Analysis of the metabolic pathways in the gas fermenting *Eubacterium limosum*

University of Nottingham

MRes Thesis

Matthew W Emerson

Bioinformatics MRes

Supervisor: Dr Thomas Millat and Prof Nigel Minton

mbxme3/4289840

September 29th, 2017

Abstract

Global warming is one of the most dangerous problems humanity faces today. The collection of C1 gases such as Carbon dioxide, Carbon monoxide and Methane, in the atmosphere is causing serious damage to the earth, with methods to remove these C1 gases ('greenhouse gases') from the atmosphere an large focus of current day research. The Wood-Ljungdahl pathway is a C1 gas fermenting pathway present in acetogenic bacteria; this uses C1 gases as a substrates to produce potentially valuable biocommodities and biofuels. *Eubacterium limosum* is a acetogenic bacteria which possesses the Wood-Ljungdahl pathway and is also identified as an acetogen of interest, but with limited literature and brief classifications there is an opportunity to analyse the Wood-Ljungdahl pathway in this bacteria further. A Wood-Ljungdahl pathway map specific for the *E.limosum* ATCC 8486 was generated using CLCBio's BLAST analysis, confirming the presence of genes known to be in the Wood-Ljungdahl pathway of *E.limosum* KIST 612, while consulting MetaCyc and BRENDA databases. A genome alignment was carried out using progressive Mauve, which allowed for a multiple genome alignment between the 3 *E.limosum* strains and then an alignment which focused on the position of the Wood-Ljungdahl pathway genes between the strains ATCC 8486 and KIST 612. Finally, ScrumPy was used to create a structural model, which allowed for an elementary mode analysis (EMA), flux balance analysis and optimal yield calculations to be conducted.

The pathway analysis of *Eubacterium limosum* 8486 produced a Wood-Ljungdahl pathway model, which included the presence of a methanol metabolism branch and a acetate CoA transferase; opposed to a butyrate kinase which was not present.

The genome alignment provided a comparison between ATCC 8486 and the other 2 strains (KIST 612 and SA11), from which it was seen to be more similar to SA11 despite widespread rearrangements throughout the middle of the genomes, with the ends conserved. In regards to the presence of the Wood-Ljungdahl genes between ATCC 8486 and KIST 612, it was seen that carbonyl and butyrate branch genes were highly conserved and shared high similarity between the genomes, with all the genes organised in the same region but translocated. The methyl branch genes were located in the same region but this region contained areas of low conservation between the 2 genomes hinting at possible mutations. None of the acetate branch genes were organised together and appeared in different regions throughout the genome, which had mostly been rearranged.

The EMA showed that 74 different modes were present; 29 modes produced iso-propanol while 27 produced butyrate. 3 elementary modes of interest were those which only produced iso-propanol and no other product (acetate was produced and then consumed). The optimal yields were calculated for single substrate and multiple substrates; iso-propanol's optimal yields were when carbon monoxide was used as a sole substrate, while acetate and butyrate's optimal yield was when carbon dioxide was used as the sole substrate. Overall, it was seen that multiple substrates provided a greater yield in acetate and butyrate, while single substrate provided a better yield for iso-propanol. The flux balance analysis showed that every gene/reaction present in the Wood-Ljungdahl pathway could be knocked out and provide a feasible flux. This was then done with multiple knockouts, which used every combination of 2 genes, which also provided a feasible flux.

Overall, the results provided an extensive mapping of the Wood-Ljungdahl pathway present in *E.limosum* 8486. While providing further understanding of the involvement of the genes and reac-

tions in this pathway. Moreover, the alignment provided a knowledge of the copious rearrangement between the *E.limosum* strains and the conservation of the Wood-Ljungdahl genes, which may be a survival response to living in a harsh environment, but may also suggest that mutation has taken place in regions where Wood-Ljungdahl genes are located as a result of this rearrangement. The structural model suggests that despite iso-propanol and butyrate having similar formation pathways, butyrate appears to be preferred, with numerous patterns emerging from the difference between the modes and yields of butyrate and iso-propanol, suggesting that the production of one or the other requires certain regulations. Moreover, all the genes in the Wood-Ljungdahl pathway can be knocked out and flux can still pass through, indicating that genetic manipulation to increasing production may be harder than first thought. However, manipulating substrates in an attempt to manipulate production could be an area of interest.

Acknowledgements

I would like to acknowledge all of the C1net and SBRC group, specifically my supervisor Dr Thomas Millat, who has helped my with every aspect of this project and being able to mentor me throughout the year. I would also liked to specifically extend my gratitude to those who have help me produce my results to help develop this project, I appreciate the time and effort which has gone into helping me over the year and further developing the project and helping to lp.Solve() many an issue;Rob, Rupert and James.

Contents

List of Figures

Chapter 1

Introduction

1.1 General Introduction

An energy crisis is currently amidst us, one which is threatening all aspects of life, not just human. We are in a position where our current demand for energy and our unrelenting harvesting of fossil fuels has led to severe changes in the earths climate system and environment, which if left to worsen could become permanent and even life threatening for humanity. This is why it is imperative that funds and research are focused towards developing alternative fuel sources and methods for sustainable carbon recycling. The ideal solution would be the development of a fuel source which is clean, renewable, sustainable and stable but is also able to meet our humongous energy demand, which is predicted to increase by an order of magnitude by 2050 [9, 10]. There have been efforts in the utilisation of renewable sources e.g solar, wind, hydro and biological fuel sources, which have become very prevalent in the last 15 years. However fossil fuels are still the dominant fuel/energy source despite the irreversible damage they might cause, meeting 80% of our energy demands while renewable energy only meets 13.5% [7]. Fossil fuel production is even predicted to increase at least until the year 2055 [1]. As a consequence, we are on the cusp of reaching the peak of fossil fuel production and will thus be on a decline thereafter, and that despite evidence showing the threat that fossil fuel induced climate change has on Earth, we appear to be continuing to deplete our reserves still. Moreover, several studies have reported that oil, coal and gas reserves could be depleted in approximately 35,107 and 37 years, respectively [2]. These results suggest that after the year 2042, coal could be the only fuel left. Despite this alarming prediction, oil production in the United States of America has increased by 3 million barrels a day over the last decade [3]. It could be debated that fossil fuels may not run out during our generation, but what can be said is that they will eventually run out and they are not a sustainable solution to our energy and fuel demand. Consequently, another century dominated by fossil fuels would be disastrous for the environment, humanity and planet Earth [3].

The use of fossil fuels is also the driving force behind climate change. The burning of fossil fuels releases the greenhouse gases: methane, carbon dioxide and carbon monoxide. Although these gases are produced naturally on Earth, the accelerated release of these greenhouse gases into the atmosphere by the ever increasing burning of fossil fuels has has triggered a man-made increase in average temperatures often referred to as global warming [4]. The greenhouse gases in the atmosphere absorb and retain radiation/heat which is reflected or released from the Earth's surface. This greenhouse effect results in an acceleration of global warming that could lead towards the rise in atmospheric gases and "dangerous anthropogenic interference with climate change" [4]. Evidence of this is already appearing, with geographical disasters such as the Arctic sea ice retreats, increase in heavy rainfall, flooding, permafrost melt, loss of glaciers, and increased intensity of hurricanes.

Furthermore, changes in temperature and rainfall may result in an alteration of the distribution of arable or even habitable land, changing sea levels that endanger low lying coastal populations and land, but also may shift the distribution range of disease vectors [8]. These changes poses severe economic costs on the existing populations recovering from damages as well as for establishing and maintaining protective measures. In worst cases, they might trigger migration with potential for conflicts on economic, cultural and even political levels. In addition, all these responses are accompanied by adverse effects on health of the affected population, including physical and mental health but also mortality [8].

Combustion of fossil fuels has another immediate effect on the environment, our quality of live and health and, thus, healthcare. It releases a mixture of air pollutants, such as carbon monoxide, sulfur dioxide, nitrogen oxides, ozone, heavy metals, and volatile organic compounds, which have a negative impact on public health [5]. The route of exposure generally occurs through inhalation and ingestion, whether inhaled through respiratory process or ingestion of contaminated food or water leading to gastrointestinal absorption. Air pollution can have adverse consequences for human health ranging from nausea and difficulty in breathing to severe illnesses like cancer and chronic obstructive pulmonary disease [5] and eventually contribute to an increased hospitalisation mortality. The World Health Organisation estimates that currently around 160,000 people die each year from side effects relating to climate change and that this number will double by 2020 [6, 7].Consequently, air pollution caused by burning fossil fuels is posing an increasing danger to humans, quality of life and health care.

In summary, it obvious that fossil fuels does not constitute a sustainable base for the generation of energy, fuels, and chemicals. Instead renewable energy sources and carbon (re)cycling could provide us with a sustainable. clean and inexhaustible supply for future generations. In addition, they might also help us to eliminate the imminent threats caused by unrestrained usage of fossil fuel, i.e. climate change, adverse impacts on health, environmental damages and last but least depletion of resources. The biotechnological application of acetogenic bacteria, such as *E. limosum*, could provide an important contribution to carbon recycling and carbon fixation due to their capability to utilise C1 gases.

1.2 Acetogens and *Eubacterium limosum*

There are over a 100 acetogenic species with 22 different genera with Clostridium, Acetobacterium and Eubacterium being the main genera [19]. However they can be found in genera which contains both acetogenic and non-acetogenic bacteria [11]. Acetogens are a type of anaerobic bacteria which are of particular interest with regards to the production of biofuel and biochemicals from renewable sources, as they can utilise C1 compounds for growth. These compounds can be derived from the gasification of renewable biomass, agricultural and municipal waste resulting in a blend of CO, CO_2 and H_2 referred as syngas or synthetic gas [17, 18]. Additionally, these bacteria represent one of the only two groups of prokaryotes, with methanogens being the other, that can grow autotrophically reducing CO_2 and CH_3COOH [11].

They are strictly anaerobes and are physiologically defined by their ability to convert the C1 compounds CO and CO_2 into acetate with acetyl-CoA as an important intermediate using the Wood-Ljungdahl pathway. In addition either H2 or water are required as additional substrate. Moreover, they are capable of growth on hexoses, pentoses, alcohols, organic compounds containing methyl groups, and formic acid, while additional products like ethanol, butyrate, and butanol could be formed depending on the acetogen and process conditions [12–15]. Interestingly ethanol production from syngas is already being commercialised [16–18] Although acetate is produced as a major product by numerous aerobic and anaerobic bacteria, acetogens have a unique characteristic which distinguishes them from other bacteria; the presence of the ACS enzyme [11]. SynGas to biofuel fermentation is an area which has its limitations however; this is because of the lack of knowledge regarding energy metabolism in acetogenic bacteria and the need for more sophisticated genetic tools for the modification of acetogens. Syngas fermenting acetogens of recent industrial interest are *Acetobacterium woodii*, *Clostridium autoethanogenum* and *Eubacterium limosum* [16–18]. The latter will be the focal point of my thesis. Due to their metabolic versatility, acetogens are able to colonise a vast variety of habitats from gastrointestinal, terrestrial, subsurface and aquatic ecosystems; they are also found to be extremophiles with respects to temperature, pH, and salinity [11, 12]. The mammalian gut is one of the main habitats harbouring numerous H_2 utilising acetogenic bacteria, in which acetate is produced via the Wood-Ljungdahl Pathway [11].

With respect to their energy metabolism, two groups of acetogens can be distinguished: 'Rnf-dependent' and 'Ech-dependent' acetogens. Rnf dependent acetogens, e.g. *A. woodii* and *C. autoethanogenum*, combine a sodium- or hydrogen-ion-dependent RNF complex and ATP synthase to generate ATP. While, Ech-dependent acetogens lack an RNF complex and use an energy converting hydrogenase (Ech) instead. Whereas the details of energy conservation in RNF-dependent acetogens well-studied [**?**], they are still uncertain for the latter group [20].

The RNF-dependent *E. limosum*, a gram-positive and obligate anaerobe, is one of the few acetogens forming butyrate when grown on syngas or methanol [21]. It is found in human and animal intestines, primarily sheep rumen [11]. Additionally, it is involved in the biotransformation of the isoflavonoids biochanin A, formononetic, and glycitein to the estrogenic metabolites genistein, daidzein, and 6,7,4'-trihydroxyisoflavone [22]. However, at this point in time knowledge of *E. limosum* characteristics, metabolism and biosynthesis pathways is limited, other than its ability to grow until high CO/CO2 concentrations and produce butyrate. Similar to *A. woodii* and *C. autoethanogenum*, it employs the Wood-Ljungdahl pathway for growth on syngas [21].

1.3 Wood-Ljungdahl pathway

For the utilisation of CO and CO_2, acetogens have developed a special metabolic network. Its core structure is often referred to as Wood-Ljungdahl pathway that mediates the formation of acetyl-CoA from two molecules of CO_2 and the subsequent production of acetate and further molecules of industrial interests. It is also involved in energy conservation in acetogens. Interestingly, the net ATP gain of the WLP is zero, but it is crucial for the maintenance of a transmembrane ion gradient. This gradient is then used to generate ATP via ion translocating ATP synthase. Moreover, acetogens can be divided into two groups dependent on which ion is translocated, those which are Na^+ dependent such as *A. woodii* and H^+ dependent acetogens such as *Moorella thermoacetica* and *C.autoethanogenum* [12]. The Wood-Ljungdahl pathway consists of two branches, the carbonyl branch and methyl branch. The carbonyl branch is folate-dependent and unique to acetogens, methanogens, and sulfate reducers, while the methyl branch is present from bacteria to humans [23]. The electrons used in the Wood-Ljungdahl pathway may be derived from molecular hydrogen 'autotrophic growth', carbon monoxide or organic donors, i.e formate, lactate, hexoses, pentoses 'heterotrophic growth' [21].

The methyl branch begins with CO_2 being reduced to formate by formate dehydrogenase. The formate is then activated and bound to tetrahydrofolate forming formyl tetrahydroformate. In the next step, water is split off resulting in a methenyl group, which is then reduced via methylene tetrahydrofolate to methyl tetrahydrofolate. The methyl group of methyl tetrahydrofolate is transferred to a corrinoid iron sulfur protein that delivers it to the bifunctional CO dehydrogenase/acetyl-CoA-synthase (CODH/ACS complex). This complex is an essential node in the WLP, where the methyl and carbonyl branch join [12].

In the carbonyl branch, the CODH subunit of the CODH/ACS complex converts a molecule of CO_2 into a molecule CO which binds to Cluster A of the subunit. The reduced oxygen is combined with two protons to form water [24]. When the methyl group from the corrinoid iron sulfur protein is transferred to CODH/ACS it forms a methyl-nickel species. A carbon-carbon bond is formed by the condensation of the methyl and carbonyl group, this forms an acetyl-metal species. Finally, the binding of a CoA molecule and synthesis of the carbon-sulfur bond through condensation of the CoA molecule and the acetyl-group, leads to the production of acetyl-CoA [24].The end product for both the methyl and carbonyl branches is acetyl-CoA.

The formed acetyl-CoA is the starting point for the formation of biomass and biocommodities such as acetate and butyrate.

Acetate formation consists of two steps. In a first step, phosphotransacetylase converts acetyl-CoA into acetyl phosphate. Subsequently, acetate kinase facilities the reaction from acetyl phosphate to acetate. In the latter step, one ATP is generated which balances the ATP required to activate formate, resulting formyl-THF, in the methyl branch [25]. An interesting feature of *E. limosum*, which distinguish it from the most other acetogens, is its capability to produce butyrate when grown on CO and CO2 as sole carbon sources. It is commonly suggested the butyrate formation in *E. limosum* is similar to solventogenic clostrida [21]. First, two molecules of acetyl-CoA are condensed to acetoacetyl-CoA by an acetyl-CoA C-acetyltransferase. In a next step, 3-hydroxybutyryl-CoA dehydrogenase reduced acetoacetyl-CoA to hydroxybutyryl-CoA. Subsequently, enoyl-CoA hydratase spilts off water resulting in crotonyl-CoA. Crotonyl-CoA is then reduced to Butyryl-CoA via the action of the ferredoxin-reducing, electron bifurcating butyryl-CoA dehydrogenase. Similar to acetate formation, phosphotransbutyrylase exchanges the CoA-moiety with a phosphate group. The final step then would be the formation of butyrate via butyrate kinase, which also would gain ATP. However, it appears that no butyrate kinase can be identified in the genome of *E. limosum*. Consequently, *E. limosum* either possesses a butyrate kinase of a yet unknown type or butyrate formation occurs via a different route in this organism [21].

However, the biofuels/ biocommodities produced as a results of this are of value and if production can be increased, could be used commercially [21]. Acetate isn't a particularly sought after commodity, with some commercial importance as salts used in dyeing and as a diuretic; it may also be found in industrial solvents. More importantly its acid form, acetic acid is the main component in vinegar. Butyrate/Butyric acid is known for its powerful odour, with its esters being used for food and perfume additives. Furthermore, butyrate can be used to produce butanol and butane, which are used as fuel additives or fuel alternative. Ultimately, this shows that there could be enormous potential for converting *E. limosum* into a viable microbial factory

1.4 Current research in *Eubacterium limosum* and the Wood-Ljungdahl pathway

Currently, research is ongoing into the potential ability of acetogens to produce high yielding biofuels from an abundant and dangerous substrate, employing the Wood-Ljungdahl pathway. Progress in the understanding and manipulation of the Wood-Ljungdahl pathway as well as searching for optimal acetogens has been made in recent years, as this area has quickly attracted a lot of attention. In 2015, Jeong et al [21] detailed the energy conservation model based on the KIST 612 strain of *E. limosum*. It provides one of the most detailed descriptions of the metabolic network in *E. limosum* and its utilisation of the Wood-Ljungdahl pathway; Additionally, it is one of a limited number of papers describing butyrate formation in acetogens. Jeong et al [21] were able to provide detailed yet clear results showing how *E. limosum* uses the Wood-Ljungdahl Pathway, showing all the genes and gene clusters which are involved in this process.

Additionally, they identified genes encoding for potential route towards butyrate formation following the common notion in solventogenic clostridia, as described in the previous subchapter. Although they did not discuss any alternative routes for butyrate formation. Furthermore, the paper identified differences in the genomic organisation of the WLP between *E. limosum* and the other more renowned acetogens, *A. woodii* and *C. autoethanogenum*. It pointed out that genes for the Wood-Ljungdahl pathway are split into 3 different clusters similar to *A. woodii*, that, however, *A.woodii* has an additional hydrogenase-encoding subunit in the first gene cluster next to the formate dehydrogenase enabling direct use of H_2 for CO_2 reduction. Moreover, it was found that the formate dehydrogenase (FDH) in *E. limosum* does not contain a second subunit, which contains a NADP-binding site, indicating that *E. limosum* does not have a NADP-dependent FDH. The second gene cluster contains genes which encode for the transformation of formate to methyl tetrahydrofolate. It was identified that *E. limosum* only has two subunits of the methylene tetrahydroformate reductase (MetF and MetV), which is similar to *Clostridium ljungdahlii*, while again *A. woodii* differs with a rnfC2 subunit [21]. The third cluster possesses genes involved with the carbon monoxide/acetyl-CoA synthase enzyme, which covers the carbon monoxide/acetyl-CoA synthase, corrinoid iron-sulfur proteins and methyltransferase. All of which possess similarities with those found in *A. woodii*, with no noticeable changes in the genes and genome, unless under specific conditions. When growing on SynGas, monomeric CODH are not present, however two genes which encode putative hydrogenases are found [21]

Acetogens are known to have either an Rnf or Ech complex; which couple the oxidation of reduced ferredoxin to the generation of either a H^+ or Na^+ membrane potential. Inspection showed that *E. limosum* uses an Rnf complex but not an Ech complex. The subunits of the RNF complex in *E. limosum* are very similar to those in *A. woodii*. Furthermore, Jeong et al's [21] were able to show that sodium ions are the coupling ions for the Rnf complex, with the electrochemical ion gradient established enough to drive ATP synthase. Further investigation gave way to an unprecedented finding, a gene cluster which encoded two different multi-subunit ATPases/synthases with no singular enzyme being similar to the bacterial F_1F_O ATP synthases [21]. Overall, Jeong et al [21] provided substantial knowledge of energy conservation in *E. limosum*, being able to clearly deduce that *E. limosum* conserves energy via a chemiosmotic mechanism, where Na+ is used as a coupling ion by an Na+ translocating Rnf complex and an Na+ translocating ATP synthase; which are the integral components of the process. The more reduced ferredoxin available for the Rnf complex, the more ATP which can be synthesised by the ATP synthase [21].

Another strain of *E. limosum* which has been the focus of recent research is SA11. In 2016 Kelly et al [26] published work detailing the research conducted on the SA11 strain of *E. limosum*, looking into energy sources and conservation for homoacetogenic bacteria; this involved an analysis of the Wood-Ljungdahl pathway in SA11. The genes in the Wood-Ljungdahl pathway, much like in KIST 612, were found in three distinct clusters. It was stated that energy conservation in the Wood-Ljungdahl pathway and acetogens is not fully understood. Although, Kelly et al were able to explain that, like in KIST 612, energy conservation in SA11 occurs through a membrane-bound Na^+ translocating Rnf complex and ATP synthases [26].

Similar to the strain KIST 612, the genes encoding for the enzymes required for the formation of acetyl-CoA and butyryl-CoA could be identified in the strain SA11. The cluster of butyrate genes included two subunits of an electron-transferring flavoprotein (EtfAB). Kelly et al [26] have proposed that butyryl-CoA dehydrogenase forms a complex with EtfAB and also uses flavin-based electron bifurcation.

It is stated that *E. limosum* does not have a butyrate kinase and uses an alternative method, which creates butyrate via the transfer of the CoA moiety from butyryl-CoA to acetate generating acetyl-CoA as an additional product [26].

Furthermore, lactate was identified as a growth supporting substrate in acetogens [26]. In SA11, a stable complex is formed between lactate dehydrogenase and two subunits of an electron-transferring flavoprotein [26]. Other *E. limosum* strains have also been the focus of recent research unrelated to the metabolic nature of the bacteria and its use of the Wood-Ljungdahl Pathway. Although, it was still an analysis of *E.limosum*'s metabolic pathways, *E.limosum* ZL-IL was the focus of work looking to gain understanding into the biotransformation of flavonolignan constituents by ZL-IL and thus the production of demethylisosilybin B, demethylisosilybin A, demethylsilybin B, and demethylsilybin A. The results indicated that the products showed significant inhibitory activities of Alzheimers beta-amyloid aggregation, which is a suspected cause of Alzheimer's disease [27].

Meanwhile, the *E. limosum* ATCC 8486 strain has had limited involvement in recent research, apart from publications of the genome; which was provided with physiological and metabolic information. Song and Cho [22] gave a genome analysis in their report, in which it was said that the ATCC 8486 strain demonstrated a NADH-dependent butanol dehydrogenase-mediated conversion of acetyl-CoA to butyryl-CoA. This finding differs from experimental observations for strain KIST 612. However, the ATCC 8486 strain is not a new strain, with the majority of the previous *E. limosum* research having been conducted on this strain, especially in regards to developing an understanding of the Wood-Ljungdahl pathway in *E. limosum*; such as which substrates are used and what the end products are [28].

The Wood-Ljungdahl pathway has received a lot more interest from current research projects, due to its presence in all acetogens and other bacteria. The research tends to focus on the products of the Wood-Ljungdahl pathway, investigating what different products can be formed naturally or via bio-engineering additional branches, methods which allow for increased production, understanding how each of the mechanisms in the pathway work, and understanding methods of energy conservation. Around 100 acetogens have been isolated, however they are metabolically, ecologically and phylogenetically diverse, which suggests that the Wood-Ljungdahl pathway may be adaptable, with differing agendas or functions dependent on differing acetogens or bacteria.

Although methods of energy conservation in the Wood-Ljungdahl pathway are known; such as the ATPase combined with an ion gradient generated in one of two ways. Recent research has indicated a third potential method of energy conservation; this method does not require any Na+ dependence nor genes needed for quinone and cytochrome synthesis. However it uses an Rnf complex; this generates a proton gradient by simultaneous oxidation of reduced ferredoxin and transfer of reduced equivalents to NAD^+. This is thought to occur in *E. limosum*, *A. woodii*, and *C. autoethanogenum* [29].

The products produced via the Wood-Ljungdahl pathway, as previously stated, are of particular interest. Although the use of them in a commercial setting is yet to truly happen. As chemicals such as acetate and 2-3-butanediol are wanted in bulk, it requires the production of these chemicals to be at an optimal level. In regards to acetate production, the most important rate limiting factors are pH level and solubility of the substrate; however experiments in recent years used to garner the greatest acetate production have altered many parameters. As of 2012, the highest recorded acetate production was produced by Demler and Weuster-Botz [30]. The anaerobic fermentation process happened in a batch-operated stirred tank bioreactor, with a maximum volumetric productivity of 7.4g acetate l-1 d-1 measured with a substrate partial pressure (pH2) of 1700 mbar and a controlled pH of 7. Overall these parameters produced 44g l-1 over 11 days with *A.woodii* as the acetogen [29, 30].

Artificial pathways are bio-engineered into the Wood-Ljungdahl pathway allowing for production of valuable bio-commodities/bio-fuels. Acetone is one of the main chemicals which can be produced via the addition of an artificially engineered pathway, it has huge potential as an alternative to the current method of acetone generation; which requires the use of crude oil. This process requires the insertion of operons containing genes needed for the synthesis of acetone, one group of genes inserted were adc, ctfA, ctfB, thla; which encode for acetoacetate decarboxylase, acetoacetyl CoA:acetate/butyrate:CoA transferase subunit A and B, and thiolase. This gives way to an acetate dependent acetone production pathway by importing genes from *C. acetobutylicum*. However, other methods which are truly synthetic use thioesterases to split off the coenzyme A-moiety from acetoacetyl-CoA, the genes te II(Thioesterase) from Bacillus subtilis and the gene ybgC (acyl-CoA thioesterase) from Haemophilus influenzae are inserted to create this pathway; which is not found naturally [29].

Butanol is another biofuel in which a production pathway can be artificially introduced into acetogens as an extension to the Wood-Ljungdahl pathway. It is a superior biofuel to ethanol; which is naturally produced in acetogens. It has a higher energy content, lower vapour pressure, is easily blended with gasoline and can be used in car engines without modifications. The approach is similar to acetone, where the butanol pathway genes in *C. acetobutylicum* are transferred into *C. ljungdahlii* via an artificial operon containing the genes thla, hbd, crt, bcd, adhE, and bdhA; which encode the butanol pathway, promotors, and terminators. This means that *C. ljungdahlii*, which naturally only produces ethanol and acetate, can artificially produce butanol [29]. Overall, between both engineered pathways it is indicated that the future work lies in optimising these pathways in preferable strains. The optimisation of these pathways will require investigations into the knockout and knock-ins of certain genes, usage of stronger promotors, genes encoding more efficient enzymes, and inactivation of genes responsible for formation of other products [29].

Another recent discovery which developed the understanding of the Wood-Ljungdahl pathway and the mechanisms at work was made by Diekert and Kryzky, who were able to predict the functions of the 'CooC' and 'Orf7' proteins. Diekert and Kryzky were able to predict that 'Orf7' was a reductive activator of the corrinoid iron sulfur protein; this was later confirmed by Hennig et al. While it was indicated that the CooC protein caused the maturation of unique metal clusters within the CODH and ACS enzymes, as well as the enzymes as well. Although, the process of this maturation, as of 2014, is yet to be concluded [31, 32].

1.5 Beneficiaries of the project

The results and research garnered from this year long endeavour will have both short and long term benefits, for numerous different parties. Initially, this work will be able to help provide extra resources into the *E. limosum* and Wood-Ljungdahl literature, with *E.limosum* having quite a limited library. By providing unique knowledge of exactly how the Wood-Ljungdahl pathway is laid out in *E. limosum* ATCC 8486 and providing flux balance and elementary mode analysis, new data and results will be available for future use. The ideal is to show that *E. limosum* is a bacteria which can be used in a commercial setting and that it is a preferable option to more well known acetogens such as *A. woodii* and *C. autoethanogenum*. While long term, a hope is that this project will be viewed beneficially as part of a collective with all the other relevant literature. Because *E. limosum* is can be easily manipulated using tools for genetic engineering, it has the potential to be applied in commercial settings for the production of many biofuels. Moreover, the Wood-Ljungdahl pathway possesses the ability to use environmentally dangerous C1 compounds as the substrates, clearly being very beneficially. Moreover, any progressions in methods to increase production of specific biofuels via knockouts and/or gene knock-ins, which are currently one of the main limiting factors in the commercial use of some acetogens.

1.6 Aims and Objective

The initial aims and objectives for the years research project were to gain a more detailed and accurate understanding of how the Wood-Ljungdahl pathway is used by *E. limosum*; specifically the ATCC 8486 strain. Then gain a further understanding through a genomic comparison between other strains of limosum and other acetogens. It is important that the knowledge of the metabolic functions of *E. limosum* and the Wood-Ljungdahl pathway are compiled and there is an understanding of the differences between each strain and reasons as to why they are different. Furthermore, once a complete mapping of the natural Wood-Ljungdahl pathway in *E. limosum* has been produced, a further mapping of the pathway including additional bio-engineered branches for the production of iso-propanol and use of methanol as a substrate should be produced. This will help provide one of the clearest and most detailed pathway mappings of the Wood-Ljungdahl pathway in *E. limosum* literature. Furthermore, the constructed pathway will be used to conduct an Elementary Mode Analysis and Flux Balance Analysis through a structural model. The aim being to investigate methods which can be used to increase production of the biofuel of interest, namely iso-propanol. This will be done by understanding how the pathway works under normal conditions and what directions are naturally favoured by the pathway. Using this information, we can make virtual alterations to the pathway for our benefit.

Chapter 2

Pathway Mapping

2.1 Introdution

Within all of *E. limosum*'s literary database, there are very few reports involving the Wood-Ljungdahl pathway and *E. limosum*'s acetogenic properties regarding the production of biofuels and biocommodities through C1 compound metabolism. The majority of the literature is focused on the determination of the growth characteristics of *E. limosum* and its mechanisms of energy conservation. This indicates that research in regards to *E. limosum* is still at a preliminary stage, with more knowledge still needing to be unearthed in order to gain an understanding similar to more renowned acetogens such as *C. autoethanogeum*, *A. woodii*, and *Moorella thermocetica*. Although, a couple of recent developments in this field have shed light upon the classification and utilisation of the Wood-Ljungdahl pathway in *E. limosum* strains. Jeong et al [21] were able to piece together a network for butyrate formation in *E. limosum*, while investigating whether this was coupled to ATP synthesis, thus establishing a complete Wood-Ljungdahl pathway and annotations of the genes involved in the strain KIST 612 [21].

Furthermore, Kelly et al [26] published an annotated genome sequence of *E. limosum* SA11. The authors have identified some genes involved in the Wood-Ljungdahl pathway and discussed several ways of energy conservation in *E. limosum*, but they were unable to establish a complete mapping of the Wood-Ljungdahl pathway for this particular strain. Such a mapping is not yet available for strain ATCC 8486. Thus, I aim to develop an accurate representation of the Wood-Ljungdahl pathway for this strain using the structure and genes identified by Jeong et al [21]. This will provide the first full annotation of the Wood-ljungdahl pathway for *E. limosum* ATCC 8486. Furthermore, I aim to clarify the route(s) to butyrate formation in *E. limosum*.

Here I will show how the genes of the Wood-Ljungdahl pathway in KIST 612 were evaluated and used to identify which of these genes are present in the Wood-Ljungdahl pathway of ATCC 8486. This will allow for the construction of a Wood-Ljungdahl pathway model specific to ATCC 8486, based on the previous findings in the strain KIST 612. Once the pathway has been established, the information will be used to produce a model of the Wood-Ljungdahl pathway in Escher (https://escher.github.io) [35], a web-based pathway building tool used to visualise the pathway.

2.2 Method

The structure of the Wood-Ljungdahl pathway as derived by Jeong at al [21] is shown in Figure 2.1. In addition, the figure also identifies the involved enzymes and their encoding genes in *E. limosum* KIST 612. The initial procedure for developing a pathway model for the Wood-Ljungdahl pathway in ATCC 8486 starts by locating the genes which make up the pathway within the genome of ATCC 8486. In order to do this, the annotated genome of KIST 612 was used alongside a paper detailing the specific genes involved in the KIST 612 Wood-Ljungdahl pathway [21, 33]. The annotated genome can be found at the NCBI Nucleotide database under the accession ID CP002273.2.

The first step of the procedure was to extract the nucleotide sequences from each of the genes associated with the Wood-Ljungdahl pathway. These nucleotide sequences had to be saved in a FASTA format to work with the BLAST analysis software, which was used in the next step.

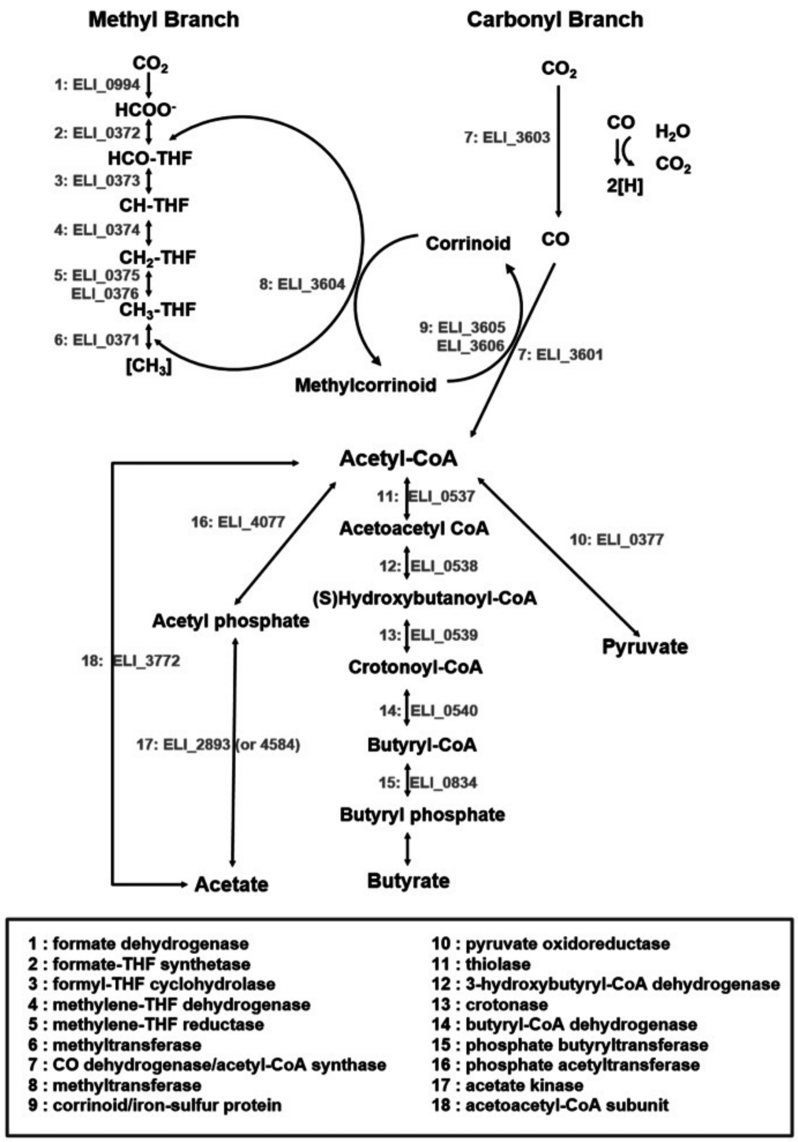

Figure 2.1: KIST 612 Wood-Ljungdahl pathway with corresponding genes

The structure of the Wood-Ljungdahl pathway derived by Jeong et al [21] is shown in Figure 2.1. Additionally, the figure identifies the involved enzymes and the their encoding genes in *E. limosum* KIST 612.

Figure 2.1 provides us with an understanding of the genes involved in the target pathway, along with a brief understanding of how all these genes are utilised by the bacteria. Using the gene names provided in Figure 2.1, the nucleotide sequence for these genes could be located in the annotated genome sequence, published on NCBI https://www.ncbi.nlm.nih.gov/nuccore/CP002273 [33]. The identification of the correct gene nucleotide sequence in the annotated genome can be accomplished in one of two ways.

Firstly, the gene names shown in figure 2.2 can be used to directly search through the genome annotation of KIST 612 for a match of gene annotations.

1 : formate dehydrogenase	**10 : pyruvate oxidoreductase**
2 : formate-THF synthetase	**11 : thiolase**
3 : formyl-THF cyclohydrolase	**12 : 3-hydroxybutyryl-CoA dehydrogenase**
4 : methylene-THF dehydrogenase	**13 : crotonase**
5 : methylene-THF reductase	**14 : butyryl-CoA dehydrogenase**
6 : methyltransferase	**15 : phosphate butyryltransferase**
7 : CO dehydrogenase/acetyl-CoA synthase	**16 : phosphate acetyltransferase**
8 : methyltransferase	**17 : acetate kinase**
9 : corrinoid/iron-sulfur protein	**18 : acetoacetyl-CoA subunit**

Figure 2.2: All of the genes which make up the Wood-Ljungdahl pathway in KIST 612 [21]

Figure 2.2 shows all of the genes involved in the Wood-Ljungdahl pathway in KIST 612, in order from the methyl and carbonyl branches to the butyrate and acetate kinases.

Secondly, each of the gene IDs, shown in Figure 2.3, identified from the pathway model, corresponded directly to a specific gene involved in each step of the pathway. The use of gene IDs provides a more accurate method of gene identification in the annotated genome; as some genes have the same functional annotation. Consequently, this was the preferential route, as incorrect naming of genes and referring to genes as 'hypothetical proteins' can lead to mistakes when identifying the Wood-Ljungdahl genes within the annotated genome.

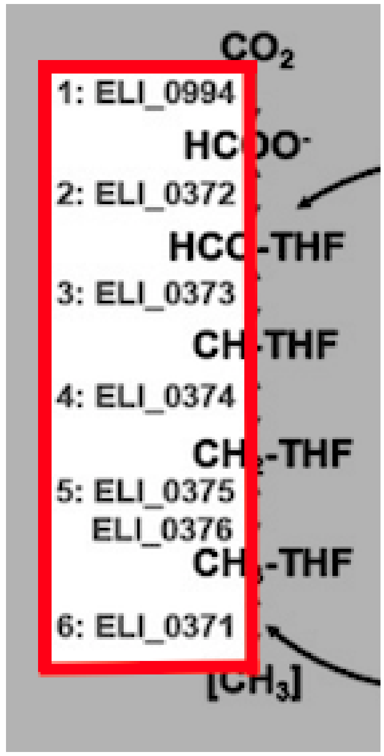

Figure 2.3: The gene IDs from the methyl branch of the KIST 612 Wood-Ljungdahl pathway [21]

Once the genes have been identified on the annotated genome, navigation to the nucleotide sequence can occur via a link from the gene annotation. This opens a page specifically for that gene, that provides information such as name, ID, position in the genome, and nucleotide sequence. The nucleotide sequence was copied into a .txt file and named by putting " >'gene name' " on the first line of the .txt file. This represents the file in a FASTA format and allows the file to be identified by the provided gene name, useable in BLAST analysis software.

After collecting all the nucleotide sequences for the genes and saving them into FASTA files, the BLAST analysis can occur. The BLAST analysis software used was CLCbio from QIAGEN [34], using the CLCbio Genomics Workbench 9 package through an academic license the University of Nottingham own.

Once all of the Wood-Ljungdahl gene FASTA files and the ATCC 8486 full genome sequence, which was created and provided to me by another colleague, have been imported into CLCbio, a BLAST can occur. The BLAST analysis can be started in CLCbio by selecting the BLAST option from Tools>Classical Sequence Analysis>BLAST. It firstly requires to select the gene one wishes to BLAST, then the genome one intends to BLAST the gene against. In this case, it would be the genes from the KIST 612 Wood-Ljungdahl pathway and the ATCC 8486 genome nucleotide sequences. Finally the programme asks to provide parameters which can be changed for the BLAST search. For this study, the default values were used. The BLAST can then be run. When the genes for the complete Wood-Ljungdahl pathway in ATCC 8486 have been identified through the BLAST analysis, they can be used to construct an ATCC 8486 Wood-Ljungdahl pathway model and visualisation of the results. Here, I used the web-based tool ESCHER which allows for the building of a pathway map/model, https://escher.github.io [35]. This tool requires the user to load a .JSON file which contains information about all the reactions and metabolites in the pathway.

In order to create the JSON file for use in ESCHER, a file detailing all the reactions in the Wood-Ljungdahl pathway was created in a Python environment; the procedure for this will be discussed in the structural model chapter. This list of all the reactions involved in the Wood-Ljungdahl pathway saved as a Python file was then run through a specialised script developed at the University of Nottingham, which allows for the conversion of Python files into JSON files.

After the .JSON file has been loaded into ESCHER, the pathway map can be constructed. The process begins by starting a new map on ESCHER and selecting Model>Load COBRA model JSON, loading the reaction data into the programme, as shown in Figure 2.4 .

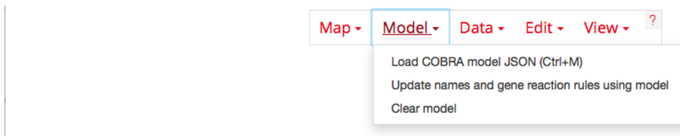

Figure 2.4: How the .JSON file is loaded into ESCHER

Then the reactions can be selected and placed onto the construction area to begin the construction of the pathway. This can be done by selecting the 'Add Reaction' tool from the toolbar and then selecting where the reaction graphic should be placed, as shown in Figure 2.5.

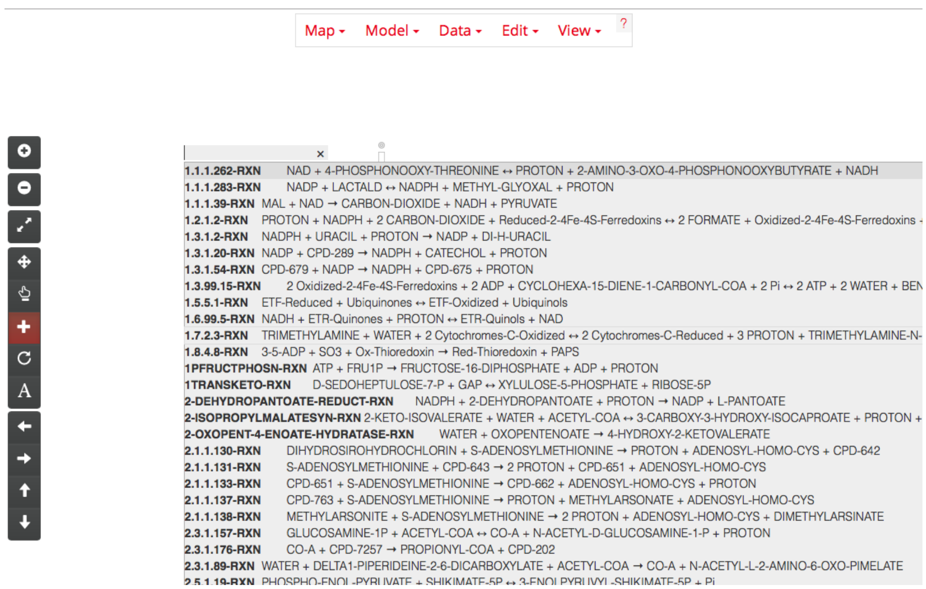

Figure 2.5: How the pathway map is constructed in ESCHER in a reaction by reaction process

The pathway map can continue to be constructed by selecting the metabolite which is involved in the next reaction in the pathway. Once selected, a list of the reactions in which this specific metabolite is involved in is provided, selecting whichever one is the next reaction in the Wood-Ljungdahl pathway, as shown in Figure 2.6.

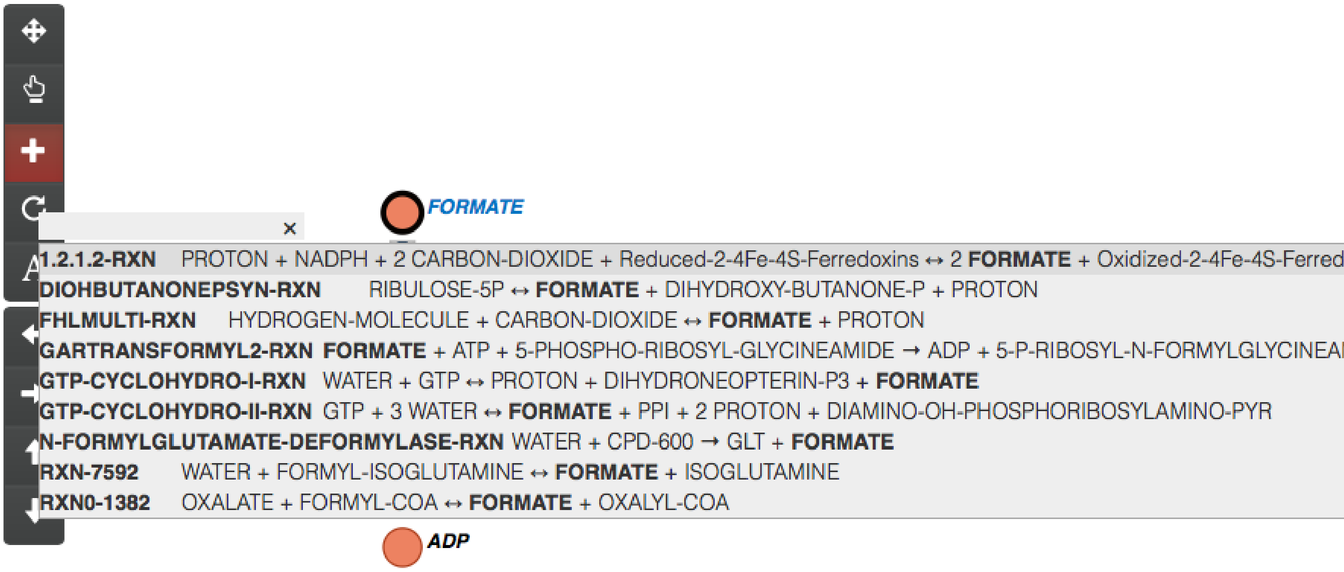

Figure 2.6: How the step by step process of adding reactions into the ESCHER map occurs

The process of developing the ESCHER map ensues until the pathway is complete, thus producing a full Wood-Ljungdahl pathway model for *E. limosum* ATCC 8486. Once the pathway map has been created, it can then be saved as an image and a JSON file. Furthermore, ESCHER can be used to visualise flux changes between alternative fluxes in the pathway for a more in depth analysis into the mechanisms of the pathway, which will be investigated during the structural modeling chapter.

Furthermore, an simplified version of the graph was also produced using Inkscape. Inkscape is a professional vector graphic design programme, which can be used to produce drawings of a professional nature; in this case a pathway map. This process required the drawing of arrows representing the Wood-Ljungdahl pathway produced in ESCHER, but allowed for a more flexible representation of the arrows, including curvature and size, which are not as flexible customisations in ESCHER. The reactions and main metabolites in the pathway were annotated with text boxes.

2.3 Results

CLCbio was used to conduct the BLAST analysis, being able to effectively and simply provide an output detailing the similarity between 2 input sequences: a gene nucleotide sequence and a full genome sequence. CLCbio not only provides an ability to conduct BLAST alignments but has a variety of genome analysis tools such as reverse translation, nucleotide to protein translation, protein BLAST, small alignments, molecular structure viewer and protein analysis among others. BLAST analysis and nucleotide to protein translation were used during this chapter, with nucleotide to protein translation converting DNA nucleotide sequences to amino acid sequences; as nucleotide sequences which appear different may be more similar when converted to an amino acid sequence, due to different codons producing the same amino acid. The BLAST analysis/protein BLAST analysis tools were both used to compare similarities of sequences and to identify their presence within a genome, the process of which is described in the methodology.

BLAST analysis is a tool which can be used to confirm whether a certain gene is located within the genome of a different bacteria/strain, through the matching of nucleotides from one sequence to an other and identifying the match with the highest identity percentage. The genes garnered from the KIST 612 pathway were input into the BLAST, with ATCC 8486 set as the target genome sequence. The BLAST finds the closest matches of the gene's nucleotide sequence within the genome sequence, providing a percentage and scoring system to determine the level of identity between the sequences. The BLAST match is assessed by three different values: percentage of identity, percentage of similarity and E-Value. The percentage identity shows the percentage of nucleotides which matched between the 2 sequences, the percentage similarity shows the similarity in the amino acid structure between the 2 sequences, determining whether the enzyme encoded for by these genes is the same or different. The percentage of similarity provides the most accurate and useful value in regards to my objective to locate the genes involved in the Wood-Ljungdahl pathway in ATCC 8486, as it is only important that the same enzyme is encoded by the gene. The comparison of two nucleotide sequences can result in a lower identity but still encode the same enzyme, because the resulting amino acid sequence shows much higher similarity. This was useful for identifying the exact sequence which encodes for the enzymes responsible for methanol metabolism, in which there were numerous possible sequences which all had similar similarities; however when translating to an amino acid sequence, one sequence presented a higher similarity than the rest.

When the BLAST was conducted using the Wood-Ljungdahl genes identified from the KIST 612 strain to understand whether they were present in the ATCC 8486, nearly all the genes which were BLASTed showed high percentage similarity. The genes which possessed a similarity of over 95% can be assumed to be a exact match, with the main function of the enzyme kept intact despite a 5% difference in the amino acid structure.

2.4 Methyl Branch

Descending through the methyl branch, the first point of interest established involved formyl-tetrahydrofolate cyclohydrolase and methylene tetrahydrofolate dehydrogenase, which were annotated as two separate enzymes and two separate genes in [21]. However, further investigation using BRENDA (https://www.brenda-enzymes.org/) a comprehensive enzyme database, was able to show that in prokaryotes such as *E. limosum*, formyl-tetrahydrofolate and methylene tetrahydrofolate were present as part of a bifunctional enzyme encoded by the same gene. This was then reiterated when the BLAST analysis was conducted, where the formyl-tetrahydrofolate cyclohydrolase and methylene tetrahydrofolate dehydrogenase sequences in the strain KIST 612 were BLAST against the ATCC 8486 genome. As a result, it could be confirmed that both function were domains of the same bifunctional gene. Thus in contrast to the annotation provided for the KIST 612 strain, the sequence analysis conducted in this project suggests that in *E. limosum* ATCC 8486 a single gene encodes for a bifunctional enzyme facilitating the two reactions.

The next point of interest in the methyl branch was the methylene tetrahydrofolate reductase. The process revealed that there were 2 different genes, annotated as methylene tetrahydrofolate reductase C terminal and 5'10 methylene tetrahydrofolate reductase, that could facilitate this reaction. However, the two enzymes differ in their cofactors. Whereas 5'10 methylene tetrahydrofolate reductase reduces/oxidises ferredoxin, methylene tetrahydrofolate reductase C-terminal is NAD(P)H dependent. These were both found to be present in ATCC 8486, although it appears that 5'10 methylene tetrahydrofolate reductase is preferred, when considering the presence of ferredoxin molecules and balancing of NAD(P)H in the Wood-Ljungdah pathway. However, it is not the only enzyme used to convert methylene tetrahydrofolate to 5-methyl tetrahydrofolate.

The next step in the methyl branch describes a 'methyltransferase'; gene 6 in Figure 2.2, which appears to be involved with another methyltransferase. Methyltransferase transfers the methyl group to 5-methyl tetrahydrofolate, while the methyltransferase labelled as gene number 8 is used to transfer a methyl group to corrinoid proteins generating methylated corrinoid proteins. There is confusion as to what the purpose is of the gene 6 'methyltransferase' in the pathway. As previously stated, this was only described as a methyltransferase in the Jeong et al [21] paper; the BLAST analysis and matching to the corresponding annotation confirmed it as a dihydropteroate synthase. Using BRENDA and MetaCyc (https://metacyc.org/); a highly curated web database for metabolic pathways, enzymes, metabolites and reactions, both describe this enzyme as 'participating in the biosynthesis pathway for folate production', showing less relation or involvement to the Wood-Ljungdahl pathway. Although, MetaCyc does further describe the enzyme to exist in a variety of types of multifunctional proteins. This combined with the knowledge that this enzyme is involved with the removal of a methyl group from (7,8-dihydropterin-6-yl)methyl diphosphate, implies a potential relation. However, this link holds little substance in the way of evidence and it can most likely be assumed that this in not a prevalent enzyme in the Wood-Ljungdahl pathway of ATCC 8486. Another suggestion could be that this is an implied methanol metabolism pathway, although the presence of a methanol metabolism branch is not provided in this pathway, it could suggest the placement of where a methanol conversion branch could be involved. The introduction of the methanol branch is an area of investigation I delve into later.

Another 'methyltansferase' was identified in the KIST 612 Wood-Ljungdahl pathway, as previously stated. This was identified as gene number 8 on Figure 2.2. Figure 2.1 suggests this 'methyltrasnferase' is involved in transferring the methyl group from 5-methyl tetrahydrofolate to corrinoid iron sulfur protein producing methylated corrinoid iron sulfur protein. A comparison to the ATCC 8486 genome results show that this methyltransferase is an 5-methyltetrahydrofolate:corrinoid iron sulfur protein Co-methyltransferase. It confirms that the function of this methyltransferase is as previous stated.

2.5 Carbonyl Branch

The carbonyl branch can be defined as the carbon monoxide dehydrogenase and the acetyl-CoA synthase enzymes; which were clearly identified and matched in our BLAST search. It is the combination of the methyl and carbonyl branch that is a potential grey area. When a BLAST search was performed of the Corrinoid iron sulfur protein genes identified in the KIST 612 pathway against the ATCC 8486 genome, they matched with 2 '5-methyltetrahydrosarcinapterin:corrinoid iron sulfur protein Co-methyltransferase'. When investigating these '5-methyltetrahydrosarcinapterin:corrinoid iron sulfur protein Co-methyltransferase' enzymes using BRENDA and MetaCyc, research conducted on *Moorella thermoacetica* suggested that the carbonyl branch was part of one complex containing all the enzymes known as acetyl-CoA decarboxylase/synthase complex. It appeared that this may be the case for *E. limosum*, due to all of these enzyme's neighbouring proximities within the genome. However research conducted on acetogens which is present on BRENDA implied that, despite the similarities in how acetyl-CoA is formed in acetogens, several individual enzymes known to be found in this complex were not present in close proximity in acetogen genomes [36].

Despite previous studies indicating that the enzymes in the carbonyl branch of acetogens are not combined in a complex, the system of reactions and key enzymes work nearly identically to how the acetyl-CoA decarboxylase/synthase complex would work.

The role of the 5-methyltetrahydrosarcinapterin:corrinoid iron sulfur protein Co-methyltransferase, which is represented as a corrinoid/iron-sulfur protein in the Jeong et al [21] paper, was briefly described in the acetogen *Clostridium thermoaceticum* as transferring a methyl group to the carbon monoxide dehydrogenase/acetyl-CoA synthase, which uses the methyl group as part of the acetyl-CoA production process [37].

The literature remains vague in regards to the use of 5-methyltetrahydrosarcinapterin:corrinoid iron sulfur protein Co-methyltransferase in the Wood-Ljungdahl pathway, other than its involvement in acetyl-CoA production through affiliation with carbon monoxide dehydrogenase and acetyl-CoA synthase. The overall reaction produces a Co(I) corrinoid iron sulfur protein and 5-methyltetrahydrosarcinapterin from methylated-corrinoid protein and tetrahydrosarcinapterin, with KIST 612 paper [21] suggesting that the methyl group from 5-methyltetrahydroasarcinapterin may then be transferred to formyl-tetrahydrofolate when generating 5'10 methyl-tetrahydrofolate, although again there is little in way of literature to support this [21].

2.6 Acetate formation

The next stage in the pathway after acetyl-CoA production is split into the different products, is acetate formation. It is usually the most prevalent product. The acetate branch was depicted to have 2 different routes of acetate production presented as a cycle as seen in Figure 2.1. The KIST 612 acetate branch seamlessly transfers across to ATCC 8486. However one difference lies in the acetoacetyl-CoA subunit (gene number 18) described in Figure 2.1 and 2.2: in the KIST 612 pathway it is proposed that there is a reversible reaction in which acetyl-CoA can be converted to acetate and vice versa in just one reaction. The initial 'standard' route of acetate production uses a phosphate acetyltransferase to produce acetyl-phosphate from acetyl-CoA, this is then used to produce acetate and ATP through an acetate kinase. Through the BLAST analysis, it was shown that this acetate kinase in KIST 612 matched to a gene(s) in ATCC 8486 which provides a similar function called propionate CoA-transferase and acetate CoA-transferase. The propionate CoA-transferase catalyses the reversible reaction of acetyl-CoA + propanoate to acetate and propanoyl-CoA. It was noted by the BRENDA database that it also uses butyrate as a acceptor of the CoA moiety replacing propanoate. This reaction which uses butyrate instead of propanoate is known as acetate CoA-transferase, adding another dimension of production formation to the Wood-Ljungdahl pathway in *E. limosum* ATCC 8486, through the interconnection of the acetate and butyrate branches.

Acetyl-CoA + Propranoate = Acetate + Propanoyl CoA

This shows another deviation between the proposed route of acetate formation in KIST 612 compared to ATCC 8486, although the route in which acetyl-CoA is converted to acetyl-phosphate, which is then converted by an acetate kinase to acetate, remains identical.

2.7 Butyrate formation

Butyrate/butyryl-CoA formation, much like acetate formation, was found to be very similar between the KIST 612 and the ATCC 8486 genomes. The same construction of the branches are apparent from the BLAST results, with 3 enzymes used to convert acetyl-CoA to butyryl-CoA. The genes in this branch aligned perfectly until butyryl CoA-dehydrogenase. However, there appeared to be a mislabeling of the crotonase and 3-hydroxybutyryl-CoA dehydrogenase, as their annotations appear to be interchanged [21]. When BLASTed against the ATCC 8486 and assessed against our corresponding annotated genes, crotonase was labeled as 3-hydroxybutyryl-CoA dehydrogenase and vice versa.

The interesting changes in the butyrate branch came after butyryl-CoA dehydrogenase, where the KIST 612 pathway implies butyrate formation occurs via a phosphate butyryltransferase and butyrate kinase. The gene encoding phosphate butyryltransferase was found to be present in the genome of ATCC 8486, meaning that butyrate phosphate can be produced. However, there is no butyrate kinase in the genome of *E. limosum* ATCC 8486. The Jeong et al [21] paper did not provide any evidence of a butyrate kinase present in the genome of KIST 612, but established that as butyrate phosphate was produced. Thus, it was probable that there would be a butyrate kinase available to convert this into butyrate.

After identifying this discrepancy, an objective became to confirm whether or not a butyrate kinase was present in *E. limosum.* Using Uniprot [38]; a web database of proteins and their sequences, a sample of FASTA nucleotide sequences of butyrate kinase was selected from different acetogens, some of which were known to be similar to *E. limosum* such as *A. woodii* and *Clostridium ljungdahlii.* A sample of 16 butyrate kinase sequences was prepared and BLASTed against the ATCC 8486 genome. Results showed that not one sequence had a remote match to the ATCC 8486 genome, with all butyrate kinase sequences having below 60% similarity on all target matches within ATCC 8486. This confirms that butyrate kinase is not present in *E. limosum* ATCC 8486, meaning that butyrate production is reliant on the acetate-CoA transferase, which transfers the CoA moiety from butyryl-CoA to acetate producing butyrate and acetyl-CoA.

2.8 Pyruvate formation

Finally, the pyruvate 'branch' which is composed of one reaction described as a pyruvate oxidoreductase in the KIST 612 pathway, was BLAST and located in the ATCC 8486 genome. This was confirmed to be described as a pyruvate synthase which catalyses a reaction which converts acetyl-CoA and CO_2 to pyruvate and oxidised ferredoxin.

2.9 Methanol formation

A colleague working on growth of *E.limosum* in the laboratory, provided information stating that it is able to grow on methanol, which has been previously reported in a 1992 paper [39]. Furthermore, laboratory results provided evidence that methanol growth can lead to the production of butyrate and acetate, assuring the presence of a methanol metabolism pathway in *E. limosum.* Research conducted by Ragsdale et al [23] on the Wood-Ljungdahl pathway stipulated that acetogenic bacteria are able to grow on a series of methyl donors including methanol. It is broken down by a methyltransferase system and the methyl group is transferred into the Wood-Ljungdahl pathway [23]. The process of methanol metabolism described by Ragsdale et al [23] is based around three enzymes: MtaA, MtaB, and MtaC, which are assumed to be encoded by 3 different genes. Overall, the methyl group from methanol is transferred to a corrinoid binding protein, becoming methylated. The methylated corrinoid protein then transfers the methyl group to a methyl acceptor, which is thought to be the acetyl-CoA synthase. Ragsdale et al did not explain the mechanism of methanol metabolism using the 3 assumed enzymes in the text, however a diagrams was used to show the general process.

The detailed process including the involved of MtaA, MtaB, and MtaC was displayed in an image, albeit not explained within the text.

Using MetaCyc and BRENDA I was able to develop an understanding of the roles MtaA, MtaB, and MtaC have in methanol metabolism. MtaA and MtaB were both genes encoding enzymes, whose reactions were listed on MetaCyc and BRENDA.

MtaA: Methyl-CoM(Coenzyme M) + Hydrogen ion + Corrinoid protein = CoM(Coenzyme M) + Methylated Corrinoid Protein

MtaB: Methanol + CoM(Coenzyme M) − > Methyl-CoM(Coenzyme M) + Water

However, MtaC was not linked to any enzyme or reaction on either of the MetaCyc and BRENDA databases; this indicated that MtaC was not an enzyme but perhaps a protein involved in the process. Studies detailing methanol reduction in *Methanosarcina barkeri* labeled MtaC as the protein to which the methyl group from methanol is transferred by MtaB [40]. When viewed in regards to the reaction of MtaB, it suggests that MtaC is a coeznyme M, which is used as an initial methyl group donor before the methyl group is transferred onto the next donor, which is the corrinoid protein. The coenzyme M is then recycled after this reaction, to be used as the initial methyl donor again.

A potential match for MtaA and MtaB were able to be located via the genome annotation of ATCC 8486 both as MtaA and MtaB. However, MtaA was labeled primarily as an 'Uroporphyinogen decarboxylase', leading to initial confusion as to whether this was in fact the MtaA gene or a mislabeling in the annotation as its corresponding reaction number - EC 4.1.1.37 - which has no relevance to the function of MtaA. After consulting the enzyme database BRENDA, I was able to establish a link between 'Uroporphyrinogen decarboxylase' and MtaA. MtaA is a methyltransferase which catalyses the transfer of the methyl group from methanol to a methyl donor, while Uroporphyinogen decarboxylase is identified as a sequential decarboxylation of four acetate group of uroporphyinogen-III to yield coproporphyrinogen-III, which does not show any surface level similarity in function.

However, upon looking on UniProt to gain further understanding about the functions of MtaA and Uroporphyrinogen decarboxylase and whether there was a possible link, it was revealed that studies on *Methanosarcina barkeri* had labeled the molecular function of the MtaA as both a uroporphyinogen decarboxylase and a methyl-Co(III) methanol-specific corrinoid protein:coenzyme M methyltransferase [38]. The indication could be that either uroporphyrinogen decarboxylase and MtaA are part of the same gene which is bifunctional or when methyltransferase activity is impaired in MtaA, due to mutations in the C-terminal end, this could mean that the enzyme then functions as uroporphyrinogen decarboxylase [41]. However, results in these same studies described that in the presence of MtaC a nitrogen/oxygen ligand was replaced by sulfur in MtaA and when MtaC is not present zinc is coordinated by a sulfur ligand and three oxygen or nitrogen ligands. Thus it may be implied that under the presence of MtaC the function of MtaA may be a methyltransferase, but without the presence of MtaC the function of MtaA may be a uroporphyinogen decarboxylase.

After developing an understanding that despite being annotated as a Uroporphyinogen decarboxylase, the enzyme appears to be bifunctional and possesses another molecular function as a methyltransferase (MtaA), confirming the presence of MtaA in *E.limosum*; despite it being primarily labelled as a Uroporphyinogen decarboxylase. A BLAST was conducted on the the two genes assumed to be MtaA and MtaB in the ATCC 8486 genome, against the NCBI genome database. The results confirmed that MtaB was identified correctly in numerous other acetogens, such as *A.woodii* and *C.autoethanogenum*. While a BLAST search of the Uroporphyrinogen decarboxylase showed matches to Uroporphyrinogen decarboxylases in *A.woodii* but MtaA methyltransferase in *E.limosum* SA11. Interestingly, it is annotated as a hypothetical protein in the *E.limosum* KIST 612 match. However, the BLAST match in SA11 represented a 92% nucleotide identity, while the match with *A.woodii* displayed a 72% identity, which indicated that the Uroporphyrinogen decarboxylase in ATCC 8486 had a similar protein structure as the MtaA methyltransferase in SA11, with a high probability of shared functional relationship.

Thereafter displaying the presence of the MtaA and MtaB genes in the ATCC 8486 genome and developing an understanding of the role of MtaC, which was not required to be confirmed in the genome within the lines of the project objectives.

2.10 Escher Map

A ESCHER pathway map was then constructed using the genes which had been confirmed to be present within *E.limosum* ATCC 8486. Overall, the pathway constructed has not been previously developed into a full pathway in *E.limosum* and does not only include the well documented methyl, carbonyl, acetate, butyrate and pyruvate branches/formation pathways, but also includes a methanol metabolising branch. Here a collective Wood-Ljungdahl pathway can be acknowledged, which shows a more accurate and higher level of detail than the Wood-Ljungdahl pathway which was previous provided for the KIST 612 strain.

The ESCHER tool provides further detail of the pathway, but not only displaying the main molecules in each reaction but all the molecules involved, detailing a much more accurate assessment of the Wood-Ljungdahl pathway. Furthermore, it is compatible with with Flux Balance Analysis results which will be used in the structural modelling chapter. The reaction data can be loaded into the ESCHER map providing a flux map for analysis. The same can be done with gene and metabolite data alongside specific settings, which allow for the visualisation of these results and the map to be customised.

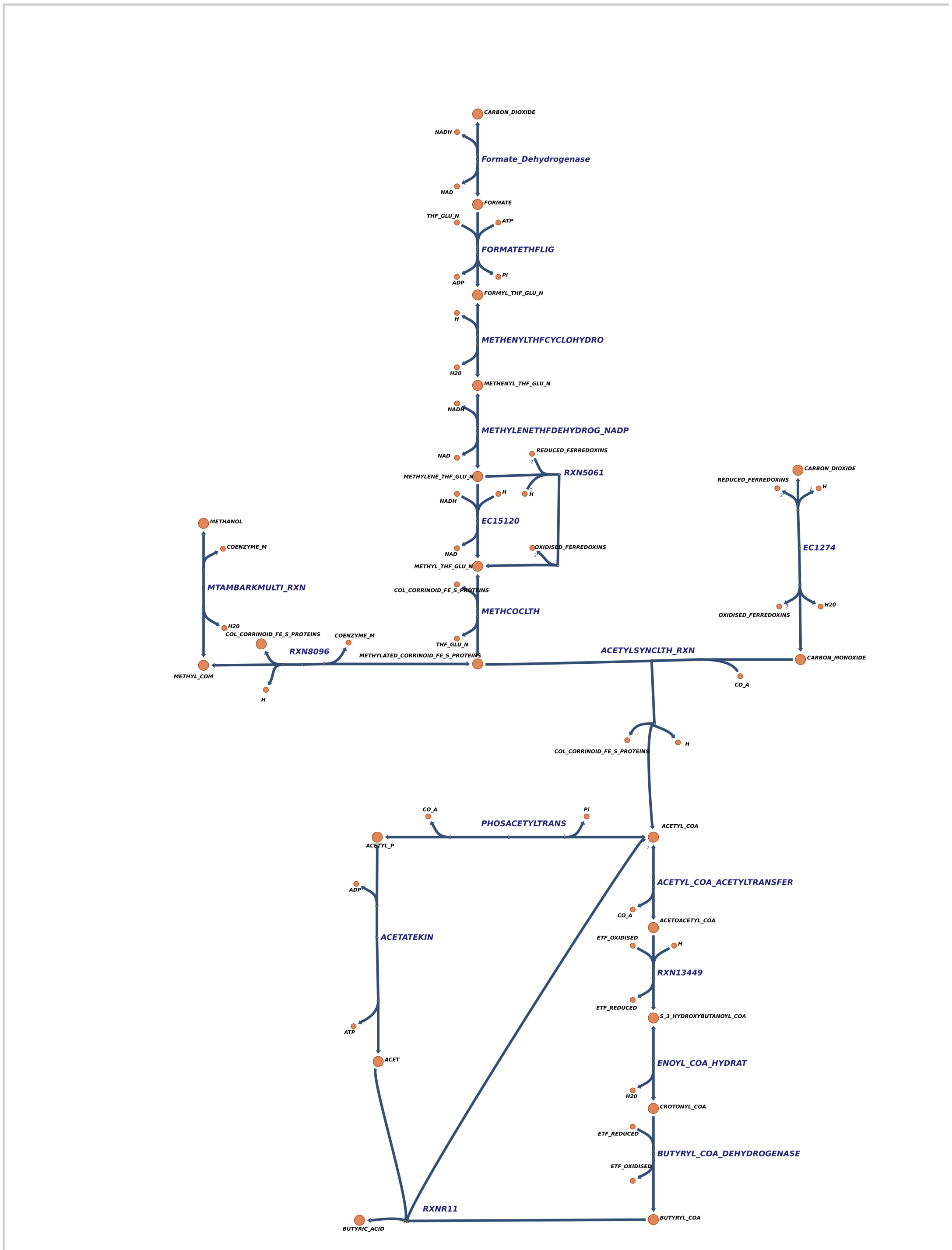

Figure 2.7: A pathway map displaying the Wood-Ljungdahl pathway present in *E.limosum* ATCC 8486

The figure showcases the Wood-Ljungdahl pathway in *E.limosum* ATCC 8486 presented in ESCHER maps, which has been confirmed throughout this chapter and applied in contrast to the

Wood-Ljungdahl pathway described in *E.limosum* KIST 612. The major distinguishing features are that the methanol metabolism pathway has been constructed and added, represented by the reactions EC 2.1.1.90 and EC 2.1.1.246. Furthermore, the reaction used to produce butyrate can be established as EC 2.8.3.8/2.8.3.1; which is an Acetate CoA transferase, opposed to the butyrate kinase suggested in Jeong et al [21]. This reaction also acts to convert acetate to acetyl-CoA, which was assumed to be a separate reaction in the KIST 612 Wood-Ljungdahl pathway, but is shown here to be part of the same reaction as butyrate production. Overall, the methyl, carbonoyl, pyruvate and the rest of the acetate and butyrate branches remain the same, with more clarification on the enzymes and reactions involved in each branch.

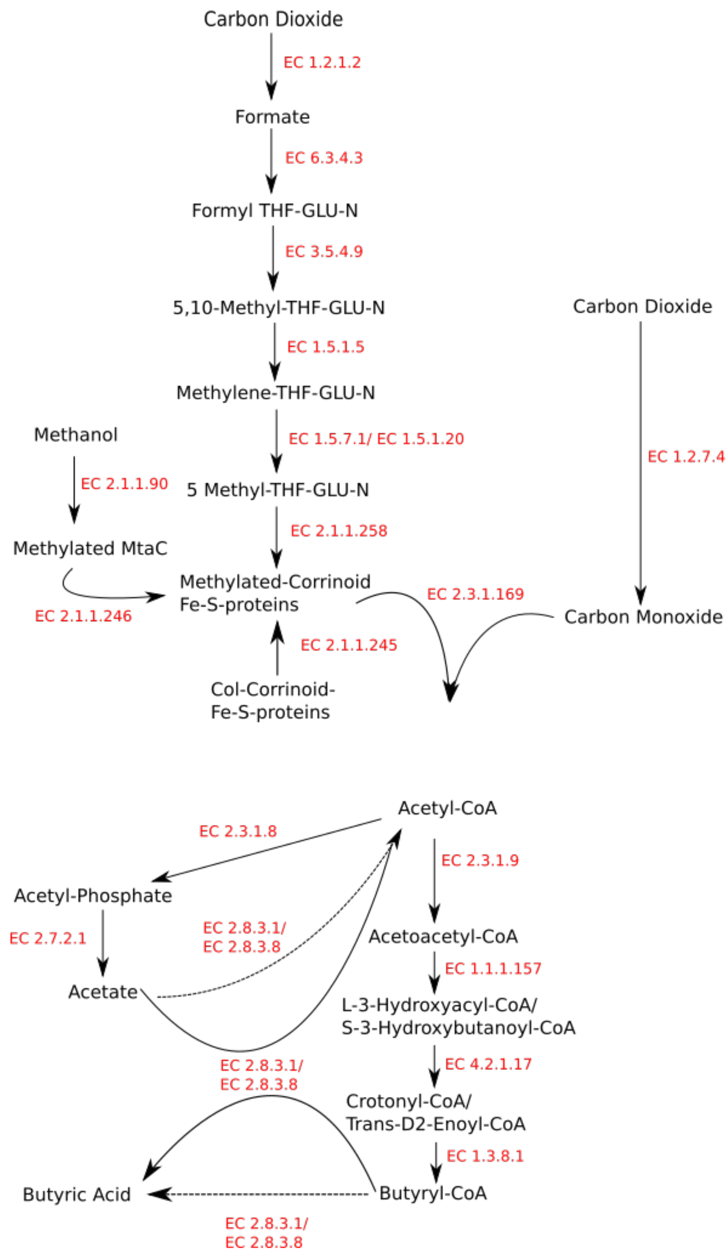

Figure 2.8: A simplified version of the Wood-Ljungdahl pathway in *E.limosum* ATCC 8486

Figure 2.8 shows a simplified version of the ESCHER maps pathway. This only contains the key substrates and products for each reaction, while being able to provide a more flexible design approach to make the pathway easier to understand.

In order to provide a more simplified version of the Wood-Ljungdahl pathway produced in ESCHER, Inkscape - a graphics editor [42]. This allowed me to produce a simplified version of the Wood-Ljungdahl pathway which only uses the main metabolites/molecules in the pathway. Furthermore, using Inkscape allows for more versatility in the structure of the pathway, meaning that the visualisation of the pathway becomes more accessible to someone looking to understand the basics of the Wood-Ljungdahl pathway. Each reaction in the pathway can be labelled by the EC number opposed to the reaction name; which is used in the ESCHER map. The EC number represents a universal identification of the reaction, however more than one reaction may have the same EC number. The reaction names which are used to label the reactions on the ESCHER pathway are more specific to the exact reaction they are labelling. Although, they are only used to identify the reactions in the MetaCyc database.

Chapter 3

Genome Alignments

3.1 Introduction

Genomic alignments are one of the primary bioinformatical tools used when analysing genomic data. It allows for a full comparison between the genomes of interest, allowing a visualisation of the changes between multiple genomes on both a large and small scale. It can be used to mark and analyse the genetic regions of interest observed during the multiple genome alignment; these may be structural, organisational, or conservational, as numerous changes are expected to occur between strains and similar bacterias due to the evolutionary process. While small scale changes tend to be nucleotide mutations such as deletions, substitutions and insertions, the larger scale changes manifest in the form of gene duplication, loss or gain, inversions, translocations and genome rearrangements.

The focus of the genome alignment in this project is to determine and understand any differences between 3 different strains of *E. limosum* (ATCC 8486, KIST 612, SA11), with a focus towards the genes involved in the Wood-Ljungdahl pathway. Any differences in the Wood-Ljungdahl genes which can be established between the ATCC 8486 and KIST 612 strains, may provide crucial information to the supposed differences experienced between the strains in a laboratory environment. Furthermore, evidence can be provided to show how conserved the Wood-Ljungdahl pathway is across the two different strains, or whether rearrangement has occurred.

There are numerous computer programs and software, which are freely available and can be used to carry out sequence/genome alignments. However, many of these programs differ tremendously in their abilities to perform alignments of certain length, or file size. Moreover, the ability of these programs to conduct a multiple whole genome alignment can also be temperamental, inconsistent, and troublesome. Ultimately, Mauve was selected as the preferred software to conduct the alignment, due to the specialised needs of the project, while being encompassed in a software package which allows for easy input, alignments and provides visual representation of the data. Overall, this meant that it contained all the requirements for this particular alignment, especially its ability to align multiple sequences from genomes to genes. It was tested for use against a range of other alignment softwares including: Peacan (https://github.com/benedictpaten/pecan/), Mugsy (http://mugsy.sourceforge.net/), Mega (http://www.megasoftware.net/) and MUMmer (http://mummer.sourceforge. Furthermore, Mauve was not only able to display the alignment with the most ease, but also with the most visually understandable representation, without losing any quality of results.

No previous study has completed a genomic alignment between these *E. limosum* strains, in particular the Wood-Ljungdahl genes within these strains. This will provide a development in the understanding of the potential differences between these genomes, along with possible insight as to what effect this may have on the bacteria and function of the Wood-Ljungdahl pathway. Due to ATCC 8486 and KIST 612 being different strains of the same bacteria, it would be assumed that genomic regions between both genomes are highly conserved with all regions present in both genomes.

3.2 Literature Review

Genome alignments are one of the main tools in genome analysis. It is used time after time to show areas of similarity or difference, which may be due to a structural, functional, or evolutionary relationships between the genomes. A brief literature review aiming to review previous studies that have used genome alignments to help classify and gain insight into acetogens. As well as those which have used Mauve to conduct the genomic alignments and the analysis these studies concluded.

One study [43] which was investigating the genome wide comparison between *C. autoethanogenum* and *C. ljungdahlii* used a genome alignment between these two acetogens. It highlighted differences which might lead to different phenotypes; in regards to more efficient ethanol and 2,3-BDO production [43]. The alignment was produced using the BowTie 2 alignment program. The results showed that the two species are remarkably similar, implying that the phentotypic differences mentioned are not a result of changes in the genome sequences [43]. Only 217 homologs representing 5% of their genome, were found in one of the species and not the other. Despite their genomic similarity the genomic similarities between the bacteria, the strains were found to behave very differently at a transcriptional level [43]. However, this paper only uses the genome alignment to provide a brief overview of similarity between the genomes, as opposed to identifying the homologs between specific genes, which is of particular interest in regards to determining the functionality of genes involved in the Wood-Ljungdahl pathway in *E. limosum*.

Another study, similar to the previous, focuses on *C. ljungdahlii*. A genome alignment was used to establish homologs to *C. ljungdahlii* genes by comparing the genomes between *C. acetobutylicum*, *C. thermocellum*, and *C. beijerenckii*. This study used a Smith-Waterman alignment algorithm. Genes with an amino acid sequence similarity of more than 60% were identified as homologs in *C. ljungdahlii*. These genes were then matched to their corresponding reactions and the reactions were then compiled into a model. The model was then used to provide insights into the metabolic response of *C. ljungdahlii* [44]. The use of genomic alignment tools to establish homologs across different *Clostridium* species has similarities to this project, whereby a BLAST analysis is used to identify the presence of Wood-Ljungdahl pathway genes across different *E. limosum* strains. Furthermore, the Smith-Waterman alignment algorithm used has limitations, not making it ideal for any large scale genomic alignment.

Other studies which have involved the alignments of acetogenic sequence appear to be mostly using alignment software/programs which allow for local sequence alignments opposed to the whole genome alignments, the latter being the objective of this project. This can be further iterated by another study, which used a sequence alignment of genes involved in energy conservation, to show the homology and conservation of these genes through different acetogens [45].

On the other hand, many different studies have used Mauve to perform whole genome alignments. One particular paper [46] used Mauve to conduct a multiple genome alignment between 3 strains of the *Neisseria gonorrhoeae*. The Mauve alignment was able to show large similarities, as well as identifying the largest variant between the strains, which was the 'Gonococcal Genetic Island'. Moreover, Mauve was used alongside a dot-plot matrix (Another form of genome alignment.) to present evidence of genetic properties which may be causing chromosomal rearrangement and homologous recombination, such as prophages [46].

Mauve was also used in another paper to show comparisons between regions which contained clusters of adhesins between two strains of *Candida glabrata*, allowing for the identification of 61 novel adhesins in one of the genomes [**?**].

While other papers again have used Mauve to identify homologous regions, numbers of inversions, insertions, deletions and SNPs. Overall, the use of Mauve to provide a whole genome alignment and show the comparison in similarity/difference in genes between the multiple genomes, shows why Mauve is the preferable software choice when considering the aim and expected results of this process. In addition, the Graphical User Interface (GUI) provided in the Mauve software also means that the results are able to be displayed clearly, more so than other GUI's used by other alignment software/programs.

3.3 Alignment method

The progressive Mauve algorithm from the Mauve package was used to undertake the alignments. The progressive Mauve provides a different approach to the algorithm used by the original Mauve software. The progressive Mauve algorithm allows for higher accuracy, higher sensitivity in conserved regions and solves the issue of that some bacterial genomes exhibit greater rearrangements than others by having more than a single choice of scoring penalty [47].

Mauve can be installed and used on various operating systems including Linux, Mac OS, and Windows, and is thus available to be compiled from source code or downloaded through an installer. However, with Windows and Mac OS being the preferable operating systems, the installation could occur with relative ease, through the Mauve website (http://darlinglab.org) [47]. It simply requires the software package to be downloaded from the website and then installed through MAC OS/Windows in-house installer, while in a Linux environment it would need to be manually compiled. In order to download and run Mauve, the system requires a 1.4 or later JAVA version installed on the computer. Once running, Mauve has a GUI which allows for ease of use, allowing for input files to be quickly selected for alignment and results shown using Mauve's own alignment display viewer. If the user is using large amounts of inputs and has strict parameters or wants to do a large scale contig reorder, it would be advised to use command lines in a terminal opposed to using the GUI. Other than this, the GUI is the preferable method of use.

Despite using Mauve in this chapter to coordinate the alignment, many other aligners were tested to find the optimal program to conduct this analysis but were unable to under the requirements of the alignment.

Mugsy, which is a linux-based multiple genome aligner, was compiled and installed onto a Linux server, which could be accessed to conduct the alignment. However, problems arose with Mugsy's ability to align multiple genomes and smaller sequences, with numerous problems needing to be troubleshooted resulting in the corruption of the alignment files, becoming a very time consuming process. Moreover, Mugsy uses an external viewer to display the alignment, which was very basic and provided none of the specialised analytical tools of Mauve's viewer. Other programs such as Clustal Omega, MUMmer, T-Coffee, and CACTUS were unable to manage inputs files as large as required. In addition, the CACTUS package was having issues being compiled onto the Linux server. Through discussions with developers of these softwares and other colleagues, the best suite of recourse was to use Mauve.

The dataset used in the alignment were in a FASTA file format, which is a format containing the nucleotide bases of the genome/sequences. The genomes used were those of the 3 strains of *E. limosum* (ATCC 8486, KIST 612, SA11). The ATCC 8486 strain was sequenced at the University of Nottingham using a combination of PacBio and Illumia reads, while the KIST 612 and SA11 nucleotide sequences were taken from the NCBI database (https://www.ncbi.nlm.nih.gov). Mauve creates numerous different output files which are used to analyse the alignments between the genomes. There is the .Mauve alignment file which contains the general alignment in a format which can be viewed in Mauve's in program viewer. A .backbone file is also produced, which contains the regions within the alignment where conservation has occurred in subsets of the genomes, while identifying possible islands of non conservation. The guide tree file and the bbcols files are used to show phylogenetic similarities between sequences in the newick tree format and show the alignment columns which are predicted to be part of a larger conserved region. Moreover, if extra analysis is conducted on permutations, orthologs, and SNPs, the output files created are tab-delimited text files.

In order to produce the alignment, the new progressive Mauve uses a novel alignment objective score known as the sum of pairs breakpoint score. This allows for the accurate detection of rearrangement breakpoints when genomes have an unequal gene content. A probabilistic alignment filtering method is also applied to remove erroneous alignments of unrelated sequences, known to occur in many other genome alignments. New methods of quantifying genome alignment accuracy based on the quality of rearrangement breakpoint measurement predictions and indel predictions, which is present in the new progressive Mauve. Results with higher accuracy are further attained through establishing 'biologically feasible' amounts of genome rearrangements, segmental gain and loss, thus preventing any improbable results.

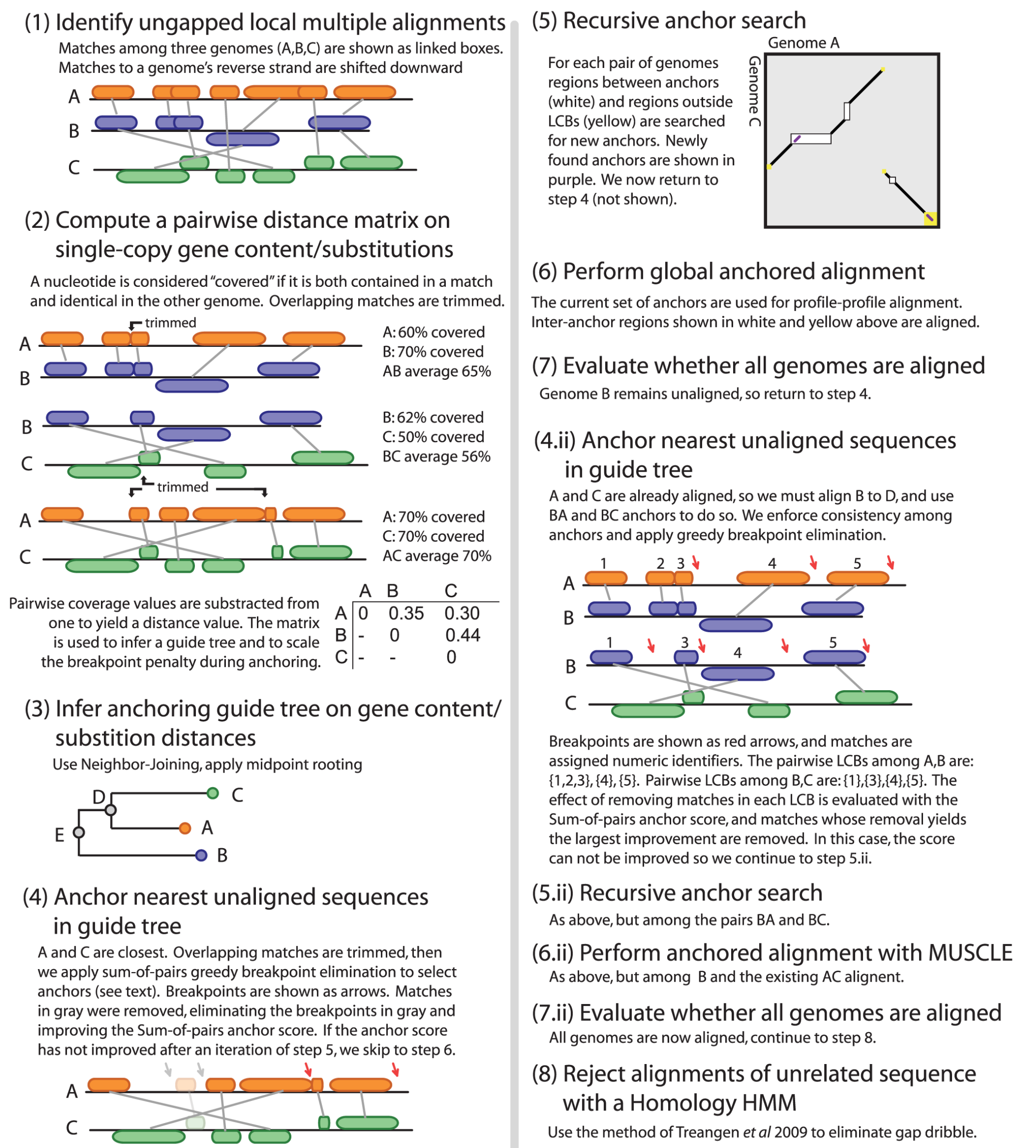

Figure 3.1: An overview of the alignment algorithm used in Mauve [47]

Figure 3.1 shows an overview of how the alignment algorithm in Mauve is used to generate the alignment. The image provides an insight into the 8 steps used by Mauve to provide an accurate multiple genome alignment. The process begins by identifying local ungapped alignments between all of the genomes, which is then reaffirmed by conducting every combination of pairwise alignments between the genomes to confirm the alignments. Anchors, which are areas of highly identical sequence matches, between the genomes are located. This process also removes breakpoints and determines a sum of pairs anchor score. Anchor searches and recursive anchor searches continue until the anchor score does not increase anymore, suggesting that all anchors present have been located between each combination of genome pairs and evaluated. The process proceeds with an alignment evaluation occurs, to conclude whether all genomes are aligned and a final rejection of alignments of unrelated sequences is conducted using a homology hidden markov model.

Parameters can then be applied allowing for personal preferences to be set to alter aspects of Mauve's alignment. A custom or default seed weight can be set to control the identification of anchors. A lower seed weight is used when a large number of genomes are being aligned or when these genomes are very different. The default setting will set a higher seed weight depending on file size. A large file size will obtain a higher seed weight and a lower file size will obtain a smaller seed weight [47].

The next option is the 'use of seed families'. The selection of this means that Mauve will use a three spaced seed pattern opposed to a singular one. The use of seed patterns allows for improved sensitivity on largely different genomes and for a higher seed weight for closely related organisms without sensitivity loss, although this comes with the cost of increased alignment time [47].

The user is also able to determine whether the alignment identifies locally collinear blocks (LCB) which are conserved segments in which there are no internal genome rearrangements. If disabled, Mauve will only identify local multiple alignments. The next parameter is 'assume collinear genomes'. It should only be used if the user is certain that there are no rearrangements between the genomes. This parameter can lead to a shortening of the alignment process.

The 'Full Alignment' option allows for Mauve to perform a recursive anchor search as displayed in figure 3.1, as well as a full gapped alignment of the genome using MUSCLE. If a full alignment is not conducted, the alignment anchors are identified and clustered into LCBs [47]. An iterative refinement option is another parameter which relates to the use of MUSCLE. MUSCLE is used in the iterative alignment to refine the initial alignment and normally improves it as shown in Figure 3.1, although this option should not be selected if biasing phylogenetic inference with a single guide tree [47]

The final parameter option, is the selection of the sum of pairs LCB scoring, which selects whether or not breakpoint penalties are applied among all pairs of extant sequences and whether penalties are applied to inferred ancestral gene ordering. It is stated that this option is only to be considered experimental and should be only disabled when aligning collinear genomes [47].

Using Mauve's internal viewer to visualise the alignment, the LCBs between the genomes are identified and shown as coloured blocks with lines connecting the aligned regions between the genomes. The created images can be exported into JPEG/PNG files. Mauve will be used to provide alignments between all of the combinations of genomes as well as a multiple alignment between all three. In order to establish an analysis in regards to the Wood-Ljungdahl pathway in ATCC 8486 and KIST 612, the genes from the Wood-Ljungdahl pathway will be split down into their respective branches, then an alignment between ATCC 8486, KIST 612 and the genes in the respective branch will be undertaken, in order to analysis the genome in regards to the Wood-Ljungdahl pathway. Breaking the Wood-Ljungdahl genes down into branches and aligning them, allows for a cleaner set of results, which can be analysed stage by stage, opposed to all at once. The ability to do this in Mauve with unannotated FASTA files, sets it apart from other alignment programs. Using this, an understanding can be established as to whether the Wood-Ljungdahl genes are highly conserved regions or whether they are in regions of rearrangement. We can also see a visualisation of the positional difference between genes in the Wood-Ljungdahl in ATCC 8486 and KIST 612.

3.4 *Eubacterium limosum* genome alignment

The initial genomes which were to be aligned were of the ATCC 8486 and KIST 612 strains, as these strains were the main point of comparison in this project. As stated, the FASTA files for each of these genomes was produced using the genome sequence produced at the University of Nottingham for the strain ATCC 8486 and the genome sequence (CP002273) from NCBI online database for KIST 612.

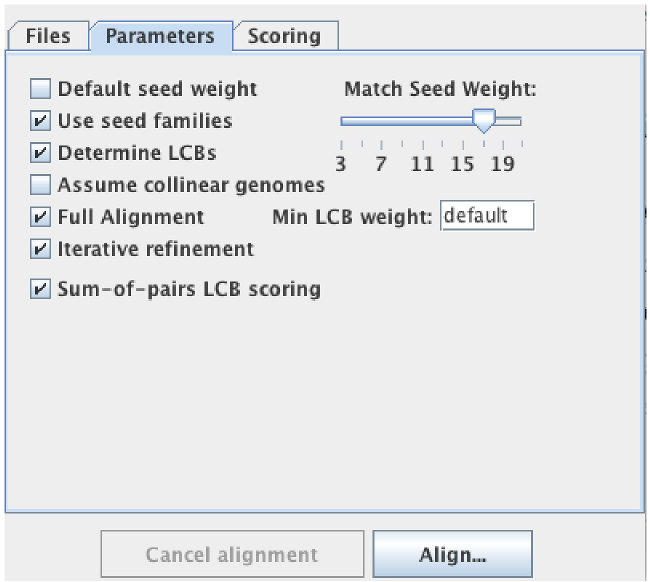

Figure 3.2: shows the Mauve alignment parameters which were used to conduct all of the genome alignments during this project

Figure 3.2 shows the specific parameters which were used for all the alignments conducted in this project. A custom match seed weight was used of 17 out of a possible 21; opposed to the default option. The use of seed families and determining LCBs were selected, with the assumption of collinear genome disabled. Both the full alignment and iterative refinement options were selected alongside the preferable sum-of-pairs LCB scoring option.

1. A match seed weight of 17 was selected out of a possible 21. The genomes being aligned are very similar and Mauve only suggests a low match seed weight when using a large number of genomes or those which are very different. The default match seed provided during all the alignments was 15, however this weighting is based on the file sizes and not the similarity of genomes, thus a higher match seed weight would be more suitable in this scenario.

2. The use of seed families was selected. It is stated that selecting the use of seed families will allow for the use of a higher match seed weight without the loss of sensitivity.

3. Determine LCBs was selected, as this is a requirement when looking to select 'full alignment', 'iterative refinement' and 'sum of pairs LCB scoring'. Moreover, using LCBs allows for a clearer and cleaner set of results when analysing, with LCBs providing a more important view of the alignment compared to just the local multiple alignments.

4. Assume collinear genomes was not selected, as there is no certainty that the genomes are free of rearrangements

5. Full alignment was selected. In order to provide a more in depth method of identifying anchors within the alignments, the use of this option allows for a more accurate and precise alignment.
6. Iterative refinement was also selected. Primarily as it improves the initial alignment and avoids phylogenetic bias
7. The sum of pairs LCB scoring was also selected. The sum of pairs LCB scoring is one of the novel inclusions in progressive Mauve, turning this option off could lead to an inaccurate alignment. Mauve developers state the disabling of this option should be considered experimental and only used when aligning collinear genomes, which can not be assumed to be the case in *E.limosum*.

The alignment was then run between all the *E. limosum* strains (ATCC 8486, KIST 612, SA11). The three way alignment used ATCC 8486 as the reference genome, allowing it to be compared to KIST 612 and SA11 on either side.

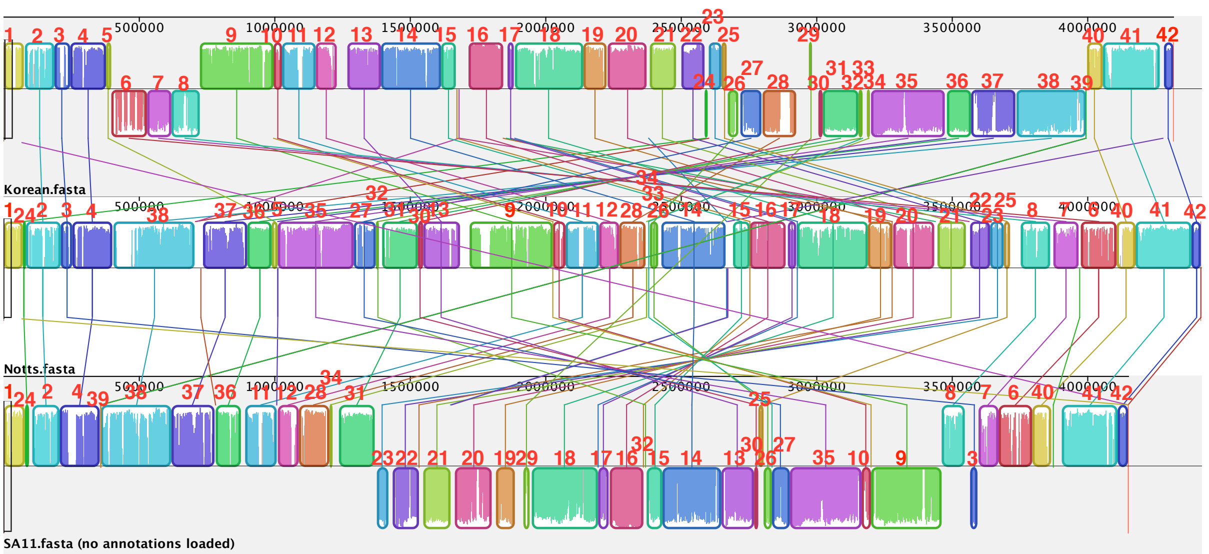

Figure 3.3: Mauve multiple alignment between the ATCC 8486, KIST 612 and SA11 strains of *Eubacterium limosum*

The alignment shows a series of 41-42 LCBs in each genome, which are labeled both numerically and by colour, each LCB is matched to their respective LCB/region in the other genomes via a line of corresponding colour. The top genome represents the strain KIST 612, the middle genome the strain ATCC 8486, and the bottom genome the strain SA11. Nearly all LCBs are present in all three genomes, indicating a high level of shared homology. Inversions can be identified through LCB blocks of each genome which appear below the line. In the KIST 612 genome, LCBs 6, 7, 8, 24, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38 and 39 are all inverted in comparison to the ATCC 8486 genome. Moreover, a large number of translocations can be observed through the order of the labeled numbers and more generally observed through the crossing of the lines matching the LCBs. From this, it could be suggested that ATCC 8486 has a more similar genome to SA11 opposed to KIST 612.

Similarity levels are established within the LCBs using the height of the lines inside them. This can be seen in block 4, where in SA11 there is a blank area which represents no similarity between the genes. While block 4 in ATCC 8486 and KIST 612 is mostly filled with colour apart from at the top, indicating high similarity.

When analysing the differences between ATCC 8486 and KIST 612 many structural and organisational differences can be observed. The initial one is inversions in the genome, demonstrated by LCBs being below the line. This shows that inversions have occurred in the 6, 7, 8, 24, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38 and 39th LCBs, which is a large proportion of the genome, being nearly half of the LCBs present in KIST 612. Furthermore, a high level of translocations are also present when comparing the genomes of ATCC 8486 and KIST 612. The LCBs of KIST 612 have been labelled in Figure 3.3 from 1-42 with their corresponding homologous LCB in ATCC 8486 labeled with the same number. As shown in the figure, ATCC 8486 does not possess the same order of LCBs as KIST 612, with a significant amount of rearrangement between the genomes present. The indication from these results is that there is a high level of homology between the genomes, apart from LCB 39 which is not found to be present in ATCC 8486. This shows that DNA shared between these genomes is very similar, with no major losses or gains of genetic data. This can be reaffirmed by looking at the similarity levels shared between the LCBs. Nearly every block is 'filled with colour' within these genomes, with the exception to a small white gap at the head of the block, showing slight differences. Occasional gaps appear inside the LCBs and outside of LCBs, showing areas in which no alignment has occurred. Although, this is expected as the strains will have sequences specific to them which are not found elsewhere. This can be observed inside the the 38th LCB in the ATCC 8486 strain and inside the 38th LCB in the KIST 612 strain. While outside of the LCB regions, gaps can be seen in-between the 15-16th and the 28-29th LCB's in KIST 612, and between the 38-37th, 13-9th and 25-8th LCB's in ATCC 8486.

Furthermore, to view the exact rearrangement present between the alignments, which may include LCBs not visible in the image without a in depth analysis, a permutation file may be generated. This provides a text file containing a numbered list of the LCBs present in each of the genomes aligned. It was used when labelling each LCB in Figure 3.3 to preventing missing smaller LCBs.

Further analysis of gaps and single nucleotide polymorphisms have also been conducted. However, due to the number of SNPs found, it is not possible to analyse them in the scope of a general alignment analysis. Although, when delving into the Wood-Ljungdahl pathway genes in the alignment between ATCC 8486 and KIST 612, the identification of SNPs could be used here to further the analysis.

What is clearly identifiable between these genomes is the structural and organisational variation which has occurred, there are numerous translocations and inversions present between each genome, which can ultimately lead to significant functional alterations between these strains phenotypically. The rearrangements between these strains of *E. limosum* could imply that these strains operate differently. Despite the fact that the genomes appear to be homologous due to the matching of LCBs, as well as these LCBs sharing high similarity and conservation between genomes, the presence of genetic rearrangement between the strains can still lead to phenotypical changes. Studies which have developed insight into structural variants between genomes further state that genome rearrangements have profound impacts on prokaryotic cells, leading to altered phenotypes [48]. The paper continues, stating that there are balanced and unbalanced forms of variation. Balanced variations retain the genetic information, while in unbalanced variations functional elements are either lost or gained. However both are difficult to interpret in terms of functional consequences [48].

It appears from Figure 3.3 that balanced structural variation has occurred, through the high level of homology seen. Moreover, further analysis would have to be conducted to evaluate the presence/level of gene loss or gain.

Inversions in the genome are known to be a major driving force behind genome evolution. In thermophobes and other types of bacteria it is has been implied that these inversions may lead to ecological advantages such as survival in extreme temperatures, stressed environments and low pH [48, 49]. Interestingly, *E. limosum* is known to be a bacteria found in the rumen and gut of mammals, and thus its retrospective environment would be in a lower pH. This suggests that rearrangement has occurred due to the different environments the three strains live in. Other research, which has analysed the genomes of other species of bacteria, particularly *Neissera*, *E. coli*, *Salmonella* and *Yersnia*, have drawn some conclusions which could be apllied to *E. limosum*. One paper [50] conducted a study whereby a comparison was made between *Shigella* and *E.coli* genomes. Between 6-8 genomes from each of the *Shigella* and *E. coli* species were analysed, focusing on the rearrangements which occurred between the genomes. In *Shigella*, 16-64 rearrangements were present alongside 549-1165 insertion sequences (IS), while the *E. coli* genomes presented 11 rearrangements and only 42-150 IS. This could indicate that rearrangement is more prevalent depending on the species of bacteria. It suggests also that *Shigella* undergoes rearrangements more commonly compared to *E. coli*. This may be indicative of *Shigella*'s environment as previously queried or it could imply that *E. limosum* like *Shigella* are more susceptible to rearrangements than bacteria like *E. coli*. Although, results in this study may be construed, as these bacteria were in conditions which allowed for a period of fast growth when the DNA was retrieved [50].

Similar work was carried out on *Neisseria* species, showing an astonishingly high number of repeats/rearrangements between different *Neisseria* species [51]. It is possible that operon structure within the genomes may have been disrupted, due the to the rearrangements and thus potentially leading to a change in the functionality of genes in disrupted operons.

It is presumed that rearrangements including translocations and inversions are caused by numerous mechanisms; some may be random. The mechanisms which are thought to cause rearrangements include the replication protein RecA. This repairs DNA by creating a replica of the damaged sequence, but RecA can also promote recombination between paralogs leading to rearrangements. The sequences involved normally in RecA mediated rearrangements are usually long repeats such as those in rRNA operons, rrn operons, transposons and IS elements [52, 53]. With this in mind, it would be interesting to see the positioning of these factors in the genomes of *E. limosum* and whether their position has any impact on the genes of interest; such as the Wood-Ljungdahl pathway genes. An in depth analysis of the operons which contain the genes present in the Wood-Ljungdahl pathway, to identify the presence of these factors associated with genomic rearrangement could be of interest further down the line. More so, an in-depth analysis of the operons of these genes in light of these factors associated with genomic rearrangement.

Although, the genomic comparison between the strains ATCC 8486 and KIST 612 are the main focus of this chapter and project, an analysis between ATCC 8486 and SA11 may help to provide further insight into the analysis of ATCC 8486. As shown in Figure 3.3, the ATCC 8486 genome is represented by the middle set of LCB blocks, while the SA11 genome is represented by the bottom row of LCB blocks. The initial observations which can be made about this comparison, is there are rearrangements, both translocation and inversion of LCBs. The inversions in SA11 can be seen to occur in the middle section of the genome only, with the vast majority of these inversions happening consecutively throughout this middle region of the genome. When observing translocation of LCB patterns in the SA11 ATCC 8486 genome alignment, it also becomes apparent that these rearrangements occur in the middle region of the SA11 genome; much like the inversions. The translocations between the genomes can clearly be seen by observing the 'X' pattern on the alignment; which is made by the lines matching the LCBs together. While the beginning and end of the SA11 genome seems to have a relatively similar organisation to ATCC 8486, with only a few rearrangements present. The inversions in SA11 have occurred in the LCB blocks 23, 22, 21, 20, 19, 29, 18, 17, 16, 32, 15, 14, 13, 30, 26, 27, 35, 10, 9, 3, which means that 22 out of 47 LCBs are inverted in SA11. Furthermore, a link can be drawn from the patterns of rearrangement between the genomes; inversions and translocations mostly happening in the same middle region of the genome. The organisation at the begin and end is retained between all 3 genomes, although this can primarily be seen between ATCC 8486 and SA11. Moreover, in regards to the structure and organisation of these genomes, Figure 3.3 shows that ATCC 8486 is more similar to SA11 than KIST 612, far more translocations occur in KIST 612 as shown by the quantity of crossing lines. The translocations in SA11 appear less random, confined mostly to the middle region; which also contains the inversions. ATCC 8486 and SA11 share the same LCBs, apart from LCB 5 which is not present in SA11. As previously stated, similarity between LCBs and genomes can be identified through the lines connecting the matching LCB, all of which share a matching colour unique to each LCB.

As established in Figure 3.3, high similarity and thus high levels of conservation of the genome are present between ATCC 8486 and SA11. Although some LCB blocks showcase areas of white, it is not expected that the genomes are completely identical, thus it can still be assumed that the LCB blocks shared between the genomes are of high similarity and highly conserved. One area of interest in regards to similarity is in LCB 4, where a full white strip can be seen in the SA11 genome. This shows that a sequence within LCB 4 in the SA11 genome does not align at all with the other 2 *E. limosum* genomes. Similar regions are also present in LCB 12 and 9 but to a lesser extent. A common reason for the presence of these unaligned gaps is presence of sequences unique to the strain within the LCB.

One of the primary interests throughout this project is the analysis of the Wood-Ljungdahl pathway in *E. limosum*. In accordance with this interest, the use of genome alignments to show the comparisons between the *E. limosum* strains ATCC 8486 and KIST 612 can be used to compare and analyse the relative positions of the genes involved in the Wood-Ljungdahl pathway. Through discussion in the previous section, it is clear that rearrangements between genomes may lead to phenotypical changes. Thus, it is of interest to see whether any patterns emerge with regards to the position of the Wood-Ljungdahl pathway genes in the genomes of the *E. limosum* strains ATCC 8486 and KIST 612.

3.5 Methyl branch alignment

In order to conduct the alignment with a focus to the Wood-Ljungdahl pathway genes, the process will be split into 4 different alignments, each one aligning the 4 branches which construct the Wood-Ljungdahl pathway: methyl branch, carbonyl branch, acetate branch, and butyrate branch. The genes involved in their respective branches will be aligned against the ATCC 8486 and KIST 612 genomes allowing for the generation of a multiple alignment, which can be used to analysis the Wood-Ljungdahl pathway on a genomic scale.

Figure 3.4: The structural organisation of the methyl branch within the genome of *E. limosum* KIST 612 (Top) and ATCC 8486 (Bottom). Red boxes indicate the LCBs where the genes of the methyl branch are organised in. A-E shows the genes which are present in the methyl branch and which box they are present in.

Each gene involved in the methyl branch of the Wood-Ljungdahl pathway is labeled A-E (A=formate dehydrogenase, B=formate tetrahydrofolate ligase, C=Formyl tetrahydrofolate cyclohydrolase/ methylene tetrahydrofolate dehydrogenase, D=Methylene tetrahydrofolate reductase 5'10, E= Methylene tetrahydrofolate C-terminal). Figure 3.4 shows that there is one line coming from A and one line coming from B, C, D, and E. This shows that the genes of the methyl branch are split into 2 different LCBs within the genome; one at the beginning and one at the end. Formate dehydrogenase (A) is shown to be in the LCB labeled as '1' by itself, also LCB 1 has been translocated between the ATCC 8486 and KIST 612 genomes. Furthermore, this LCB possesses a few white gaps which are known to represent sequences which have not aligned at all; one of these gaps are particularly large. The overall conservation and similarity within this LCB appears to be mostly retained. The rest of the methyl branch (B,C,D,E) is located in LCB 2 at the beginning of the genome. Between ATCC 8486 and KIST 612, this LCB has undergone minimal rearrangement, only moved position slightly between the genomes. The similarity and conservation shared of LCB 2 between these genomes is relatively low, with most of the sequences in this block unable to be aligned to those in same LCB of the other genome, although some areas of high similarity are present.

Formate dehydrogenase is located in an LCB separate from the other genes, while all the other genes are present together in a different LCB. Despite the separation of formate dehydrogenase, it is expected that the rest of the genes which work together in the methyl branch would be located within the same region of the genome. This may suggest their presence in an operon together, but ultimately the close proximity of these genes within the genome suggests a relationship or reliance between them. Furthermore, the fact that this LCB has not experienced any major rearrangement indicates the importance in the functionality of these genes. As previously stated, the similarity between this LCB leads to more questions, with the presence of a gap in the LCB 1 of ATCC 8486 and also less internal similarity in LCB 2 between the genomes. Whether the dissimilarities shown are occurring in the methyl branch genes, if not, what are these areas of sequences which could not be aligned and do they have any implication on the functionality of the methyl branch genes. These areas of different sequences could be other genes within the same operon(s) as the methyl branch genes, the presence of different genes could lead to changes in how the methyl branch genes function or could even have effects on how the Wood-Ljungdahl pathway functions.

3.6 Carbonyl branch alignment

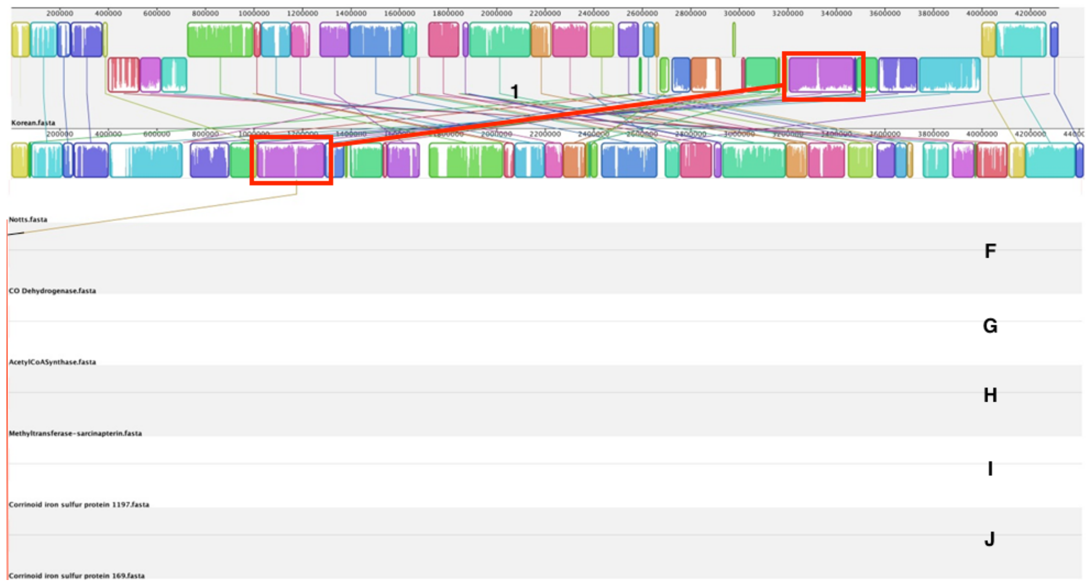

Figure 3.5: The structural organisation of the carbonyl branch within the genome of *E. limosum* KIST 612 (Top) and ATCC 8486 (Bottom). The red box indicates the LCBs the genes are organised in. F-J shows the genes which are present in the carbonyl branch and which box they are in.

A multiple alignment of the genes involved in the carbonyl branch between the *E. limosum* KIST 612 and ATCC 8486 strains is shown in Figure 1.5. Each gene involved in the Carbonyl branch of the Wood-Ljungdahl pathway is labeled F-J (F=Carbon monoxide dehydrogenase, G=Acetyl CoA synthase, H=Methyltransferase-sarcinapterin, I,J=Corrinoid iron sulfur proteins). Figure 3.5 shows a single line from these genes which is matched to a pair of LCBs labelled as '1', showing that all the genes of the Carbonyl branch are all located in the same LCB/region of the genome. As established through Figure 3.5, this LCB has undergone rearrangement between the ATCC 8486 and KIST 612 genomes, being both translocated and inverted in the KIST 612 strain compared to the ATCC 8486 strain. The similarity and conservation of this LCB between the genomes is very high, the block is mostly filled with only a few gaps of unaligned sequences symbolised by the white gaps. Overall the LCB is clearly conserved and highly similar between the genomes.

These results go along with the general consensus that all the genes which make up the Carbonyl branch would be closely positioned in the genome together. Due to the interlinking nature of the Carbonyl branch it is definitely expected that these genes would be grouped together. In regards to acetyl-CoA production, it is known that these genes combine to encode for one large reaction in methanogens, opposed to a series of smaller separate reactions as it appears to be the case in *E. limosum*. However, it is expected that these genes have a strong link which is confirmed by the results shown in Figure 3.5. Furthermore, the LCB in which the Carbonyl branch is present is very highly conserved and highly similar between the genomes. This in line with the notion that acetyl-CoA production is one of the most important processes in these bacteria. It is clear to see why conservation appears higher in this block opposed to others, as it is an integral function of the cell.

Due to this fact, one would assume that the inversion and translocation of this LCB block would have little effect on the functionality of the genes within, due to their importance in the bacteria.

3.7 Acetate branch alignment

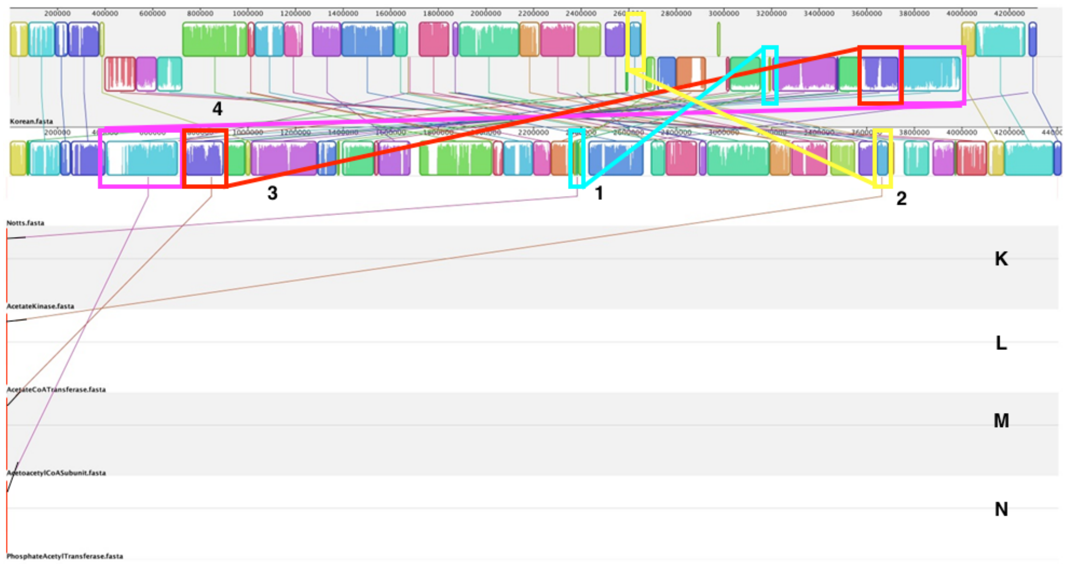

Figure 3.6: The structural organisation of the acetate branch within the genomes of *E. limosum* KIST 612 (Top) and ATCC 8486 (Bottom). The red box indicates the LCBs the genes are organised in. K-N shows the genes which are present in the acetate branch and which box they are in.

A multiple alignment of the genes involved in the Acetate branch between *E. limosum* KIST 612 and ATCC 8486 strains. Each gene involved in the Acetate branch of the Wood-Ljungdahl pathway is labelled K-N (K= Acetate Kinase, L=Acetate CoA Transferase, M = Acetoacetyl CoA Subunit/Propionate CoA transferase, N = Phosphate Acetyl Transferase). Figure 3.6 shows that each of these genes involved in the Acetate branch of the Wood-Ljungdahl pathway is present in a separate LCB and separate areas of the genome. Acetate kinase 'K' is shown to be present in the LCB '1' symbolised by the light blue annotations. It is an exceptionally small LCB which is shown to be inverted and translocated. However, the similarity and conservation of this block between the two genomes is very high, as it does not appear to possess any white gaps. Secondly, Acetate CoA Transferase, is present in the LCB '2' annotated by the yellow boxes/lines. Note that a slight translocation of the LCB is present which may be a result of the translocation of the inverted region next to it. Moreover, high similarity and conservation appears to be present in this LCB, as shown by the block being mostly filled by colour with very little white areas being possessed. The genes encoding for Acetoacetyl CoA Subunit/Propionate CoA transferase and Phosphate Acetyl Transferase are located in are neighbouring LCBs, with the first annotated in red and the latter annotated in pink. Both of these regions have undergone major rearrangements, including being translocated to the other side of the genome from ATCC 8486 to KIST 612, while also being part of a cluster of inverted LCBs. When looking into the similarities and conservation shared between these LCB regions across the genomes, there isn't the high level of similarity and conservation which has been seen for LCBs containing other genes. LCB '3', which contains the gene encoding for Acetoacetyl CoA Subunit/Propionate CoA transferase, shows two peaks of high similarity with areas of low similarity either side. While LCB '4', which contains the gene encoding for Phosphate Acetyl

Transferase, shows a large area of sequence which is not conserved between the genomes; among a few other areas of dissimilar sequences in the LCB.

Only the acetoacetyl CoA subunit/propionate CoA transferase and phosphate acetyl transferase genes were in relative proximity to each other, although it was these genes which underwent the biggest rearrangements in the genome. It would have been expected that all the genes involved in Acetate branch would be positioned close together or in close proximity, due to the involvement each reaction (encoded for by these genes) has with each other in the Acetate branch. Surprisingly, the opposite result opposite is shown in Figure 3.6. Albeit, the acetate CoA transferase could be seen as involved in the butyrate production pathway and thus grouped in close proximity with the genes involved in this branch; as it converts butyryl CoA to butyrate as well as acetate to acetyl-CoA. However regardless of this, the other genes in this branch are not located in the same LCBs/regions either. It could be assumed that phosphate acetyl transferase and acetate kinase would have been located together, due to there reliability on each other to generate ATP and acetate through the Acetate branch, however they are found to be on the opposite sides of the genome in ATCC 8486. The acetate kinase and acetate CoA transferase LCBs appear to be more similar between genomes than the similarity shared between the acetoacetyl CoA subunit/propionate CoA transferase. These two genes may be located in more conserved regions which is a shared similarity between both genomes, with acetate kinase not only producing acetate but also ATP. This ATP produced here is extremely important in terms of balancing ATP used previously in the pathway. While acetate CoA transferase provides the only way of producing butyrate and another method of producing acetate, both useful products within the cell. This could explain the levels of conservation in the LCBs where these genes are situated, opposed to those of the other 2 genes which are known to be useful but not integral to the pathway. Another interesting pattern emerging from Figure 3.6 is that the LCBs 1, 3, 4 are all present in the same cluster of inversions in the KIST 612 genome. While acetate CoA transferase, which could be considered part of the butyrate pathway the only gene not inverted, this will be an area of interest in the results in Figure 3.7.

3.8 Butyrate branch alignment

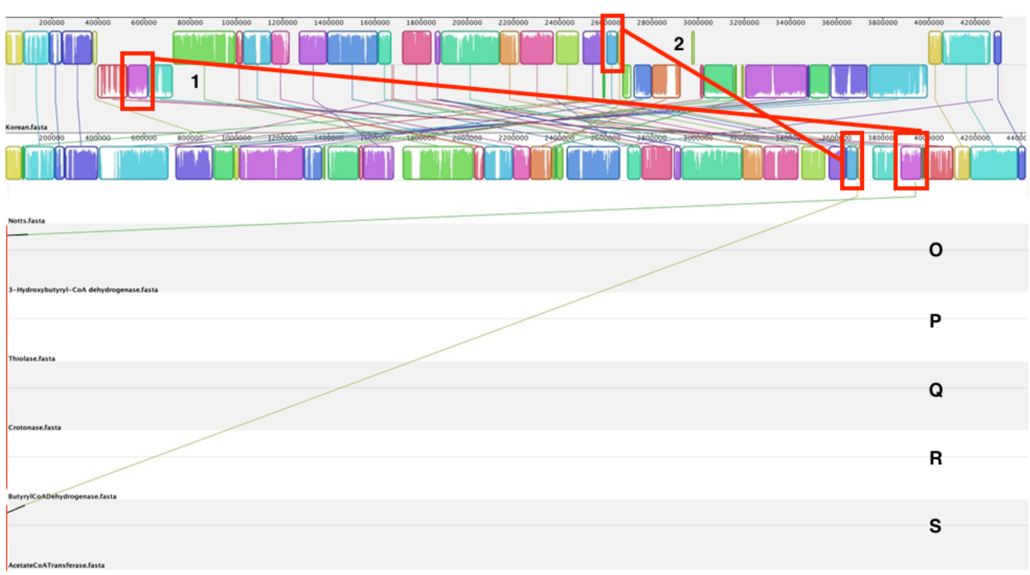

Figure 3.7: Genome alignment between *E. limosum* ATCC 8486 and KIST 612, focusing on the butyrate branch in the Wood-Ljungdahl pathway

Figure 3.7 shows a multiple alignment in Mauve between the *E. limosum* genomes ATCC 8486 and KIST 612 and genes which are involved in the Butyrate branch of the Wood-Ljungdahl pathway. Each gene involved in the Butyrate branch of the Wood-Ljungdahl pathway is labelled O-S (O= 3-Hydroxybutyryl-CoA dehydrogenase, P= Thiolase, Q= Crotonase, R= Butyryl CoA dehydrogenase, S= Acetate CoA transferase). Figure 3.7 shows all these genes are present in the same LCB block of the genome. It is shown that this LCB has undergone rearrangement, being translocated from the end of ATCC 8486's genome to the beginning of KIST 612's genome. Moreover, this LCB '1' has been inverted along with a small cluster of other LCBs between the genomes. As previously seen in Figure 3.6, the LCB containing acetate CoA transferase is not present with the rest of the genes in the butyrate branch, although in ATCC 8486 they are in relatively close proximity of each other at the end of the genome. Furthermore, the similarity of the LCB in which the butyrate genes reside in can be seen to share a high similarity and conservation between the genomes. No full white streaks appear to be present in this block insinuating any sequences which are not conserved. In fact, it can be seen that all the sequences within this block are conserved up to a certain value, and thus it can be assumed that the functionality provided by this LCB is retained.

The butyrate branch is part of a small cluster of LCBs which can be viewed on Figure 3.7 as LCB '2'. LCB '2' and the LCBs either side have undergone translocation and inversion to the beginning of the genome in KIST 612, compared to being at the end of the genome in ATCC 8486. Acetate CoA transferase is present in both Figure 3.6 and 3.7 representing the alignments of the acetate and butyrate branch respectively, it can be seen that this gene is found to not share LCBs with either the genes in the acetate or butyrate branch. As shown in Figure 3.6, acetate CoA transferase is located in close proximity to the butyrate genes in ATCC 8486, however in KIST 612 the genes have been completely rearranged. Overall, Figure 3.7 suggests that the genes involved in butyrate pathway labelled O-R are highly interlinked, most likely as their only purpose is to produce butyrate, and thus are highly reliant on each other.

3.9 Discussion

The results which have been garnered throughout this chapter can be used to provide an insight into *E. limosum* and the Wood-Ljungdahl pathway shared between the different strains. Initially it can be seen from Figure 3.3 that the genomes from the 3 strains have undergone some major rearrangements. Experimental observations suggest that environmental factors, such as temperature and pH, have a significant impact on the frequency of genomic rearrangements in bacteria. Especially bacteria living in harsh environments like *E. limosum*, which resides in the rumen and gut of mammals, exhibit very frequent rearrangements of their genomic structure indicating that it may be a survival trait evolved by bacteria. However, in light of this, it is important to observe whether this may have any profound effect on the gas fermenting metabolism of *E. limosum*. The alignment which included the genes involved in the methyl branch, showed that formate dehydrogenase was located in a different LCB compared to the other methyl branch genes. The reasons for this are most likely due to formate dehydrogenase being involved in numerous different roles within the bacteria and that it is not used solely in the Wood-Ljungdahl pathway. This could also imply why the similarity and conservation are much lower in the LCB where formate dehydrogenase is located, compared to the LCB's where the other methyl branch genes are present. With formate dehydrogenase being used in other pathways, it is reasonable to assume that formate dehydrogenase could have different neighbouring genes depending on its roles and functions in each strain. The other methyl branch genes were shown to be collectively a part of the same large LCB which has been translocated between the genomes. These genes which are more specific to the Wood-Ljungdahl pathway are closely linked and thus expected to be positioned closely in the genome; which they are. The LCB they are in shows areas of low or no similarity, it would be interesting to see whether these areas of unaligned sequences in the genome have any affect on methyl branch genes. Although, due to the large size of this LCB the areas of low/no similarity may not even be in close proximity to the methyl branches genes, let alone in the same operon; which is something that could have an implication on how they function.

The carbonyl branch shows an example of a highly linked and conserved set of genes between both of the *E. limosum* strains. Although, rearrangement such as translocation and inversion has occurred, it is clear to see very high similarity and conservation in the LCB which contains these genes in both genomes. As previously stated, the presence of the acetyl-CoA synthase gene in the carbonyl branch and the fact that these genes are thought to be linked very heavily and reliant on each other, it is no surprise to observe the high levels of conservation that is shown to be present.

The acetate branch provided a completely different set of results to the other branches. Whereby all the genes which make up the acetate branch are separated and distributed throughout the genome. The initial implication could be that these genes are all involved in additional processes and are found to be among genes which are collectively used in different pathways. Of the patterns which emerged from these results, it appears that genes such as acetate kinase and acetate CoA transferase, both of which are the genes which produce the 'end stage product' of the Wood-Ljungdahl pathway, are in LCBs which are more highly conserved and similar between the strains. While the opposite was to be said for the genes phosphate acetyl transferase and acetoacetyl CoA subunit/propionate CoA transferase, which are used in the pathway but not the main component of producing acetate and butyrate. Although, into order to create a balanced Wood-Ljungdahl pathway ATP must be produced, this occurs via acetate kinase catalysing acetyl-phosphate into acetate, which would not be possible without the presence of phosphate acetyl transferase and should be considered to be as integral as acetate kinase. A consideration should be made to the genes or sequences which neighbour each of the genes involved in the acetate branch. The relationship/impact of these neighbouring genes could lead to alternative methods of ATP production or unfamiliar extensions of the Wood-Ljungdahl pathway; specific to each strain/bacteria. This could answer why acetate kinase and phosphate acetyl transferase are not located near each other in the genomes of *E.limosum*, in contrast to other acetate forming bacteria and despite having a major relationship in terms of acetate production; an integral function of an acetogen.

The genes which encode for the butyrate branch are present together in the same LCB within both the ATCC 8486 and KIST 612 genomes. The presence of these genes together in the same relatively small LCB suggests that they might be organised in an operon together or neighbouring operons which are strongly linked. This would be a fair assumption to make due to these genes primarily and most likely only used in butyrate formation, therefore the grouping of these genes into an operon or neighbouring operons which are highly conserved could be what figure 3.7 is showing.

The results from Chapter 1 suggest that the only known method of butyrate formation in *E. limosum* uses an Acetate CoA transferase, which produces acetyl-CoA as part of the same reaction; Acetyl-CoA being an extremely valuable molecular building block to the bacteria. Seeing as butyrate formation only appears to be possible in *E. limosum* through Acetate CoA transferase, the gene which encodes for this reaction is not present alongside the other genes which make up the butyrate branch or even the acetate branch; as this reaction uses acetate to produce butyrate and acetyl-CoA. All the other genes which together form the butyrate branch are located in the same highly conserved LCB, showing the presence of a strong link in terms of butyrate formation, which makes Acetate CoA tranferase's omission even more interesting. This could suggest the presence of a butyrate kinase which is located alongside the other butyrate branch genes in the genome. However, a thorough search and BLAST analysis of multiple different butyrate kinases was conducted to test this, the butyrate kinase would have to have an extremely different sequence to the ones tested; making this an unlikely scenario. It could also be possible that the main objective of the butyrate branch in *E. limosum* is to produce Butyrate-Phosphate and/or Butyrate-CoA instead. There could be an unknown value of these products, making them more favourable to produce over butyrate in the bacteria, such as them being used in other reactions to produce other preferred products. Although, looking at what reactions Butyrate-Phosphate and Butyrate-CoA are involved in shows a narrow and unlikely selection. Another suggestion is that Butyrate-Phosphate and Butyrate-CoA could be preferred in order to help balance aspects of the Wood-Ljungdahl pathway; such as the reduction-oxidation side. Where if butyrate production was the main objective the pathway may becoming unbalanced, not to say that butyrate is not produced, just that it would be produced depending upon the whether other more important objectives/criteria are met within the pathway first. It would be beneficial to compare laboratory growth results with the model results, which will be established in the structural model chapter, to gain a better understanding of how the pathway may work.

Chapter 4

Structural Model

4.1 Introduction

Metabolic modelling is a bioinformatical technique that it is used by developing a reconstruction of a metabolic network within an organism to conduct virtual simulations. This computational tool provides an in-depth insight into molecular mechanisms with or without alteration in the organism of choice. The importance of metabolic modelling is that an experiment can be simulated computationally and the resulting predictions can be used guide new experimental strategies. Thus, metabolic models provide additional insight into an organism's metabolism, however experimental observations are required to ensure its validity and support its improvement.

This chapter will report simulations and computational results from the metabolic model for the Wood-Ljungdahl pathway in *E. limosum* established throughout this project. Furthermore, the original model is compared to mutants that contain an additional bioengineered isopropanol pathway or gene-knockouts.

At the University of Nottingham an iso-propanol pathway has been bioengineered into *E. limosum*, using the genes from other clostridia such as *C. acetobutylicum* and *C. beijerinckii* which possess the genes for iso-propanol formation. In order to develop an understanding for how well this branch is integrated into the Wood-Ljungdahl pathway, an analysis using the metabolic model is appropriate. The metabolic model can be used to gain a number of insights into the Wood-Ljungdahl pathway that has been established throughout this project and with the additional iso-propanol pathway. The model can be used to determine the maximal theoretical yields of products, test whether gene knockouts are feasible, what effect they have on production, and provides a better understanding of the networks activity. Furthermore, knockouts can be simulated on any of the genes present in this network, to predict if productivity can be increased and to see which genes are integral to the network. Finally, the model can be used to provide a flux change analysis, whereby the fluxes through the different networks can be compared and analysed.

In this study, the focus is on changes to the metabolic network with and without the addition of the iso-propanol pathway and potentially knockouts to both networks that might result in the formation of acetone. Using this analysis, differences between the natural metabolic network in *E. limosum* and the iso-propanol bio-engineered network can be established, with an aim to see production of iso-propanol or whether it can be increased through alterations. In Figure 4.1, the metabolic network with the addition of an Isopropanol pathway is shown.

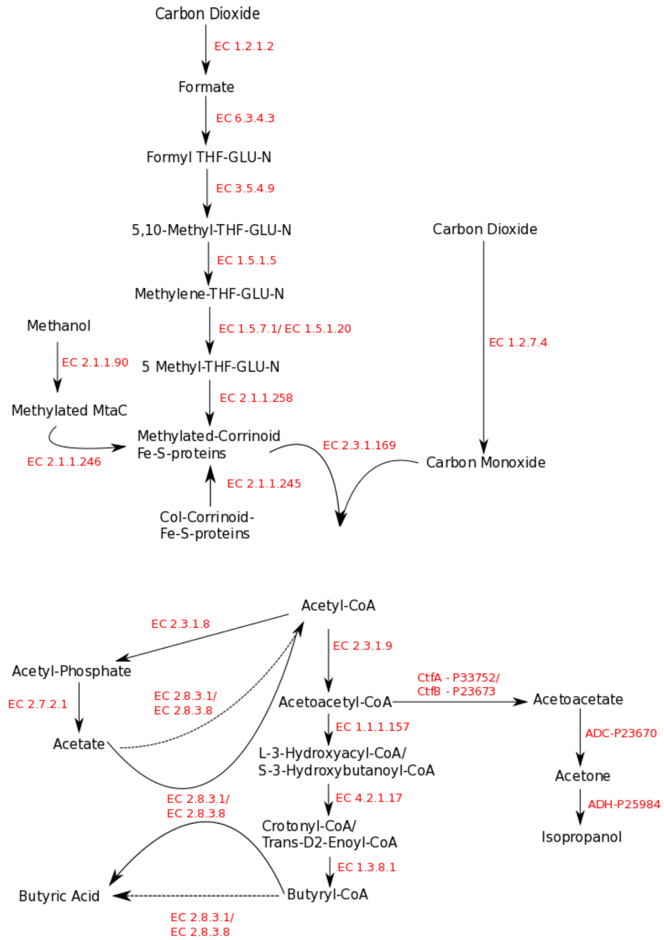

Figure 4.1: The Wood-Ljungdahl metabolic network with the addition of an iso-propanol pathway. The dotted line showing what is produced from the EC 2.8.3.1/2.8.3.8 reaction.

As shown in this figure, isopropanol formation starts at acetoacetyl-CoA, which is part of the butyrate formation pathway. Acetoacetyl-CoA is converted to into acetoacetate by an acetoacetyl CoA transferase, much like the acetate CoA transferase used in butyrate formation, whereby the CoA moiety is transferred from acetoacetyl-CoA onto acetate producing acetyl-CoA and acetoacetate. The two genes, ctfA Uniprot ID P33752 and ctfB P23673, encoding for the CoA-transferase (CtfAB) originate from *C. acetobutylicum*, and are also represented by the EC number 2.8.3.9. In a next step, acetoacetate is converted into acetone by acetoacetate decarboxylase (Adc) which is also from *C. acetobutylicum* (Uniprot ID P23670), this is also represented by the EC number of 4.1.1.4. Finally, acetone is transformed into iso-propanol by a secondary alcohol dehydrogenase (SAdh) from *C. beijerinickii* NRRL B593 (Uniprot ID P25984), it is also represented by the EC number 1.1.1.80.

In order to fulfil these objectives, a structural metabolic model had been created in ScrumPy. ScrumPy is a powerful and versatile metabolic modelling tool which uses the Python programming language for implementation [**?**]. It can be used on Windows and Linux, with Mac OS installation soon to be introduced. ScrumPy provides all the tools needed for the intended metabolic analysis. Moreover, ScrumPy uses a simplified primary user interface language derived from Python and is the preferred tool within the Synthetic Biology Research Centre (SBRC) at the University of Nottingham. Thus, the decision was made to follow the choice of my colleagues. ScrumPy, requires the user to create a model of the network which then allows for an extensive analysis including knockouts, Flux Balance Analysis (FBA), and Elementary Mode Analysis (EMA). An introduction and tutorial is available online providing an overview of how to carry out the analysis (http://mudshark.brookes.ac.uk/ScrumPy/Doc).

Although such a model is seen as a good representation of the metabolic network, it should be noted that there is an element of inaccuracy in these models and many limiting factors are not taken into account. As a consequence, the simulation results should be interpreted as an approximation generating an estimation of the networks capabilities. Due to the limitations of this project, such effectors as environmental conditions, interactions with other pathways and thermodynamic feasibility of metabolic reactions was not considered. However, they might taken into account in more detailed and larger models as well as more thorough theoretical analyses.

The results for the structural model will be particularly interesting, due to the addition of the iso-propanol pathway which has not been added into *E. limosum* ATCC 8486 in a model or laboratory yet. Due to this inclusion, it could be assumed that butyrate production will decrease to accommodate for the iso-propanol production, while acetate production should remain high due to its involvement in ATP generation. It is not known whether the inclusion of the iso-propanol pathway will have an affect on the involvement of butyrate in the recycling of acetate to acetyl-CoA.

Furthermore, the inclusion of a homologous thiolase gene could mean that more flux is directed towards Iso-Propanol and butyrate formation. However, this is unfavourable in terms of acetate formation and its ability to produce ATP, which should clearly be favourable within the organism. This could not be able to be investigated in a structural model as kinetic properties are not present, only if the homologous thiolase possessed a different reaction mechanism could it work.

In terms of knockouts, it could be assumed that both the butyrate and iso-propanol formation pathways can be activated and deactivated through the knockout of 3-hydroxybutyryl-CoA dehydrogenase (butyrate formation) and acetoacetyl-CoA transferase (Iso-propanol formation). This could lead to increased activity of the only viable pathway (acetate production), while the acetate formation genes remain integral to the pathway. The expected yield would most likely show acetate to have the highest yield, with a presumed split between iso-propanol and butyrate which favours butyrate.

4.2 Structural modelling literature review

Although there are numerous types of models, in context of this project a structural metabolic model will be used. Such a model can be described as a list of reactions and associated properties which are assumed to be present in the organism under investigation. Furthermore, it contains a description of the environment the organism resides in. The information about the reactions in this model can then be used to generate results using different software packages.

There are different types of metabolic models, with kinetic models and structural models the most frequently used metabolic models. In kinetic models, every reaction is defined in terms of its stoichiometry and rate equation. It requires various kinetic parameters and initial metabolite concentration before the analysis can be carried out. Once assembled, several investigations can be conducted including: Values of initial metabolite concentrations or parameters may be varied, current set of reaction rates which are determined by the parameters and metabolite values, determination of reaction rates and metabolite concentrations after the model system has evolved for some time period, and reaction rates and metabolite concentrations at steady state [54].

Structural modelling can be used as an alternative to kinetic modelling; which has its faults. Kinetic modelling relies on kinetic equations and parameters being applied to the model, which can come in the form of huge variations leading to significant computational and times expenses. While structural models are formed just using the information of the reactions and their stoichiometry, which is a much more simplified approach opposed to kinetic modelling [54]. Structural modelling is a preferable technique for drawing more general conclusions, whereby a kinetic model allows for conclusions to be developed under specific constraints or scenarios.

The structural model focuses on the structure of the system which is displayed as a stoichiometry matrix **N**. Its columns represent the reactions and its rows the intracellular metabolites. A stoichiometry value is assigned to each element within the matrix, with 0 indicating that the metabolite is not involved in the reaction, a negative value indicates that the metabolite is consumed in this reaction, and a positive value means that a metabolite is produced in this reaction [54]. This convention focuses on intracellular metabolites. As a consequence, transporter activity becomes positive when a metabolite is taken up and negative when it is excreted. Assuming that the metabolic network approaches a steady state, the structural model must fulfill the following matrix equation.

$$\mathbf{N} \cdot \nu \,, \tag{4.1}$$

Where ν is a vector whose elements corresponds to reaction rates or fluxes at steady state.This equation shows the foundation of structural modelling and its techniques.

For this project, structural modelling will be applied to provide conclusions on the Wood-Ljungdahl pathway. However, it was imperative to acknowledge which would be the best modelling approach to take, particularly between kinetic and structural modelling, as well as weighing up the advantages and disadvantages of the modelling packages available.

Due to the lack of literature and research on the Wood-Ljungdahl pathway and extended metabolic networks in *E. limosum*, kinetic information such as enzymatic rate equations and associated parameter values are not available. Moreover, the development of these *in vivo* or *in vitro* experiments requires an extended amount of time and research [55], this would not be possible within the scope of this project. Overall, a model based purely on the structure of the metabolic network would be a more suitable approach to take in light of an assessment of the literature. Further research could use kinetic models allow parameters and initial concentrations to be set, with changing or steady factors such as pH, temperature, light etc, which are then simulated over a time period to see how the flux pathway operates in real time.

The structural model requires relatively little information in order to conclude a general insight. The requirements being, the reactions within a pathway or metabolic network and some general information such as reactions which allow for the steady state to be achieved. For instance, in *E. limosum* the RNF complex, ATP synthase and a transhydrogenase are not strictly reactions associated in the Wood-Ljungdahl pathway, but are required to create a steady state in the model. The information required for this model is available from the literature for *E. limosum* KIST 612 and was adapted for *E. limosum* ATCC 8468. This structural model provides a basis for future more complex structural and kinetic models.

The software used to carry out the metabolic model was ScrumPy (http://mudshark.brookes.ac.uk), which was developed at Oxford Brookes University. The team that developed this tool have a relationship with the University of Nottingham SBRC team and as such it is the primary metabolic modelling tool used among my colleagues; thus there was no need to seek out alternative software. ScrumPy can be used for both structural and kinetic modelling as well as linear programming, which are all used to carry out a FBA in a structural model. As stated in the introduction, ScrumPy uses Python along with commands relevant to metabolic modelling allowing for results to be gathered. The advantages of using ScrumPy opposed to other modelling softwares like Pathway Tools (https://bioinformatics.ai.sri.com/ptools/), merlin (http://www.merlin-sysbio.org/), RAVEN toolbox (http://biomet-toolbox.org/index.php?page=downtools-raven), was that due to ScrumPy being an established modelling tool at the University of Nottingham, specialists and guides in ScrumPy were easily at hand. Moreover, the use of the same modelling tool allows for consistency and reliability in future and present research, as there could be differences in the results acquired from different modelling tools.

The requirements for the use of ScrumPy are relatively small. It can be installed on Windows and Linux. Additionally, it is planned to make the software package available for Mac OS in the near future. The installation on Linux requires the downloading and compiling of a tarball file, which installs the software package in a new directory. It can be operated through typing 'ScrumPy' in the terminal and using the ScrumPy commands on a selected model. The simplicity of installation as well as the commands used to analyse the model, which will be shown in the method section, both show advantages for the use of ScrumPy. The model of the pathway can be created at ease and the ScrumPy site possesses an extremely useful installation and user guide, providing examples of how to use structural modelling, kinetic modelling, and linear programming to gain results from a model (http://mudshark.brookes.ac.uk/ScrumPy/Doc).

A disadvantage of ScrumPy could be the lack of a GUI and representation of results, as ScrumPy is operated in a plain terminal and the analysis must be conducted through a command line interface. Thus, ScrumPy is not as easily usable as other GUI modelling programmes, but it does allow the user to add specific parameters and functions which are not restricted by a limited GUI. A knowledge of Python may be needed alongside a small learning curve on how to use some ScrumPy specific techniques, albeit ScrumPy still aims for simplification and ease of use despite the lack of GUI. The representation of the results is usually presented on screen in the form of a matrix, which may not be the most visually pleasing or best way to analyse the results, opposed to other GUI modelling tools. ScrumPy does allow users to produce graphs (used in kinetic modelling) or save matrices or fluxes to Excel or Notepad, which can then be used to construct a preferred representation of the results. This can also be seen as a positive as other software may provide a better representation of the results than the modelling tools. For instance, in this chapter results of flux changes in the model will be taken from ScrumPy and represented in ESCHER maps, for a more appropriate visualisation of flux change.

In ScrumPy, the structural model can be used to provide an EMA and FBA of the Wood-Ljungdahl pathway and extended metabolic networks. EMA provides a way of extracting biologically meaningful pathways from a metabolic network [56]. The EMA calculates all solutions in the admissible flux space by solving the equation (4.1), each possible solution which can achieve this steady state is an elementary (flux) mode. A non-decomposability constraint makes sure that each elementary mode is unique, thus each elementary mode can be described as a unique, minimal set of enzymes which support the steady-state operation of a metabolic network [56].

The FBA is another method of analysis which can be used to analysis a structural model. It can predict fluxes using linear programming, the reaction stoichiometry and additional constraints; such as set flux rates, uptake rates and excretion rates. FBA can be and has been used to predict the fluxes in numerous models as well as the viability of knockouts with in metabolic network [57]. FBA is based on laws of mass conservation and the application of optimisation principles to predict the optimal distribution of metabolic resources within a network [58]. The analysis under steady-state conditions only requires the information about the stoichiometry of the metabolic pathway, the results then show the flux through each reaction under particular constraints, in particular the flux of the output metabolites. Constraints can then be changed to view whether production can be increased. The constraints applied to change this production can either be: setting a total flux of input/output compound to be used/produced in the network or a maximum or minimum amount of flux which is allowed to pass through a specific reaction [59].

In the literature these tools have been used to analyse numerous structural models of metabolic networks. One study used an EMA and FBA to investigate whether sucrose production could be increased in sugar cane [54]. Reactions carrying a high flux, then were targeted for knockout, as a large flux implies the reactions importance for the metabolic network. The ability to knockout this reaction could lead to an increase in sucrose production as a means to maintain the steady state. The use of the model allowed for a quicker and cheaper investigation as to whether these knockouts were viable, then using this information so that experimental investigations could be conducted [54]. The same paper also mentions another study which used an EMA in a structural model to develop a better understanding for how the Calvin cycle works. The results from this EMA produced a testable hypothesis 'The response of starch synthesis flux to changes in the external free phosphate is due to the sensitivity of ADPG-pyrophosphorylase to phosphoglycerate and free phosphate' [54].

Furthermore, the EMA showed that all the modes contained starch degradation and presence of Rubisco, establishing a connection between net starch degradation and activity of Rubisco [54].

A FBA was used to identify drug targets in the mycolic acid pathway found in *Mycobacterium tuberculosis* [60]. The investigated metabolic network consisted of 219 reaction catalysed by 28 enzymes. The FBA was able to provide insights into the metabolic capabilities of the pathway, but also identified essential reactions/enzymes and, thus, potential targets for drugs. This analysis resulted in 5 new enzyme targets for drug design (AccD3, Fas, FabH, Pks13, DesA1/2 and DesA3), while previously only InhA was known as potential target [60].

Finally, other groups have used EMA and FBA as a method to computationally predict alterations to metabolic networks for increased production of compounds of interest. One paper aimed to increase succinate production in *E. coli*, the theoretical maximum yields of succinate under aerobic and anaerobic conditioners were both 1.0 mol/mol gylcerol. Based on an EMA, conditions for increasing succinate production were identified as overexpression of phosphoenolpyruvate carboxylase or heterogonous pyruvate carboxylase. While it was determined that keeping the molar fraction of oxygen under 0.65 mol/mol would be beneficial for increasing the succinate production [61].

4.3 Structural model method

ScrumPy was installed on a specialised ScrumPy linux VirtualBox, which was accessed through the University of Nottingham. The VirtualBox creates a 'virtual machine' which in this case was running a Debian 7 (Wheezy) Linux system, where ScrumPy had previously been installed and used by other colleagues. All that was required was to install VirtualBox on Mac OS and log into the 'University of Nottingham SBRC Wheezy VirtualBox'. This was a preferable option as it had been tried and tested, eliminating the need to troubleshoot any errors in the installation.

A model of the reactions in the metabolic network was needed as the dataset for this analysis. This was a text file which contained a list of all the reactions in the metabolic network, including the iso-propanol formation branch. The text file begins with 'Structural()' providing a description of what the file will be used for and how it will be processed, being subject to a structural analysis and not a kinetic analysis. Furthermore, another directive 'External(H2O,H) is added at the top of the file, this labels water and protons (H) as external metabolites that are present in the environment of the organism. Finally, each reaction was written in a specific format in order to be readable in ScrumPy, for example:

FORMATE DEHYDROGENASE:
1 CARBON DIOXIDE + 1 NADH = 1 NAD + 1 FORMATE
~

This line states the reaction name or enzyme name at the top. The next line displays the actual chemical reaction and how many molecules of each metabolite are used. The final line contains a tilde which replaces the kinetic statement. This is then repeated for every reaction in the network. The information of each reaction can be acquired from the MetaCyc database (https://metacyc.org), which uses EC numbers to identify enzymatic reactions along with specific reaction names. These two standardised properties were used throughout this project to present reactions in the network.

Transporters in the network are also listed in this file and define metabolites which are exchanged with the environment. These metabolites include the substrates CO_2, CO, methanol and the products acetate, butyrate and iso-propanol. These were formatted as such within the file:

METHANOL_tx:
x_ METHANOL = METHANOL
~

This symbolises *E. limosum*'s capability to exchange methanol between the environment and the cytosol.

Lastly, in order for the model to work in a structural analysis it must be balanced, thus other reactions which are used in the cell to balance the pathways were added into the model. These include an RNF complex, known to be present in *E. limosum* [21], ATP synthase, and a transhydrogenase. Although, these have not been explicitly stated as present within *E. limosum*, results from growth experiments in the laboratory suggest that transhydrogenase activity must be present. However, the mechanisms are not yet known.

Once the file has been created, it can be loaded into ScrumPy using the command ' ScrumPy.Model('Name of the Model.py'), an EMA, knockout analysis or FBA can be conducted, which will be described later in the subchapter. The results of the analysis will be displayed in the ScrumPy terminal and from there can be copied out into an excel or saved as a CSV file to construct a table. The FBA results can be saved into a CSV file format showing all reactions with their corresponding flux.

Once the text file containing the model of the Wood-Ljungdahl pathway has been loaded into ScrumPy, the metabolic analysis can begin using the ScrumPy specific commands. These are shown in a tutorial on the ScrumPy website, a step by step method of the commands used during this analysis will be provided.

In the EMA, the model is loaded into ScrumPy and assigned under a variable name. The elementary modes of a model can then be analysed using the command:

elmo=Model.ElModes()

Which takes the name of the model 'Model' and determines the elementary modes and stores them in the variable elmo. The elementary modes are then converted into numerical modes, which provide a better format and representation of the modes for analysis, among other things. This was achieved using the command:

nelmo=elmo.GetNumericalModes(elmo.mo.cnames)

Breaking down this command, 'elmo.mo.cnames' is a matrix of the elementary modes. The 'mo' part is used to create a matrix showing the relationship between reactions and modes, with 'cnames' meaning that the matrix will use the reaction names (column names) from the model alongside their respective modes. This is subject to the command 'GetNumericalModes()' which converts the values of the modes in the matrix to numerical values opposed to binary values. The 'elmo' contains the elementary modes, which are converted into numerical modes and stored in the variable 'nelmo'

The numerical modes in nelmo are represented in a matrix, which shows the numerical mode flux values for each reaction in every elementary mode. Using this command:

nelmo.mo

A matrix of the reactions in the model and their respective numerical modes across all the present elementary modes is displayed. This command can be applied to generate yields and analyse patterns in the elementary modes of the network. Yields can be calculated using the numerical modes by taking a value from a substrate transporter (i.e. CO_2, CO, methanol) and the product transporter (butyrate, acetate, iso-propanol) from the same elementary mode.

The next metabolic analysis' are conducted using linear programming, which allows for knockouts and FBA to be generated. The process begins by generating a linear program associated with the model.

```
lp=Model.GetLP()
```

This command generates a linear program 'lp' which is set to minimise a set of fluxes subject to the steady-state and irreversibility constraints, investigations of the model are then able to be carried out by manipulating the 'lp'. In regards to attempting knockouts in the pathway, the method is as follows.

```
substratelist=['COtx', 'CO2tx','METHANOLtx']
def knockoutanalysis(Model, knockoutreaction):
    lp=Model.GetLP()
    lp.SetObjective(Model.sm.cnames)
    for i in Model.sm.cnames:
        if 'tx' in i and i not in substratelist:
            lp.SetFluxBounds(i:(None,0))
    lp.SetFixedFlux('ATP SYNTHASE':1.0)
    lp.SetFixedFlux(knockoutreaction:0.0)
    lp.Solve()
    sol=lp.GetPrimSol()
    return sol
```

By starting the function 'knockout analysis()' with the arguments of the pathway model and the reaction that should be knocked out. ScrumPy determines whether or not the pathway is functional with the specified reaction knocked out and provides the fluxes through the network should there be a feasible solution. A substrate list is set defining which external metabolites are the substrates for the metabolic network and can be taken up by the organism.

The objective function is set using 'lp.SetObjective(Model.sm.cnames)'. For this project, an objective function was chosen which minimises the total flux in the network. ATP SYNTHASE is set to a flux of 1 indicating that ATP must be produced in a living organism. Consequently, the knockout would be lethal if the remaining network is incapable of producing ATP for cellular growth.

The parameter 'knockoutreaction' identifies the reaction that has been chosen to be knocked out of the pathway. Its flux is set to 0 and, thus, this reaction does not carry any flux.

Finally, the function 'lp.Solve()' tries to find a feasible solution for the mutated network. The result of this operation is indicated by either 'Optimal', when a feasible solution could be found, or 'No feasible' when the knockout is lethal. For a feasible solution, the flux through the network can be determined using the command 'sol=lp.GetPrimSol()' and returned by the command 'return sol'.

This ties in with FBA, as the flux can be analysed by printing 'sol=lp.GetPrimSol()' using 'print sol'. When conducting a general FBA the constraints applied to create the knockout are not needed, although the ATP SYNTHASE should still be set to 1, as it provides the flux for the basic state of the network.

When conducting a flux change, a FBA with multiple constraints can be compared to the original FBA, for example, knockout constraints and constraints which pass a fixed flux through the substrate transporter to see how much of a certain product is produced under a specific substrate intake; in a similar process to how ATP SYNTHASE has a fixed flux in the knockout analysis.

Flux results generated under constraints (including the iso-propanol branch) and alternative reactions can be compared for changes using ESCHER maps.

ESCHER provides a tool which allows FBA results to be loaded into an assembled pathway map and visualised, as previously explained in this chapter. In regards to a flux change comparison there should be two columns of flux results, for instance: One column which has the flux results of the normal Wood-Ljungdahl pathway and another column which has the flux results when an iso-propanol formation branch is added. The CSV file with the flux changes in can be loaded into ESCHER, through the option 'Load reaction data' under the 'Data' option in the toolbar. Beforehand, the ESCHER pathway map should be loaded into ESCHER; as generated in the pathway map chapter. This allows for a visualisation of the FBA results and flux changes on the pathway map.

Overall, ScrumPy allows for an in-depth analysis of the extended Wood-Ljungdahl pathway, by providing results such as optimal yields and elementary modes. These can be used to identify patterns and show how this metabolic network operates, which products are formed, how much of the products it produces, and what are the feasible modes/routes in the network. In regards to using this network to produce valuable bio-commodities such as iso-propanol, a invaluable insight into the understanding and effects on the Wood-Ljungdahl pathway when a synthetic iso-propanol branch is added can be attained. Furthermore, FBA and knockout analysis can help determine whether the pathway can be manipulated into positive changes, for instance; an increased production in iso-propanol as a result of a knockout. More flux change comparisons can also provide an insight into the constraints which may or may not lead to increased iso-propanol production, as well as further insight to how the Wood-Ljungdahl pathway functions in the *E. limosum* wildtype.

4.4 Elementary Mode Analysis

In this chapter the aim was not only to develop an understanding of the Wood-Ljungdahl pathway in *E. limosum*, but also focusing on the incorporation of the iso-propanol pathway. Iso-propanol is a biofuel/biocommodity of particular interest in the biotechnology industry, with current objectives focused on increasing production within organisms through genetic manipulation.

The elementary mode is an irreducible pathway consisting of a minimal set of biochemical reactions and transport processes at steady state, that cannot be decomposed into simpler subnetworks without loss of functionality. The metabolic activity at steady state is then represented as a linear combination of those elementary modes [62, 63]. Allowing for carbon monoxide, carbon dioxide and methanol as carbon sources, the EMA indicated the presence of 74 elementary modes all of which were unique. While the naturally found Wood-Ljungdahl pathway resulted in 45 elementary modes, the addition of the Isopropanol pathway added 31 new elementary modes to the network.

The initial insights made from the EMA without an Isopropanol pathway, is that carbon monoxide appears to be the preferred substrate. Interestingly, the methanol branch also appears to act in a reversed manner, where the methyl group from the methylated corrinoid proteins is transferred onto coenzyme M. This could suggest an importance in the recycling of corrinoid proteins using the methyltransferase involved in methanol metabolism, which is incorporated into a potential recycling mechanism of (methylated) corrinoid proteins. There are 3 other processes within the pathway which require the methyl group from the methylated corrinoid proteins; acetyl-CoA synthase, 5-methyltetrahydrofolate:corrinoid/iron-sulfur protein Co-methyltransferase (METHCOCLTH-RXN/ EC 2.1.1.258), and an independent methyltransferase (RXN12908/EC 2.1.1.245). This could suggest that methylated corrinoid proteins produced in the methyl branch are recycled back to corrinoid proteins using the methanol branch as well as through acetyl-CoA synthase. With the independent methyltransferase (EC 2.1.1.245) potentially used alongside one of these reactions. The resulting methanol production could be used as a potential fuel source for the bacteria, although this is speculative.

Another point of interest from the EMA was which methylene tetrahydrofolate reductase was preferential in *E. limosum*, both catalyse methylene tetrahydrofolate GLU-N to 5 methyl tetrahydrofolate GLU-N using different cofactors. Ultimately, both are used in the pathway, however the preferred reaction was catalysed by Ga0111251_113114 a methylenetetrahydrofolate reductase [NAD(P)H] (EC 1.5.1.20) which also converts NADH to NAD^+, while methylenetetrahydrofolate reductase (ferredoxin) (EC 1.5.7.1),Ga0111251_11346, oxidises reduced ferredoxin.

In regards to the products, there were 16 modes which did not contain acetate and 18 which did not contain butyrate. Butyrate was produced in every mode it was present in and never consumed, while acetate was shown to be primarily consumed. It is not surprising to see this, as butyrate production requires acetate and acetate production leads to ATP generation balancing the ATP requirement from formate dehydrogenase. Both of which are of importance in the Wood-Ljungdahl pathway. Moreover, there appears to be no patterns present between acetate and butyrate consumption/production in the elementary modes, other than what was described in the previous statement.

When the iso-propanol pathway is included in the model and the EMA conducted, the results show there are 74 elementary modes present. The EMA shows that carbon dioxide and carbon monoxide are the preferred substrates, being consumed in 28 and 42 elementary modes respectively. Although, carbon dioxide is produced via carbon monoxide and utilised in 26 modes. Furthermore, carbon dioxide is at a net production in all modes it is present in (63), while carbon monoxide is always consumed. Methanol is is present in 30 modes. In 14 of these modes it is consumed, while it is produced in 16.

Moreover, when looking at the products, there are some interesting points to note. Firstly, iso-propanol and butyrate are never produced in the same elementary mode, both products are mutual exclusive. Secondly, more elementary modes which lead to iso-propanol production (29) are present than those which lead to butyrate production (27), while acetate is only shown to be produced in 11 modes but consumed in 35 modes. Importantly, acetate is always consumed in butyrate and isopropanol-producing modes, a link that is to be expected. In all modes neither producing butyrate or isopropanol, acetate is either produced or not present. The latter occurs in only 10 modes. These findings clearly indicate the importance of acetate formation for the production of butyrate and isopropanol, respectively.

The focus of this investigation is to assess iso-propanol formation when it is bio-engineered into the Wood-Ljungdahl pathway found in *E. limosum.* The EMA presented some interesting modes in regards to iso-propanol, where it was the only liquid metabolite produced.

Reaction name	Mode 1	Mode 2	Mode 3
RXN5061	0	4/5	0
CO_tx	5/3	13/5	11/7
RXN8096	4/3	0	0
ACETOACETATE DECARBOXYLASE	2/3	2/5	2/7
EC15120	0	0	4/7
PHOSACETYLTRANS	2/3	2/5	2/7
PYRUVATE_tx	0	0	0
FORMATETHFLIG	0	4/5	4/7
BUTYRATE_tx	0	0	0
METHENYLTHFCYCLOHYDRO	0	4/5	4/7
ATPSYNTHASE	1/3	1	1
RXNR11	0	0	0
FORMATE DEHYDROGENASE	0	4/5	4/7
BUTYRYL COA DEHYDROGENASE	0	0	0
ENOYL CO HYDRAT	0	0	0
EC1274	-1/3	-9/5	-1
ACETYLSYNCLTH RXN	4/3	4/5	4/7
METHANOL_tx	4/3	0	0
METHYLENETHFDEHYDROGENASE_NADP	0	4/5	4/7
ACETATEKIN	2/3	2/5	2/7
ISO PROP	-2/3	-2/5	-2/7
ACETOACETYL COA TRANSFER	2/3	2/5	2/7
RXN13449	0	0	0
CO2_tx	-1	-7/5	-5/7
METHCOCLTH	0	4/5	4/7
TRANSHYDROG	2/3	2/5	2/7
RNF	2/3	2	2
MTAMBARKMULTI_RXN	4/3	0	0
ISOPROPANOL DEHYDROGENASE NADP 2B	2/3	2/5	2/7
ACETATE_tx	0	0	0
PYRUFLAVREDUCT_RXN	0	0	0
ACETYL COA ACETYLTRANSFER	2/3	2/5	2/7

Figure 4.2: Table of the elementary modes present in which only iso-propanol is produced, the reaction names are labelled in the left most column, while the other 3 columns provide the flux through each of the reactions for all 3 modes.

In Mode 1, carbon monoxide and methanol are the substrates, whereas the other two modes use only carbon monoxide. For example in Table 4.2 in mode 3 'CO_tx' possesses a positive flux (as a transporter this means it is being consumed), it can then be seen that the reaction 'EC1274', anaerobic carbon-monoxide dehydrogenase, has a negative flux indicating that it is producing carbon dioxide from carbon monoxide. The reaction 'ACETYLSYNCLTH RXN' has a positive flux which shows that carbon monoxide is also used to produce acetyl-CoA as well. Interestingly, it appears that some of the carbon dioxide produced by carbon monoxide is used in the methyl branch, as there is a positive flux passing through methyl branch reactions such as 'FORMATETHFLIG' and 'METHENYLTHFCYCLOHYDRO'.

Once acetyl-CoA is produced the modes show that iso-propanol is generated in the same procedure as normal, with acetyl-CoA split between acetate and acetoacetyl-CoA production; which are both required to produce iso-propanol. Acetate is produced through the 'PHOSACETYLTRANS' and 'ACETATEKIN' reactions and the ACETYL_COA_ACETYLTRANSFER' reaction. All of the acetate and acetoacetyl-CoA is then used to produce acetoacetate through 'ACETOACETYL_COA_TRANSFER', which leads to a net flux of 0 through the acetate transporter 'ACETATE_tx' as the amount produced was then consumed in iso-propanol formation. Lastly, iso-propanol is formed from acetoacetate through the reactions 'ACETOACETATE DECARBOXYLASE' and 'ISOPROPANOL DEHYDROGENASE NADP 2B'.

The differences between each mode only occur before acetyl-CoA was produced, in the methyl, carbonyl and methanol branches, most likely due to there being only one way of producing iso-propanol. It can be seen in Mode 1 that the methanol branch is used to generate methylated corrinoid proteins and thus acetyl-CoA, while the methyl branch is not used at all. Moreover, the ATP synthase and RNF complex carry a small flux in Mode 1 (1/3,2/3) compared to Mode 2 and 3 (1,2). The reduced dependency on ATP generation via RNF and ATP synthase in Mode 1 in comparison to Modes 2 and 3 allows this mode to produce more isopropanol. This is caused by the fact that the use of methanol means that the ATP-dependent formate-tetrahydrofolate ligase in the methyl branch. The only other difference is that Mode 2 uses the NADH-dependent methylene tetrahydrofolate reductase reaction ('RXN5061'), while Mode 3 employs the ferredoxin-dependent version of this reaction.

The EMA results did not show any presence of the independent tetrahydrosarcinapterin dependent methyltransferase 'EC 2.1.1.245', which was included in the model. This is most likely due to its dependency of tetrahydrosarcinapterin, which is not produced in this pathway and thus any use of this methyltransferase would lead to a tetrahydrosarcinapterin imbalance. Although, its presence in the Wood-Ljungdahl pathway could be involved with the folate dependent methyltransferase (EC 2.1.1.258) or acetyl-CoA synthase (EC 2.3.1.169), previously mention in this section. This methyltransferase could be a subreaction of either of these methyl group utilising reactions and due to these reactions being defined under one larger 'lumped' reaction, thus the presence of a singular subreaction is unnecessary, unless all subreactions are present. Another suggestion is that when considering 2 molecules of methanol as the only substrate present. One molecule may be converted into methylated corrinoid-Fe-S-Proteins, through the methanol branch. The other may be used to produce carbon dioxide through the methyl branch, which is then utilised by the carbonyl branch to produce carbon monoxide and thus acetyl-CoA. This split of the methanol branch into the methyl branch may use a different methyltransferase to EC 2.1.1.258, which might be EC 2.1.1.245. Although, this is again speculative suggestions.

4.5 Yields

The EMA results can also be used to find the optimal yields of all the products. This provides an insight into whether products in the network can be produced in significant amounts and any patterns between the ratios of production. The carbon conversion yields were calculated in order to compare yields of products in which the substrates and products have different sizes and carbon contents, thus a Carbon Conversion Yield (CCY) is a more accurate and comparable yield. The CCY is calculated by first multiplying each substrate and product by the number of carbon atoms present in each molecule. The sum of all consumed carbon is calculated, which is essentially the sum of all the substrates in the modes. Finally, the CCY is then calculated for each product of interest as shown in the formula below. This is done for every mode and every combination of products and source compounds.

$$\mathrm{CarbonConversionYield} = \frac{\mathrm{product(producedcarbon)}}{\sum \mathrm{substrates(consumedcarbon)}}$$

An Excel file provide the method with annotations of how exactly the optimal carbon conversion yields were calculated in each mode is available in the appendix.

Products	Optimal carbon conversion yields (CCY)
Iso-propanol	0.667
Butyrate	1
Acetate	1

Figure 4.3: A table which shows the optimal carbon conversion yields of products of interest and how many elementary modes they are present within the Wood-Ljungdahl pathway

The CCY's were calculated for the products of interest in the Wood-Ljungdahl pathway, shown in Table 4.3. The CCY shows how much of the carbon consumed is used to produce a certain product, represented by a value in the range [0, 1], with 0 showing no production and 1 showing all carbon consumed is converted into the product.

The optimal CCY for iso-propanol was 0.667 and was present in the mode which used carbon monoxide and methanol as substrates. It was the lowest CCY compared to acetate and butyrate's CCY, but this still showed a majority (two thirds) of carbon consumed was converted into iso-propanol in this elementary mode. Out of the 25 modes 11 showed iso-propanol production, which was more than acetate and butyrate, which were produced in 3 and 9 modes respectively. The optimal CCY for acetate was 1, which states that all the carbon consumed in this elementary mode was converted into acetate and is the optimal conversion, this was the same with butyrate which also had a optimal CCY of 1, much like iso-propanol both were produced from the substrates carbon monoxide and methanol. The high optimal CCY for acetate may be due to the fact that ATP generation coincides with acetate production, which suggests the importance of carbon flow passing through the acetate branch. Interestingly, the optimal CCY for butyrate is larger than that of Iso-Propanol, implying that more butyrate may be produced and done so more efficiently compared to iso-propanol, despite both being produced in a similar process. The difference may relate to the production of carbon dioxide and presence of flux through ATP synthase, RNF complex, and transhydrogenase in the iso-propanol optimal CCY mode, while not in the butyrate mode. The flux through ATP synthase, RNF complex, and transhydrogenase, alongside carbon dioxide production, could act to balance the network when iso-propanol is produced, thus some of the carbon consumed is diverted away from iso-propanol production.

The elementary modes which produced the optimal CCY for acetate, butyrate, and iso-propanol were analysed to provide further insight into the rest of the pathway. In all of these elementary modes, there was no utilisation of the methyl branch, as methylated corrinoid-Fe-S-proteins were generated through the methanol branch, without consuming ATP. Although carbon dioxide was not present in the acetate and butyrate modes, it was present in the iso-propanol mode alongside other reactions as previously described.

Furthermore, a colleague Rob Mansfield who was experimentally investigating growth of *E. limosum* ATCC 8486 shared some results, which included the production (mM) of the products mentioned in this chapter over time intervals. Using the results from both the optimal CCY's and the laboratory yield ratios, a comparison between results from the model and laboratory can be conducted, with the results provided in the appendix.

The results provided a ratio between butyrate and iso-propanol production, which allows for a comparison of production and changes in cellular behaviours over time. While the comparison between butyrate and iso-propanol can identify any differences between butyrate which is naturally produced in this pathway and iso-propanol in which production is a result of a bio-engineered pathway. The laboratory results were taken over 11 time intervals on 2 different occasions, in this comparison results were used from a 3 time intervals: 28-31hrs, 100-102hrs, 194.5-243.5hrs.

Iso-Propanol:Butyrate ration	Laboratory production results (June 2017)	Laboratory production results (April 2017)
	1.93:1(28hrs)	2:1(31hrs)
	0.18:1(102hrs)	0.97:1(100hrs)
	0.15:1(243.5hrs)	0.04:1(194.5hrs)

Figure 4.4: A table which shows the different ratios between iso-propanol and butyrate found in the laboratory results over time

Firstly, it can be seen in the laboratory results that over the duration of the experiment, iso-propanol to butyrate ratio decreases. This is an interesting pattern to observe as it appears that there might be a reason for *E. limosum* to produce iso-propanol during its initial growth and switch to butyrate production later on. It can be seen that the CCY ratios obtained from the model would match the ratio of the laboratory results at a time interval in-between 30-100 hours. When comparing this to the elementary modes, in the optimal CCY mode for iso-propanol flux can be seen passing through ATP synthase, generating ATP. This could show why iso-propanol production is preferred to butyrate production during the initial growth phase.

It is clear that iso-propanol is not produced as much as butyrate within *E. limosum* and as such it appears butyrate production may be more favoured in the pathway, as a means of converting acetate to acetyl-CoA. Overall, the model could be seen as accurately representative the iso-propanol to butyrate ratio, as it appears this ratio would be present in the laboratory grown strain between the 30-100 hour time interval, as the ratio is shown to decline. Moreover, at the later time intervals the ratio indicates that iso-propanol production becomes near insignificant at 0.15:1 and 0.04:1, showing that iso-propanol formation is incredibly inefficient in *E. limosum* during later stages of the experiments.

4.6 Flux Balance Analysis

The next structural modelling technique that was carried out was the Flux Balance Analysis (FBA). This method was able to provide an understanding of the genes/reactions that could be potentially knocked out and what effect these may have on the pathway. In order to simulate the knock out of genes/reactions in the model, the flux constraint on the knockout target was set to 0 as describe in the methods. While the other constraint was set so that exactly 1 ATP must be produced in order to recreate the environment in the organism that unless 1 ATP is generated the pathway will not be utilised.

The results from the knockout analysis displayed that every single gene/reaction in the pathway was able to be knocked out. Although the pathway is still able to function after all the knockouts have occurred, when crucial pathways are disrupted alternative means are used to compensate this, such as if the methyl branch is disrupted because of a knockout, the methanol branch will act to compensate. There were 3 different flux balances which occurred in the pathway as a result of single knockouts.

Reaction name	Flux Balance 1	Flux Balance 2	Flux Balance 3
RXN5061	0	0	0
CO_tx	1	3	1.66
RXN8096	1	-2	0
ACETOACETATE DECARBOXYLASE	0	0	0
EC15120	0	2	0.66
PHOSACETYLTRANS	1	0	0.66
PYRUVATE_tx	0	0	0
FORMATETHFLIG	0	2	0.66
BUTYRATE_tx	0	0	0
METHENYLTHFCYCLOHYDRO	0	2	0.66
ATPSYNTHASE	1	1	1
RXNR11	0	0	0
FORMATE DEHYDROGENASE	0	2	0.66
BUTYRYL COA DEHYDROGENASE	0	0	0
ENOYL CO HYDRAT	0	0	0
EC1274	0	-3	-0.99
ACETYLSYNCLTH RXN	1	0	0.66
METHANOL_tx	1	-2	0
METHYLENETHFDEHYDROGENASE_NADP	0	2	0.66
ACETATEKIN	1	0	0.66
ISO PROP	0	0	0
ACETOACETYL COA TRANSFER	0	0	0
RXN13449	0	0	0
CO2_tx	0	-1	-0.33
METHCOCLTH	0	2	0.66
TRANSHYDROG	0	0	0
RNF	0	6	0
MTAMBARKMULTI_RXN	1	-2	0
ISOPROPANOL DEHYDROGENASE NADP 2B	0	0	0
ACETATE_tx	-1	0	-0.66
PYRUFLAVREDUCT_RXN	0	0	0
ACETYL COA ACETYLTRANSFER	0	0	0

Figure 4.5: A table which shows the different fluxes in Wood-Ljungdahl pathway and extended metabolic networks dependent on the gene/reaction which has been knocked out

The 3 different fluxes can be seen depending on which gene/reaction is knocked out. The flux in 'Flux Balance 2' is a result of the 'ACETYLSYNCLTH RXN', 'PHOSACETYLTRANS' and 'ACETATEKIN' genes/reactions being knocked out, which are involved in acetyl-CoA and acetate production respectively. The flux in 'Flux Balance 3' is a result of the 'MTAMBARKMULTI RXN' and 'RXN8096' genes/reactions of the methanol branch being knocked out. While the flux in 'Flux Balance 1' is the result of the remaining genes/reactions being knocked out, which includes the butyrate, Iso-Propanol, carbonyl and methyl branch except for the 'ACETYLSYNCLTH RXN' reaction/gene.

It can be seen that any knockout in the pathway under the set constraints leads to neither butyrate or iso-propanol being produced, while in 'Flux Balance 1' and 'Flux Balance 2' it is shown that acetate is produced. Under the constraints 'Flux Balance 1' the methanol and carbonyl branch are utilised in order to produce acetyl-CoA and then using the acetate branch in order to produce acetate. Although, it has been seen that acetate may be recycled into acetyl-CoA as part of butyrate or iso-propanol production, this is not present in any of these knockout fluxes, where no flux passes through either the butyrate or iso-propanol branch.

'Flux Balance 3' is similar to 'Flux Balance 1' except that the methanol branch is knocked out and, thus, cannot be utilised anymore. Instead acetyl-CoA is generated through the methyl and carbonyl branches, where carbon monoxide is used to produce carbon dioxide and acetyl-CoA, as one would from the native Wood-Ljungdahl pathway. Although most of the carbon dioxide is used to produce acetyl-CoA through the methyl branch, some is excreted as a product. Acetate is then generated in the same fashion as in 'Flux Balance 1', however the flux in 'Flux Balance 3' is lower as a result of some carbon dioxide being excreted as a product. This implies that methanol and the methanol branch provide a greater flux/production as a source when compared to carbon dioxide and the methyl branch as a source. This might be caused by the reduced ATP requirement to generate acetyl-CoA.

In 'Flux Balance 2' the genes/reactions necessary to produce acetate were knocked out, thus flux only passes through the methyl, carbonyl and methanol branches. Interestingly, carbon monoxide is utilised as the main substrate, due to 'ACETYLSYNCLTH RXN' being knocked out carbon monoxide can only be used to produced carbon dioxide. Although some carbon dioxide is excreted, the majority is consumed through the methyl branch leading to methanol production, thus methanol and carbon dioxide are used as production outlets. Furthermore, despite ATP being consumed in the 'formate dehydrogenase' reaction in this set of fluxes, it is not produced, due to 'ACETATEKIN' being knocked out. This appears to leave the pathway in an ATP deficit, while if the methyl branch was not utilised and only carbon dioxide produced, it would not be the case. Consequently, while such network might be interesting for the production of methanol, it seemingly does not support cellular growth.

Multiple knockouts were then conducted in a similar procedure to the single knockouts, except with 2 different reactions set to 0 flux. A loop was used in the knockout script to check the feasibility of every combination of multiple (2 genes) gene knockouts. The results showed that every combination was feasible, although there were too many combinations for an in-depth analysis of the fluxes to be carried out. The fact that all multiple knockout combinations resulted in a feasible pathway is a positive outcome for genetic manipulation of the pathway.

The final step of the FBA was to compare flux changes when the model is altered, in this case the model was altered by removing the iso-propanol branch. The intent of the flux change comparison is to analysis the affect of bio-engineering a iso-propanol formation branch in *E. limosum* and any other factor which might effect products of interest.

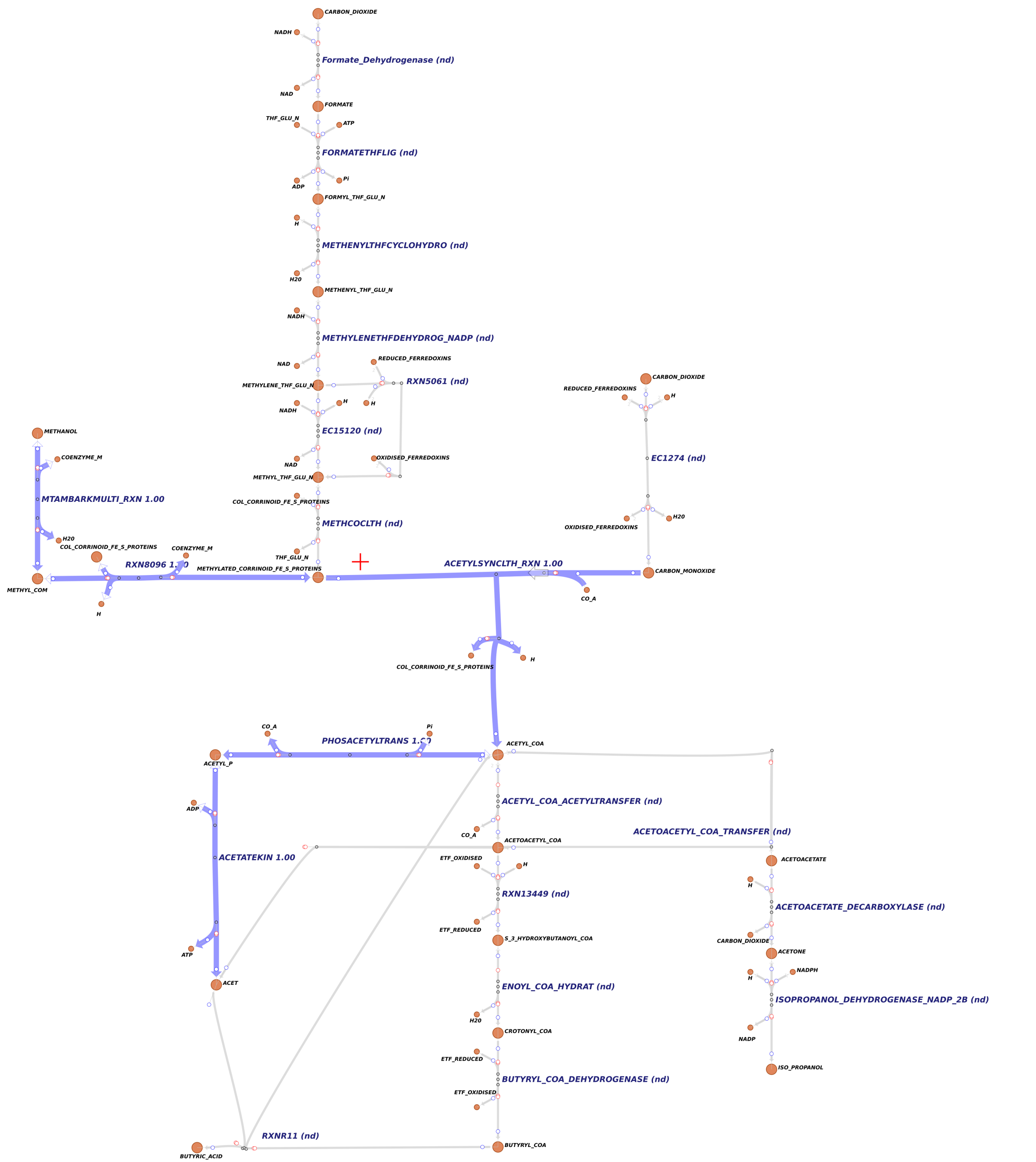

Figure 4.6: A figure which shows a flux map representing the flux through the pathway when both the iso-propanol branch is and is not included in the FBA.

Figure 4.6 showed that when carbon dioxide, carbon monoxide and methanol are set substrates and the flux through ATP Synthase was set to 1.0, the presence of the iso-propanol formation branch has no effect on the flux in the FBA, thus only one flux map is present in the figure. The flux map shows that the preferred substrates are methanol and carbon monoxide, with carbon monoxide only used to produce acetyl-CoA, despite being known to produce significant amounts of carbon dioxide; which is not present at all. While methanol appears to be the preferred source for methylated corrinoid protein production under the current constraints, again opposed to carbon dioxide. This is likely due to ATP consumption in the methyl branch when carbon dioxide is used, while the constraints in the model aim to generate ATP. However, it seems the methanol branch is shorter than the methyl branch and only requires 1 proton and a corrinoid protein in order to generate methylated corrinoid proteins. Meanwhile, the methyl branch requires ATP, NADH, THF GLU N, H, and reduced ferredoxins, which in comparison may lead to a less efficient branch for generating methylated corrinoid proteins than the methanol branch.

The acetate branch is then used as the only method of ATP generation in the pathway satisfying the pathway constraints set. Seemingly, the butyrate and iso-propanol formation branches are not required to satisfy the constraints, thus such any alterations to the model which do not affect the acetate branch, 'ACETYLSYNCLTH RXN' or the methanol branch, will not change the flux. The flux map is able to reinforce the similarities between the iso-propanol and butyrate branch, despite the iso-propanol branch not being naturally present in *E. limosum*, with neither of the branches appearing crucial in the pathway. It seems that butyrate formation is not required in a basic survival/ATP generating state of the bacteria and thus not hugely important.

4.7 Discussion

The results attained from the structural modelling can provide an insight into the functioning of the Wood-Ljungdahl pathway within *E. limosum*, as well as investigating some initial genetic manipulations in the model. The overall aim being to understand what effects iso-propanol production has on the metabolic activity and how the Wood-Ljungdahl pathway functions. An objective throughout this chapter was to gain further insight into iso-propanol formation in *E. limosum*. It was seen in the EMA that there were only 3 modes where iso-propanol was solely produced, all these modes had the similar routes but differing fluxes. The route which possessed the largest flux through the iso-propanol transporter used methanol as the substrate, opposed to carbon dioxide which was used in the other 2 modes. This could indicate that the use of the methanol branch instead of the methyl branch could increase production and flux of a product. Moreover, the methyl branch is known to consume ATP, which could suggest why a larger flux is present in the mode which does not use it, although, it still remains a highly used branch implying that there may be another reason for its presence and use.

The EMA showed that butyrate and iso-propanol were never produced together in the same mode, even in the EM's which possessed the greatest ATP generation these products. This most likely suggests that a combination of elementary modes is active in *E. limosum*, allowing it to produce butyrate and iso-propanol simultaneously under appropriate constraints and objectives. Ultimately, the butyrate and iso-propanol branches share many similarities which could support them being simultaneously used, the EMA shows that there are 3 EM each which only produce butyrate or iso-propanol, while overall there are 29 modes in which iso-propanol production is present and 27 modes in which butyrate production is present. The even split in EM patterns indicates that there is no preference between the butyrate or iso-propanol formation, despite the butyrate branch naturally occurring in *E. limosum* it does not appear to provide a substantial benefit or integral product under the studied conditions, as less iso-propanol production modes would be expected. It would also be of interest to see the effect of acetate production on butyrate and iso-propanol production, as both require acetate to be produced. It could be debated whether duplicating the acetate branch or adding a different acetate formation branch could lead to increased production of iso-propanol, butyrate or both.

Another point of interest, is the route acetate takes when either butyrate or iso-propanol is produced. The elementary modes show that acetate may be consumed in one of three ways to produce butyrate and iso-propanol. Firstly, acetate is produced through the acetate branch and converted into either butyrate or iso-propanol through the acetate CoA transferase reaction. Secondly, external acetate is taken up into the cell and converted into butyrate or iso-propanol. Lastly, external acetate is taken up into the cell, converted into butyrate or iso-propanol, and converted into acetyl-CoA passing through the acetate branch in a reverse manner. It is also interesting to note that the last of the three methods is present when the methyl, carbonyl and methanol branches are not used and particularly when butyrate is produced, while in terms of iso-propanol production it seems the carbonyl and methanol branches are utilised. An understanding on whether different methods of acetate consumption or production may alter iso-propanol production, which could potentially suggest why its optimal CCY is lower than butyrates. This would be an interesting area for future work.

When calculating the optimal CCY of products, iso-propanol was shown to have a lower optimal yield than acetate and butyrate. Firstly, this indicated that the iso-propanol branch is active in the bacteria, which can be further backed up by the laboratory results showing that flux is directed towards iso-propanol production. The fact that acetate and butyrate have a higher CCY than iso-propanol may be expected. When considering the modes which generated these optimal CCY's and the substrates used, flux is only present through the carbonyl and methanol branches, with no utilisation of the methyl branch. Although carbon dioxide is produced alongside iso-propanol, it is not produced alongside butyrate, leading to a lower CCY.

When comparing the experimentally observed iso-propanol to butyrate ratios, it could be seen that the production ration decreased over time. This implies that during the initial growth period formation of iso-propanol is favoured, whereas during later growth stages butyrate formation becomes preferred. Interestingly, the objective of the cell differs through different growth periods, with ATP generation being the main objective during initial growth and redox and acid formation balance the main objectives during stationary growth. Whether there is a link between ATP generation being a main objective during initial growth and iso-propanol production being preferred, would be a further point of interest to look into. Furthermore, the link between redox balance objective and butyrate production at stationary growth could be further elucidated. Although, when assessing the production ratio's it is important to state that the values are only comparative of one another, not necessarily that large amounts of iso-propanol is produced in the initial growth phase, just more than butyrate. Whether or not more or less iso-propanol or butyrate was produced during the initial growth phases compared to stationary growth phases cannot be stated from these results.

A comparison between the production ratios from the structural model and laboratory results could have allowed for a more accurate insight into the model. However, the model would have to include kinetic and biological regulations which were not considered within the scope of the project such as; product inhibition, thermodynamic regulation, pH, activity of elementary modes. Instead the model was a structural model, which only represents the steady state solution and is not based on any dynamic constraints such as time.

The FBA was used to conduct a knockout analysis and flux change comparisons. The knockout analysis found that every gene/reaction involved in the pathway could be knocked out, although this resulted in 3 different sets of pathway fluxes, none of which contained iso-propanol or even butyrate. One suggestion may be that the knockouts lead to the pathway acting in a 'survival' manner, whereby it accomplishes the bare essential only producing acetate and generating ATP, rather than acetate then being used to produce iso-propanol or butyrate. On the other hand, this may be a repercussion of the model constraints, as once the knockout constraint and the ATP synthase set flux of 1 have been satisfied, the FBA does not showcase any further flux activity which may have led to production of iso-propanol or butyrate. At face value, the FBA results suggest that any single gene knockouts under the assumption that *E. limosum* has an objective of generating ATP, acetate production is preferential to butyrate or iso-propanol production. While, the EMA shows that 'normally' the pathway shows a preference to iso-propanol and butyrate production above acetate, which is heavily consumed. Moreover, the observation that when acetate production is obstructed via the knockout, methanol was seen to be produced instead. Without the ability to produce acetate, it would be assumed that the pathway would not be functional. However, it appears that *E. limosum* may have a reason for producing methanol when in this state, opposed to carbon dioxide and carbon monoxide. This questions whether the pathway would still be functional if methanol and acetate branches had a reaction knocked out which prevented the production of either. All combinations of multiple knockouts were feasible, which suggests an even more simplified functionality of the pathway could be present, but due to the amount of possible combinations it was a large task to assess this. The knockouts can also provide an understanding of substrate manipulation, such as which knockouts prevent certain substrates being used and what the alternative substrates are. This could provide knockouts the ability to the manipulate the preferred substrate route, such as preventing carbon monoxide being used for carbon dioxide production, thus moving flux towards acetyl-CoA production and potentially increase product of interest yields. Furthermore, by knocking out genes which are essential in substrate branches (methyl, carbonyl and methanol), the disruption of acetyl-CoA production might require the bacteria to break down acetate in order to produce acetyl-CoA, which could use the butyrate/iso-propanol branches. However, the questions are whether the bacteria can be forced into taking up acetate due to knockouts disrupting acetyl-CoA production, whether this would prevent them taking up carbon dioxide or carbon monoxide. Which is a major benefit of the Wood-Ljungdahl pathway.

This also questions issues with structural modelling. The model is based upon the reactions stored on the MetaCyc database, including the information provided about the direction of the reaction. Firstly, these directions may be incorrect as they depend on the kinetic and thermodynamic laws used to determine the feasibility of chemical reaction direction. As this is not a kinetic model kinetic or thermodynamic information specific to these reactions was not provided, thus the direction on MetaCyc must relied upon. However, the MetaCyc database is backed by evidence from other studies, although there can be no certainty in these results or whether they transfer to *E. limosum*. Overall, this could lead to inaccuracies throughout the model, EMA and FBA. Although, this is what was expected from the model, it is used to provide a general insight opposed to accurate assessment.

The FBA was used to generate a flux map showcasing the flux present through the reactions in the pathway, which was to be compared to flux maps with pathway alterations. The results showed that the alterations made all led to the same flux map generation, as seen in figure 4.6, which shows a basic flux through the pathway which enables ATP generation. While an objective of this chapter was to understand the impact genetic manipulation could have on iso-propanol production, it appears as though there is a limited impact based on these results. However, this depends on how the flux is determined in the FBA, as the flux could represent the minimal state after the constraints have been satisfied. As the constraints are satisfied when ATP is generated by acetate kinase, the flux would end with acetate production. Whether an organism whose main objective is ATP generation would then use the acetate to produce iso-propanol or butyrate, which do not lead to ATP generation, shows a lack of motive for the organism. Overall, it appears that the blocking of the acetate branch or acetate production as means to impact iso-propanol or butyrate production may be redundant. As under these constraints only acetate production appears favourable.

Consequently, the constraints set during this FBA may not accurately represent the organism, due to the lack of kinetic and thermodynamic information available. Which results in the inability to conduct a detailed assessment of the pathway. Once *E. limosum* enters a stable growth phase and does not prioritise ATP production, a resultant change in the flux of the pathway may led to increased iso-propanol and butyrate production. Although, as previously stated, whether or not the amount of iso-propanol or butyrate production increases between the initial growth phase and stable growth phase cannot be stated, just that the preference switches from iso-propanol to butyrate during this transition. It can be seen from the laboratory results a colleague generated that functionality and pathway usage changes as the bacteria grows. Initially, iso-propanol production was larger than that of butyrate, over numerous hours of growth this then completely reversed. This means that the flux in the pathway is dynamic and will undergo changes depending on the objectives of the bacteria such as different growth phases. The fact that ATP conservation and generation is not considered a priority once *E. limosum* is established can be shown by the presence of the methyl branch. The methyl branch is a large highly conserved group of reactions, which consume ATP in order to produce methylated corrinoid proteins. A branch this large would not provide any benefit to the pathway when a suitable alternative of the methanol branch is present, unless the priorities of the pathway change, where using carbon dioxide as a source compound is more important to the bacteria than ATP conservation.

Although, even though iso-propanol and butyrate are not present in the FBA and knockout solutions, they still may be produced in sub-optimal or alternative solutions, which seems a possibility due to the amount of elementary modes present. In this project the FBA and knockout analysis were deriving a solution for the minimisation of total flux in the network, where butyrate and iso-propanol do not provide the minimum flux.

The addition of a second thiolase gene could also allow for a change in pathway flux. Adding a duplicate thiolase gene into the genome may act to pull flux towards butyrate and iso-propanol production and away from acetate. However, for this to be studied it must be conducted in a kinetic model in order to consider the kinetic regulations.

Another point of discussion, although it appears that the methanol branch is preferred to the methyl branch when generating methylated corrinoid proteins, this may be the result of the 'minimal' route to satisfy the constraints. The methanol branch is 4 reactions less than the methyl branch, which may explain its prevalence on the flux map and preference. Although, it is more likely to be the presence of ATP consumption in the methyl branch, which leads to methanol being the preferred source under the constraints.

Chapter 5

Conclusion

In conclusion, in regards to the title of the project 'Analysis of the metabolic pathways in the gas fermenting *Eubacterium limosum*' a broad and general understanding of *E. limosum* ATCC 8486 strain has been accomplished. Previous literature on *E. limosum* was very limited, let alone literature on the ATCC 8486 strain which only had a few papers including it. The project began by assessing the paper written by Jeong et al which classifies energy conservation methods and the Wood-Ljungdahl pathway(a gas fermenting pathway) in *E. limosum* KIST 612. The Wood-Ljungdahl pathway has been an area of interest due to the biocommodities and biofuels it produces, while consuming noxious C1 compounds like carbon dioxide, carbon monoxide and methanol. Alongside the relative ease at which *E. limosum* can be genetic manipulated compared to other acetogens, there was tremendous interest in developing an understanding of the Wood-Ljungdahl pathway in *E. limosum* ATCC 8486 at the University of Nottingham.

The pathway had already been produced by Jeong et al for *E. limosum* KIST 612, which was used as a template for defining the pathway in ATCC 8486. Using techniques such as BLAST analysis, DNA to amino acid translation as well as databases like BRENDA and MetaCyc. I was able to confirm the genes present in ATCC 8486 and develop a Wood-Ljungdahl pathway specific to this strain, while providing clarification for lesser detailed parts and identifying new metabolic networks as part of the pathway. The main findings confirmed that butyrate formation occurs through a acetate CoA transferase, opposed to a butyrate kinase which was confirmed to not be present in *E. limosum* ATCC 8486. Moreover, a presumed acetate to acetyl-CoA reaction was found to be part of the butyrate forming reaction, opposed to an separate reaction. Thus acetate must be present in order to generate butyrate. Furthermore, a methanol branch added to the Wood-Ljungdahl pathway, which was identified in papers detailing how methanol is metabolised in other acetogens. The process of methanol metabolism was then clarified and described in the text, while previously the literature only displayed it as an image. The identification of a gene which encodes for a bifunctional enzyme was found to be present, it functions as a formyl tetrahydrofolate cyclohydrolase and methylene tetrahydrofolate dehydrogenase, this was previously described as 2 separate genes. This led to the production of a pathway map representing the Wood-Ljungdahl pathway in *E. limosum* ATCC 8486 including extended metabolic networks such as the methanol branch. This was the first description of the Wood-Ljungdahl pathway in *E. limosum* ATCC 8486 and provided an insight into the pathway which was not present in literature.

A comparative genomic alignment was carried out using progressive Mauve, between the 3 strains of *E. limosum*, SA11, KIST 612 and ATCC 8486. While another alignment focused on the Wood-Ljungdahl pathway genes between ATCC 8486 and KIST 612 was also carried out. Overall, the genomes of the 3 strains were shown to have undergone heavy rearrangements including translocation and inversions. Although, it was apparent that SA11 and ATCC 8486 were more similar as they possessed fewest rearrangements between strains. The Wood-Ljungdahl pathway genes of the methyl, butyrate and carbonyl branch were conserved in there own separate region, which had shown elements of translocation but ultimately high similarity between genomes (This could be argued for the methyl branch genes). None of the genes in the acetate branch were positioned near each other in the genome, all located in different regions, which was interesting due to their involvement together to produce acetate. It would be assumed that they would be organised similarly to the butyrate or methyl branches. Ultimately, this could lead to mutations within the Wood-Ljungdahl pathway and thus a different phenotypical function. However, it has been reported that bacteria which live in harsh environments such as *E. limosum* are prone to rearrangements, which is a developed survival trait.

Iso-propanol has been identified as a valuable biocommodity and production is of interest in the biotechnology sector. An iso-propanol formation branch was bio-engineered into *E. limosum* in order to allow for iso-propanol production. In order to provide analysis of the iso-propanol branch in *E. limosum*, a structural model was created and used to analyse elementary modes, potential pathway knockouts, and flux balance/maps in regards to the iso-propanol branch's addition to the Wood-Ljungdahl pathway. Numerous metabolic modelling tools were used including elementary mode analysis, optimal carbon conversion yields (CCY) and knockout analysis. The EMA and optimal CCY were able to show that iso-propanol was part of an active pathway within the Wood-Ljungdahl pathway, with similarities between how iso-propanol and butyrate are represented throughout the EMA. Butyrate and iso-propanol were never present in the same elementary modes, but represented similarly across all the elementary modes, suggesting that numerous modes are used simultaneously in the bacteria. Iso-propanol's optimal CCY was lower than butyrates and acetates, which achieved the maximum CCY of 1. The optimal CCY provided an insight into the preferred substrate combination, preferred products, and an understanding of the flux present in an optimal CCY pathway, allowing for production patterns to be found. Although every reaction/gene could feasibly be knocked out, it did not provide any results of interest for genetic manipulation. Other than a possible method to manipulate which substrates are consumed.

Bibliography

[1] N. Khan, N. Abas, and A. Kalair, "Review of fossil fuels and future energy technologies," *Futures*, vol. 69, pp. 31–49, May 2015.

[2] S. Shafiee and E. Topal, "When will fossil fuel reserves be diminished?" *Energy Policy*, vol. 37, no. 1, pp. 181–189, Janurary 2009.

[3] D. Hem, "The future of fossil fuels - is it the end?" *Oxford Review of Economic Policy*, vol. 32, no. 2, pp. 191–205, Janurary 2016.

[4] S. Soloman, G.-K. Plattner, R. Knutti, and P. Friedlingstein, "Irreversible climate change due to carbon dioxide emissions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 6, pp. 1704–1709, February 2009.

[5] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environmental Pollution*, vol. 151, no. 2, pp. 362–367, Janurary 2008.

[6] B. Brunekreef and S. T. Holgate, "Air pollution and health," *The Lancet*, vol. 360, pp. 1233–1242, October 2002.

[7] M. Asif and T. Muneer, "Energy supply, its demand and security issues for developed and emerging economies," *Renewable and Sustainable Energy Reviews*, vol. 11, no. 7, pp. 1388–1412, September 2007.

[8] A. Haines, R. S. Kovats, D. Campbell-Lendrum, and C. Corvalan, "Climate change and human health: Impacts, vulnerability and public health," *Public Health*, vol. 120, no. 7, pp. 585–596, July 2006.

[9] I. Dincer, "Renewable energy and sustainable developmetn: a crucial review," *Renewable and Sustainable Energy Reviews*, vol. 4, no. 2, pp. 157–175, June 2000.

[10] N. L. Panwar, S. C. Kaushik, and S. Kothari, "Role of renewable energy sources in environmental protection: A review," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 3, pp. 1512–1524, April 2011.

[11] F. Imkamp and V. Müller, "Acetogenic bacteria," *eLS Wiley Online Library*, Janurary 2007.

[12] V. Müller, "Energy conservation in acetogenic bacteria," *Applied and Enviornmental Microbiology*, vol. 69, no. 11, pp. 6345–6353, November 2003.

[13] M. Köpke, C. Mihalcea, F. Liew, J. H. Tizard, M. S. Ali, J. J. Conolly, B. Al-Sinawi, and S. D. Simpson, "2,3-butanediol production by acetogenic bacteria, an alternative route to chemical synthesis, using industrial waste gas," *Applied and Enviornmental Microbiology*, vol. 77, no. 15, pp. 5467–5475, August 2011.

[14] J. Shin, Y. Song, and Y. Jeongm Byung-Kwan Cho, "Analysis of the core genome and pan-genome of autotrophic acetogenic bacteria," *Frontiers in Microbiology*, vol. 7, September 2016.

[15] K. Schuchmann and V. Müller, "Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria," *Nature Reviews Microbiology*, vol. 12, pp. 809–821, November 2014.

[16] J. Bertsch and V. Müller, "Bioenergetic constraints for conversion of syngas to biofuels in acetogenic bacteria," *Biotechnology for Biofuels*, vol. 8, no. 210, December 2015.

[17] S. De Tissera, M. Köpke, S. D. Simpson, C. Humphreys, N. P. Minton, and P. Dürre, "Syngas biorefinery and syngas utilization," *Advances in Biochemical Engineering/Technology*, vol. 10, June 2017.

[18] A. M. Henstra, J. Sipma, A. Rinzema, and A. J. M. Stams, "Microbiology of synthesis gas fermentation for biofuel production," *Current Opinion in Biotechnology*, vol. 18, no. 3, pp. 200–206, June 2007.

[19] H. L. Drake, A. S. Gößner, and S. L. Daniel, "Old acetogens, new light," *Annals of the New York Academy of Sciences*, vol. 1125, pp. 100–128, March 2008.

[20] K. Schuchmann and V. Müller, "Energetics and application of heterotrophy in acetogenic bacteria," *Applied and Enviornmental Microbiology*, vol. 82, no. 14, pp. 4056–4069, July 2016.

[21] J. Jeong, J. Bertsch, V. Hess, S. Choi, I.-G. Choi, I. S. Chang, and V. Müller, "Energy conservation model based on genomic and experimental analyses of a carbon monoxide-utilizing, butyrate-forming acetogen, *Eubacterium limosum* KIST612," *Applied and Enviornmental Microbiology*, vol. 81, no. 14, pp. 4782–4790, June 2015.

[22] Y. Song and B.-K. Cho, "Draft genome sequence of chemolithoautotrophic acetogenic butanol-producing *Eubacterium limosum* ATCC 8486," *Genome Annoucements*, vol. 3, no. 1, February 2015.

[23] S. W. Ragsdale, "Enzymology of the Wood-Ljungdahl pathway of acetogenesis," *Annals of the New York Academy of Sciences*, vol. 1125, pp. 129–136, March 2008.

[24] ——, "The eastern and western branches of the Wood-Ljungdahl pathway:how the east and west were won," *BioFactors*, vol. 6, pp. 3–11, 1997.

[25] S. W. Ragsdale and E. Pierce, "Acetogenesis and the Wood-Ljungdahl pathway of Co2 fixation," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1784, no. 12, pp. 1873–1898, August 2008.

[26] W. J. Kelly, G. Henderson, D. M. Pacheco, D. Li, K. Reilly, G. E. Naylor, P. H. Janssen, G. T. Attwood, E. Altermann, and S. C. Leahy, "The complete genome sequence of *Eubacterium limosum* SA11, a metabolically versatile rumen acetogen," *Standards in Genomic Sciences*, vol. 11, no. 26, March 2016.

[27] Y. Zhang, D.-H. Yang, Y.-T. Zhang, X.-M. Chen, L.-L. Li, and S.-Q. Cai, "Biotransformation on the flavonolignan constituents of silybi fructus by an intestinal bacterial strain *Eubacterium limosum* ZL-IL," *Fitoterapia*, vol. 92, pp. 61–71, Janurary 2014.

[28] B. R. S. Genthner, C. L. Davis, and M. P. Bryant, "Features of rumen and sewage sludge strains of *Eubacterium limosum*, a methanol and H2-CO2 utilising species," *Applied and Enviornmental Microbiology*, vol. 42, no. 1, pp. 12–19, July 1981.

[29] B. Schiel-Bengelsdorf, "Pathway engineering and synthetic biology using acetogens," *FEBS letters*, vol. 586, no. 15, pp. 2191–2198, July 2012.

[30] M. Demler and D. Weuster-Botz, "Reaction engineering analysis of hydrogenotrophic production of acetic acid by *Acetobacterium woodii*," *Biotechnology and Bioengineering*, vol. 108, no. 2, pp. 470–474, February 2010.

[31] J.-H. Jeoung, S. Goetzl, S. E. Hennig, J. Fesseler, C. Wörmann, J. Dendra, and H. Dobbek, "The extended reductive acetyl-CoA pathway: ATPases in metal cluster maturation and reductive activation," *Biological Chemistry*, vol. 395, no. 5, May 2014.

[32] S. E. Hennig, J.-H. Jeoung, S. Goetzl, and H. Dobbek, "Redox-dependent complex formation by an ATP-dependent activator of the corrinoid/iron-sulfur protein," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 14, pp. 5235–5240, April 2012.

[33] H. Roh, H.-J. Ko, D. Kim, D. G. Choi, S. Park, S. Kim, I. S. Chang, and I.-G. Choi, "Complete genome sequence of a carbon monoxide-utilizing acetogen, *Eubacterium limosum* KIST612," *Journal of Bacteriology*, vol. 193, no. 1, pp. 307–308, Janurary 2011.

[34] (2017, September). [Online]. Available: https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/

[35] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson, "Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways," *PLOS Computational Biology*, vol. 11, no. 8, 2015.

[36] S. Gencic, E. C. Duin, and D. A. Grahame, "Tight coupling of partial reactions in the acetyl-CoA decarbonylase/synthase (acds) multienzyme complex from *Methanosarcina thermophila*: acetyl c-c bond fragmentation at the a cluster promoted by protein conformational changes," *The Journal of Biological Chemistry*, vol. 285, no. 20, pp. 15 450–15 463, May 2010.

[37] J. Maupin-Furlow and J. G. Ferry, "Characterization of the cdhd and cdhe genes encoding subunits of the corrinoid/iron-sulfur enzyme of the co dehydrogenase complex from *Methanosarcina thermophila*," *American Society for Microbiology*, vol. 178, no. 2, pp. 340–346, Janurary 1996.

[38] The, UniProt, and Consortium, "Uniprot: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 45, no. D1, pp. 158–169, Janurary 2017. [Online]. Available: http://www.uniprot.org

[39] P. Loubiére, E. Gros, V. Paquet, and N. D. Lindley, "Kinetics and physiological implications of the growth behaviour of *Eubacterium limosum* on glucose/methanol mixtures," *Journal of General Microbiology*, vol. 138, pp. 979–985, February 1992.

[40] P. v. d. Meijden, C. v. d. Lest, C. v. d. Drift, and G. D. Vogels, "Reductive activation of methanol: 5-hydroxybenzimidazolylcobamide methyltransferase of *Methanosarcina barkeri*," *Biochemical and Biophysical Reseach Communications*, vol. 118, no. 3, pp. 760–766, February 1984.

[41] M. Krüer, M. Haumann, W. Meyer-Klaucke, R. K. Thauer, and H. Dau, "The role of zinc in the methylation of the coenzyme M thiol group in methanol:coenzyme M methyltransferase from *Methanoasarcina barkeri*," *European Journal of Biochemistry*, vol. 269, no. 8, pp. 2117–2123, April 2002.

[42] (2017, September). [Online]. Available: https://inkscape.org/en/

[43] E. Marcellin, J. B. Behrendorff, S. Nagaraju, S. DeTissera, S. Segovia, R. W. Palfreyman, J. Daniell, C. Licona-Cassani, L.-e. Quek, R. Speight, M. P. Hodson, S. D. Simpson, W. P. Mitchell, M. Köpke, and L. K. Nielsen, "Low carbon fuels and commodity chemicals from

waste gases – systematic approach to understand energy metabolism in a model acetogen," *Royal Society of Chemistry*, vol. 18, pp. 3020–3028, Janurary 2016.

[44] H. Nagarajan, M. Sahin, J. Nogales, H. Latif, D. R. Lovely, A. Ebrahim, and K. Zengler, "Characterizing acetogenic metabolism using a genome scale metabolic reconstruction of *Clostridium ljungdahlii*," *Microbial Cell Factories*, vol. 12, no. 118, November 2013.

[45] V. Hess, A. Poehlein, M. C. Weghoff, R. Daniel, and V. Müller, "A genome-guided analysis of energy conservation in the thermophilic, cytochrome-free acetogenic bacterium *Thermoanaerobacter kivui*," *BMC Genomics*, vol. 15, no. 1139, December 2014.

[46] R. Spencer-Smith, E. M. Varkey, M. D. Fielder, and L. A. S. Synder, "Sequence features contributing to chromosomal rearrangements in *Neisseria gonorrhoeae*," *PLOS One*, vol. 7, no. 9, September 2012.

[47] A. C. E. Darling, B. Mau, and N. T. Perna, "progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement," *PLOS One*, vol. 5, no. 6, June 2010.

[48] V. Periwal and V. Scaria, "Insights into structural variations and genome rearrangements in prokaryotic genomes," *Bioinformatics Review*, vol. 31, no. 1, pp. 1–9, September 2015.

[49] B. Kumwenda, D. Litthauer, and O. Reva, "Analysis of genomic rearrangements, horizontal gene transfer and role of plasmids in the evolution of industrial important thermus species," *BMC Genomics*, vol. 15, no. 1, p. 1, September 2014.

[50] O. Skovgaard, M. Bak, A. Løbner-Olesen, and N. Tommerup, "Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing," *Genome Research*, vol. 21, no. 8, pp. 1388–1393, August 2011.

[51] G. Achaz, E. Coissac, P. Netter, and E. P. Rocha, "Associations between inverted repeats and the structural evolution of bacterial genomes," *Genetics*, vol. 164, no. 4, pp. 1279–1289, August 2003.

[52] D. Hughes, "Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes," *Genome Biology*, vol. 1, no. 6, December 2000.

[53] W. Qian, Y. Jia, S.-X. Ren, Y.-Q. He, J.-X. Feng, L.-F. Lu, Q. Sun, G. Ying, D.-J. Tang, H. Tang, W. Wu, P. Hao, L. Wang, B.-L. Jiang, S. Zeng, W. Y. Gu, G. Lu, L. Rong, Y. Tian, Z. Yao, G. Fu, B. Chen, R. Fang, B. Qiang, Z. Chen, G.-P. Zhao, J.-L. Tang, and C. He, "Comparative and functional genomic analyses of the pathogenicity of phytopathogen *Xanthomonas campestris* pv. campestris," *Genome Research*, vol. 15, pp. 757–767, May 2005.

[54] M. G. Poolman, H. E. Assmus, and D. A. Fell, "Applications of metabolic modelling to plant metabolism," *Journal of Experimental Botany*, vol. 55, no. 400, pp. 1177–1186, May 2004.

[55] R. Steuer, T. Gross, J. Selbig, and B. Blasius, "Structural kinetic modeling of metabolic networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 32, pp. 11 868–11 873, August 2006.

[56] C. T. Trinh, A. Wlaschin, and F. Srienc, "Elementary mode analysis: A useful metabolic pathway analysis tool for characterizing cellular metabolism," *Applied Microbiology and Biotechnology*, vol. 81, no. 5, pp. 813–826, Janurary 2009.

[57] N. R. Boyle and J. A. Morgan, "Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*," *BMC Systems Biology*, vol. 3, no. 1, p. 4, August 2009.

[58] C. H. Schilling, J. S. Edwards, D. Letscher, and B. Ø. Palsson, "Combining pathway analysis with flux balance analysis fro the comprehensive study of metabolic systems," *Biotechnology and Bioengineering*, vol. 71, no. 4, pp. 286–306, 2001.

[59] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis," *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, March 2010.

[60] K. Raman, P. Rajagopalan, and N. Chandra, "Flux balance analysis of mycolic acid pathway: Targets for anti-tubercular drugs," *PLOS Computational Biology*, vol. 1, no. 5, October 2005.

[61] Z. Chen, H. Liu, J. Zhang, and D. Liu, "Elementary mode analysis for the rational design of efficient succinate conversion from glycerol by *Escherichia coli*," *Journal of Biomedicine and Biotechnology*, vol. 3, no. 1, p. 14, July 2010.

[62] S. Schuster, T. Dandekar, and D. A. Fell, "Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering," *Trends Biotechnology*, vol. 17, no. 2, pp. 53–60, 1999.

[63] R. Heinrich and S. Schuster, "The modelling of metabolic systems. structure, control and optimality," *Biosystems*, vol. 47, no. 1-2, pp. 61–77, 1998.