle.com/us/products/applications/health-sciences/e-clinical/data-hub/index.html.

Ouzounis, C.A. & Valencia, A. (2003). Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics,* 19(17), pp. 2176-2190.

Paredes-Sabja, D., Shen, A. & Sorg, J.A. (2014). Clostridium difficile spore biology: sporulation, germination, and spore structural proteins. *Trends Microbiol,* 22(7), pp. 406-416.

Petersen, S.P. & Ahring, B.K. (1991). Acetate Oxidation in a Thermophilic Anaerobic Sewage-Sludge Digester - the Importance of Non-Aceticlastic Methanogenesis from Acetate. *Fems Microbiology Ecology,* 86(2), pp. 149-157.

Petrosino, J.F., Highlander, S., Luna, R.A., Gibbs, R.A. & Versalovic, J. (2009). Metagenomic pyrosequencing and microbial identification. *Clin Chem,* 55(5), pp. 856-66.

Pevzner, P.A., Tang, H.X. & Waterman, M.S. (2001). An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A,* 98(17), pp. 9748-9753.

Pitluck, S., Yasawong, M., Munk, C., Nolan, M., Lapidus, A., Lucas, S., Del Rio, T.G., Tice, H., Cheng, J.F., Bruce, D., Detter, C., Tapia, R., Han, C., Goodwin, L., Liolios, K., Ivanova, N., Mavromatis, K., Mikhailova, N., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y.J., Jeffries, C.D., Rohde, M., Spring, S., Sikorski, J., Goker, M., Woyke, T., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C. & Klenk, H.P. (2010). Complete genome sequence of Thermosediminibacter oceani type strain (JW/IW-1228P(T)). *Standards in Genomic Sciences,* 3(2), pp. 108-116.

Pop, M., Phillippy, A., Delcher, A.L. & Salzberg, S.L. (2004). Comparative genome assembly. *Briefings in Bioinformatics,* 5(3), pp. 237-248.

Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L. & Hammond, W.E. (1997). Medical data mining: knowledge discovery in a clinical data warehouse. *Proc AMIA Annu Fall Symp*, pp. 101-5.

Ramsay, I.R. & Pullammanappallil, P.C. (2001). Protein degradation during anaerobic wastewater treatment: derivation of stoichiometry. *Biodegradation,* 12(4), pp. 247-57.

Reams, A.B. & Neidle, E.L. (2004). Selection for gene clustering by tandem duplication. *Annual Review of Microbiology,* 58, pp. 119-142.

Romero, D. & Palacios, R. (1997). Gene amplification and genomic plasticity in prokaryotes. *Annual Review of Genetics,* 31, pp. 91-111.

Roos, D.S. (2001). Computational biology - Bioinformatics - Trying to swim in a sea of data. *Science,* 291(5507), pp. 1260-+.

Rui, J.P., Qiu, Q.F. & Lu, Y.H. (2011). Syntrophic acetate oxidation under thermophilic methanogenic condition in Chinese paddy field soil. *Fems Microbiology Ecology,* 77(2), pp. 264-273.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics,* 16(10), pp. 944-945.

Saier, M.H., Reddy, V.S., Tamang, D.G. & Vastermark, A. (2014). The Transporter Classification Database. *Nucleic Acids Research,* 42(D1), pp. D251-D258.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol,* 4(4), pp. 406-25.

Salzberg, S.L., Delcher, A.L., Kasif, S. & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research,* 26(2), pp. 544-548.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74(12), pp. 5463-7.

Sasaki, D., Hori, T., Haruta, S., Ueno, Y., Ishii, M. & Igarashi, Y. (2011). Methanogenic pathway and community structure in a thermophilic anaerobic digestion process of organic solid waste. *Journal of Bioscience and Bioengineering,* 111(1), pp. 41-46.

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. & Sander, C. (1994). GeneQuiz: a workbench for sequence analysis. *Proc Int Conf Intell Syst Mol Biol,* 2, pp. 348-53.

Schink, B. (1997). Energetics of syntrophic cooperation in methanogenic degradation. *Microbiology and Molecular Biology Reviews,* 61(2), pp. 262-&.

Schink, B. (2002). Synergistic interactions in the microbial world. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology,* 81(1-4), pp. 257-261.

Schnurer, A. & Nordberg, A. (2008). Ammonia, a selective agent for methane production by syntrophic acetate oxidation at mesophilic temperature. *Water Science and Technology,* 57(5), pp. 735-740.

Schnurer, A., Schink, B. & Svensson, B.H. (1996). Clostridium ultunense sp nov, a mesophilic bacterium oxidizing acetate in syntrophic association with a hydrogenotrophic methanogenic bacterium. *International Journal of Systematic Bacteriology,* 46(4), pp. 1145-1152.

Schnurer, A., Zellner, G. & Svensson, B.H. (1999). Mesophilic syntrophic acetate oxidation during methane formation in biogas reactors. *Fems Microbiology Ecology,* 29(3), pp. 249-261.

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods,* 5(1), pp. 16-18.

Schwarz, J.I.K., Lueders, T., Eckert, W. & Conrad, R. (2007). Identification of acetate-utilizing Bacteria and Archaea in methanogenic profundal sediments of Lake Kinneret (Israel) by stable isotope probing of rRNA. *Environmental microbiology,* 9(1), pp. 223-237.

Scow, K.M. & Hicks, K.A. (2005). Natural attenuation and enhanced bioremediation of organic contaminants in groundwater. *Current Opinion in Biotechnology,* 16(3), pp. 246-253.

Shigematsu, T., Tang, Y.Q., Kobayashi, T., Kawaguchi, H., Morimura, S. & Kida, K. (2004). Effect of dilution rate on metabolic pathway shift between aceticlastic and nonaceticlastic methanogenesis in chemostat cultivation. *Applied and Environmental Microbiology,* 70(7), pp. 4048-4052.

Shimada, T., Morgenroth, E., Tandukar, M., Pavlostathis, S.G., Smith, A., Raskin, L. & Kilian, R.E. (2011). Syntrophic acetate oxidation in two-phase (acid-methane) anaerobic digesters. *Water Science and Technology,* 64(9), pp. 1812-1820.

Sieber, J.R., McInerney, M.J. & Gunsalus, R.P. (2012). Genomic Insights into Syntrophy: The Paradigm for Anaerobic Metabolic Cooperation. *Annual Review of Microbiology, Vol 66,* 66, pp. 429-452.

Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L. & Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic Acids Research,* 41(D1), pp. E344-E347.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. & Kasprzyk, A. (2009). BioMart - biological queries made easy. *BMC genomics,* 10.

Smith, T.F. & Waterman, M.S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology,* 147(1), pp. 195-197.

Snel, B., Huynen, M.A. & Dutilh, B.E. (2005). Genome trees and the nature of genome evolution. *Annual Review of Microbiology,* 59, pp. 191-209.

Sobieraj, M. & Boone, D. (2006). Syntrophomonadaceae.In The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community. *New York: Springer-Verlag*, pp. 1041–46.

St. Pierre, S.E., Ponting, L., Stefancsik, R., McQuilton, P. & the FlyBase, C. (2014). FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic Acids Research,* 42(D1), pp. D780-D788.

Staden, R. & Mclachlan, A.D. (1982). Codon Preference and Its Use in Identifying Protein Coding Regions in Long DNA-Sequences. *Nucleic Acids Research,* 10(1), pp. 141-156.

Stadtman, T.C. (1996). Selenocysteine. *Annual Review of Biochemistry,* 65, pp. 83-100.

Stams, A.J.M. (1994). Metabolic Interactions between Anaerobic-Bacteria in Methanogenic Environments. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology,* 66(1-3), pp. 271-294.

Stein, L. (2001a). Genome annotation: from sequence to biology. *Nat Rev Genet,* 2(7), pp. 493-503.

Stein, L. (2001b). Genome annotation: from sequence to biology. *Nature Reviews Genetics,* 2(7), pp. 493-503.

Stephanie (2011). Beneficial Bacteria: 12 Ways Microbes Help The Environment | WebEcoist. URL: http://webecoist.momtastic.com/2011/09/26/beneficial-bacteria-12-ways-microbes-help-the-environment/.

Stoesser, G., Sterk, P., Tuli, M.A., Stoehr, P.J. & Cameron, G.N. (1997). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research,* 25(1), pp. 7-13.

Subramanian, A.R., Weyer-Menkhoff, J., Kaufmann, M. & Morgenstern, B. (2005). DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics,* 6.

Sun, L., Müller, B. & Schnürer, A. (2013). Biogas production from wheat straw: community structure of cellulose-degrading bacteria. *Energy, Sustainability and Society,* 3(1), pp. 1-11.

Sun, L., Muller, B., Westerholm, M. & Schnurer, A. (2014). Syntrophic acetate oxidation in industrial CSTR biogas digesters. *Journal of Biotechnology,* 171, pp. 39-44.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biology,* 2(6).

Tatusov, R.L., Koonin, E.V. & Lipman, D.J. (1997). A genomic perspective on protein families. *Science,* 278(5338), pp. 631-637.

Thingstad, T.F. & Lignell, R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology,* 13(1), pp. 19-27.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research,* 22(22), pp. 4673-4680.

Topel, T., Kormeier, B., Klassen, A. & Hofestadt, R. (2008). BioDWH: a data warehouse kit for life science data integration. *J Integr Bioinform,* 5(2).

Vallender, E.J. (2011). Expanding whole exome resequencing into non-human primates. *Genome Biology,* 12(9).

Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Le Fevre, F., Longin, C., Mornico, D., Roche, D., Rouy, Z., Salvignol, G., Scarpelli, C., Smith, A.A.T., Weiman, M. & Medigue, C. (2013). MicroScope-an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Research,* 41(D1), pp. E636-E647.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & Médigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Research,* 34(1), pp. 53-65.

Van Domselaar, G.H., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A.C., Dong, X.L., Lu, P., Szafron, D., Greiner, R. & Wishart, D.S. (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Research,* 33, pp. W455-W459.

Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E. & Stein, L. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biology,* 8(3), p. R39.

Venter, J.C. (2006). Shotgunning the Human Genome: A Personal View. In: John, W. & Sons, L. (eds) *Encyclopedia of Life Sciences*. Chichester, UK: John Wiley & Sons, Ltd[2014/04/04/09:45:13].

Visel, A., Rubin, E.M. & Pennacchio, L.A. (2009). Genomic views of distant-acting enhancers. *Nature,* 461(7261), pp. 199-205.

Voelkerding, K.V., Dames, S. & Durtschi, J.D. (2010). Next Generation Sequencing for Clinical Diagnostics-Principles and Application to Targeted Resequencing for Hypertrophic Cardiomyopathy A Paper from

the 2009 William Beaumont Hospital Symposium on Molecular Pathology. *Journal of Molecular Diagnostics,* 12(5), pp. 539-551.
Vonheijne, G. (1986). A New Method for Predicting Signal Sequence Cleavage Sites. *Nucleic Acids Research,* 14(11), pp. 4683-4690.
Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10(1), pp. 57-63.
Watson, J.D. & Cookdeegan, R.M. (1991). Origins of the Human Genome Project. *Faseb Journal,* 5(1), pp. 8-11.
Wei, Y., Zhou, H., Zhang, L., Zhang, J., Wang, Y., Wang, S., Zhou, Z. & Yan, X. (2014). Draft Genome Sequence of Clostridium ultunense Strain BS (DSMZ 10521), Recovered from a Mixed Culture. *Genome Announc,* 2(1).
Weiland, P. (2010). Biogas production: current state and perspectives. *Applied Microbiology and Biotechnology,* 85(4), pp. 849-860.
Westerholm, M., Dolfing, J., Sherry, A., Gray, N.D., Head, I.M. & Schnürer, A. (2011a). Quantification of syntrophic acetate-oxidizing microbial communities in biogas processes. *Environmental Microbiology Reports,* 3(4), pp. 500-505.
Westerholm, M., Leven, L. & Schnurer, A. (2012). Bioaugmentation of Syntrophic Acetate-Oxidizing Culture in Biogas Reactors Exposed to Increasing Levels of Ammonia. *Applied and Environmental Microbiology,* 78(21), pp. 7619-7625.
Westerholm, M., Roos, S. & Schnurer, A. (2010). Syntrophaceticus schinkii gen. nov., sp nov., an anaerobic, syntrophic acetate-oxidizing bacterium isolated from a mesophilic anaerobic filter. *Fems Microbiology Letters,* 309(1), pp. 100-104.
Westerholm, M., Roos, S. & Schnurer, A. (2011b). Tepidanaerobacter acetatoxydans sp. nov., an anaerobic, syntrophic acetate-oxidizing bacterium isolated from two ammonium-enriched mesophilic methanogenic processes. *Systematic and Applied Microbiology,* 34(4), pp. 260-266.
Whitman, W., Bowen, T. & Boone, D. (2006). The Methanogenic Bacteria. In: Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E. (eds) *The Prokaryotes* Springer New York, pp. 165-207.
Wold, B. & Myers, R.M. (2008). Sequence census methods for functional genomics. *Nature Methods,* 5(1), pp. 19-21.
Yada, T., Totoki, Y., Takagi, T. & Naka, K. (2001). A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Research,* 8(3), pp. 97-106.
Ye, Q., Roh, Y., Carroll, S.L., Blair, B., Zhou, J.Z., Zhang, C.L. & Fields, M.W. (2004). Alkaline anaerobic respiration: Isolation and characterization of a novel alkaliphilic and metal-reducing bacterium. *Applied and Environmental Microbiology,* 70(9), pp. 5595-5602.
Zhang, W.Y., Chen, J.J., Yang, Y., Tang, Y.F., Shang, J. & Shen, B.R. (2011). A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLoS One,* 6(3).

Zhang, Y., Romero, H., Salinas, G. & Gladyshev, V.N. (2006). Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. *Genome Biology,* 7(10).

Zhaoli (2012). Machine Learning in Bioinformatics. *2011 International Conference on Computer Science and Network Technology (Iccsnt), Vols 1-4*, pp. 582-584.

Zhilina, T.N., Zavarzina, D.G., Kolganova, T.V., Tourova, T.P. & Zavarzin, G.A. (2005). "Candidatus contubernalis alkalaceticum," an obligately syntrophic alkaliphilic bacterium capable of anaerobic acetate oxidation in a coculture with Desulfonatronum cooperativum. *Microbiology,* 74(6), pp. 695-703.

Zhou, Y., Liang, Y.J., Lynch, K.H., Dennis, J.J. & Wishart, D.S. (2011). Phast: A Fast Phage Search Tool. *Nucleic Acids Research,* 39, pp. W347-W352.

Zinder, S.H. (1984). Microbiology of anaerobic conversion of organic wastes to methane: recent developments. *ASM News,* 50, pp. 294-298.

Zinder, S.H. & Koch, M. (1984). Non-Aceticlastic Methanogenesis from Acetate - Acetate Oxidation by a Thermophilic Syntrophic Coculture. *Archives of Microbiology,* 138(3), pp. 263-272.