2011

Computational Discovery of Phenotype Related Biochemical Processes for Engineering

Andrea M. Rocha
University of South Florida, amrocha@mail.usf.edu

Computational Discovery of Phenotype Related Biochemical Processes for Engineering Bacterial Biohydrogen

by

Andrea M. Rocha

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Civil & Environmental Engineering
College of Engineering
University of South Florida

Major Professor: James R. Mihelcic, Ph.D.
Nagiza Samatova, Ph.D.
Daniel Yeh, Ph.D.
Shekhar Bhansali, Ph.D.
Kathleen Scott, Ph.D.

Date of Approval:
April 28, 2011

Keywords: Bioenergy, Dark Fermentation, Metabolic Processes, Systems Biology,
Clostridium acetobutylicum

DEDICATION

I would like to dedicate this dissertation to my loving family. Throughout my graduate studies, they supported me, encouraged me to follow my dream, and provided me with unending support.

ACKNOWLEDGEMENTS

There are many people I would like to thank for their contribution to my dissertation. First and foremost, I would like to thank Dr. James Mihelcic and Dr. Nagiza Samatova for their guidance, support, and encouragement in completing this work. I would also like to thank my committee members, Dr. Daniel Yeh, Dr. Kathleen Scott, and Dr. Shekhar Bhansali for their time and assistance with my project. I also want to thank my research group and collaborators from North Carolina State University for their help and assistance with this work. Specifically, I would thank Matt Schmidt, William Hendrix, Kanchana Padmanabhan, Kevin Wilson, Wenbin Chen, and Katie Shpanskaya.

In addition, I would like to say thank you to my family and friends who have supported me during my doctoral studies. I would like to give a special ‘thank you’ to my parents and sister for their unending love and support throughout the years. I would like to thank Mr. Bernard Batson for his support and assistance throughout my time at USF. Last, I would like to thank Debbie McCoy and her staff for providing me the opportunity to participate in the RAMS programs at Oak Ridge National Lab.

Support for this work has been funded by the NSF Bridge to the Doctorate Fellowship, USF Delores Auzenne Diverse Student Success Fellowship, Alfred P. Sloan Fellowship, GEM Fellowship, State of Florida 21st Century World Class Scholar Program, and Oak Ridge National Laboratory (ORNL) Research Alliance in Math & Science (RAMS) internship program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Application of bioengineering technologies for enhanced biological hydrogen production is a promising approach that may play a vital role in sustainable energy. Due to the ability of several naturally occurring microorganisms to generate hydrogen through varying metabolic processes, biological hydrogen has become an attractive alternative energy and fuel source.

One area of particular interest is the production of biological hydrogen in organically-rich engineered systems, such as those associated with waste treatment. Despite the potential for high energy yields, hydrogen yields generated by bacteria in waste systems are often limited due to a focus on microbial utilization of organic material towards cellular growth rather than production of biogas. To address this concern and to improve upon current technological applications, metabolic engineering approaches may be applied to known hydrogen producing organisms. However, to successfully modify metabolic pathways, full understanding of metabolic networks involved in expression of microbial traits in hydrogen producing organisms is necessary.

Because microbial communities associated with hydrogen production are capable of exhibiting a number of phenotypes, attempts to apply metabolic engineering concepts have been restricted due to limited information regarding complex metabolic processes and regulatory networks involved in expression of microbial traits associated with biohydrogen production.

To bridge this gap, this dissertation focuses on identification of phenotype-related biochemical processes within sets of phenotype-expressing organisms. Specifically, through co-development and application of evolutionary genome-scale phenotype-centric comparative network analysis tools, metabolic and cellular components related to three phenotypes (i.e., dark fermentative, hydrogen production and acid tolerance) were identified. The computational tools employed for the systematic elucidation of key phenotype-related genes and subsystems consisted of two complementary methods. The first method, the Network Instance-Based Biased Subgraph Search (NIBBS) algorithm, identified phenotype-related metabolic genes and subsystems through comparative analysis of multiple genome-scale metabolic networks. The second method was the multiple alignments of metabolic pathways for identification of conserved metabolic sub-systems in small sets of phenotype-expressing microorganisms. For both methodologies, key metabolic genes and sub-systems that are likely to be related to hydrogen production and acid-tolerance were identified and hypotheses regarding their role in phenotype expression were generated. In addition, analysis of hydrogen producing enzymes generated by NIBBS revealed the potential interplay, or cross-talk, between metabolic pathways.

To identify phenotype-related subnetworks, three complementary approaches were applied to individual, and sets of phenotype-expressing microorganisms. In the first method, the Dense ENriched Subgraph Enumeration (DENSE) algorithm, partial “prior knowledge” about the proteins involved in phenotype-related processes are utilized to identify dense, enriched sets of known phenotype-related proteins in *Clostridium acetobutylicum.* The second approach utilized a bi-clustering algorithm to identify phenotype-related functional association modules associated with metabolic controls of phenotype-related pathways. Last, through comparison of hundreds of genome-scale networks of functionally associated proteins, the *α, β*-motifs approach, was applied to identify phenotype-related subsystems.

Application of methodologies for identification of subnetworks resulted in detection of regulatory proteins, transporters, and signaling proteins predicted to be related to phenotype-expression. Through analysis of protein interactions, clues to the functional roles and associations of previously uncharacterized proteins were identified (DENSE) and hypotheses regarding potentially important acid-tolerant mechanisms were generated (*α, β*-motifs). Similar to the NIBBS algorithm, analysis of functional modules predicted by the bi-clustering algorithm suggest cross-talk is occurring between pathways associated with hydrogen production.

The ability of these phenotype-centric comparative network analysis tools to identify both known and potentially new biochemical process is important for providing further understanding and insights into metabolic networks and system controls involved in the expression of microbial traits. In particular, identification of phenotype-related metabolic components through a systems approach provides the underlying foundation for the development of improved bioengineering technologies and experimental design for enhanced biological hydrogen production.

CHAPTER 1: INTRODUCTION

1.1 Problem Statement and Motivation

Over the past few decades, there has been an increase in the number of comparative genomic studies aimed at identifying potential relationships between metabolic networks and microbial environments [28]. Specifically, there is a large interest in understanding and accurately predicting phenotype-specific microbial traits for the advancement of genetic engineering and industrial technology [27]. Information provided by these studies can be used to improve gene drug delivery, microbial processes involved in environmental restoration (e.g. bioremediation), and regulation of biological processes, such as production of biogas for bioenergy [29-31]. Specific to this research and of particular interest to the discipline of environmental engineering, is the potential to redirect metabolic pathways and networks to express specific microbial traits necessary to enhance biological hydrogen production in engineered systems (e.g. wastewater treatment) [27, 32, 33]

Biological hydrogen production is becoming a more popular alternative as a renewable energy source for biofuels and bioenergy due to the potential for high energy yields [11]. This is due to the ability of a large number of naturally occurring microorganisms to generate biohydrogen using different biological processes, thus. providing for the development of several different biohydrogen technologies [33]. One promising approach is the application of resources found in wastewater and other waste materials for biohydrogen production [2, 11]. Production of biohydrogen in association with a wastewater treatment facility can be seen as a viable option that allows for simultaneous generation of two renewable resources—water and bioenergy [34]. Depending on the organic substrate and the microbial community present, it is possible to remove a majority of the organic material from wastewater, thus producing clean water for reuse or discharge to aquatic ecological systems. The biohydrogen generated can then be used directly in fuel cells [35] or transported to other facilities where it can be used as a biofuel resource.

To generate hydrogen gas, hydrogen producers utilize one of three main metabolic processes. The three processes are light fermentation, dark fermentation of organic matter, and decomposition of water by photosynthesizing microorganisms (e.g. bio-photolysis) [22, 36, 37]. In light fermentation, organisms utilize simple organic compounds as a carbon source (e.g., glucose and sucrose) and a light source (e.g. sunlight) to generate hydrogen [6, 38]. Dark fermentative bacteria differ from the other two hydrogen-producing methods in that hydrogen evolving reactions are carried out without light energy by a number of heterotrophic bacteria [4, 22]. In this process, hydrogen is produced from dark fermentation reactions when organic substrates are utilized by heterotrophic bacteria as both the carbon and energy source for heterotrophic growth [4, 22]. In the last method, bio-photolysis, photosynthetic organisms can breakdown water molecules into hydrogen gas and oxygen [7, 22, 38]. Production of hydrogen through this process can be carried out either directly by exposure to solar radiation or indirectly under dark (fermenting) conditions [7]. A detailed description of each metabolic process is presented in Chapter 2.

Of the hydrogen producing organisms associated with wastewater and waste materials, a majority of these species appear to utilize dark fermentation metabolic processes to produce hydrogen. During this process waste materials (e.g., sludge) are collected and placed in an anaerobic chamber where they can be degraded by fermentation processes. This study focuses on dark fermentative hydrogen producing processes. A general overview of dark fermentation hydrogen production in wastewater treatment is provided in Figure 1.1.

1.2 Systems Biology to Enhance Understanding of Biohydrogen Production

Although the full maximum potential for biohydrogen produced in association with wastewater treatment has not been reached, many systems biology and molecular biology studies are currently being conducted to identify metabolic and cellular networks necessary to increase biohydrogen production by hydrogen–producing microorganisms [6, 39, 40]. These studies include metabolic flux studies [41], molecular characterization of key enzymes involved in hydrogen production [42, 43], identification and reconstruction of metabolic networks within model organisms for hydrogen production [44]. Examples of metabolic pathways of particular interest for these studies include central metabolism and amino acid biosynthesis.

While central metabolic pathways have been well-characterized for a number of organisms (e.g., *Escherichia coli*), the production of specific metabolites within central metabolism, and flow of carbon within these pathways with respect towards hydrogen production remain unclear [41]. To address this problem, Noguez et al. [41] recently evaluated the pathways involved in: (1) production of glutamate and alpha-ketoglutarate, (2) incorporation of nutrients into glycoylsis, TCA, and PPP, as well as, amino acid

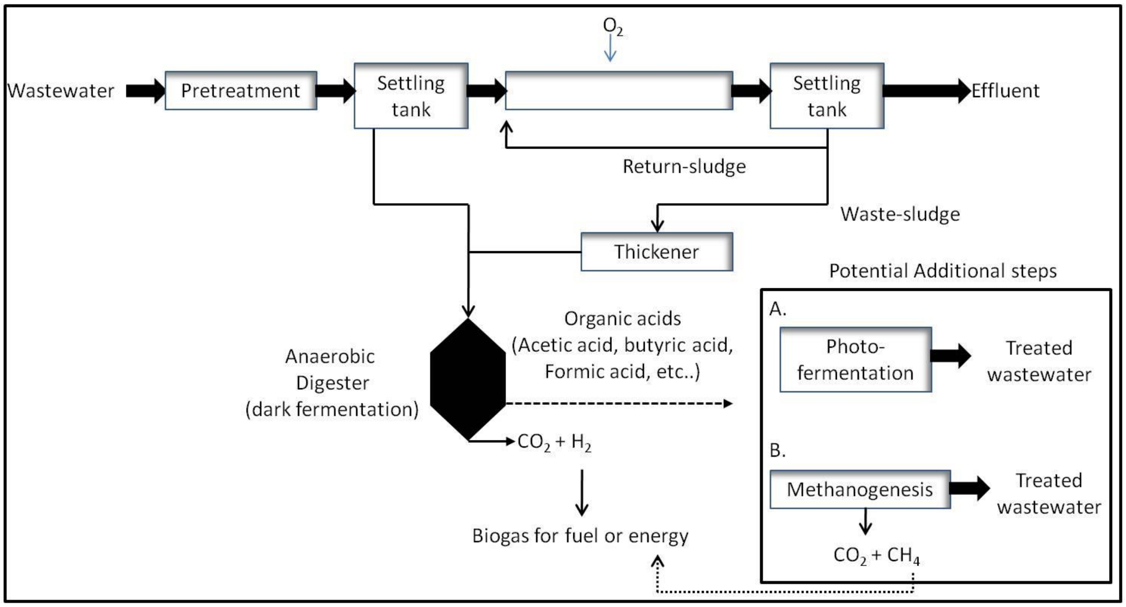


Figure 1.1 Schematic diagram of biological hydrogen production from conventional wastewater treatment.

biosynthesis, and (3) the flux of nutrients through these pathways during metabolic shifts. Pathways evaluated in that study were for the hydrogen producer, *Clostridium acetobutylicum.* From these studies, the authors were able to demonstrate utilization of bifurcated TCA within *C. acetobutylicum* [41]. In addition, through isotope studies, they were able to identify which pathways were utilized and demonstrated the ability of *C. acetobutylicum* to direct most of its glycolytic flux towards production of acetyl-CoA and towards amino acid biosynthesis [41]. Utilization of isotopes provides valuable insights into the flow of carbon within organisms and the activity of biochemical enzymes. In addition, the application of flux studies allows us to understand the role of regulatory or signaling proteins within specific pathways. For example, these studies were able to determine the flow of carbon in central metabolism when *C acetobutylicum* was producing solvents instead of acetate. In addition, they were able to show changes in production of amino acids as a result of metabolic shift. Understanding the factors regulating or involved in metabolic shifts pertaining to hydrogen production is of great interest for sustaining hydrogen production in wastewater systems.

In addition to the *C. acetobutylicum* studies, other studies have focused on central metabolism in relation to hydrogen production. One example is a study by McKinlay and Harwood [45], where central metabolic fluxes were evaluated to determine flow of electrons and utilization of carbohydrates in *Rhodopsuedomonas palustris*, a light fermentative species. Within that study, transcriptional controls for electron shifts from the Calvin cycle to hydrogen production were identified. Understanding of transcriptional regulation is particularly important for redirecting the flow of electron and carbon in engineered organisms.

In addition to metabolic pathways, there have been a number of studies focused on understanding the structures and role of key enzymes associated with hydrogen production [42, 43, 46, 47]. Examples of enzymes include nitrogenase and hydrogenase enzyme complexes. Hydrogen producing organisms capable of fixing nitrogen contain enzyme complexes termed nitrogenase. Within nitrogenase complexes, nitrogen gas is converted to ammonia, inadvertently resulting in the production of hydrogen gas as a byproduct [6, 39, 46]. Hydrogenase complexes are responsible for either hydrogen production or consumption [47].

Analysis of hydrogenase enzymes have identified three different types, each associated with a number of accessory proteins necessary for activation [47, 48]. These include the [NiFe]-hydrogenase, [FeFe]-hydrogenase, and non-metal containing hydrogenase enzyme [47]. Further details on the roles of these enzyme complexes are provided in Chapter 2. Understanding complexes like nitrogenase and hydrogenase enzymes, is important in deciphering regulatory mechanisms and activity of these key enzymes. For example, in studies evaluating accessory proteins present in [NiFe]-hydrogenase complexes, HypCDEF proteins are described as regulators for maturation of uptake hydrogenase through participation in development of the active center [47, 49]. If one of the Hyp proteins is missing, the entire complex is inactivated.

While systems and molecular biology studies, such as the ones described above, provide further understanding of metabolic pathways related to hydrogen production, a complete understanding of metabolic processes by hydrogen producers is still lacking. In most studies, experimental analysis is often limited to one individual organism or one specific pathway. In the case of species level evaluation, information generated is specific to the target organism rather than understanding of the phenotype as a whole. In the case of hydrogen production, understanding of phenotype expression will provide further understanding for the enhancement of hydrogen yields in mixed communities.

Similarly, from the studies described, we see the importance of a systems approach towards understanding metabolic and cellular networks involved in hydrogen production. Given that pathways are inter-related and can cross-talk between one another, analysis of entire networks, rather than individual pathways is necessary if one is to improve upon hydrogen yields.

1.3 Metabolic Engineering for Hydrogen Production

Application of metabolic engineering approaches to enhance biological hydrogen production in engineered systems (e.g., wastewater treatment) is a promising approach that is currently being explored by many researchers [27, 32, 50]. Due to the importance of total hydrogen yields from starting material (e.g., organic matter), a number of metabolic engineering techniques have been employed. These studies focus on improving production rates in single organisms through modification of metabolic pathways or regulatory systems of individual model microorganisms [6, 27, 39]. Such regulation includes up and down regulation of transcriptional genes, as well as insertion or deletions of genes and mutations to direct metabolism toward hydrogen production and away from hydrogen consumption [6, 27, 40]. Examples of current metabolic engineering approaches for addressing hydrogen production are presented in Table 1.1.

To successfully modify metabolic or regulatory pathways, it is important to have a full understanding of all metabolic networks involved. This is especially true for hydrogen producers exposed to a number of environmental stresses (e.g., pH and anaerobic conditions). In this case, the organisms must be able to exhibit more than one phenotype to generate hydrogen and increase cellular biomass. In order to enhance the expression of one phenotype, such as hydrogen production, consideration of the role other phenotypes may have on hydrogen producing metabolic networks is necessary. As such, metabolic engineers require a deciphering of which genes, pathways, and metabolic networks are related to the expression of a given phenotype.

1.4 Objective and Proposed Research Approach

The overall objective of this research is to use a systems-biology approach to discover genes, metabolic pathways, and subnetworks related to specific phenotypes essential for optimal microbial production of hydrogen using dark fermentation processes. To accomplish the objective, a joint computational and engineering framework was used. The engineering framework can then be used to evaluate current approaches to biohydrogen production using dark fermentation methods encountered at

Table 1.1 Examples of current metabolic engineering studies and techniques for enhancement of biological hydrogen production.

Methodology	Study Description	Reference
Developed NAD(P)H:H_2 synthetic pathways in *Escherichia coli* BL21	Evaluated the thermodynamic limitations of the pathway in relation to hydrogen production	[51]
Inactivated uptake hydrogenases in *E. coli* K12	Suppression of enzymes known to remove hydrogen and reduce hydrogen yields	[27]
Increased expression of formate hydrogen lyase (upregulation of genes)	Upregulated expression of an enzyme involved in hydrogen production	[27]
Developed mutant strains of *Rhodopseudomonas palustris*	Enhanced hydrogen production by making the organism dependent on the process; Hydrogen production was linked to ATP synthesis and nitrogen fixation	[6]
Designed a synthetic operon for hydrogen production by converging *hyd* genes into plasmids	Developed an operon for expression of genes, such as hydrogenase, involved in hydrogen formation	[52]

wastewater treatment plants to determine ideal operating parameters (e.g., temperature, pH, redox conditions, and oxygen requirement) necessary for optimized microbial hydrogen production (Chapter 2).

Information about microbial growth parameters and understanding of system requirements are necessary to identify phenotypes ideal for optimizing hydrogen production by microorganisms. The framework entails development of methodology for statistical comparison and identification of phenotype-related genes (Chapter 3), metabolic pathways and subnetworks (Chapter 4), and the discovery of known and novel regulatory and uncharacterized proteins (Chapter 5) in groups of microorganisms that express the same phenotype. A detailed description of the computational methodologies used in this research is provided in Chapter 6.

An overview of the research aims is provided in Figure 1.2. A list of technologies co-developed and their capabilities is provided in Table 1.2. Information obtained from these methodologies not only complemented experimental approaches, but provides insights into metabolic networks and system controls involved in the expression of microbial traits. Such information is necessary for advancing engineering approaches that result in more efficient biohydrogen production. The following sections provide a brief overview of each Chapter of this dissertation.

1.5 Development of Statistical Analysis Techniques for Identification of Phenotype-Related Genes (Chapter 3)

The first component of this research is aimed at identifying key genes related to the dark fermentative hydrogen producing, anaerobic versus aerobic respiration, and acid-tolerance phenotypes. The goal of this Chapter is to develop a comparative graph-theoretical algorithm of a target phenotype. To accomplish this goal, two complementary approaches are pursued: (1) Genotype phylogenetic profiling (GPP); and (2) Phylogenetic phenotype profiling (PPP).

The proposed genotype phylogenetic profiling (GPP) approach relies on statistical enrichment analysis of orthologs present in microorganisms expressing the target phenotype and microorganisms that do not express the target phenotype. Phylogenetic profiling, such as GPP, is a technique based on the inclusion or exclusion of a gene in an organism's genome. Using this technique, one can correlate sets of orthologous genes present in phylogenetic profiles with phenotype-expressing organisms. Based on ortholog distribution, a Student's *t*-test was used to determine which orthologs are related

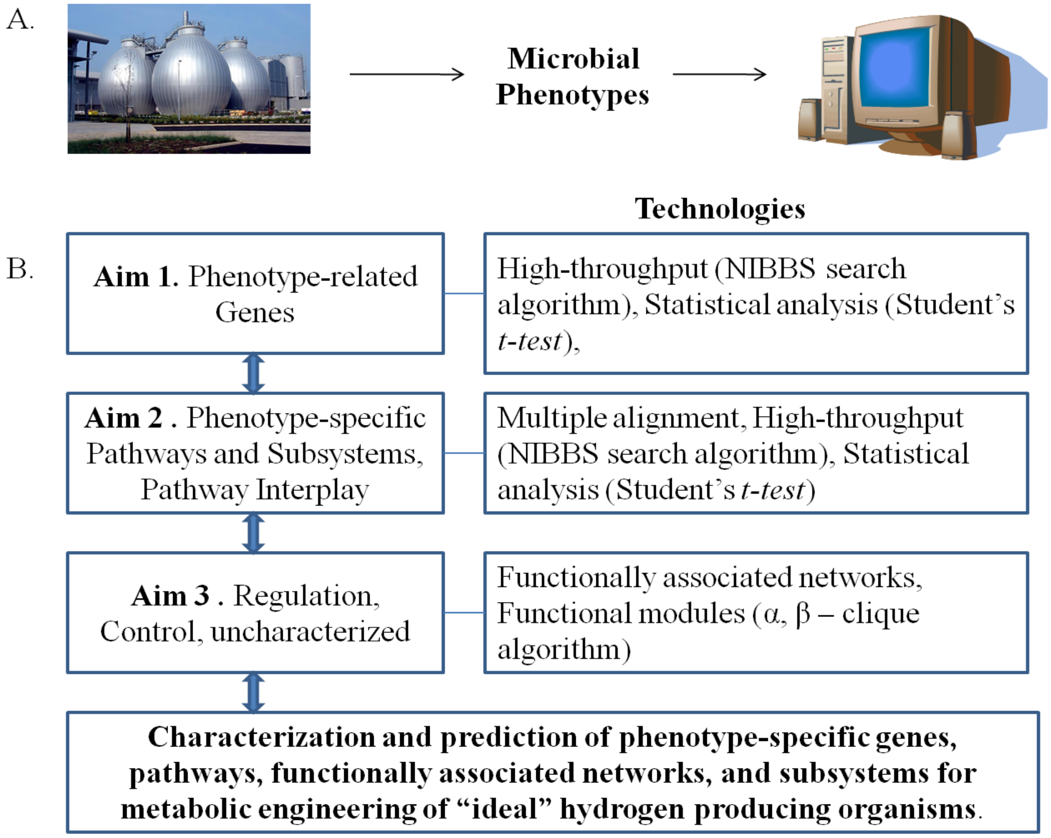

A.
Microbial Phenotypes
B.
Technologies
Aim 1. Phenotype-related Genes
High-throughput (NIBBS search algorithm), Statistical analysis (Student's t-test),
Aim 2 . Phenotype-specific Pathways and Subsystems, Pathway Interplay
Multiple alignment, High-throughput (NIBBS search algorithm), Statistical analysis (Student's t-test)
Aim 3 . Regulation, Control, uncharacterized
Functionally associated networks, Functional modules (α, β – clique algorithm)
Characterization and prediction of phenotype-specific genes, pathways, functionally associated networks, and subsystems for metabolic engineering of "ideal" hydrogen producing organisms.

Figure 1.2 The overall proposed study. (A) Microbial phenotypes will be determined based on characteristics of wastewater treatment facilities, such as anaerobic digesters (left). Microbial phenotypes will be used in computational studies. (B) Specific aims and technologies used to address the aims

to the target phenotype. This method provides the basis for a quick and robust analysis to identify potentially important phenotype-related genes. Functional enrichment analyses of these genes were performed to create a functional profile for a given phenotype.

The phylogenetic phenotype profiling (PPP) approach relies on co-development of an algorithm for the discovery of sets of genes that are statistically significant in microorganisms expressing the phenotype compared to microorganisms not expressing the phenotype. To generate sets of genes or seed sets, comprehensive generic metabolic

Table 1.2 List of co-developed technologies and their respective capabilities.

Technology	Capabilities	Description
Student's T-test	Identifies phenotype-related genes	Utilizes list of genes; Considers copy number; Robust statistical approach
Network Instance-Based Biased Subgraph (NIBBS) Algorithm	Identifies phenotype-related genes and sub-systems	Utilizes network context with respect to genes; Specific to metabolic enzymes; Utilizes binary notation
Multiple Alignment Algorithm without Abstraction	Identifies conserved metabolic networks across multiple organisms	Aligns metabolic networks across multiple phenotype expressing organisms
Dense Enriched Subgraph Enumeration (DENSE) Algorithm	Identifies genes that are functionally associated to a set of known phenotype-related proteins; Provides insight into functional roles of uncharacterized proteins	Incorporates researchers' prior knowledge; Enumerates dense and enriched subgraphs in genome-scale networks of functionally associated proteins
α,β-Clique Algorithm	Identifies conserved functional modules with at least α networks of phenotype expressing organisms and β networks of organisms that do not exhibit the phenotype	Enumerates α,β-cliques related to phenotype expressing organisms and analyzes them using sets of phenotype-related genes and modules
Bi-Clustering Motif Algorithm	Identifies network motifs conserved across sets of phenotype-expressing organisms; Potential to identify cross-talk between pathways	Utilizes a bi-clustering algorithm; Identifies clusters of orthologous groups and develops a bipartite graph

network maps from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database were used to help identify sets of phenotype related genes. Phylogenetic profiles of phenotype-expressing and non-phenotype expressing organisms were created using orthologous genes. Using a comprehensive generic metabolic network from the KEGG database and search algorithm, sets of phenotype related genes were identified. In this process, genes that are significantly more conserved (i.e., show bias) in microorganisms expressing the target phenotype were identified using binomial probability distribution. Conserved genes identified were used to create a seed set of genes, which is expanded upon in Chapter 4 to discover phenotype-specific metabolic pathways. In addition, phenotype-related genes were used to evaluate functional profiles of groups of related microorganisms and to map metabolic networks.

1.5.1 Impact

The research explained in Chapter 3 impacts the current base of scientific knowledge as follows.

- Development and application of the techniques developed in this Chapter allow for the identification of known and novel phenotype-related genes across sets of phenotype-expressing organisms.
- Application of the Network Instance-Based Biased Subgraph Search (NIBBS-Search) algorithm allows for the identification of non-specific, phenotype-related enzymes. These enzymes are considered important for expression of pathways, such as central metabolism, that interact with hydrogen producing pathways.

- Information provided by the Student’s T-test and NIBBS algorithm provides the foundational building blocks necessary for designing molecular studies and engineered organisms for the development of new technologies in biohydrogen production.

1.6 Identification of Phenotype-Specific Metabolic Pathways (Chapter 4)

In this Chapter, research focused on identifying phenotype-specific metabolic pathways for each of the target phenotypes. To classify which metabolic pathways are specific to the phenotypic trait, a methodology was developed to statistically compare the average pathway enrichment across groups of phenotype-expressing microorganisms. Classification of metabolic pathways using the statistical approach was based on the assumption that metabolic pathways enriched with phenotype-related genes are also related to the expressed phenotypic trait.

Identification of phenotype specific metabolic pathways was carried out using three complementary approaches: (1) Pathway enrichment, (2) Network Instance-Based Biased Subgraph Search (NIBBS) algorithm, and (3) Multiple alignments of metabolic pathways.

The pathway enrichment approach uses only the phenotype-related orthologs discussed in Section 1.4.2 of Chapter 1 rather than all the orthologs listed in the National Center for Biotechnology Information (NCBI) cluster of orthologous groups (COGs) of proteins COG database. In phenotype expressing microorganisms, several different pathways containing different genes and enzymes necessary for the expression of the phenotype may be present. As such, it is important to consider enrichment of enzymes and their corresponding genes when determining phenotype-specific pathways. Pathway enrichment for each microorganism is defined as the total number of genes enriched by the total number of enzymes in the pathway. The pathway enrichment for each organism was used to determine the average enrichment of the phenotype. A Student's T-test was used to determine if a metabolic pathway is specific to the presence of phenotypic trait or if it is important to its absence.

The second approach utilizes a heuristic algorithm, NIBBS, for discovery of phenotype-specific metabolic pathways similar to the one described in Section 1.4.2 of Chapter 1. In this method, the NIBBS algorithm was used to identify a set of reactions that both forms subnetworks in a comprehensive generic metabolic network and is statistically biased towards phenotype expressing microorganisms. Using binomial distribution, the bias of a reaction set was determined and the bias value was used to determine the reaction set's phenotypic trait and non-phenotypic trait coverage. As a general rule, the larger the bias value, the more a seed set is related to a phenotype. Since some pathways may be related to the expression of the phenotype, but are not specific to expression of the phenotype, it is necessary to determine if phenotype-related and non-phenotype related pathways overlap. This is accomplished by further expanding upon identified seed sets to include reactions that are related, but not necessarily specific to expression of the phenotypic trait. Information generated was used to provide insight into the role of metabolic pathways in expression of specific phenotypic traits.

The last approach is based on co-development of a multiple alignment algorithm for identification of conserved networks, particularly metabolic pathways in phenotype-expressing microorganisms. Unlike other alignment algorithms, this method allows for the identification of conserved metabolic enzymes, compounds, and reactions without abstraction across multiple species. Identification of phenotype-specific subnetworks is based on the assumption that pathways that are conserved primarily in phenotype-expressing organisms are more likely to be related to the expression of that phenotype.

1.6.1 Impact

The research explained in Chapter 4 impacts the current base of scientific knowledge as follows.

- Through application of the NIBBS search algorithm and Student's T-test, known and novel phenotype-related pathways across entire metabolic networks, within sets of phenotype-expressing organisms, were identified. The predicted pathways and sub-networks identified by the NIBBS algorithm provide further information regarding metabolic processes involved in hydrogen production. Such information can be used to develop testable hypotheses about the role of pathways within model organisms.
- While the phenotype acidophilic has been studied in a number of organisms, information regarding acid-tolerance in *C. acetobutylicum* and other hydrogen producing organisms is not well known. In this study, metabolic pathways and subnetworks related to organisms capable of acid-tolerance are reported. Such information provides for further understanding of acid response metabolic mechanisms in bacteria.
- The multiple alignment algorithm is the first algorithm capable of aligning multiple pathways based on the similarity between reactions, compounds, and enzymes present in the pathway.

1.7 Identification of Functional Network Associations for Specific Phenotypes (Chapter 5)

The last component of this research focuses on identifying phenotype-related functional association networks and modules in sets of phenotype and non-phenotype expressing microorganisms. The focus of the research in Chapters 2 and 3 is to identify individual phenotype-related genes and phenotype-specific metabolic pathways in order to understand which metabolic processes and entities are associated with expression of target phenotypes. While knowledge of phenotype related genes and metabolic pathways is important for designing genetically modified organisms, expression of a phenotype by microorganisms is usually the result of interactions between multiple pathways rather than individual pathways. In biological systems, phenotype-related genes encode for a number of functionally associated proteins which may be found across a number of different metabolic, regulatory, and signaling pathways. Together these pathways form a functionally associated network that is responsible for the expression of a particular phenotype.

In this Chapter the goal is to identify phenotype-related functional association networks to provide further understanding of regulation of phenotype expression and predict controls on metabolic pathways and networks. To accomplish this goal, three complementary approaches were used to identify significantly conserved subnetworks: (1) comparative analysis of multiple functional association networks, (2) comparative analysis of multiple functional modules using the α, β- clique algorithm, and (3) identification of enriched subgraphs using knowledge priors (Dense ENriched Subgraph Enumeration (DENSE) algorithm).

In the first approach, functional graph modules of phenotype-specific metabolic pathways that were previously identified using the Student's T-test and NIBBS search algorithm (Chapter 4) were created and aligned to identify association between graph modules. In each graph modules, nodes represented phenotype-related ortholog genes and edges represented a connection or relationship between nodes. At the present time, this relationship has not been identified, but potential relationships may include interactions and regulations of genes. Using a parallel algorithm, the maximum clique size was determined for each module to identify sub-graphs that are densely connected. To identify phenotype related functional association networks, modules were aligned and scored to measure the similarity between metabolic pathways and identify conserved functionally associated subnetworks. Subnetworks identified, were used to evaluate metabolic controls and regulation of pathways responsible for expression of phenotypes.

In the second approach, *α,β*-motifs allow for identification of *functional modules* that, in addition to metabolic subsystems, could include their regulators, sensors, transporters, and even uncharacterized proteins that are predicted to be related to the target phenotype. By comparing hundreds of genome-scale networks of functionally associated proteins, this method identifies those functional modules that are enriched in at least *α* networks of phenotype-expressing organisms but may still appear in no more than *β* networks of organisms that do not exhibit the target phenotype. Using *α,β*-motifs approach, for instance, clusters of genes responsible for synthesis, metal insertion, or regulation of hydrogenase and nitrogenase enzymes complexes were identified. Within hydrogen producers, these two complexes play important roles in the production of hydrogen.

In the last approach, a clustering algorithm capable of utilizing *prior* knowledge to identify dense and enriched biological subnetworks was co-developed. Identification of dense subnetworks allows one to identify groups of proteins that interact with each other and are likely to be functionally related. To identify these subnetworks, proteins known to be present in specific organisms were used to created biological networks. Applying the clustering algorithm, dense and enriched subnetworks were identified. The density and enrichment of a subnetwork was tested to identify biologically relevant parameters. Information provided by identified subnetworks provides knowledge regarding proteins associated with potentially important species for biological hydrogen production.

1.7.1 Impact

The research explained in Chapter 5 impacts the current base of scientific knowledge as follows.

- Identification of conserved functional networks related to target phenotypes provides insight on how metabolic pathways may regulate expression of phenotypes.
- Identification of cross-talk between interacting sub-networks to fully understand and predict key metabolic networks involved in phenotype expression (e.g., hydrogen production). To the best of the author's knowledge, there are currently no computational methods available for predicting cross-talk between metabolic pathways and networks.
- Dense subnetworks obtained using knowledge priors, can be used to improve functional annotations of proteins within target organisms. Because proteins

within the clusters are assumed to interact with one another and likely to have similar functions, predictions regarding the function of uncharacterized proteins can be made.

- Proteins identified using the DENSE algorithm can be used to provide clues to genes responsible for synthesis, metal insertion, or regulation of hydrogenase enzymes. Identification of key regulatory mechanisms involved in either activation or inactivation of various hydrogenases allows metabolic engineers to focus on individual accessory proteins involved in regulation of these enzymes.

CHAPTER 2: IDENTIFICATION OF IDEAL MICROBIAL PHENOTYPES FOR BIOLOGICAL HYDROGEN PRODUCTION FROM WASTEWATER AND WASTE MATERIALS

2.1 Biohydrogen Production as a Renewable Energy Resource

As the demand for fossil fuels and cost for oil increases, there is a growing sense of urgency to identify alternative renewable resources to meet U.S. energy and fuel demands on a national and global level [26, 38]. One promising alternative is biological production of hydrogen as a renewable energy source for bioenergy and biofuel [53, 54]. Consideration of hydrogen as an alternative energy source is particularly important in terms of green energy and addressing global climate changes—including emission of greenhouse gases from fossil fuels. Hydrogen itself is considered an environmentally friendly alternative fuel since it is able to burn clean without emission of carbon dioxide into the atmosphere [4, 11, 25]. Similarly, biological hydrogen production is considered a green technology since it does not rely upon the energy intensive processes commonly seen in abiotic and traditional approaches, such as electrolysis of water [22, 26, 38].

In addition to being a more environmentally friendly source of fuel, biohydrogen production is favored over other alternative fuels due to a number of factors. These include: (1) utilization of hydrogen for a number of applications (e.g. synthesis of ammonia, fuel) [2]; (2) the ability to generate high energy yields [20]; (3) the ability of microorganisms to generate biohydrogen using different metabolic processes [26, 38]; (4) ability of microorganisms to degrade organic compounds, including some organic pollutants [2, 26]; and (5) availability of large quantities of biomass from biosolids or other renewable resources, such as wastewater [2, 26, 55]. Of these factors, the potential energy yield produced from hydrogen makes biohydrogen one of the leading candidates for bioenergy today.

Examples of energy yield from various processes including hydrogen are shown in Table 2.1. Based on reported literature values hydrogen has a thermal value of 118 kJ/g of substrate (glucose) [12]. The highest potential energy yields are reported as ranging from 122 kJ/mol of substrate (glucose) to 186 kJ/mol substrate compared to ethanol and methane energy yields of 18 and 145 kJ/mol substrate, respectively [2, 4, 11, 26]. Variations in energy yield from hydrogen are partially due to the specific biological processes employed for generating biohydrogen and reactor-specific operational parameters. These factors and their impact on hydrogen production yields will be discussed throughout the following sections.

2.2 Microorganisms Capable of Hydrogen Production

To date a number of hydrogen-producing microorganisms have been identified, spanning across different phenotypes and phylogenies. A list of well-studied hydrogen-producing microorganisms is provided in Table 2.2. Phenotypes include, but are not limited to thermophiles, mesophiles, aerobes, fermentative species, photosynthetic species, and acidophiles [20, 22, 26, 56]. Of the species, *Escherichia coli, Clostridium, Anabaena*, and *Enterobacter* are some of the most widely studied organisms for hydrogen production. This is due to their potential for higher production rates compared to other species, comprehensive understanding of metabolic systems, and/or their ability to survive chemical and environmental operational conditions (e.g. temperature and pH) for a specific system [11, 16].

Table 2.1 Energy yield from various processes.

Renewable resource	Potential energy yield (kJ energy)	Reference
Ethanol	148 kJ/L	[4]
Methane	145kJ/L	[4]
Hydrogen	185 kJ/L	[2]
Hydrogen	186 kJ	[4]
Hydrogen	143 GJ/L	[22, 23]

Despite the number of hydrogen-producing organisms reported in literature, complete understanding of the metabolic processes and factors influencing hydrogen production for each organism is still lacking. Depending on a number of factors, such as environmental conditions (e.g., temperature, exposure to toxins) and internal regulation (e.g., metabolic pathway available, gene expression, transcriptional regulation), the amount of hydrogen produced will vary among organisms. In particular, the type(s) and supply of carbon source present plays a significant role in the ability of the organisms to produce hydrogen. This is particularly important since not one microorganism can utilize all types of organic compounds available.

An example of how utilization of organic compounds can be limiting is in *Zymomonas mobilis*, a microorganism potentially important for production of solvents (e.g., ethanol) and biohydrogen [35]. Application of this organism in hydrogen producing systems is often limited since it is not able to directly utilize cellulosic compounds but rather intermediates such as glucose, sucrose, and fructose [53]. To understand how different organisms can utilize various organic compounds, a short

Table 2.2 List of well-studied hydrogen producing organisms. The National Center for Biotechnology Information (NCBI) was used to verify completely sequenced organisms.

Biological Method	**Organism**	**Sequenced**	**Reference**
Bio-photolysis	*Anabaena cylindrica*	no	[11]
	Anabaena sp. PCC 7120	yes	[17]
	Anabaena variabilis	yes	[11]
	Chlamydomonas reinhardtii	no	[11]
	Chlorella fusca	no	[11]
	Chlorococcum littorale	no	[11]
	Oscillatoria sp.	no	[11]
	Scenedesmus obliquus	no	[11]
	Synechococcus sp.	no	[11]
Light Fermentation	*Chloroflexus aurantiacus*	yes	[2]
	Rhdopseudomonas palustris	yes	[2, 11, 17]
	Rhodobacter capsulatus	no	[12, 26]
	Rhodobacter sphaeroides	yes	[12, 26]
	Rhodovulum sulfidophilum W-1S	no	[11, 27]
	Rhosospirillum rubruum	yes	[26]
Dark Fermentation	*Bacillus licheniformis* ATCC 14580	yes	[20]
	Caldicellulosiruptor saccharolyticus DSM 8903	yes	[2]
	Clostridium acetobutylicum	yes	[2, 16]
	Clostridium beijerinckii NCIMB 8052	yes	[16]
	Clostridium butyricum	no	[12]
	Clostridium paraputrificum M-21	no	[11]
	Clostridium pasteurianum	no	[11]
	Clostridium perfringens ATCC 13124	yes	[16]
	Clostridium thermocellum DSM 1313	yes	[16]
	Clostridium thermolacticum	no	[11]
	Desulfotomaculum geothermicum	no	[20]
	Enterobacter aerogenes	no	[11]
	Enterobacter cloacae ITT-BY 08	no	[11]
	Escherichia coli	yes	[2]
	Hafnia alvei	no	[11]
	Methanobacterium sp.	no	[12]
	Thermoanaerobacterium thermosaccharolyticum	no	[20]
	Thermoanaerobacterium thermosaccharolyticum (band B-1)	no	[11]
	Thermotoga martima	no	[2]
	Thermotoga neapolitana DSM 4359	yes	[20]

review of the different types of methods for biological hydrogen production are described below.

2.3 Biological Methods for Hydrogen Production

One of the benefits of utilizing biohydrogen production techniques is the ability to utilize a number of different hydrogen-producing microorganisms under varying environmental constraints (e.g., presence or absence of oxygen, temperature, and pH). In most biohydrogen studies, these organisms are often classified into three main categories based on the method of hydrogen production. These three categories are: (1) decomposition of water by photosynthesizing microorganisms (i.e., bio-photolysis); (2) photo-fermentation of organic matter (i.e., light fermentation), and (3) dark fermentation of organic matter [22, 26, 36, 37, 53].

2.3.1 Bio-Photolysis

Bio-photolysis is a process which involves the conversion of water molecules into hydrogen and oxygen by photosynthetic microorganisms, such as cyanobacteria or microalgae [7, 22, 38]. The process of bio-photolysis can be described through the following reaction:

$$H_2O \rightarrow H_2 + 1/2O_2$$

Production of hydrogen through this process can be carried out either directly by exposure to solar radiation or indirectly under dark (fermenting) conditions [7]. During direct bio-photolysis hydrogen production is generally mediated by nitrogenase, an essential enzyme for nitrogen-fixation in cyanobacteria [6, 7]. In cyanobacteria, such as *Anabaena*, nitrogenase(s) is located within the heterocyst of the organism, where nitrogen-fixation or the conversion of nitrogen gas to ammonia occurs and hydrogen is evolved [6, 39]. Figure 2.1 shows during hydrogen evolution, pigments present in both photosystem I and photosystem II absorb solar radiation and ferredoxin (FD) reduces protons present to produce hydrogen.

Indirect bio-photolysis utilizes stored energy, such as carbohydrates, produced during photosynthesis. Unlike the direct process, indirect bio-photolysis involves hydrogenase—an important enzyme involved in hydrogen uptake and evolution in dark fermenting microorganisms [22, 57]. During this process, energy is released during dark fermentation (absence of light) and reducing equivalents are oxidized by hydrogenase rather than nitrogenase to generate hydrogen gas [7]. Since hydrogen evolution occurs as a by-product and is generally not considered beneficial to the organisms, in general hydrogen conversion efficiency is <1% [6, 7]. As such, direct and indirect bio-photolysis is generally not considered as the ideal method for biological hydrogen production.

2.3.2 Light Fermentation

Similar to bio-photolysis, light fermentation relies on photosynthetic microorganisms to carry out biological hydrogen production through reduction of N_2 (gas) by nitrogenase enzymes [6, 8, 38, 46]. During this process, organisms are able to utilize simple organic compounds as a carbon source (e.g., glucose and sucrose) and sunlight (or an artificial light source) to produce energy required to carry out metabolic and growth requirements [6, 38]. The overall hydrogen-producing reaction (nitrogenase reaction) is:

$$N_2 + 8e^- + 10\ H^+ + 16MgATP \rightarrow 2NH_4^+ + H_2 + 16\ MgADP + 16\ Pi$$

In this reaction Pi represents orthophosphate [8, 26]. A schematic drawing of the metabolic process for hydrogen production is shown in Figure 2.2

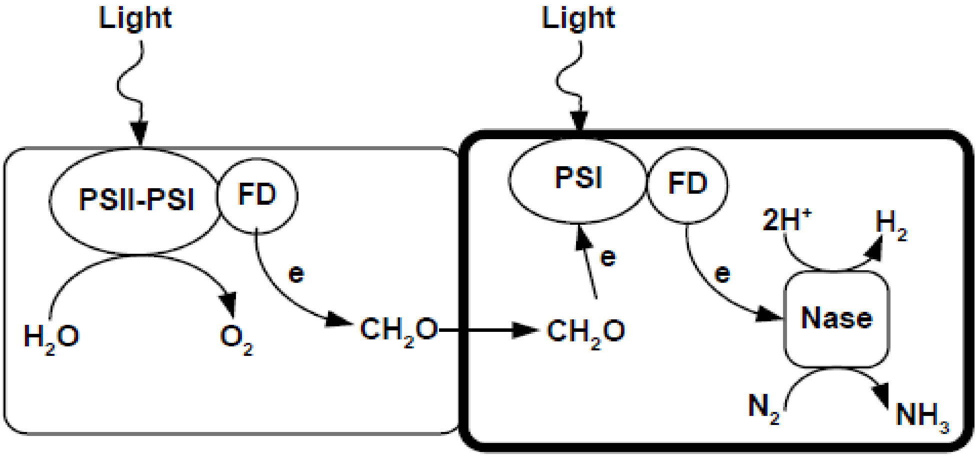


Figure 2.1 Example of hydrogen formation from nitrogenase and the reducing equivalent ferredoxin in nitrogen-fixing cyanobacteria. Picture taken from Yu and Takahashi [7]

One advantage of using light fermentation reactions is the ability of these photosynthetic organisms to degrade and utilize a number of different organic substrates, including those typically found in wastewater [6, 26]. In light fermentation, hydrogen production is generally carried out by purple, non-sulfur photosynthetic bacteria, such as *R. palustris* [26]. As mentioned earlier, nitrogenase enzymes present in these bacteria are key enzymes for hydrogen production. However, within these photosynthetic organisms hydrogenases –enzymes responsible for hydrogen evolution and hydrogen uptake within prokaryotic cells [39, 57]—may be present. Hydrogenase catalyzes the following reversible reaction:

$$H_2 \rightarrow 2H^+ + 2e^-$$

In *R. palustris*, these enzymes are generally involved in hydrogen uptake. As such, the presence of hydrogenase uptake enzymes in these organisms results in lower hydrogen yields.

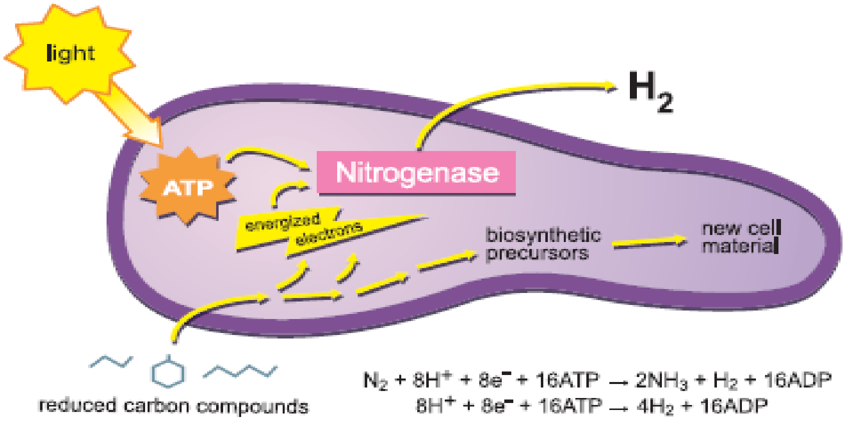


Figure 2.2 Overview of the metabolic process for hydrogen production via light fermentation. Image taken from Rey et al [6]

In addition to the presence of hydrogen uptake proteins, another important factor impacting hydrogen yields is the availability of light energy to carry out photosynthetic fermentative and light harvesting reactions [26]. Application of light fermentation to hydrogen production is not widely used because of the costs associated with supplying light and problems with maintaining adequate light penetration from sunlight as microbial biomass increases [26].

2.3.3 Dark Fermentation

Dark fermentative bacteria differ from the previous two hydrogen-producing methods in that hydrogen evolving reactions are carried out without light energy by a number of heterotrophic bacteria [4, 22]. A detailed list of 21 dark fermentative bacteria was provided previously in Table 2.2. In general, hydrogen is produced from dark fermentation reactions when organic substrates (or biomass resources) are utilized by heterotrophic bacteria as both the carbon and energy source for heterotrophic growth [4, 22]. During this process, organic substrates are hydrolyzed and processed through the glycolytic pathway through a series of reactions to generate pyruvate. From here pyruvate is converted to acetyl-CoA and then to acetate to form CO_2 and H_2 (gas) [2, 22]. To generate H_2 (gas) two enzymes, pyruvate-ferredoxin oxidoreductase and hydrogenase, are required. Ferredoxin catalyzes the reaction to oxidize acetyl-CoA and hydrogenase converts reduced ferredoxin to oxidized ferredoxin and H_2 (gas) [2, 8]. A schematic depicting these reactions is shown in Figure 2.3.

Although hydrogen can be produced aerobically, there is growing attention on anaerobic and facultative anaerobic bacteria due the ability of these organisms to utilize a wider variety of complex organic compounds and produce a number of active hydrogenase enzymes. Hydrogenases are key enzymes for hydrogen evolution in dark fermentative bacteria and are easily inactivated by various factors, including the presence of oxygen [57]. Due to the importance of hydrogenase in both hydrogen production and hydrogen uptake, several studies have examined the role of hydrogenase enzymes in a number of different hydrogen-producing organisms [50, 58]. These studies have found many microorganisms, including *Clostridium acetobutylicum*, are capable of having both hydrogen uptake (e.g. [FeFe]-hydrogenase) and hydrogen evolving enzymes (e.g. [NiFe]-hydrogenase). Although much work has been done on characterizing a number of hydrogenase enzymes in several bacteria, clear understanding of the regulation of these enzymes and their interplay with various metabolic pathways is still lacking.

Although dark fermentative bacteria are capable of utilizing a number of organic compounds and contaminants, maximum hydrogen yields are 4 moles of H_2 per mole of glucose [2, 16, 22] for anaerobic bacteria and 2 moles of H_2 per mole of glucose for facultative anaerobes [22]. The reason for low yields is mainly due to utilization of glucose towards cellular growth [2]. One possibility currently being considered for

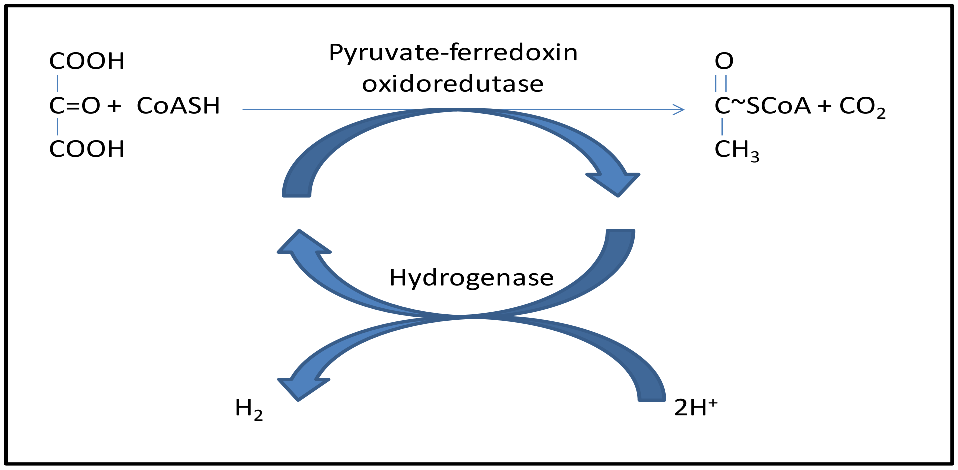


Figure 2.3 Schematic of the pyruvate-ferredoxin oxidoreductase (1) and hydrogenase (2) reactions involved in dark fermentative hydrogen production**.** (1) Pyruvate is oxidized to generate acetyl-CoA and CO2. (2) Ferredixon (Fd) is oxidized by hydrogenase to produce H2. Schematic was reconstructed following the figure from D. White [8].

enhanced hydrogen yields is genetic modification or re-direction of metabolic pathways away from biomass production and towards hydrogen production [34].

Due to the ability of light and dark fermentative bacteria to evolve hydrogen using ambient temperatures and reduce a number of organic substrates, the research described in Chapters 3-5 focuses primarily on fermentative microorganisms rather than those involved in bio-photolysis. In particular, emphasis will be placed primarily on dark fermentation reactions.

2.4 Wastewater and Waste Materials as Biomass

One of the most fundamental problems towards successful application of bio-energy technologies is the selection of the biomass or biologically derived materials used by organisms to generate renewable energy [59]. A number of studies have shown that biomass from cellulose and other plant derivatives, animal wastes, fats, industrial wastes, and wastewater can be utilized by a number of different microorganisms to produce bioenergy and biofuels [3, 11, 38, 53, 59]. Of these biomass resources, wastewater and other municipal waste materials (e.g., solid waste, wastewater biosolids) are currently being considered as a potentially important renewable resource for biological hydrogen production [2, 11, 16, 56].

Consideration of wastewater for hydrogen production is due to the following factors. First, wastewater and its waste materials contains organic substrates available to serve as carbon and energy sources for fermentative microorganisms [2]. This is particularly important since a number of dark fermentative organisms can break down and utilize various types of simple and complex organic compounds. Second, utilization of wastewater does not require the use of potentially important agricultural crops or other competitive biomass [11]. Wastewater is also abundant (and its generation will increase as the world urbanizes and its population increases) and can be converted into hydrogen gas, as well as other biogases at a relatively fast rate.

Another advantage to using wastewater as a biomass, particularly in dark fermentation reaction, is the presence of established technologies at wastewater treatment facilities [60]. With established technologies in place, the cost for building a suitable facility can be significantly reduced. During current wastewater treatment, organic matter is decomposed and fermented to generated organic acids, CO_2, and H_2 (gas) through dark fermentation reactions [2, 25]. To remove the remaining organic acids from wastewater, a second step involving methanogenesis is carried out, resulting in the production of CH_4 and $CO_{2.}$ Often acidogenesis and methanogenesis are carried out together in an anaerobic digester by two different communities of bacteria. Hydrogen production is conducted by acidogenic bacteria while methanogenesis is carried out by hydrogen-consuming methanogens [2]. Although methane gas can be used as an alternative fuel source, hydrogen may be preferred due to its large range of industrial uses, its use in fuel cells, and its ability to burn cleanly [2, 61].

To generate renewable energy sources such as biohydrogen and treated wastewater, methanogenesis must either be carried out in a separate reactor to avoid hydrogen consumption and to remove remaining organic acids, or it must be removed from the wastewater treatment process and replaced by an alternative method. One potential method for continued treatment of wastewater materials is a secondary step involving photo-fermentation. In this process, organic acids produced during the dark fermentation stage would be further degraded, thus removing organic wastes from the water and lowering the oxygen demand. As more research is carried out, other potential wastewater treatment reconfigurations and designs will be developed and tested for optimal performance and cost.

Lastly, wastewater itself is an important renewable resource and there is now much discussion about viewing it a resource, not as a waste [34]. Conventional wastewater treatment facilities generally have high costs associated from operational requirements (e.g., electricity, chemical additions for treatment) [2]. Through simultaneous wastewater and biological hydrogen production there is a potential to address some of the costs by direct utilization of hydrogen generated while recovering valuable water. Development of microbial fuel cells for direct utilization of energy is one future goal for wastewater treatment facilities and transportation (e.g., combustible engines) [35].

2.5 Current Application and Studies Using Wastewaters From Various Sources

To evaluate potential hydrogen yields and the impact of different types of wastewater on hydrogen production, laboratory and bench-scale studies have been conducted using a variety of wastewater and waste materials from industrial, agricultural, and municipal sources. Wastewater sources considered here include (but are not limited to) rice winery wastewater [25], food wastes [20], tofu wastewater [12], swine wastewater [15], and municipal wastewater and solid wastes [2, 56].

Results from these studies show differences in hydrogen production yield due to variation in basic operational parameters (e.g., hydraulic retention times and temperature), variation in organic substrates (e.g., presence of glucose or sucrose), and types of microorganisms present in the system. In wastewaters containing high concentrations of carbohydrates, such as those in rice winery and food wastes, higher hydrogen yields were generally obtained. Examples of hydrogen yields obtained from different wastewaters are provided in Table 2.3.

2.6 Current Challenges to Hydrogen Production in Wastewater Treatment

Based on literature reviews of studies evaluating potential biological hydrogen production from varying wastewaters, a set of biological and operational factors impacting the rate and yield of hydrogen production have been identified [54]. Understanding these biological and abiotic factors upon hydrogen production is essential, particularly when designing efficient systems for hydrogen evolution or restructuring current wastewater treatment systems for bioenergy production.

Table 2.3 Examples of hydrogen production yields using different wastewater types as biomass.

System	Organic compounds or carbohydrates present	Operational parameter	Type of organisms present	Reported hydrogen yields	Reference
Tofu wastewater	Sucrose and starch were the primarily sources		Anoxygenic phototrophs	1.9 ml ml $^{-1}$ wastewater (in 120hr)	[12]
Swine wastewater	Organic compounds in wastewater; glucose addition in some samples	pH = 4.8-5.9	Mixed group of acidogens	0.9-1.0 m^3 m^{-3} day^{-1}	[15]
Food waste	Organic compounds from food in a 1:3 food to water diluted solution	pH= 4.5-6.5	Acidogens	Thermophilic H_2 yields ranged from 0.6-0.9 mol mol-1 hexose Mesophilic H_2 yields ranged from 0.03-0.1 mol mol-1 hexose	[20]
Rice winery wastewater	Raw wastewater; carbohydrates	pH=4.5-6.0	Acidogens	1.37-2.14 mol mol^{-1} hexose	[25]

2.6.1 Microbial Composition for Hydrogen Production

Amongst the factors listed above, one of the most critical factors impacting hydrogen production yields in dark fermentation reactions is the composition of the microbial community in the anaerobic digester and the community coming in to the digester from the influent. Studies have shown that in order to enhance hydrogen production, the microbial community in the anaerobic digester should consist mainly of hydrogen producers. Potential hydrogen consumers (e.g., methanogens) present in the systems should either be limited or removed from the system [2].

Currently, microbial communities in anaerobic digesters consist of a mixed community of anaerobic or facultative anaerobic bacteria capable of both hydrogen production and consumption [2]. Overall species composition in wastewater and in treatment systems may contain *E. coli*, *Clostridia* sp., methylotrophs and methanogens [2] with the target organisms for wastewater treatment being methanogens. However, for hydrogen production this type of microbial community commonly observed during waste treatment would actually result in decreased hydrogen yields since many methylotrophs and methanogens are also hydrogen-consumers [54]. Based on literature reports, *Clostridium* and *Enterobacter* are two of the most widely considered genera for hydrogen production due to their ability to generate high yields of hydrogen, utilize a number of organic compounds, their presence in natural environments, and their ability survive in a number of environments [2, 11, 16, 54]. Therefore, ideally one or both of these organisms should be present in a mixed community of hydrogen-producing organisms.

Although several operational parameters can be altered to select for and maintain hydrogen-producing in anaerobic digesters, influent coming from industrial, agricultural, or municipal sources will contain both naturally occurring hydrogen-producing and hydrogen-consuming microorganisms. Microorganisms introduced from the influent may not only take-up hydrogen, but may out-compete important hydrogen producers for essential nutrients.

One approach to address this problem is to treat the sludge entering the anaerobic digester to remove any unwanted microorganisms and target others. Pretreatment methods may be range from chemical additions (e.g., inhibitors) [62], heat shock treatments [63], and altering of the pH to conditions unfavorable to hydrogen-consumers [2]. However, one must be cautious when applying pretreatments to incoming sludge as treatments may have direct and/or indirect impacts on hydrogen producers.

To provide an understanding of how these factors impact biological hydrogen production, each factor will be discussed briefly. In addition, a general overview of the different factors and how they impact hydrogen production is provided in Table 2.4. A detailed review on biological and operational concerns has be performed previously by Li and Fang [2, 26] and Kapdan and Kargi [11].

2.6.2 Biological Challenges: Impacts on Hydrogen Production

In biological systems, whether in anaerobic digester or culture flasks, if conditions are not favorable the organism will not grow or will not express desired phenotypic traits. The same is true for culturing microorganisms for wastewater treatment and biological hydrogen production. Factors that impact both biohydrogen production and removal of organics from wastewater includes: availability of organic compounds, additional nutrients, toxicity, endospore formation, and pH (shown in Table 2.4) [2, 11].

Table 2.4 Biological factors impacting hydrogen production in anaerobic digesters.

Factor	Impact on system	Potential Solution(s)	References
Availability of organic compounds	As organic loading decreases, hydrogen production will decrease	Supplement of sludge or wastewater Decrease retention time to introduce new sludge material	[13, 14]
Phosphorus	Lack of basic nutrients such as nitrogen and phosphorus will inhibit growth	Supplement with nutrients	[18]
Iron and Magnesium	Trace metals such as iron and magnesium are important for hydrogen production due to their role in hydrogenase enzymes Iron limitation may result in enhanced solvent production and decreased hydrogen production	Supplement with iron and other trace metals	[18, 24]
pH	Fluctuations in pH aids in selection for microbial communities and helps to regulate the type of organic acids produced. pH 4.5 – 6.0 is the range needed for limiting hydrogen consumers and for production of acetate as a by-product	Adjust pH chemically or adjust flow hydraulic retention times to reduce build up of organic acids and acidification of the system	[2, 11, 18]
Temperature	Higher temperatures result in higher hydrogen yields (> 35°C)	Incorporate short periods of heat treatment. However, this would provide additional costs to the operation	[2, 11]

In addition to the selection of the types of microorganisms present in anaerobic digestion or dark fermentation reactions, the type of available organic substrates and concentration of the substrates plays a large role in hydrogen production. In wastewater lacking high concentrations of carbohydrates or carbon-rich organic matter, hydrogen yields are generally lower than in water and waste materials rich in degradable organic matter [2, 13]. For example, in studies by Okamoto et al. [13] and Lay et al. [14], wastewaters containing carbohydrates reported yields 10-20 times higher than wastewater and waste materials which contained mostly fats and other organic content.

When designing hydrogen producing biological systems, another factor to consider is the availability of essential nutrients and trace metals necessary for growth and carrying out metabolic functions (e.g., hydrogen evolution). These factors include nitrogen, phosphorus, iron, magnesium, and zinc [16, 18]. Iron concentration in particular play a critical role in the maturation and activation of hydrogenase enzymes responsible for hydrogen evolution [50]. In a study by Lee et al. [64], the optimal iron concentration was reported as 353 mg Fe^{3+} L^{-1}. In addition, they found that under iron-limitation, a metabolic shift in the production of fermentation end-products would occur, resulting in production of solvents rather than acetate.

A potential problem with the generation of hydrogen production is the simultaneous production of organic acids as by-products. Depending on pH, temperature, and other environmental conditions, the types of fermentation end-products produced may shift, resulting in the formation of undesirable acids (e.g., lactate and propionate) or solvents (e.g., ethanol) [16, 20]. In a study by Shin et al. [20], a distinct shift in end products was noticed when the pH changed from pH 4.5 to 6.5. As the pH increased, acetate production increased [20]. However, at pH = 4.0 propionate and ethanol have been reported as the predominant fermentation end-products [25].

While the type of fermentation end-product present in a system may not seem important, the end-product allows identification of the type of metabolic fermentation processes occurring in the cell. It also provides information on whether or not the system(s) is efficiently producing hydrogen. For enhanced hydrogen production, acetate is the desired fermentation end-product because of its higher hydrogen yield compared to other by-products, such as butyrate [16]. Specific differences in conversion efficiencies can be observed by comparing the two chemical reactions below:

$C_6H_{12}O_6 + 2H_2O_2 \rightarrow 2CH_3COOH + 2CO_2 + 4H_2$ glucose into acetate

$C_6H_{12}O_6 \rightarrow CH_3CH_2CH_2COOH + 2CO_2 + 2H_2$ glucose into butyrate

The first reaction shows that the maximum theoretical hydrogen yield is 4 H_2 per mol of glucose produced when acetate is the end product [2, 54] compared to a maximum theoretical hydrogen yield of 2 H_2 with butyrate as the end product [2, 8, 16].

While this may seem like a limiting factor in hydrogen production, research is currently underway to increase maximal yields by changing environmental parameters (e.g., increased temperatures) and re-directing metabolic pathways (e.g. metabolic engineering). Furthermore, as hydrogen and fermentation end-products are introduced into the system, build up of organic acids will result in a decrease in pH [54]. Change in pH has been reported to not only change metabolic shifts, but to also directly impact hydrogenase activity [65]. To overcome this problem, research has been conducted to assess whether changes in operational standards may help in maintain pH levels. In fact some studies have suggested to have a continuous flow to remove build up of organic acids as new waste material enters the system [2, 16].

2.6.3 Operational Parameters: Temperature, HRT, Partial Pressure

Other factors impacting hydrogen production yields in wastewater treatment systems include operational parameters set by the system design. These include, but are not limited to: temperature, hydraulic retention time, and partial pressure. A general overview of operational parameters important for hydrogen production is provided in Table 2.5. A brief description of key parameters and their impacts on hydrogen yields is discussed below. An overview of potential operational and biological challenges for dark fermentative hydrogen production is depicted in Figure 2.4.

To increase hydrogen production rates, several studies have shown that increased temperatures are capable of producing hydrogen yields greater than the theoretical limit of 4 mols of H_2 per mol of substrate (e.g., glucose) [2, 25, 54]. In fact, a study by Classen et al. [9] demonstrated that production of hydrogen is thermodynamically unfavorable at mesophilic temperature ranges. This is due to the high positive Gibbs free energies for hydrogen production from fermentation products, such as acetate [9]. Table 2.6 provides a comparison of Gibb's free energies for the fermentation reactions involved in hydrogen production. Similarly, a review of differences in hydrogen yields conducted by Li and Fang [2] indicates hydrogen yields increase from an average of 173 mL H_2 g^{-1} hexose to 191 mL H_2 g^{-1} hexose when temperature ranges increased from 15-30ºC to 32-39ºC. While other studies have reported significant increases between mesophilic and thermophilic temperature range [1], Li and Fang report values that are comparable for these temperature regimes. .

Table 2.5 Operational factors impacting hydrogen production in anaerobic digesters.

Factor	Impact on system	Potential Solution(s)	References
Temperature	High temperature are associated with hydrogen production yields	Conduct fermentation reactions at high temperatures (< 35^{o}C)	[1-3]
	High temperatures may inhibit hydrogen production by mesophilic bacteria	Use thermophilic organisms or run the reactor a mesophilic temperatures	
Partial pressure	Build up of partial pressure is shown to inhibit hydrogen production	Release of gas build up by addition of inert gas (e.g. nitrogen) Remove biogas(es) to a separate container	[2]
Hydraulic retention time (HRT)	Hydrogen production rates tend to increase with higher HRT values	Evaluation of optimal HRT values to determine optimal HRT for individual systems	[16]

Table 2.6 Gibbs free energy values for hydrogen producing reactions. Values represent key reactions involved in biological hydrogen production using dark fermentative processes. Fd = Ferredoxin; Red = Reduced; Ox = Oxidized.

Reaction	$\Delta G'_0$ kJ/mol	Reference
Acetate → Hydrogen		
$CH_3COOH + 2H_2O \rightarrow 2H_2 + 2\ CO_2$	+104.6	[9]
Butyrate → Hydrogen		
$CH_3CH_2CH_2COO^- + 2H_2O \rightarrow 2H_2 + 2\ CHO_3COO^- + H^+$	+48.1	[8]
Formate→Hydrogen		
$HCOO^- + H_2O \rightarrow H_2 + HCO_3$	+1.1	[19]
Formate: Hydrogen lyase		
$H_2 + HCO_3^- \rightarrow HCO_2^- + H_2O$	-1.3	[21]
Hydrogenase		
$2Fd^{Red} + 2H^+ \rightarrow 2Fd^{Ox} + H_2O$	-1.3	[21]

While higher temperatures may result in enhanced hydrogen yields from fermentation reactions, it is not an ideal option due to costs associated with energy inputs to increase temperature. In order to justify hydrogen production from a wastewater source, the goal would be to create a system that is more cost efficient in the long run—not more material or energy intensive. A second downfall to applying high temperatures to complete fermentation reactions is its potential to decrease or inhibit growth of key hydrogen-producing microorganisms present [2, 11, 54]. For example, some hydrogen-producing microorganisms form spores or vegetative cells until environmental conditions are favorable [2, 16, 54]. However, not all hydrogen-producers are capable of spore-formation and those that cannot tolerate high temperatures will not survive.

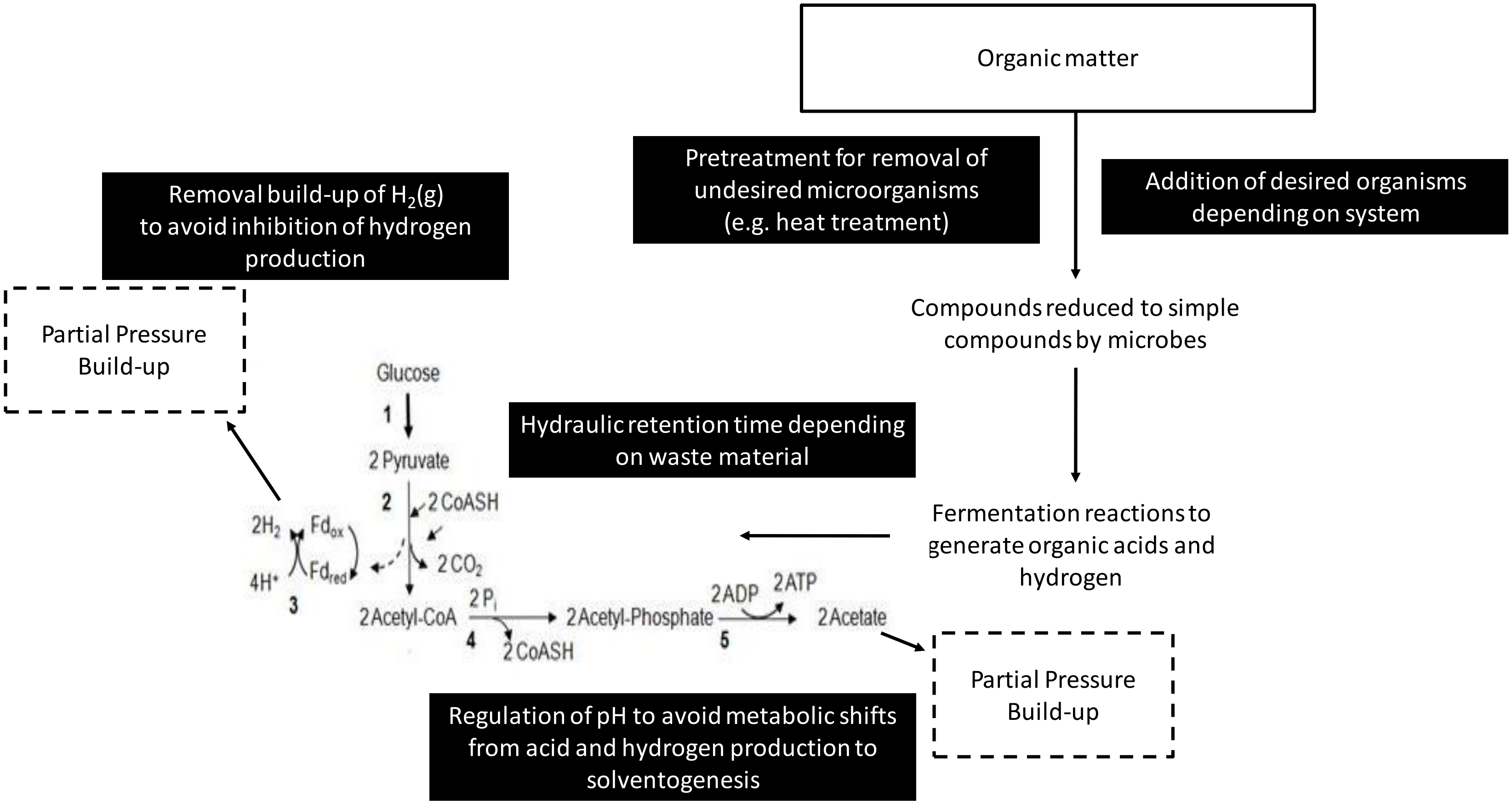

Organic matter
Pretreatment for removal of undesired microorganisms (e.g. heat treatment)
Addition of desired organisms depending on system
Compounds reduced to simple compounds by microbes
Fermentation reactions to generate organic acids and hydrogen
Removal build-up of H_2(g) to avoid inhibition of hydrogen production
Partial Pressure Build-up
Hydraulic retention time depending on waste material
Glucose
1
2 Pyruvate
2
2 CoASH
2 CO_2
$2H_2$
Fd_{ox}
Fd_{red}
$4H^+$
3
2 Acetyl-CoA
2 P_i
4
2 CoASH
2 Acetyl-Phosphate
2ADP
2ATP
5
2 Acetate
Partial Pressure Build-up
Regulation of pH to avoid metabolic shifts from acid and hydrogen production to solventogenesis

Figure 2.4 Schematic overview of dark fermentative hydrogen production from glucose. Potential operational problems (dashed boxes) and solutions (dark boxes) are provided. Metabolic reaction is representative of acetic acid fermentation.). Enzymes for *E. coli*: 1, glycolytic enzymes; 2, pyruvate formate lyase (E.C. 2.3.1.54); 3, formate hydrogen lyase (E.C. 1.1.99.33); 4, phosphotransacetylase (E.C. 2.3.1.8); 5, acetate kinase (E.C. 2.7.2.1). Arrows with larger width indicate a series of reactions. Arrows with narrow width indicate individual reactions.

Another factor that regulates the rate of hydrogen production is the build-up of partial pressure during dark fermentation. In the last steps of hydrogen production H_2 and CO_2 are both produced [2, 22]. As the amount of biogas builds up, particularly H_2, hydrogenase activity is inhibited and hydrogen production decreases [2]. To avoid build up of hydrogen gas or biogas in general, there are two possible solutions—sparge the reactor with nitrogen or remove hydrogen from the reactor at the rate it is produced [16].

A number of studies have attempted to identify the optimal hydraulic retention time (HRT). Review of these studies indicates that the optimal HRT fluctuates with wastewater type and a number of environmental conditions [2, 16, 25, 56]. For example, Hawkes et al. [16] reports the optimal HRT value was reported as 8h, while Lay [63] reports the optimal HRT value as 17h. While in general there seems to be an increase in hydrogen production with increased HRT values, there is no precise value currently accepted for all systems.

2.7 Phenotypes for Dark Fermentative Hydrogen Production in Wastewater Treatment Facilities

From the above review of hydrogen production studies using various wastewater sources, a number of biological and operational factors controlling hydrogen yields and rates have been identified. In each study, researchers tested a number of operational controls in order to identify the correct combination of factors to maximize conversion of carbohydrates or organic matter to H_2. While efforts to design efficient wastewater treatment facilities are necessary to reduce costs and improve water quality standards, it is not the only approach towards generating high yields of biohydrogen.

In recent years, focus has shifted towards development of metabolically engineered organisms capable of expressing desired phenotypic traits. For the case of hydrogen production using wastewater, this concept could be applied to create a mixed microbial community “ideal” for enhancing hydrogen production. This is particularly important since a number of phenotypes are necessary to optimize overall hydrogen yields and not a single hydrogen-producing microorganism has been identified that is capable of expressing all these phenotypes. However before steps can be completed to create an ideal hydrogen producing microbial community, fundamental questions regarding metabolic processes involved in expression of phenotypes must be addressed. These questions include:

- Q1: What genes are important for expression of desired phenotypic traits?
- Q2: What phenotypic traits are expressed by what biochemical pathways?
- Q3: Which metabolic networks are associated with expression of desired phenotypes?
- Q4: What regulatory mechanisms are present and how do they control interplay between metabolic pathways?

To address these four questions, phenotypes important for hydrogen production using wastewater and waste materials will be evaluated using computational biology approaches. Phenotypes selected for the study include:

- Hydrogen Production and Dark Fermentative Hydrogen Production:
  In order to understand metabolic and cellular processes involved in the expression and regulation of the phenotype hydrogen production, microorganisms and representative of all three types of hydrogen production will be used to identify phenotype-related genes and pathways. Further, individual and sets of microorganisms known to produce hydrogen using dark

fermentation methods will be evaluated in detail. In wastewater facilities, dark fermentation reactions are considered the most feasible in terms of hydrogen production. Therefore, emphasis in this study will be placed on hydrogen production using dark fermentative bacteria.

- Facultative Anaerobiosis: To carry out hydrogen production using fermentative bacteria in anaerobic digesters, microorganisms need to be either anaerobic or facultative anaerobic. To date a number of anaerobic species, such as *C. acetobutylicum*, have been identified. Due to the possibility of accidental exposure to oxygen when sludge is brought into the reactor or during mixing, it is important to have microorganisms, such as facultative anaerobes, that can survive in aerobic and anaerobic conditions present within the mixed community.
- Acidophilic (pH= 4.5-6.5): Acid-tolerant organisms are those capable of growing in slightly acidic and acidic conditions (pH 4-6). Similar to acidophiles, acid-tolerant organisms have developed metabolic and cellular acid tolerance response (ATR) systems to protect themselves when exposed to acid environments [66].

For hydrogen producers, the presence of ATR systems is extremely important, particularly in respect to acidogenesis. During acidogenesis, organic acids (e.g. butyrate and acetate) are produced, thus lowering the pH level in the medium. In solventogenic organisms, such as *C. acetobutylicum*, the change in pH results in a metabolic shift from acidogenesis to solventogenesis. As a result, the organism will stop producing acetate and butyrate and generate solvents (e.g. acetone and butanol). To prevent metabolic shifts and maintain conversion of glucose (or other sugar compound) to hydrogen at maximum yields, organisms need to be able to tolerate acidic pH conditions.

CHAPTER 3: IDENTIFICATION OF PHENOTYPE-RELATED GENES: AN INTEGRATED APPROACH TO CATALOGING GENES IMPORTANT FOR HYDROGEN PRODUCERS

3.1 Motivation

Throughout the last decade, the availability and number of completely sequenced microbial genomes has substantially increased due to advances in experimental and computational approaches [28]. Sequence data generated through numerous large-scale genome and metagenome sequencing projects is growing at an exponential rate. This is illustrated by the amount of stored data that has doubled at least every 18 months from 1998 to 2008 (Figure 3.1). As a result, computational analysis of microbial genomes and proteomes using microbial sequences has led to further understanding of cellular structures, metabolic networks, and gene functions in model microorganisms that could not otherwise be identified from experimental data alone [28, 67, 68].

In particular, much work has been performed in the area of comparative genomic analysis to identify metabolic processes related to and responsible for the expression of specific microbial phenotypes or traits (e.g., anaerobic respiration) [69]. Examples of these works include phylogenetic profiling studies and experimental molecular biology studies, such as those described in Chapter 1. An example of a computational phylogenetic profiling study includes that by Kastenmuller et al. [70] , which compares phylogenetic profiles for a target phenotype to entire sets of enzymes. From the enzyme sets identified, subsets of enzymes correlated to known enzymes present within metabolic

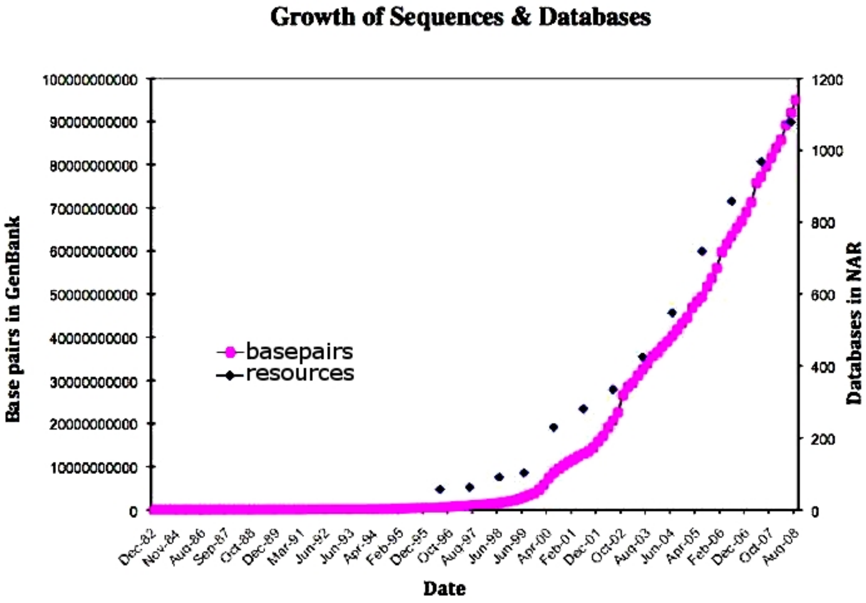


Figure 3.1 The phenomenal growth of sequence data from 1998 to 2008. Image taken from Lathe et al. 2008 [5].

pathways are used to identify phenotype-related pathways [70]. While computational and experimental approaches have aided in the discovery of essential enzymes and pathways involved in phenotype expression, these studies often target individual species or specific metabolic networks (e.g. central metabolism). Thus, information regarding interactions between sub-networks and regulation of key pathways may be missing.

Understanding genotype-phenotype associations is important not only for furthering our knowledge of internal cellular processes, but is also essential in providing the foundation necessary for genetic engineering of microorganisms for industrial use (e.g., production of bioenergy or biofuels) [27]. Modification of phenotypic traits, such as hydrogen production or acid tolerance, is carried out through the alteration of genes or metabolic routes within an organism’s genome. In terms of hydrogen production, the desired result is enhanced hydrogen yields through concurrent production of organic acids (e.g. acetate and butyric acid) or from upregulation of hydrogenase enzyme complexes [27, 32, 71]. In order to modify an organism's phenotype, understanding of essential genes and how interacting biochemical pathways realize specific phenotypic traits (e.g. acid tolerance, hydrogen production) is critical. While some pathways are directly involved in the expression of a phenotype, it is the interaction and communication of multiple pathways that ultimately influence phenotype expression. An example of this is seen in aerobic respiration. In this phenotype, the citric acid cycle (i.e., TCA cycle) is directly related to the expression of aerobic respiration. However, other pathways, such as those involved in synthesis of Acetyl-CoA, may be rate limiting and therefore regulate TCA [8]. As such, identification of genotype-phenotype associations across an organism's entire metabolic system is necessary for engineers to improve upon phenotype expression. Unfortunately, a number of experimental and computational approaches that can be used to identify such associations generally rely upon evaluation of enzymes within individual pathways rather than across entire metabolic networks.

3.2 Aims and Contributions of Research

The work presented in this Chapter enables the identification of sets of genes related to expression of microbial phenotypes important for biological hydrogen production. Identification of phenotype-related genes is the first step necessary for designing an overall integrative systems biology approach that will span three different "omics" levels–genomics, metabolomics, and transcriptomics (Figure 3.2). Through linking genes to microbial traits, the foundation for designing a genetically modified or engineered microorganism can be set. In addition, sets of phenotype-related

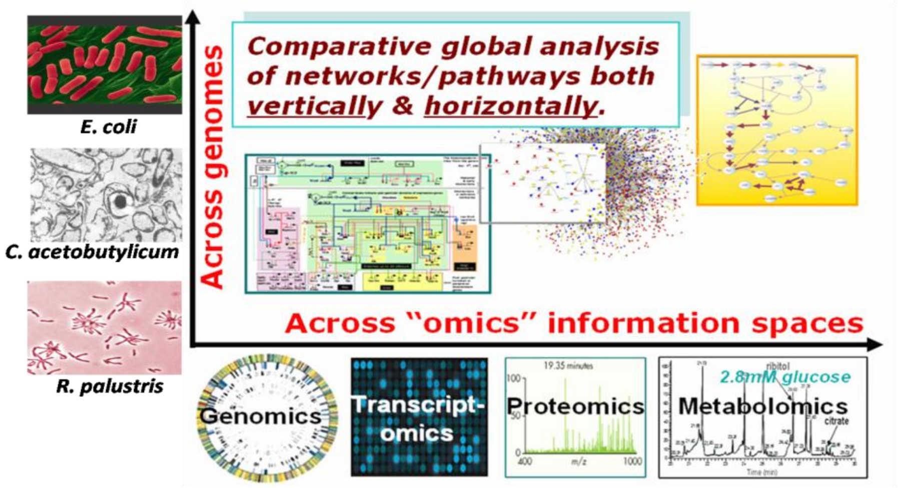


Figure 3.2 General overview of the integrative systems approach used in this research to evaluate hydrogen-producing microorganisms. The proposed approach allows for identification of metabolic genes and pathways related to specific phenotypes across genomes and across different information spaces or “omics” levels (left)

genes can be used as “knowledge priors” to identify potential functional relationships between proteins in groups of related microorganisms (Chapter 5).

The scope of this work is tailored to three phenotypes: (1) hydrogen production, (2) dark fermentative, hydrogen production (e.g. via acetate and butyrate production), and (3) acid tolerance. This research is comprised of the two specific aims. The first aim is to enable the discovery of phenotype-related genes through the development of a large-scale statistical approach. Genes identified in the first aim are used for: (1) quick analysis to identify potentially important phenotype-related genes in phenotype expressing microorganisms, (2) discovery of phenotype-related metabolic pathways (Chapter 4), and (3) identification of functional association and potential cross-talks between metabolic networks and other functional modules (Chapter 5).

The second aim is identification of phenotype-related genes through co-development of (1) a network-based, small-scale computational approach based on utilization of knowledge priors with (2) a high-throughput, metabolic genome-scale network analysis. Application of the network-based approach was used to identify small sets of functionally associated genes for individual organisms expressing a given phenotype. Identification of genes for specific microorganisms provides information and further understanding of potentially important associations between genes (e.g., regulation of expression). The high-throughput method was used to generate seed sets of enzymes, which were then expanded upon in Chapter 4 to discover phenotype-specific metabolic pathways.

3.3 Results

In this section, predicted phenotype-related enzymes obtained using both the statistical approach and NIBBS algorithm are presented. For each method a set of experiments was conducted to predict phenotype-related enzymes and to assess the ability of each method in predicting the enzymes. In each experiment a set of phenotype expressing microorganisms was selected as our positive instance and set of non-phenotype expressing microorganisms as our negative instance. A list of all organisms used in the experiments is provided in Tables B.1 – B.3 of Appendix B.

Predicted enzymes were reviewed to determine whether the enzymes were part of a system known to be related to phenotype-expressing organism. Descriptions of the predicted phenotype-related enzymes for each of the three phenotypes are presented in the following sections. Also, enzymes predicted were also assessed to determine if the NIBBS algorithm could identify enzymes that are not specific to phenotype expressing organisms but are required for pathway expression. Descriptions of these findings are presented for the example pathways TCA versus rTCA. A complete summary of the enzymes identified using the T-test and NIBBS algorithm for each phenotype are provided in Appendix C.

In addition to predicting phenotype-related enzymes, a comparative analysis of the two methods was conducted to demonstrate the potential for the NIBBS algorithm to incorporate and identify *knowledge priors.* In this study, *knowledge priors* represent phenotype related enzymes that are statistically biased towards a given set of phenotype expressing microorganisms. Phenotype related enzymes were identified based on statistical analyses (2-sample T-test; p-value 0.05) of the ortholog distribution in phenotype and non-phenotype expressing organisms. Phenotype related enzymes identified from the T-test represent *knowledge priors,* which can be incorporated into the NIBBS algorithm.

3.3.1 Validation of Methodology Using Known Aerobic and Anaerobic Organisms

Two experiments were performed, one for aerobic respiration phenotype, and the other for anaerobic respiration phenotype, in order to assess the ability of both approaches to identify phenotype-related enzymes. Using 36 aerobic organisms and 36 anaerobic organisms, analysis of the NIBBS enzymes ($\varphi < 0:05$) shows 86% and 75% recall, respectively when one or the other are used as positive instances. The results showed that NIBBS enzymes for aerobic respiration contained 261 enzymes and for anaerobic respiration contained 93 enzymes, while the Student's T-test identified 131 enzymes for aerobic respiration and 64 enzymes for anaerobic respiration.

Examination of the enzymes found by the Student's T-test but missed by NIBBS-Search shows that they are typically present in most of the phenotype-expressing and non-expressing organisms. The reason some enzymes are identified as phenotype-related by the statistical analysis is due to the fact that they typically have a higher copy number in phenotype-expressing organisms. Since NIBBS-Search uses binary data (i.e., whether at least one copy of the enzyme is present in the organism), these enzymes are not identified by NIBBS-Search as biased. In addition, because the NIBBS algorithm does not rely on the enzyme distribution across entire sets of organisms, it is capable of identifying subgroups of organisms among the list of given species. As such, it is not expected that NIBBS contain identical sets of enzymes as those identified with the Student's T-test approach.

3.3.1.1 Enzymes Predicted by NIBBS for Validation Experiment

Evaluation of the phenotype-related enzymes identified for aerobic organisms show the NIBBS algorithm was able to discover a small set of known enzymes associated with pathways commonly associated with aerobic and autotrophic carbon fixation. In Table 3.1, enzymes identified as aerobic related contained enzymes which make up components of the TCA cycle and the glyoxylate bypass. Other enzymes identified as phenotype-related are present due to phenotype associations with sub-groups of organisms in our dataset. These include organisms with similar fatty acid metabolism, amino acid metabolism, and photosynthetic organisms. Enzymes predicted as related to anaerobic organisms included 2-oxoglutarate synthase and ATP-dependent citrate lyase, which are related to the reductive TCA (rTCA) cycle (Table 3.1). The enzyme results associated with the anaerobic organisms is counter intuitive since rTCA is an autotrophic carbon fixation pathway and not associated with the phenotype anaerobic. The finding of rTCA related enzymes are likely related to a subset of organisms or subphenotypes present in the dataset.

Table 3.1 Known aerobic related enzymes that make up the TCA cycle and the glyoxylate bypass that are present (+) or absent (-) in the data set identified by the NIBBS algorithm and T-test approach.

EC Number	Enzyme Name	Aerobic Pathway	NIBBS	T-test
2.3.3.1	citrate (Si)-synthase	TCA, glyoxylate bypass	+	+
1.2.4.2	oxoglutarate dehydrogenase (succinyl-transferring)	TCA	+	+
1.3.99.1	succinate dehydrogenase	TCA	+	+
1.1.1.37	malate dehydrogenase	TCA, glyoxylate bypass	+	+
4.1.3.1	isocitrate lyase	glyoxylate bypass	+	+
2.3.3.9	malate synthase	glyoxylate bypass	+	+
6.2.1.5	succinate---CoA ligase (ADP-forming)	TCA	+	-
4.2.1.2	fumarate hydratase	TCA	+	-
1.1.1.42	isocitrate dehydrogenase (NADP+)	TCA	+	-
4.2.1.3	aconitate hydratase	TCA, glyoxylate bypass	+	-

3.3.2 Hydrogen Related Enzymes

To identify enzymes related to biological hydrogen production, two sets of experiments were performed using the hydrogen producing phenotype and the dark fermentative hydrogen producing phenotype. Similar to the anaerobic experiment, enzymes identified were evaluated to assess the ability of the NIBBS algorithm to predict phenotype-related enzymes.

To determine the enzymes involved in the overall metabolic processes involved in hydrogen production, 14 hydrogen producing and 11 non-hydrogen producing organisms were selected. Hydrogen producing organisms were representative of the three methods for biological hydrogen production—bio-photolysis, light fermentation, and dark fermentation. Enzymes involved in dark fermentative hydrogen producing reactions were identified using 8 dark fermentative microorganisms and 11 non-hydrogen producing organisms. Additional experiments were run for the phenotypes light fermentation and bio-photolysis, however, these results are not described below since the organisms representative of these methods are not the focus of this study. Results for these experiments are presented in Appendix C.

Analysis of the NIBBS enzymes compared to the statistical enzymes ($\varphi < 0:05$) shows that NIBBS was able to identify or recall 65% hydrogen producing enzymes identified using the Student's T-test (see Appendix C). When a subset of 8 dark fermentative hydrogen producing microorganisms were used as a positive instance, 50% NIBBS enzymes ($\varphi < 0:05$) were recalled. The NIBBS enzymes for hydrogen production utilizing all three hydrogen methods contained 363 enzymes compared to dark fermentative hydrogen production which contained 206 enzymes. The Student's T-test identified 52 enzymes for hydrogen producing and 33 enzymes for dark fermentation hydrogen producing.

Similar to aerobic and anaerobic respiration, a small set of enzymes were predicted by the Student's T-test but missed by NIBBS for the phenotype dark fermentation. Examination of these enzymes shows that they are generally present in both phenotype and non-phenotype expressing organisms. As with anaerobic respiration, these enzymes contained a higher copy number in phenotype-expressing organisms. As such, they were identified as phenotype-related.

3.3.2.1 Hydrogen-Related Enzymes

Using *Clostridium acetobutylicum* as a model organism for dark fermenting hydrogen producers, the key metabolic pathways for hydrogen production, shown in Figure 3.3, were examined for the presence or absence of enzymes involved in each pathway. Analysis was conducted using predicted enzymes by the NIBBS method using the seed set generation process and the knowledge priors provided by the T-test. The two

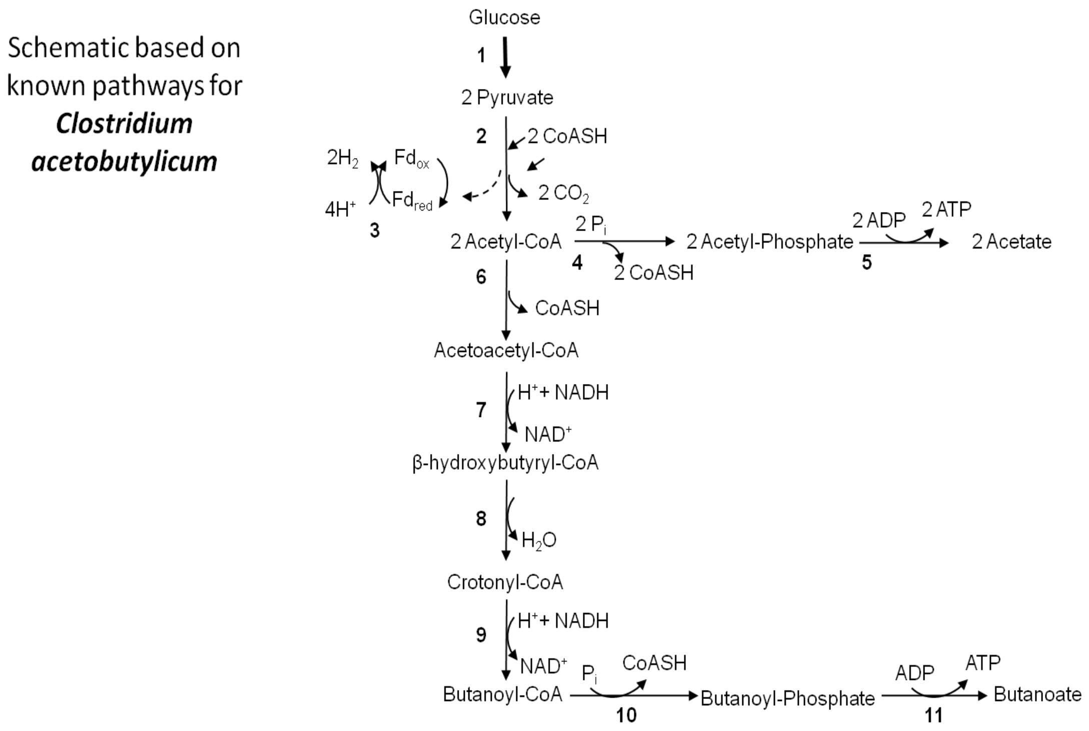
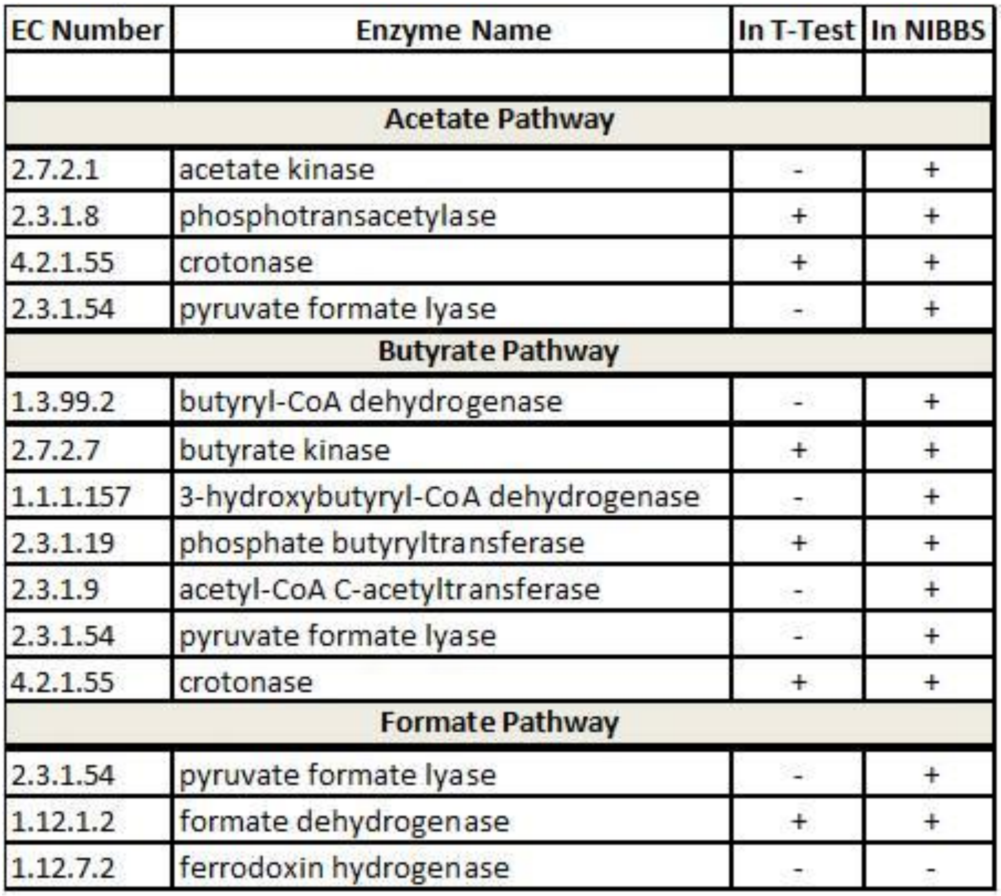


Figure 3.3 Schematic of key metabolic pathways for hydrogen production in *Clostridium acetobutylicum***.** Arrows with larger width indicate a series of reactions. Arrows with narrow width indicate individual reactions. Enzymes: 1, glycolytic enzymes; 2, pyruvate ferredoxin oxidoreductase (E.C. 1.2.7.1); 3, hydrogenase (E.C.1.12.7.2); 4, phosphotransacetylase (E.C. 2.3.1.8); 5, acetate kinase (E.C. 2.7.2.1); 6, acetyl-CoA acetyltransferase (thiolase) (E.C. 2.3.1.9); 7, β-hydroxybutyryl-CoA dehydrogenase (E.C. 1.1.1.157); 8, crotonase (E.C. 4.2.1.55); 9, butyryl-CoA dehydrogenase (E.C. 1.3.99.2); 10, phosphotransbutyrylase (E.C.2.3.1.19); 11, butyrate kinase (E.C. 2.7.2.7). Abbreviations: Ferredoxin (Fd); Coenzyme A (CoASH)

pathways, acetate and butanoate (i.e., butyrate), were selected as specific pathways for hydrogen production based on their potential hydrogen yield (see Chapter 2).

Table 3.2 shows that within the acetate pathway, NIBBS identified all of the constituent enzymes, pyruvate formate lyase (E.C. 2.3.1.54), acetate kinase (E.C. 2.7.2.1), and phosphotransacetylase (E.C. 2.3.1.8), as present within *C. acetobutylicum*. Whereas, the T-test only identified E.C. 2.3.1.8, all seven enzymes active in the butyrate pathway were found by the NIBBS method. The component enzymes for this pathway

Table 3.2 The presence (+) or absence (-) of enzymes in *Clostridium acetobutylicum*, in the dark fermentative, hydrogen producing experiments.



EC Number	Enzyme Name	In T-Test	In NIBBS
Acetate Pathway			
2.7.2.1	acetate kinase	-	+
2.3.1.8	phosphotransacetylase	+	+
4.2.1.55	crotonase	+	+
2.3.1.54	pyruvate formate lyase	-	+
Butyrate Pathway			
1.3.99.2	butyryl-CoA dehydrogenase	-	+
2.7.2.7	butyrate kinase	+	+
1.1.1.157	3-hydroxybutyryl-CoA dehydrogenase	-	+
2.3.1.19	phosphate butyryltransferase	+	+
2.3.1.9	acetyl-CoA C-acetyltransferase	-	+
2.3.1.54	pyruvate formate lyase	-	+
4.2.1.55	crotonase	+	+
Formate Pathway			
2.3.1.54	pyruvate formate lyase	-	+
1.12.1.2	formate dehydrogenase	+	+
1.12.7.2	ferrodoxin hydrogenase	-	-

are butyryl-CoA dehydrogenase (E.C. 1.3.99.2), phosphate butyryltransferase (E.C. 2.3.1.19), butyrate kinase (E.C. 2.7.2.7), 3-hydroxybutyryl-CoA dehydrogenase (E.C. 1.1.1.157), acetyl-CoA C-acetyltransferase (E.C. 2.3.1.9), pyruvate formate lyase (E.C. 2.3.1.54) and crotonase (E.C. 4.2.1.55). Among these, only three were found by the T-test.

In addition to the above pathways, the formate pathway was also reviewed. A general overview of formate production is shown in Figure 3.4. While it is not reported in the literature that *C. acetobutylicum* utilizes a formate pathway, it is possible that *C. acetobutylicum* may contain genes encoding some enzymes necessary for formate

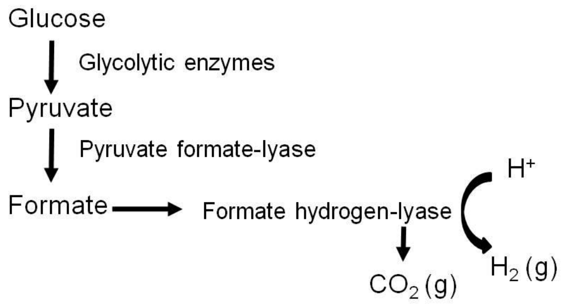


Figure 3.4 General overview of hydrogen production through the formate pathway

production. Of the three key enzymes described in Figure 3.4, NIBBS was able to identify only two of them. These are pyruvate formate lyase (E.C. 2.3.1.54) and formate dehydrogenase (E.C. 1.12.1.2). The second enzyme, along with formate dehydrogenase, forms the formate hydrogen lyase complex called ferredoxin hydrogenase (E.C. 1.12.7.2) [71]. This enzyme is common in many organisms and is not a phenotype specific toward dark fermentation.

Other enzymes identified using the NIBBS algorithm, include those involved in glycolysis and nitrogen fixation. In this study, a large number of enzymes involved in glycolysis were predicted as conserved across hydrogen producing organisms, but not conserved across non-hydrogen producing organisms [40]. This is mostly a result of the ability of the dark fermentative organisms to utilize organic compounds, such as glucose, for their carbon source. In terms of hydrogen production, glycolysis is a preliminary step needed for acetate or butyrate production as was depicted previously in Figure 3.3. In addition, glycolysis provides the energy sources necessary for biological hydrogen production to occur.

The predicted enzyme associated with both nitrogen fixation and hydrogen is nitrogenase. While this enzyme is typically present in photosynthetic organisms such as

Rhodopsuedomonas palustris, it is also present in *C. acetotbutylicum* [72]. Nitrogenase is responsible for conversion of ammonia during nitrogen-fixation. It is during this process that H_2 (g) is produced as a by-product [6, 72]. In *C. acetobutylicum*, nitrogen-fixation occurs as a response to nitrogen availability. When ammonium concentration or nitrogenous compounds are abundant, nitrogenase activities will decrease [72]. Due to the regulation of nitrogenase by nitrogenous compounds, nitrogenase is not a main source of hydrogen production in organisms such as *C. acetobutylicum.*

3.3.3 Phenotype-Related Enzymes: Acid-Tolerant

In order to predict the enzymes related to a microorganism's ability to tolerate low pH conditions, 10 acid-tolerant organisms and 8 alkaliphiles were evaluated using the Student's T-test and NIBBS algorithm. Analysis of the NIBBS enzymes ($\varphi < 0.05$) shows 73% of the acid-tolerant enzymes were recalled when acid-tolerant organisms were used as positive instance (see Appendix C). NIBBS enzymes predicted 164 enzymes, while the Student's T-test identified only 17 as phenotype-related. Enzymes identified by the Student's T-test and missed by NIBBS included enzymes involved in central metabolism, amino acid metabolism, and lactic acid metabolism.

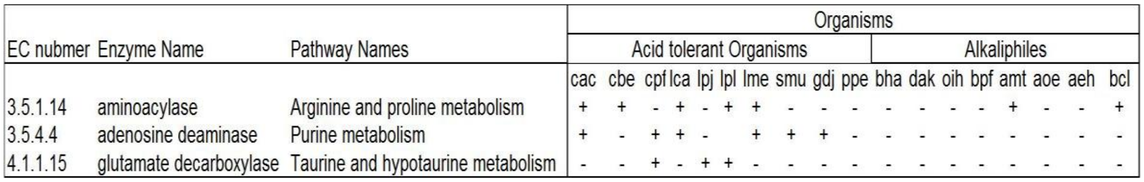
3.3.3.1 Acid-Tolerant Enzymes Predicted by NIBBS

To identify acid-tolerant enzymes, *C. acetobutylicum* was used as our model organism. In many fermentative, hydrogen producing experiments and in natural systems, acetogenic *Clostridium* species are often present. Review of the literature indicated that *C. acetobutylicum* and many other hydrogen producing species can tolerate and maintain hydrogen production in acidic pH ranging from 4.5-6 [73]. To survive, these organisms have developed metabolic and cellular acid tolerance response (ATR) systems to protect themselves when exposed to acid environments [66]. While a few acid-tolerant or resistant systems have been described in organisms such as Lactobacilli, little is known about metabolic pathways involved in acid tolerance, particularly in *Clostridium* species.

Analysis of the predicted enzymes for *C. acetobutylicum* did not reveal a distinct acid resistant metabolic system. However, review of the predicted enzymes across other hydrogen producers revealed the potential for an acid resistant system. Identified enzymes included glutamate decarboxylase (E.C. 4.1.1.15; Gad), a known enzyme involved in acid-resistance in some microorganisms, including *Clostridium perfringens,* a known hydrogen producer (Table 3.3). In *Escherichia coli, C. perfringens*, and some *Lactobacilli* the internal pH can be neutralized by a decarboxylase system—glutamate and arginine decarboxylase [66, 74, 75]. In *Lactobacilli,* glutamate decarboxylase converts glutamate to γ-amino butyric acid (GABA), which is quickly removed and replaced by another glutamate molecule [66]. While glutamate decarboxylase plays a vital role in this decarboxylase system, other proteins and antiporters are required for neutralization of the internal pH to occur.

Table 3.3 shows that glutamate decarboxylase was only present in 3 of our 10 acid-tolerant organisms. They are *Lactobacillus plantarum* JDM1, *Lactobacillus plantarum* WCFS1, and *Clostridium perfringens* ATCC 13124. Prediction of glutamate decarboxylase by NIBBS was due to the presence of the enzyme in a small subset of organisms within the dataset and the absence of the enzyme in non-phenotype expressing dataset. Based on the absence of glutamate decarboxylase in many of the organisms, including hydrogen producing *C. acetobutylicum* and *C. beijerinckii*, glutamate decarboxylase can be classified as not specific for, but rather related to acid tolerance.

Table 3.3 Comparison between the presence (+) or absence (-) of enzymes in acid-tolerant organisms and alkaliphilic (non-acid-tolerant organisms), in acid-tolerant experiments. Each row represents enzymes identified by NIBBS and their corresponding pathways. Acid-tolerant organisms: *Clostridium acetobutylicum* (cac), *Clostridium beijerinckii* (cbe), *Clostridium perfringens* (cpf), *Lactobacillus casei* (lca), *Lactobacillus plantarum* JDM1 (lpj), *Lactobacillus plantarum* WCFS1 (lpl), *Streptococcus mutans* (smu), *Gluconacetobacter diazotrophicus* (gdj), *Pediococcus pentosaceus* (ppe). Non-acid-tolerant (alkaliphiles) included: *Bacillus* halodurans (bha), *Desulfurivibrio alkaliphilus* (dak), *Ocenobacillus iheyensis* (oih), *Bacillus pseudofirmus* (bpf), *Alkaliphilus metalliredigens* (amt), *Alkaliphlius oremlandii* (aoe), *Alkalilimnicola ehrlichei* (aeh), *Bacillus clausii* (bcl)



			Organisms																	
			Acid tolerant Organisms										Alkaliphiles							
EC nubmer	Enzyme Name	Pathway Names	cac	cbe	cpf	lca	lpj	lpl	lme	smu	gdj	ppe	bha	dak	oih	bpf	amt	aoe	aeh	bcl
3.5.1.14	aminoacylase	Arginine and proline metabolism	+	+	-	+	-	+	+	-	-	-	-	-	-	-	+	-	-	+
3.5.4.4	adenosine deaminase	Purine metabolism	+	-	+	+	-		+	+	+	-	-	-	-	-	-	-	-	-
4.1.1.15	glutamate decarboxylase	Taurine and hypotaurine metabolism	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-

The presence in *C. perfringens* and the absence within other Clostridium species does not necessarily indicate *C. acetobutylicum* is incapable of similar mechanisms. In fact, incorporation of a decarboxylase system similar to that of *C. perfringens* and *L. plantarum* into hydrogen producers, such as *C. acetobutylicum*, may be necessary to maintain hydrogen production and acidogenesis.

Table 3.3 also shows that enzymes predicted as related to acid tolerance in the organism *C. acetobutylicum* included those involved in purine metabolism and arginine/proline metabolism. This is shown in the presence (+) of the enzyme adenosine deaminase (E.C. 3.5.4.4). In fact, this enzyme was also present in 6 of our 10 acid-tolerant organisms and absent in our non-phenotype expressing bacteria. The absence of this enzyme (-) in alkaliphilic bacteria suggests that adenosine deaminase is either inhibited at high pH values or is specific toward acidophilic bacteria. In organisms, adenosine deaminase, an enzyme of the purine salvage pathway, is responsible for deaminating adenosine to form inosine [76]. As a member of the purine salvage pathway, adenosine deaminase plays an important role in recycling intermediates (e.g. guanosine, adenosine) generated during degradation of DNA and RNA [76]. This is particularly important in respect to degradation of DNA due to acidic conditions [75]. In organisms that are adapted to acidic environments, such as extreme acidophiles, degradation of DNA may not be a concern. In acid tolerant organisms, degradation of DNA and RNA may or may not be a concern depending on whether the organism is adapted to a particular environment and then suddenly exposed to a pH change. The presence of enzymes associated with purine and arginine metabolism are uncertain. Further investigation of these metabolic pathways through experimental analysis is necessary to determine if they play a role in acid response.

Another enzyme predicted by the NIBBS-search algorithm is aminoacylase or N-acylamino acid amido hydralase (E.C. 3.5.1.14). Aminoacylases are responsible for the hydrolysis of L-acyl-amino acids to a corresponding amino acid [77]. In hydrogen producing organisms exposed to acidic conditions, production of amino acids such as glutamate could potentially be beneficial for regulation of intracellular pH. However, the exact role for this enzyme in expression of acid-tolerance in hydrogen producing organisms such as *C. acetobutylicum* is unclear. To fully understand the role of aminoacylases in acid-tolerance, molecular studies evaluating the role of aminoacylases present in organisms exposed to acidic environments or various organic acids are necessary.

3.4 Discussion

One of the advantages of the NIBBS algorithm is its ability to utilize entire metabolic networks for identification of conserved metabolic sub-paths in phenotype expressing microorganisms. In each network, metabolic reactions and enzymes that catalyze each reaction are represented and utilized by the algorithm to search through all potential pathways that may be conserved in phenotype expressing microorganisms. This is particularly important since microbial metabolic reactions may utilize more than one pathway to generate a compound or metabolite necessary to carry out other metabolic or cellular functions.

Examples of this include the metabolic routes for production of the biosynthetic precursor 2-oxoglutarate (or alpha ketoglutarate). One of the most common routes for production of 2-oxoglutarate in aerobic organisms is the two-step TCA reaction catalyzed by isocitrate dehydrogenase (E.C. 1.1.1.42). In this reaction isocitrate dehydrogenase catalyzes the oxidation of isocitrate to oxalosuccinate and followed by the decarboxylation of oxalosuccinate to form 2-oxoglutarate [8]. A second reaction leading to the formation of 2-oxoglutarate utilizes isocitrate dehydrogenase (NAD^+; E.C. 1.1.1.41) in a one-step reaction to convert isocitrate to 2-oxoglutrate dehydrogenase.

Because the NIBBS algorithm considers metabolic compounds, it has the capability to identify reactions for both TCA and rTCA. Previously, Table 3.1 for the TCA subsystem data showed the NIBBS algorithm was able to identify reactions for these two pathways. However, if the NIBBS algorithm relied solely on metabolic enzymes to identify conserved pathways, potentially important metabolic pathways may be missed. For example, if we studied aerobic respiration and included the enzyme, isocitrate dehydrogenase (E.C. 1.1.1.42), the algorithm would only provide information on reactions involving this enzyme. As such, potentially important metabolic predictions may be overlooked.

3.4.1 Additional Applications for NIBBS Algorithm

The NIBBS employs a two step process to identify phenotype related systems. The first step is identification of seed sets or *knowledge priors*. Identification of these seed enzymes is necessary because most phenotypes are differentiated by only a few enzymes. For example, the TCA and rTCA phenotypes differ only in a few edge labels. Thus identifying these edges reduces our search space in terms of identifying the set of target organisms that could possibly contain the subsystem, and reduce the number of organism networks that have to be aligned. It also reduces the number of computations the algorithm has to perform because the sub-graphs the algorithm has to enumerate are only those that contain some part of the seed set. Using the seed sets generated, these enzymes can then be used to identify phenotype-related pathways (Chapter 4) and phenotype-related networks (Chapter 5).

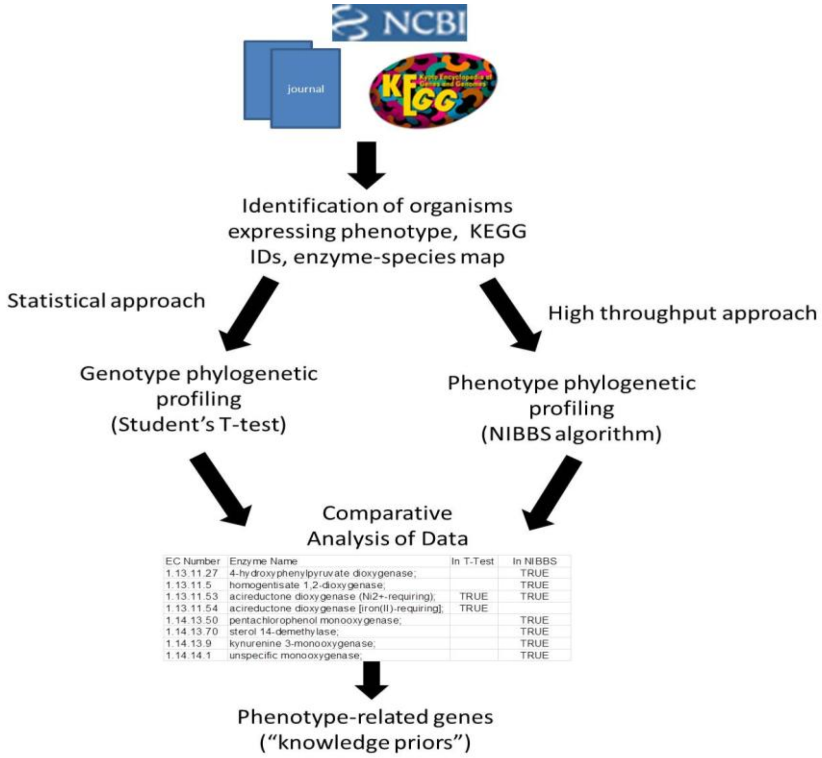
3.4.2 Limitation on Acid-Tolerant Results

To identify phenotype-related genes associated with an important phenotype for hydrogen production, sets of acid-tolerant and non-acid-tolerant (or alkaliphiles) were evaluated and the results reported in this Chapter. The data reported is only representative of a small group of acid-tolerant bacteria consisting of 9 Firmicutes and 1 Proteobacteria. As such, results obtained are biased towards acid-tolerant Firmicutes, rather than the phenotype acid-tolerant. While a large diverse dataset is ideal, the number of completely sequenced organisms representative of this phenotype present in the NCBI and KEGG database was limited at the time of study.

3.5 Approach

To achieve these aims, two different complementary techniques were used in parallel to generate a comprehensive set of phenotype-related genes related to the three target phenotypes—hydrogen production via dark fermentation, hydrogen production (all three methods), and acid tolerance. The first technique is a statistical approach that relies on using genotype phylogenetic profiling (GPP) to identify phenotype-related genes. The second method, Network Instance-Based Biased Subgraph search (NIBBS) algorithm, is a new heuristic algorithm that utilizes phylogenetic profiling (PPP) to identify gene sets in phenotype expressing and non-phenotype expressing microorganisms.

To identify phenotype-related genes, the NIBBS algorithm capitalizes on high-throughput comparative analysis of multiple genome-scale metabolic networks. For both methods, classification of genes as phenotype-related is based on the assumption that for a given phenotypic trait, there is a subset of key genes related to this trait that are evolutionarily conserved across multiple genomes with the same trait. As such, if a gene is evolutionarily conserved across a group of related microorganisms exhibiting the same phenotypic trait, and not conserved in a set of non-phenotype expressing microorganisms, then the gene is likely related to the phenotype. An overview of the three approaches is provided in Figure 3.5 and a brief introduction to each method is provided below. A detailed description for each method is provided in Chapter 6 of this dissertation.


NCBI
journal
KEGG
Identification of organisms expressing phenotype, KEGG IDs, enzyme-species map
Statistical approach
High throughput approach
Genotype phylogenetic profiling (Student's T-test)
Phenotype phylogenetic profiling (NIBBS algorithm)
Comparative Analysis of Data
Phenotype-related genes ("knowledge priors")

EC Number	Enzyme Name	In T-Test	In NIBBS
1.13.11.27	4-hydroxyphenylpyruvate dioxygenase;		TRUE
1.13.11.5	homogentisate 1,2-dioxygenase;		TRUE
1.13.11.53	acireductone dioxygenase (Ni2+-requiring);	TRUE	TRUE
1.13.11.54	acireductone dioxygenase [iron(II)-requiring];	TRUE	
1.14.13.50	pentachlorophenol monooxygenase;		TRUE
1.14.13.70	sterol 14-demethylase;		TRUE
1.14.13.9	kynurenine 3-monooxygenase;		TRUE
1.14.14.1	unspecific monooxygenase;		TRUE

Figure 3.5 Overview of the two approaches for identification of phenotype-related genes

3.5.1 Genotype Phylogenetic Profiling Using Student's T-Test

One approach to predicting which genes are associated with particular phenotypes, is through statistical analysis using genotype phylogenetic profiling (GPP). Phylogenetic profiling is a concept that was initially used to functionally annotate related genes [78, 79], but has more recently been applied toward identifying phenotype-related genes [78]. In GPP, the profile is based upon the inclusion or exclusion of a gene in an organism's genome. Using this technique, one can correlate sets of orthologous genes present in phylogenetic profiles with phenotype-expressing organisms.

To identify phenotype-related genes, genotype phylogenetic profiling was used in development of a statistical approach, for identification of a core set of genes related to the target phenotypes as shown in Figure 3.6. Using GPP, the statistical enrichment of

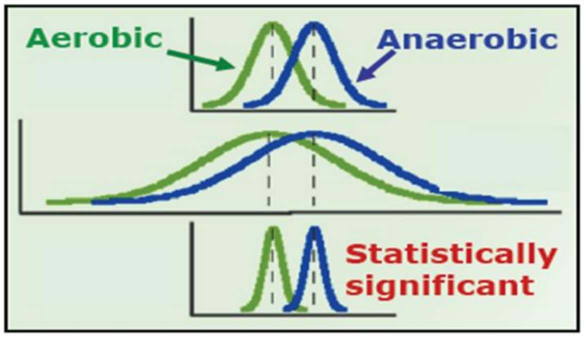


Figure 3.6 Schematic of the approach to assess statistical significance of gene enrichment.

orthologs present in microorganisms expressing the target phenotype and microorganisms that do not express the target phenotype is calculated. Based on ortholog distribution in the GPP, a Student’s T-test is used to determine which orthologs are related to the target phenotype. Application of GPPs to this research allows for quick and robust statistical analysis and prediction of sets of genes related to specific phenotypes.

3.5.2 NIBBS Algorithm: High Throughput Comparative Analysis of Multiple Genome-Scale Metabolic Networks

The NIBBS approach is a high throughput computational approach that utilizes two algorithms for identification of conserved sets of genes in phenotype-expressing organisms. Unlike the previous approach, which evaluates all genes present in each species, this method is complex and limited to identification of genes in metabolic pathways. Here, a phenotype-driven comparative analysis of genome-scale metabolic networks from the KEGG is used to predict phenotype-related genes across phylogenetic profiles of phenotype expressing and non-phenotype expressing organisms.

Development and application of comparative techniques, such as NIBBS, for the discovery of phenotype-related orthologs allows for (1) expansion of previously identified sets of genes shown in literature, and (2) identification of potential COG groups related to each phenotype. Together this will result in the creation of a comprehensive database of orthologs, which can be used in further metabolic studies (e.g., functional association of proteins).

CHAPTER 4: IDENTIFICATION OF PHENOTYPE-SPECIFIC METABOLIC PATHWAYS: APPLICATION TO HYDROGEN PRODUCTION

4.1 Motivation

Production of biological hydrogen (bio-hydrogen) is potentially a feasible technology for generation of alternative energy and fuels. The ability of several naturally occurring microorganisms to generate hydrogen using various metabolic processes, such as dark fermentative metabolic routes, makes bio-hydrogen production a favorable option for addressing bioenergy needs [22]. Unfortunately, application of bio-hydrogen technologies is often limited due to overall low hydrogen production and yields generated by hydrogen-producing microorganisms [22, 40]. This is primarily due to the flow of carbon toward cellular growth rather than toward pathways involved in the production of hydrogen [22, 32]. To help overcome this problem, one approach currently under consideration is bioengineering the direction of microbial metabolic pathways away from biomass production, and toward the expression of phenotypic traits related to hydrogen production [40].

Although the concept of metabolically engineering microbial organisms is not new, the application of engineering organisms for enhanced biohydrogen is still in its early stages. This is primarily because of the time and cost necessary to conduct extensive biochemical studies in order to fully understand metabolic networks involved in expression of microbial traits [31]. An overview of steps taken in engineering studies is presented in Figure 4.1.

Examples of biochemical studies currently being conducted was presented in Chapter 1. From these studies, it is observed that only a small number of metabolic and cellular networks in bacteria have been extensively evaluated in model organisms. Similarly, with metabolic engineering studies, most work has focused on the regulation of hydrogenase and accessory genes in model species, such as *Clostridium* and *Escherichia coli* [39, 52]. In some cases, hydrogenase operons have been created with the hope of inserting the new operon into a host organism, thus creating an ideal hydrogen producing organism [52]. In addition, a few studies have focused on constructing hydrogen producing pathways not traditionally present in the host organism in hope of increasing hydrogen yields [80]. One example of this is a study by Veit et al. [80] who created a ferredoxin-dependent NAD(P)H:H2 anaerobic pathway and inserted it into organisms that traditionally do not have this pathway. Unfortunately, Veit et al. [80] were not able to successfully enhance and maintain high yields of hydrogen production.

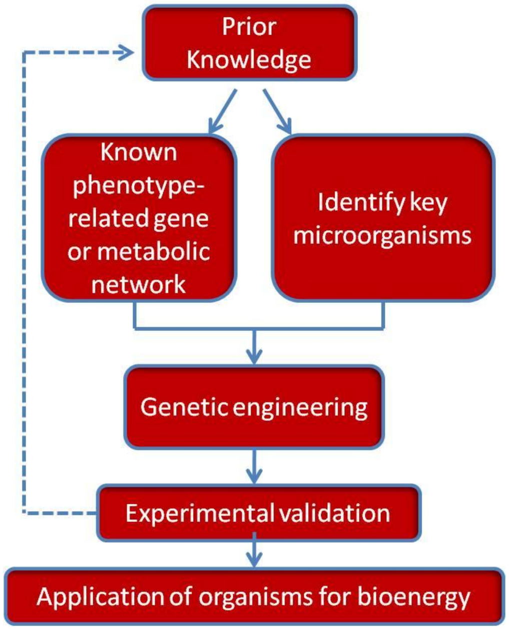

Prior Knowledge
Known phenotype-related gene or metabolic network
Identify key microorganisms
Genetic engineering
Experimental validation
Application of organisms for bioenergy

Figure 4.1 Flow chart demonstrating the overall process for genetic engineering studies

Due to the fact that most hydrogen related studies are focused on specific organisms rather than the phenotype at hand, information obtained to date is species specific rather than representative of the overall phenotype—hydrogen production.

Because most hydrogen producers are found in mixed communities, such as those found by communities that are fed complex waste materials, evaluation of phenotype-related pathways and sub-networks could be used to formulate hypotheses and improve the experimental design of organisms.

One approach to identifying phenotype-related metabolic pathways and processes is through the application of computational biology approaches. As demonstrated in Chapter 3, high throughput technologies can provide fast and accurate predictions regarding individual phenotype-related metabolic genes. Predictions such as these provide the underlying foundation toward understanding phenotype-related metabolic processes and eventually predicting other metabolic processes. However, methodologies that rely on comparative genome analysis and phylogenetic profiling still struggle to identify phenotype-related metabolic pathways. One reason is that the identification of phenotype-related metabolic pathways from entire metabolic networks is time intensive. In addition, if only a subset of known enzymes is studied, such as with Kastenmuller's approach [70], there is a potential that the accuracy of predictions could be compromised. For example, because many enzymes are present in a number of pathways, including non-phenotype expressing pathways, potentially important metabolic predictions may be overlooked if categorized as not biased towards phenotype expression. Accurate predictions of metabolic processes are particularly important when conducting comparative analysis across a group of organisms.

4.2 Research Aims and Contribution

The work presented in this Chapter allows for prediction of metabolic pathways and sub-networks related toward expression of phenotypes important for biological hydrogen production. Identification of phenotype-related pathways and sub-networks arise from co-development and application of two complementary computational approaches. The first approach utilizes the Network Instance-Based Biased Subgraph Search (NIBBS-Search) algorithm. Here, the NIBBS algorithm expands upon the seed set of enzymes identified in Chapter 3 to identify common subgraph networks. Through comparative analysis of multiple genome-scale metabolic networks, phenotype-type related pathways are predicted. The second methodology is a multiple alignment algorithm, which aligns metabolic pathways based on similarities among reactions, compounds, enzymes, and pathway topology. Information derived from both approaches was used to develop hypotheses on pathway requirements, such as potential pathway interactions and cross-talk necessary for phenotype expression. Description of each method is presented in Section 4.6 and in Chapter 6 of this dissertation.

The scope of this work was tailored to the following phenotypes: dark fermentative hydrogen production, and acid tolerance. The work conducted in this Chapter consists of two specific aims. The first aim is to enable identification of known and novel phenotype-related pathways through application of the metabolic genome scale network approach titled NIBBS algorithm. Through co-development of a high-throughput method, conserved pathways across entire metabolic maps can be identified. Evaluation of entire metabolic pathways allows for the discovery of potential interplay between pathways important for the expression of key phenotypes.

The second aim is to identify conserved metabolic sub-networks within metabolic pathways of phenotype-expressing organisms through application of a comparative analysis pathway tool. Identification of conserved sub-networks present in pathways known to be important in phenotype expression provides further understanding in the role pathways play in expression of target phenotypes.

4.3 Results

In this section, predicted phenotype-related pathways and sub-networks obtained using both the NIBBS search algorithm and the multiple alignment algorithm are presented. For each method, a set of experiments was conducted to (1) assess the ability of each approach to predict phenotype-related metabolic pathways and sub-networks; and (2) identify pathways both related to and specific to expression of the target phenotype.

4.3.1 High Throughput Analysis of Phenotype Related Pathways: NIBBS Algorithm

To identify pathways using the NIBBS search algorithm, sets of phenotype expressing and non-phenotype expressing organisms were selected and used as a positive and a negative instance, respectively. Similar to the phenotype-related gene studies, phenotypes representative of hydrogen production and acid tolerance were analyzed to predict pathways. In order to identify pathways representative of organisms in wastewater and hydrogen producing systems, only dark fermentative hydrogen producing organisms were evaluated. To validate the methodologies ability to differentiate between pathways, a comparison between organisms known to utilize the TCA pathway and rTCA pathway was established. A list of the organisms used in this study is presented in Appendix B.

4.3.1.1 TCA versus rTCA Pathway

Due to the ability of the NIBBS-Search algorithm to predict phenotype-related enzymes through the utilization of phenotype-related metabolic systems, the algorithm is capable of identifying phenotype-related enzymes that are not necessarily specific to phenotype-expressing organisms. To demonstrate this feature of NIBBS-Search, two experiments were conducted comparing the two well-characterized metabolic networks tricarboxylic acid (TCA) cycle and the reverse TCA (rTCA) cycle. Sets of organisms known to utilize the TCA and rTCA cycle were selected and analyzed. Selection of these two metabolic systems was based on the ability of these pathways to utilize the same set of metabolites and that they have common enzymes. An overview of the enzymes identified by NIBBS for TCA and rTCA is provided in Table 4.1.

Using sixteen organisms that utilize the TCA cycle and six organisms that utilize the rTCA cycle, the NIBBS algorithm was able to identify all but one TCA enzyme, malate dehydrogenase (EC 1.1.1.37), in the top ranking subsystems (φ-value of 0.002). Malate dehydrogenase is part of another system with a larger φ-value of 0.006, which also includes seven of the eight TCA enzymes (isocitrate dehydrogenase is not included). All eight of the TCA enzymes are, therefore, part of at least one statistically significant system (φ-value less than 0.05) identified in the TCA experiment. To ensure the sensitivity of the algorithm to identifying key enzymes characteristic for each pathway, we reviewed the results to determine if key rTCA were present in any of the positive instances. In this study, we did not identify any of the three key enzymes unique to rTCA. Thus, suggesting that the NIBBS algorithm was able to properly predict the TCA pathway for phenotype-expressing organisms.

Similar results were obtained in the rTCA experiment, when rTCA-utilizing organisms were used as positive instances. A top ranking system identified in the rTCA

Table 4.1 The presence (+) or absence (-) of enzymes in the organisms used in the TCA and rTCA experiments**.** rTCA organisms: *Chlorobaculum tepidum* (cte), *Chlorobium limicola* (cli), *Sulfurimonas denitrificans* (tdn), *Aquifex aeolicus* (aae), *Hydrogenobacter thermophilus* (hth), *Nautilia profundicola* (nam). *TCA organisms: Bordetella broniseptica* (bbr), *Staphylococcus saprophyticus* (ssp), *Myxococcus xanthus* (mxa), *Leptospira interrogans serovar lai* (lil), *Helicobacter pylori* (hpa), *Listeria innocua* (lin), *Escherichia coli* (eco), *Shewanella oneidensis* (son), *Anaplasma marginale St. Maries* (ama), *Bdellovibrio bacteriovorus* (bba), *Bordetella parapertussis* (bpa), *Bordetalla bronchiseptica* (bbr), *Geobacillus kaustophilus* (gka), *Legionella pneumophila Lens* (lpf), *Neisseria gonorrhoeae* (ngo), *Sinorhizobium meliloti* (sme).

		rTCA microorganisms						TCA microorganisms															
E.C. number	**Enzyme description**	**cte**	**cli**	**tdn**	**aae**	**hth**	**nam**	**bbr**	**ssp**	**mxa**	**lil**	**hpa**	**lin**	**eco**	**son**	**ama**	**bba**	**bpa**	**bbr**	**gka**	**lpf**	**ngo**	**sme**
Common Enzymes																							
1.1.1.37	malate dehydrogenase	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	-	+
4.2.1.3	aconitase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1.1.1.42	isocitrate dehydrogenase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
6.2.1.5	succinate thiokinase	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+
4.2.1.2	fumarase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
TCA specific enzymes																							
1.3.99.1	succinate dehydrogenase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1.2.4.2	alpha ketoglutarate dehydrogenase	-	+	-	-	-	-	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+
2.3.3.1	citrate (Si)-synthase	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rTCA specific enzymes																							
1.2.7.3	2-oxoglutarate synthase	+	+	+	-	+	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-
2.3.3.8	ATP citrate synthase	+	+	+	-		+	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-
1.3.1.6	fumarate reductase	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

experiment (φ-value = 0.002) contained seven of the eight rTCA enzymes, including all the five enzymes that the rTCA cycle shared with the TCA cycle (Table 4.1). The rTCA-related enzyme, fumarate reductase (EC 1.3.1.6) was not found in any system identified in the rTCA experiment.

In the rTCA experiment, systems identified by the NIBBS algorithm include two enzymes, citrate synthase (EC 2.3.3.1) and succinate dehydrogenase (EC 1.3.99.1) that are typically associated with the TCA pathway [30]. This is because these two enzymes are not only present in all of the rTCA expressing organisms in the experiment but also in most, if not all, of the TCA expressing organisms in the experiment. This makes them likely to be included in the set of expansion edges, since they do not decrease the φ-value of the system. The presence of these TCA-related enzymes in rTCA related systems may represent an additional functionality of these enzymes outside of their involvement in the TCA cycle. While these two enzymes are not required by rTCA expressing organisms to utilize the rTCA pathway, a possible hypothesis is that they serve other necessary functions in organisms that utilize the rTCA pathway. This hypothesis is based on the fact that these enzymes are present in almost all of the rTCA organisms. In comparison, a third key TCA enzyme, alpha ketoglutarate dehydrogenase (EC: 1.2.4.2), is not typically present in rTCA expressing organisms.

4.3.1.2 Pathways Related to Dark Fermentative Hydrogen Production

From analysis of the enzymes identified previously in Chapter 3, the NIBBS algorithm was able to identify the most relevant metabolic pathways for dark fermentative hydrogen production. While these pathways are important for hydrogen production, additional metabolic pathways present within organisms may also play an important role in impacting hydrogen yields.

Using NIBBS, the following pathways were identified as top ranking metabolic pathways significant (p-value = 0.05) for *C. acetobutylicum* with respect to dark fermentative hydrogen production. They are: fatty acid biosynthesis (KEGG pathway ID ec00061**),** purine metabolism (KEGG pathway ID ec00230), arginine and proline metabolism (KEGG pathway ID ec00330), and cysteine and methionine metabolism (KEGG pathway ID ec00270). An overview of these pathways and their relation to hydrogen production is presented in the following sections. A complete listing of the pathways with their rankings is presented in Table 4.2.

4.3.1.2.1 Fatty Acid Biosynthesis

Fatty acids are methylene carbon chains with a carboxyl group that are generally associated with formation of structural membranes and maintenance of the membrane's fluidity [8]. Within bacteria, fatty acids may be present in different forms such as branched, long chain, short chain fatty acids, volatile, or hydroxylated [8]. Formation or synthesis of fatty acids are generally initiated through the carboxylation of the acetyl-CoA [44]. In dark fermentative bacteria, such as *C. acetobutylicum,* acetyl-CoA is an important intermediary which leads to the formation of acetate, butyrate, solvents, and fatty acids. As such, redirection of metabolic pathways away from fatty acid formation and toward acidogenesis (e.g. acetate formation) is vital for enhanced hydrogen production.

Analysis of results showed fatty acid biosynthesis was the highest ranking metabolic pathway for *C. acetobutylicum* in both the phenotype and its sub-phenotype—

Table 4.2 List of top ranking pathways and their enrichment score for the phenotype dark fermentative hydrogen production.

Pathway ID	Pathway description	P-value
cac00061	Fatty acid biosynthesis	1.54E-33
cac00230	Purine metabolism	2.64E-17
cac00330	Arginine and proline metabolism	6.28E-12
cac00520	Amino sugar and nucleotide sugar metabolism	3.00E-11
cac00270	Cysteine and methionine metabolism	5.58E-11
cac00030	Pentose phosphate pathway	1.67E-09
cac00040	Pentose and glucuronate interconversions	1.48E-08
cac00400	Phenylalanine, tyrosine and tryptophan biosynthesis	4.26E-08
cac00051	Fructose and mannose metabolism	1.81E-07
cac00260	Glycine, serine and threonine metabolism	1.59E-06
cac00860	Porphyrin and chlorophyll metabolism	4.51E-06
cac00250	Alanine, aspartate and glutamate metabolism	1.13E-05
cac00920	Sulfur metabolism	2.15E-05
cac00500	Starch and sucrose metabolism	2.51E-05
cac00480	Glutathione metabolism	3.69E-05
cac00300	Lysine biosynthesis	0.000258
cac00910	Nitrogen metabolism	0.000831
cac00010	Glycolysis / Gluconeogenesis	0.001192
cac00052	Galactose metabolism	0.00143

hydrogen producing organisms and dark fermentative hydrogen producing organisms, respectively. The presence of this pathway in both categories suggests that fatty acid biosynthesis may play a key role in regulating metabolic routes for hydrogen formation, specifically in dark fermentation. Findings in this study are similar to previous reports on the role of fatty acids in acetate and butyrate formation. In a study by Huang et al. [73], the presence of short-chained fatty acids during acidogenesis were linked to initiation of solventogenesis to form butanol and acetone in fermenting bacteria [73]. This is a resultant of fatty acid build up within the cells. As the short chain fatty acids accumulate, bacterial cells form a transmembrane pH gradient leading to induction of solvent production [73].

While many times bioengineers focus on re-direction of electron and carbon flow in the key hydrogen producing pathways, the ability of hydrogen producers to respond to pH changes both internally and externally while maintaining hydrogen production is essential. The response to both formation and uptake of fatty acids present in waste materials could potentially act as a key regulator in metabolic shifts in *C. acetobutylicum* and potentially across other dark fermentative hydrogen producing bacteria.

4.3.1.2.2 Purine Metabolism

Purines are nucleotide bases which can be found either in free forms or attached to ribose 5-phosphate to form nucleotides and nucleic acids [8]. Organisms may synthesize purine nucleotides for use in the structural make up of nucleic acids or use in ATP metabolism [81]. During purine synthesis, amino acid donors are utilized to form purine rings and other purine structural components. Examples of amino acid donors include glutamine and aspartic acid [8]. In free form, purine nucleotide bases are harmful and toxic to the organisms. Therefore, they must be removed or transformed to a non-toxic compound. As such, many organisms have mechanisms to anaerobically degrade purine compounds through fermentation of xanthine into intermediates, which could potentially form acetate and formate [82]. One such organism capable of purine degradation is *Clostridium ljungdahlii* [83]. In *C. ljungdahlii,* purine compounds are degraded to form intermediates such as glycine and betaine. These intermediates in turn are reduced, resulting in acetate formation by the enzyme acetate kinase [83, 84].

Depending on the respiration requirement of the organisms (e.g. aerobic versus anaerobic) the degradation pathway used by microorganisms will vary. In this study dark fermentative hydrogen producers was selected. Within this phenotype, the facultative anaerobic phenotype and the anaerobic phenotype were included. Therefore, an extensive review of metabolic reactions is necessary to determine which degradation pathways, if any, are utilized. However, based on the high ranking of this pathway in our study for *C. acetobutylicum*, it is predicted that purine metabolism (degradation and synthesis) plays a minor role generation of acetate in dark fermentative bacteria.

4.3.1.2.3 Arginine and Proline Metabolism

L-Proline and L-arginine are two amino acids commonly found within both eukaryotic and prokaryotic organisms [85, 86]. In bacterial cells, L-proline is synthesized from L-glutamate by the enzyme glutamate kinase [86, 87]. In addition to biosynthesis of proline, some bacteria have been reported to take up and utilize proline as either a carbon or nitrogen source for metabolic growth [88]. In *Escherichia coli*, proline and proline betaine have been linked to increased osmotolerance and protection in cells [89]. Such protection would be beneficial in dark fermentation species for microbial response to induce water stress.

L-arginine is also an important precursor in nitrogen metabolism and protein synthesis in bacterial cells [85]. It can be metabolized by cells to produce other amino acids, including proline, or it can be utilized by the cell as either a carbon or nitrogen source. In addition, L-arginine may serve as an energy source for anaerobic bacteria. This is done through ATP production from L-arginine in the arginine deiminase pathway [85]. L-arginine biosynthesis occurs similar to L-proline in requiring L-glutamate as a precursor to biosynthesis. In this process, L-glutamate is deaminated through the enzyme glutamate dehydrogenase.

In this study, arginine and proline metabolism was identified as a potentially important pathway for *C. acetobutylicum* with respect to both hydrogen producing organisms and the sub-phenotype dark fermentative hydrogen production. In addition to identifying arginine and proline metabolism in an individual species, evaluation of hydrogen producing related enzymes shows this KEGG pathway as significant and likely related hydrogen production.

4.3.1.2.4 Cysteine and Methionine Metabolism

Methionine is a sulfur-containing amino acid which is used for biosynthesis of cysteine [90]. In general, most organisms can either take-up methionine or synthesize it to form other amino acids and help initiate protein synthesis [91]. Cysteine, another sulfur-containing amino acid important for the production of glutathione, is a compound that aids in protecting the cell from oxidative stress [91, 92]. In hydrogen producing organisms, cysteine ligands and residues play an important role in the structure of [Fe-S] clusters and hydrogenase enzymes [42, 47, 91]. Additionally, cysteine ligands aid in the binding of [Fe-S] clusters together within nitrogenase enzymes [93]. Nitrogenase enzymes are typically found in nitrogen fixing bacteria and are considered key enzymes to hydrogen production in light fermentative bacteria [39]. However, studies on nitrogen fixation have found that many dark fermentative species, such as Clostridium, are capable of utilizing nitrogenase enzymes [72]. However, in this study, we do not consider hydrogen production through nitrogenase a key metabolic route. This is mainly due to the energy expense needed for nitrogen-fixation by organisms such as *C. acetobutylicum.*

The role of cysteine and methionine in formation of [Fe-S] clusters for both hydrogenase and nitrogenase activity demonstrates the relationship of this cysteine and methionine metabolism in hydrogen producing organisms. From the NIBBS analysis, the cysteine and methionine KEGG pathway is predicted as a significant metabolic route in both *C. acetobutylicum* and in the set of organisms expressing the hydrogen producing phenotype (see Table 4.2). In addition to the NIBBS algorithm, the cysteine and methionine pathways were identified as a significant pathway using differing approaches, such as mutual information (Table 4.3). The presence of this pathway in mutual information supports our findings that cysteine and methionine are bio-hydrogen related pathways.

4.3.2 Acid-Tolerant Pathways

Metabolic pathways related to the expression of acid-tolerance vary across organisms and sub-sets of organisms, as shown by the analysis of phenotype-related enzymes (Chapter 3). This is particularly true between Gram negative and Gram positive organisms [66], which contain different response mechanisms for acid exposure. In this study, the acid-tolerant organisms selected consisted mainly of Gram positive, acid-tolerant bacteria from the phylum Firmicutes. As such, results reflect metabolic pathways present to a small group of bacteria capable of acid-tolerance rather than across a diverse set of organisms capable of expressing the phenotype acid-tolerant.

Using the NIBBS-search algorithm, 7 enriched pathways (p-value = 0.05) were identified. Of these pathways, the following metabolic pathways were predicted as top

Table 4.3 Comparison of pathway enrichment for three top ranking pathways based on the NIBBS and T-test enzymes, T-test enzymes only, and a hypergeometric test when using *C. acetobutylicum* as a model species.

Pathways	NIBBS + T-test enzymes (p-value =0.08)	T-test enzymes only (p-value 0.05)	Hypergeometric test in *C. acetobutylicum* (p-value
Cysteine and methionine metabolism	+	+	+
Purine metabolism	-	-	+
Arginine and proline metabolism	-	+	+

ranking with respect to acid-tolerance based on enzyme enrichment. They are purine metabolism (KEGG pathway ID ec00230) and arginine and proline metabolism (KEGG pathway ID ec00330). A list of pathways and their enrichment score is presented in Table 4.4. Since the basic role of purine metabolism and arginine/proline metabolism was described in detail in the previous section, this section focuses mainly on the relationship of the pathway with respect to acid-tolerance.

4.3.2.1 Purine Metabolism

Similar to the phenotype hydrogen producing, the NIBBS-search algorithm predicted purine metabolism as a potentially significant pathway for organisms expressing acid-tolerance. Purine metabolism encompasses biosynthesis, degradation, and salvage of purines within microorganisms. Together these pathways are necessary for survival and growth of organisms. Purines, along with pyrimidines, make-up vital components of nucleic acids (e.g. DNA and RNA), and are involved in the synthesis of many vitamins and coenzymes (e.g. ATP) [8]. As such, the high ranking of purine metabolism is likely a result of its role in nucleic acid synthesis (and growth) rather than specificity to the phenotype acid-tolerant. However, as shown in Chapter 3, individual enzymes present within purine metabolism may play a role in maintaining purine and

Table 4.4 List of top ranking pathways and their enrichment score for the phenotype acid tolerance.

Pathway ID	Pathway Name	P-Value
cac00230	Purine metabolism	2.56E-14
cac00330	Arginine and proline metabolism	1.75E-11
cac00520	Amino sugar and nucleotide sugar metabolism	1.02E-10
cac00260	Glycine, serine and threonine metabolism	3.43E-10
cac00270	Cysteine and methionine metabolism	1.27E-09
cac00400	Phenylalanine, tyrosine and tryptophan biosynthesis	1.77E-09
cac00240	Pyrimidine metabolism	1.90E-08
cac00860	Porphyrin and chlorophyll metabolism	1.99E-06
cac00760	Nicotinate and nicotinamide metabolism	7.82E-06
cac00500	Starch and sucrose metabolism	1.02E-05
cac00040	Pentose and glucuronate interconversions	1.67E-05
cac00561	Glycerolipid metabolism	5.57E-05
cac00051	Fructose and mannose metabolism	6.37E-05

nucleic acids during periods of acid stress (National Center for Biotechnology Information). In fact, studies evaluating acid resistance have noted the potential of purine genes deoB and guaA which encodes for phosphopentomutase and GMP synthase respectively, in assisting with acid tolerance [66]. Proteins associated with these genes are involved in the salvage pathway. In some Lactobacillus species, organisms can utilize nucleobases, such as guanine and adenine, generated during DNA and RNA degradation to synthesize nucleotides [94]. The salvage of these purine nucleobases is particularly important during dark fermentative hydrogen production when organic acid (e.g. butyrate) accumulation lowers pH in the medium. If the internal pH value is not regulated and decreases, DNA and purine bases present are subject to degradation [75]. The presence of salvage pathway enzymes, such as adenosine deaminase, allows organisms to utilize the degraded bases to regenerate nucleotides and nucleic acids.

Therefore, it is predicted that sub-pathways within the purine salvage are related to the expression of acid-tolerance and resistance. Experimental analysis is needed to determine the exact role of purine salvage in bacterial response to low pH.

4.3.2.2 Arginine and Proline Metabolism

In hydrogen producing organisms, decarboxylation and deamination of amino acids (e.g. arginine) have been linked to osmotolerance and protection of cells in the presence of environmental stress [66]. One amino acid in particular, which exhibits this trait, is arginine. While arginine can be an important source of nitrogen and energy for bacteria, it is also considered an alkaline amino acid, thus making it an important component in combating acid stress. One mechanism involving arginine is decarboxylation of glutamate and arginine in *Lactobacilli.* In this process, arginine is decarboxylated, then the decarboxylated product removed and another arginine product transported into the cell [66]. Another mechanism is the arginine deiminase pathway (ADI). This pathway is responsible for the conversion of arginine to orthine, ammonium, and carbon dioxide. The ammonium produced is then used to increase the internal pH [66].

From the predicted NIBBS results, the presence of the ADI or decarboxylation pathways was not predicted in our model organism, *C. acetobutylicum*. However, key enzymes involved in these pathways for *C. perfringens* were shown as present (Chapter 3; Figure 3.8), thus suggesting these pathways may be utilized by this organism in response to acid stress. For the first pathway, the NIBBS algorithm was only able to predict the presence of glutamate decarboxylase (E.C.4.1.1.15) and did not identify arginine decarboxylase. This suggests *C. perfringens* may not utilize this route for acid tolerance.

In the ADI pathway, only two of the three essential enzymes associated with this pathway were identified. They were arginine deiminase (E.C. 3.5.3.6) and ornithine transcarbamylase (E.C. 2.1.3.3). In addition, we noted the presence of agmatine deiminase (3.5.3.12), an enzyme responsible for the conversion of agmatine to N-carbamoylputrescine and ammonia. Based on the presence of agmatine deiminase, we predict *C. pefringens* utilize this enzyme in arginine metabolism in response to acid stress. While it does not appear that *C. acetobutylicum* utilizes these two pathways, there has been reports that it is capable of utilizing similar mechanisms through activation of homologous genes [75]. However, review of these types of genes has not been well characterized to date. As such, analysis of genes present in the hydrogen producing *C. pefringens* can be used to provide clues to the expression of acid-tolerance.

4.4 Multiple Alignment for Identification of Conserved Metabolic Subnetworks

Multiple alignments of pathways, such as those identified in Section 4.3, enable identification of conserved metabolic components (e.g. enzymes) and further refinement of phenotype-related sub-networks within smaller sets of organisms. Through analysis of these conserved metabolic components, scientists can gain insights into metabolic processes, structural information, and evolutionary relationships (e.g. identification of homologous proteins) [95]. Such information is of particular interest to biological engineers. Information derived by alignments can be used to identify conserved metabolic sub-pathways related to phenotypes involved in production of biofuels, such as biological hydrogen production.

In this section, two multiple alignment studies were conducted. In the first study, pathways in sets of phenotype-expressing organism are aligned to assess the performance of the multiple alignment algorithm in identifying sub-pathways or sub-networks. In the second study, metabolic pathways for hydrogen production were aligned to identify conserved sub-pathways related to dark fermentative hydrogen production. For both studies, KEGG Pathway maps corresponding to target pathways were aligned in order to identify similarities across pathway topologies. Metabolic pathways and corresponding KEGG maps for both studies are presented in Table 4.5. Acid tolerance was not included in the multiple alignment results due to the complexity and limited information regarding acid resistant mechanisms in metabolic pathways.

4.4.1 TCA Cycle for Validation of Alignment Results

To evaluate the performance of the algorithm, the tricarboxylic acid (TCA) cycle (KEGG MAP 00020) was aligned across 5 aerobic organisms and analyzed with respect to published microbial physiological data. Organisms in this study included: *Bordetella bronchiseptica* RB50 (bbr), *Staphylococcus saprophyticus subsp. saprophyticus* ATTC 15305 (ssp), *Myxococcus xanthus* DK 1622 (mxa), *Leptospira interrogans serovar Lai* str. 56601(lil), and *Helicobacter pylori* HPAG1 (hpa).

The alignment results generated for the TCA cycle pathway demonstrate the ability of the algorithm to identify key enzymes associated with phenotype-related metabolic pathways. In this study, TCA enzymes (Table 4.6) were correctly aligned in all the organisms except in hpa. As demonstrated by literature, bbr, ssp, mxa, and lil are described as having complete TCA cycles. While the presence of the TCA cycle is often considered an indicator of aerobic respiration, it is not always present or complete in

Table 4.5 KEGG Pathway maps selected for aerobic respiration and hydrogen production studies.

Phenotype	KEGG ID	KEGG Pathway
Aerobic respiration	map00020	TCA cycle
Hydrogen production	map00010	Glycolysis
	map00620	Pyruvate metabolism
	map00630	Glyoxylate and dicarboxylate metabolism
	map00520	Amino sugar

some aerobic species [8]. In this study, TCA enzymes succinate thiokinase and malate dehydrogenase were not aligned in *hpa,* thus suggesting this organism lacks a complete TCA cycle. However, based on literature findings by Hoffman et al. [96], *H. pylori* does contain complete TCA cycle. Thus the absence of succinate thiokinase and malate dehydrogenase is an error associated with using the KEGG database. In the case of *H. pylori*, Hoffman et al. [96], suggest that the operation of the TCA may be dependent on the growth environment and availability of metabolites.

In addition to identifying reactions associated with TCA, our algorithm was able to identify the presence of an alternate enzyme, such as 2-oxoglutarate synthase, which is used in the formation of succinyl-CoA and reduced ferredoxin. 2-oxoglutarate synthase is an important enzyme that can be present in both the TCA cycle and the reductive TCA cycle. Identification of this enzyme in *H. pylori* compared to the other species is a result of KEGG including 2-oxoglutarate synthase as an alternative to oxoglutarate: ferredoxin oxidoreductase.

4.4.2 Dark Fermentative Hydrogen Production

To further investigate and identify conserved metabolic sub-networks associated with dark fermentative hydrogen production, the KEGG pathway map for pyruvate

Table 4.6 Results for aligning the metabolic components of KEGG Pathway 00020 (TCA cycle) across five aerobic organisms**.** Only the alignments for enzymes are presented in this table. Each row in the table below represents an alignment of a single enzyme across the organisms*: Bordetella bronchiseptica* RB50 (bbr), *Staphylococcus saprophyticus* subsp. saprophyticus ATCC15305 (ssp), *Myxococcus xanthus* DK1622 (mxa), *Leptospira interrogans* serovar Lai STR.56601 (lil), and *Helicobacter pylori* HPAG1 (hpa). The + entries in the table indicate which enzyme was identified as being aligned for that organism, whereas ◊ indicates that no aligned enzymes were identified for that organism. Enzymes marked with * are part of the TCA cycle.

Enzyme commission	Enzyme name	bbr	ssp	mxa	lil	hpa
1.1.1.42	Isocitrate dehydrogenase*	+	+	+	+	+
1.3.99.1	Succinate dehydrogenase *	+	+	+	+	+
2.3.3.1	Citrate synthase*	+	+	+	+	+
4.2.1.2	Fumarase*	+	+	+	+	+
4.2.1.3	Aconitase*	+	+	+	+	+
1.2.4.1	Pyruvate dehdyrogenase	+	+	+	+	
1.2.7.1	Pyruvate synthase					+
1.2.4.2	Alpha-ketoglutarate dehydrogenase*	+	+	+	+	
1.2.7.3	2-oxoglutarate synthase					+
1.1.1.37	Malate dehydrogenase*	+	+	+	+	◊
1.8.1.4	Dihydrolipoyl dehydrogenase	+	+	+	+	◊
2.3.1.12	Dihydrolipoyl lysine-residue acetyltransferase	+	+	+	+	◊
2.3.1.61	Dihydrolipoyl lysine-residue succinyltransferase	+	+	+	+	◊
6.2.1.5	Succinate thiokinase*	+	+	+	+	◊
6.4.1.1	Pyruvate carbvoxylase	+	+	+		◊
4.1.3.6	Citrate(pro-3S)-lyase				+	◊
4.1.1.32	Phsophoenolpyruvate carboxykinase (GTP)	+		+		◊
4.1.1.49	Phsophoenolpyruvate carboxykinase (ATP)		+		+	◊
4.1.3.6	Citrate(pro-3S)-lyase	+				◊
1.2.7.3	2-oxoglutarate synthase		+			◊
1.1.141	Isocitrate dehydrogenase (NAD+)			+	+	◊

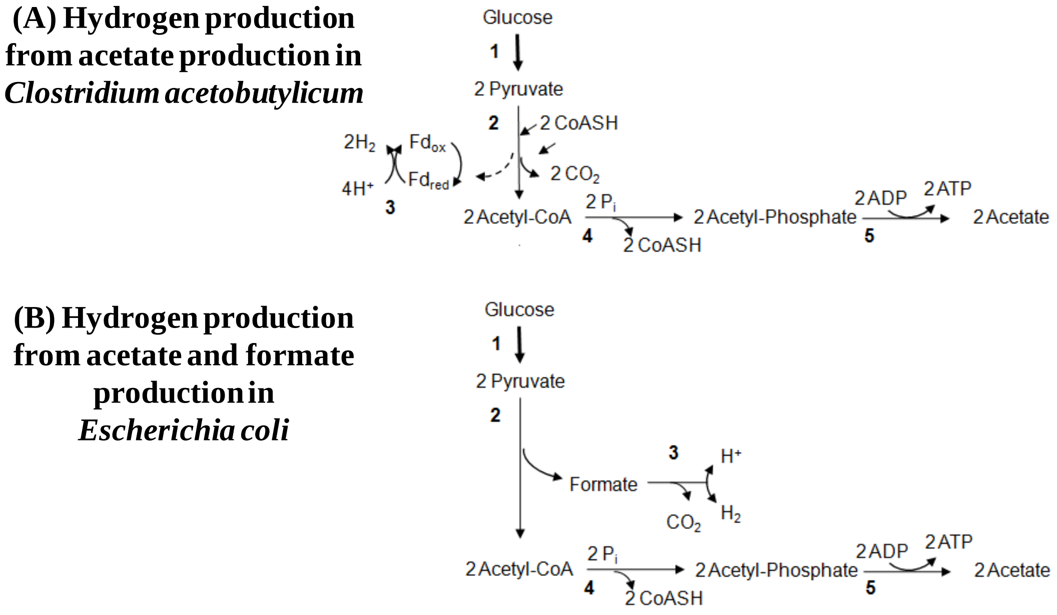
metabolism (KEGG MAP 00620) was aligned across 8 dark fermentative hydrogen producing species. Organisms in this study include *Bacillus licheniformis* (bli), *Caldicellulosiruptor saccharolyticum* (csc), *Clostridium acetobutylicum* (cac), *Clostridium beijerinchkii* (cbe), *Clostridium perfrigens* (cpf), *Clostridium thermocellum* (cth), *Escherichia coli* (eco), and *Thermotoga neapolitana* (tna).

Selection of the pathway maps were based on the presence or absence of known hydrogen producing, dark fermentative routes. To generate hydrogen, three metabolic routes associated with hydrogen production are often targeted for molecular studies. They are the metabolic routes for production of acetic acid, butyric acid, and formic acid. In *C. acetobutylicum*, hydrogen is often generated in association with production of acetic and butyric acid (see Chapter 3, Figure 3.5).

In other dark fermentative bacteria, such as *E. coli*, hydrogen is produced in association with butyric acid, acetic acid, and formate production (Figure 4.2). Due to the differences in hydrogen production in these two organisms (Figure 4.2), alignment of pyruvate metabolism was selected to identify which of the two routes were conserved across 8 known hydrogen producers.

The alignment results generated for pyruvate metabolism, identified enzymes associated with acetate fermentation pathways present in Figure 4.2. Review of the alignment results indicate the algorithm was able to align the two enzyme involved in conversion of Acetyl-CoA to acetate, across *C. acetobutylicum* and *E. coli* (Table 4.7). These are phosphate acetyltransferase (2.3.1.9) and acetate kinase (2.7.2.1). The alignment results also indicate the presence of these enzymes within other organisms, such as *C.* saccharolyticum, but are poorly aligned.

Review of the sub-pathway for conversion of pyruvate to Acetyl-CoA, is consistent with literature findings for *E. coli*. In this study, the multiple alignment algorithm predicts *E. coli*, *C. beijerinchkii*, and *B. licheniformis* (Table 4.2). While this enzyme has been reported in *E. coli*, it is not reported as being utilized by both *Clostridium* species and *B. lincheniformis.* In fact, literature reports suggest that *C. acetobutylicum* utilize pyruvate ferredoxin oxidoreductase (PFOR) to generate Acetyl-CoA [32, 44]. However, in this study, PFOR is not predicted as conserved across a set of



(A) Hydrogen production from acetate production in *Clostridium acetobutylicum*

(B) Hydrogen production from acetate and formate production in *Escherichia coli*

Figure 4.2 Comparison of routes for acetate production in the two dark fermentative organisms, Clostridium acetobutylicum (A) and Escherichia coli (B). Enzymes for C. acetobutylicum: 1, glycolytic enzymes; 2, pyruvate formate lyase (E.C. 2.3.1.54); 3, hydrogenase (E.C.1.12.7.2); 4, phosphotransacetylase (E.C. 2.3.1.8); 5, acetate kinase (E.C. 2.7.2.1). (B) Enzymes for E. coli: 1, glycolytic enzymes; 2, pyruvate formate lyase (E.C. 2.3.1.54); 3, formate hydrogen lyase (E.C. 1.1.99.33); 4, phosphotransacetylase (E.C. 2.3.1.8); 5, acetate kinase (E.C. 2.7.2.1). Arrows with larger width indicate a series of reactions. Arrows with narrow width indicate individual reactions.

organisms and is not indicated as present in *C. acetobutylicum* (Table 4.2). The absence of PFOR in *C. acetobutylicum* could be the result of one of two things—complete absence of the enzyme in the KEGG database or missed by KEGG due to heterogeneity in the enzyme sequence. The latter is most likely to be the cause since PFORs have been shown to be divergent in sequences and structure.

Another enzyme predicted as conserved across a number of species was methylglyoxal synthase (4.2.3.3). In this study, methylgyoxal synthase was present across all organisms except *C. perfringens.* In *C. acetobutylicum*, this enzyme is responsible for production of methylglyoxal and inorganic phosphate from

Table 4.7 Results for aligning the metabolic components of KEGG Pathway 00620 (Pyruvate metabolism) across eight organisms. Only the alignments for enzymes are presented in this table. Each row in the table below represents an alignment of a single enzyme across the organisms: *Bacillus licheniformis* (bli), *Caldicellulosiruptor saccharolyticum* (csc), *Clostridium acetobutylicum* (cac), *Clostridium beijerinchkii* (cbe), *Clostridium perfrigens* (cpf), *Clostridium thermocellum* (cth), *Escherichia coli* (eco), and *Thermotoga neapolitana* (tna). The + entries in the table indicate which enzyme was identified as being aligned for that organism, whereas ◊ indicates that no aligned enzymes were identified for that organism. Enzymes marked with * are part of one of two pathways for acetate production.

Enzyme Comission	Enzyme Description	cac	eco	cpf	cbe	tna	bli	cth	csc
6.4.1.2	acetyl-CoA carboxylase	+	+	+	+	◊	+	◊	◊
1.1.1.37	malate dehydrogenase	+	+	◊	◊	◊	+	+	◊
1.1.1.38	malate dehydrogenase (oxaloacetate-decarboxylating)	◊	◊	◊	+	◊	◊	◊	◊
1.1.1.38	malate dehydrogenase (oxaloacetate-decarboxylating)	+	+	◊	◊	◊	+	+	+
1.1.1.77	lactaldehyde reductase	◊	◊	◊	+	◊	◊	◊	◊
1.2.1.10	acetaldehyde dehydrogenase (acetylating)	+	+	◊	+	◊	◊	+	◊
1.2.3.3	pyruvate oxidase	◊	◊	◊	◊	◊	+	◊	◊
2.3.1.54	formate C-acetyltransferase *	+	+	◊	+	◊	+	◊	◊
2.3.1.8	phosphate acetyltransferase*	◊	◊	◊	◊	◊	◊	◊	+
2.3.1.8	phosphate acetyltransferase*	+	+	◊	+	◊	◊	◊	◊
2.3.1.12	dihydrolipoyllysine-residue acetyltransferase	◊	◊	◊	◊	◊	+	◊	◊
2.7.2.1	acetate kinase*	◊	◊	◊	◊	◊	◊	◊	+
2.3.1.9	acetyl-CoA C-acetyltransferase	+	+	◊	+	◊	◊	◊	◊
1.8.1.4	dihydrolipoyl dehydrogenase	◊	◊	◊	◊	◊	+	◊	◊
1.2.7.1	pyruvate synthase*	◊	◊	◊	◊	◊	◊	◊	+
2.3.3.13	2-isopropylmalate synthase	+	+	◊	+	+	+	+	+
2.3.3.14	homocitrate synthase	+	◊	◊	+	◊	◊	◊	◊
2.3.3.9	malate synthase	◊	+	◊	◊	◊	+	◊	◊
2.7.9.1	pyruvate, phosphate dikinase	◊	◊	◊	◊	◊	◊	◊	+
2.7.2.1	acetate kinase*	+	+	◊	◊	◊	+	◊	◊
2.7.9.2	pyruvate, water dikinase	◊	◊	◊	+	◊	◊	◊	◊
2.7.1.40	pyruvate kinase	◊	◊	◊	◊	◊	◊	◊	+
2.7.9.2	pyruvate, water dikinase	+	+	◊	◊	◊	+	◊	◊
2.7.2.1	acetate kinase*	◊	◊	◊	+	◊	◊	◊	◊
2.3.1.12	dihydrolipoyllysine-residue acetyltransferase	◊	◊	◊	◊	◊	◊	◊	+
3.6.1.7	acylphosphatase	+	+	+	+	+	+	◊	◊
4.2.3.3	methylglyoxal synthase	+	+	◊	+	+	+	+	+
6.4.1.1	pyruvate carboxylase	+	◊	◊	+	◊	+	◊	◊
1.2.1.22	lactaldehyde dehydrogenase	◊	+	◊	◊	◊	◊	◊	◊
1.1.1.27	L-lactate dehydrogenase	+	◊	◊	+	◊	◊	◊	◊
1.1.1.40	malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+)	◊	+	◊	◊	◊	◊	◊	◊
1.2.1.3	aldehyde dehydrogenase (NAD+)	◊	◊	◊	◊	◊	+	◊	◊
2.7.1.40	pyruvate kinase	+	+	◊	+	◊	+	◊	+
1.1.1.28	D-lactate dehydrogenase	+	+	◊	+	◊	◊	◊	◊
1.1.1.27	L-lactate dehydrogenase	◊	◊	◊	◊	◊	+	◊	+
6.2.1.1	acetate-CoA ligase	◊	+	◊	◊	◊	◊	◊	◊
2.3.1.12	dihydrolipoyllysine-residue acetyltransferase	◊	+	◊	◊	◊	◊	◊	◊
1.8.1.4	dihydrolipoyl dehydrogenase	◊	+	◊	◊	◊	◊	◊	◊
1.1.1.77	lactaldehyde reductase	◊	+	◊	◊	◊	◊	◊	◊
1.2.7.1	pyruvate synthase*	◊	◊	◊	◊	◊	◊	+	◊
1.1.1.79	glyoxylate reductase (NADP+)	◊	+	◊	◊	◊	◊	◊	◊
1.1.2.3	L-lactate dehydrogenase (cytochrome)	◊	+	◊	◊	◊	◊	◊	◊
1.1.99.16	malate dehydrogenase (acceptor)	◊	+	◊	◊	◊	◊	◊	◊
1.2.2.2	pyruvate dehydrogenase (cytochrome)	◊	+	◊	◊	◊	◊	◊	◊
3.1.2.6	hydroxyacylglutathione hydrolase	◊	+	◊	◊	◊	◊	◊	◊
4.1.1.31	phosphoenolpyruvate carboxylase	◊	+	◊	◊	◊	◊	◊	◊
4.1.1.49	phosphoenolpyruvate carboxykinase (ATP)	◊	+	◊	◊	◊	◊	◊	◊
4.4.1.5	lactoylglutathione lyase	◊	+	◊	◊	◊	◊	◊	◊
1.2.4.1	pyruvate dehydrogenase (acetyl-transferring)	◊	+	◊	◊	◊	◊	◊	◊

dihydroxyacetone. Studies by Huang et al. [97] indicate that methylgyoxal synthase is potentially an important enzyme for converting sugars to diols for solvent production. If the main goal is production of hydrogen via acetogenesis, then down regulation of enzymes involved in solventogenesis, such as methylgyoxal, need to be considered. As such, the multiple alignment algorithm not only identifies potentially important hydrogen-related sub-networks, but also identifies other networks to consider when engineering organisms.

4.5 Discussion

In summary, the NIBBS search algorithm and the multiple alignment algorithm were able to identify phenotype-related metabolic pathways and sub-networks across sets of phenotype-expressing microorganisms. Specifically, through co-development and application of the NIBBS algorithm, both pathways specific to and those related to dark fermentative, hydrogen production and acid-tolerance were presented. From those identified pathways, scientists are able to gain insight into the potential role some pathways, such as fatty acid metabolism, on metabolic shifts between hydrogen production and solvent formation.

In addition, through comparison of multiple phenotypes deemed important for hydrogen production in wastewater, pathways responsible for expression of more than one phenotype were identified. Specifically, pathways for purine metabolism and the pathways for proline and arginine metabolism were predicted as related to dark fermentative hydrogen production and acid-tolerance. Due the continued presence of these two pathways, engineers and scientists can experimentally test the role of the pathways as survival mechanisms for acid response and hydrogen production.

Identification of these shared pathways for the two phenotypes is due to the ability of the multiple organisms to express multiple phenotypes. For example, *Clostridium acetobutylicum* and *Clostridium perfringens* ATCC 13124 are both dark fermenting organisms, but they also share other common phenotypes like anaerobicity and tolerance to acid. These phenotypes, if analyzed together, may provide us more information about the phenotype systems in these two organisms than looking at each phenotype individually.

4.5.1 NIBBS for Identification of Cross-Talk

In addition to identifying conserved metabolic pathways, results from the NIBBS algorithm suggest that this method can potentially identify metabolites common to different metabolic pathways. One example of such a metabolite is acetyl-CoA. Acetyl-CoA is generated from pyruvate during glycolysis and can be utilized by differing pathways, including the aerobic TCA cycle and the anaerobic formate hydrogen lyase pathway. In the aerobic TCA pathway, the enzyme, pyruvate dehydrogenase, catalyzes the decarboxylation of pyruvate to CO_2 (g) and acetyl-CoA. Acetyl-CoA generated using this process can then be incorporated into the TCA cycle to produce important biosynthetic precursors for other metabolic pathways and energy for microorganisms [8, 98]. In the anaerobic pathway, pyruvate formate lyase is used to convert pyruvate into acetyl-CoA and formate. Formate produced can then be oxidized by formate hydrogen lyase (FHL) to form CO_2 (g) and H_2 (g). In the hydrogen studies, the NIBBS algorithm predicts the presence of both pyruvate formate lyase (E.C. 1.1.99.3) and pyruvate dehydrogenase (E.C. 1.2.4.1) when dark fermentative hydrogen producing organisms are compared to non-hydrogen producing organisms. The presence of both pathways may be due to the fact that some dark fermentative microorganisms are capable of utilizing both pathways and the degree to which they utilize each pathway may be dependent on the “cross-talk” between both pathways. However, depending on environmental conditions the bacteria are grown under, the organism may express one over the other. To understand the role of these pathways, further experimental analysis is required.

Identification of common metabolites and potential cross-talk between metabolic pathways is a key step towards understanding metabolic processes, networks, and regulation of phenotype expression in organisms such as hydrogen producing organisms. While numerous genetic and experimental studies have been conducted to understand the metabolic processes involved in hydrogen production, there is still little understanding of the cross-talk between key hydrogen producing pathways. To help close this gap, biologists could potentially use the NIBBS algorithm to provide hypothesis driven work. One way would be to identify phenotype related pathways, such as the two pathways for acetyl-CoA production, and then conduct molecular studies to review these pathways in organisms shown to be positive for both pathways. Since both related pathway information and identification of individual species containing these pathways are identified by the NIBBS algorithm, biologists could use this information to develop and direct their work.

4.5.2 Reliance of the Multiple Alignment Algorithm on KEGG Data

Similar to the NIBBS algorithm, the alignment algorithm was able to identify conserved sub-networks across phenotype-expressing microorganisms. This was demonstrated in the correct alignment of the TCA pathway in aerobic organisms. When applied to a set of organisms, such as dark fermentative species, that contain more than one route for production of Acetyl-CoA, the alignments did not appear to be as accurate. For example, if we consider the enzyme formate C-acetyltransferase, the algorithm incorrectly predicts this as the main route for conversion of pyruvate to Acetyl-CoA in *C. acetobutylicum*. In addition, the algorithm did not predict the main enzyme reported for this reaction (pyruvate ferredoxin oxidoreductase) as being present. The lack of prediction for pyruvate ferredoxin oxidoreductase is most likely to due to inaccuracies in the KEGG database.

The presence of formate C-acetyltransferase in *C. acetobutylicum* does not necessarily indicate the organism utilizes this route. Further molecular analysis is needed to measure the activity and main function of this enzyme. The presence simply suggests this enzyme is present and potentially could function in formation of Acetyl-CoA. In general, the multiple alignment algorithm is a potentially powerful tool in identifying conserved regions within target organisms. However, the accuracy in prediction is reliant upon the data present in the KEGG database.

4.6 Approach

4.6.1 NIBBS Algorithm for Prediction of Phenotype-Related Pathways

To enable the prediction of phenotype-related metabolic pathways, the network's instance-based subgraph search algorithm or NIBBS algorithm was employed. The NIBBS algorithm is a phenotype-driven comparative analysis tool that is capable of searching genome-scale metabolic networks across phenotype exhibiting organisms to identify phenotype related metabolic enzymes and sub-systems. Because it is computationally intractable to consider all possible enzyme sets, some previous approaches have considered only those enzyme sets that correspond to a known metabolic pathway [99]. However, such a methodology is limited because it cannot detect an interplay, or cross-talk, among phenotype-related metabolic pathways. This calls for approaches that could compare networks at the genome-scale rather than at the individual pathway level.

From a graph-theoretical perspective, application of NIBBS addresses a computational problem that aims to identify connected subgraphs that are both conserved and statistically enriched across metabolic networks of phenotype expressing organisms. To date, the heuristic NIBBS algorithm is the only known algorithm that can identify the desired subgraphs across tens and even hundreds of metabolic networks in practical time. NIBBS "grows" the desired subgraphs and limits the number of subgraphs grown [100]. NIBBS runs in a matter of seconds, yet the subgraphs identified by the NIBBS are highly correlated to the subgraphs that are the most likely to represent phenotype-related metabolic pathways. A detailed description of the NIBBS search algorithm is provided in Chapter 6.

To evaluate the ability of the NIBBS algorithm to identify pathways specific to expression of a phenotype, a comparative analysis of the TCA and rTCA pathway in sets of organisms known to utilize these two pathways was analyzed. For identification of known and new pathways related to phenotype expression, the phenotypes dark fermentative hydrogen production and acid tolerance were evaluated across sets of phenotype expressing and non-phenotype expressing organisms. A complete list of organisms is provided in Appendix B.

4.6.2 Multiple Alignment Without Abstraction

To identify common sub-networks among multiple pathways in phenotype expressing organisms, a multiple alignment approach without extraction was co-developed and employed. Unlike previous multiple alignment methodologies, this approach does not rely on the alignment of individual entities, such as enzymes or compounds. Instead, multiple pathways are aligned based on the similarities among reactions, compounds, and enzymes present in the pathway topology. A schematic depicting an alignment of pathways is provided in Figure 4.3. Using this approach, common sub-networks present in sets of phenotype expressing organisms can be identified, thus providing clues to pathways involved in phenotype expression. A detailed description of the algorithm is provided in Chapter 6.

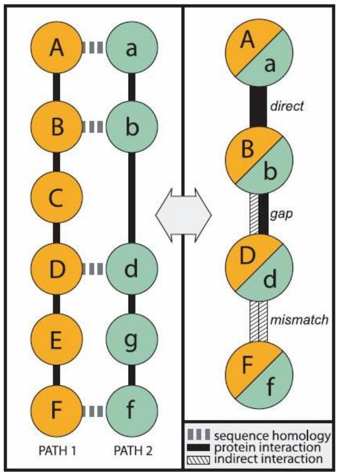


Figure 4.3 Alignment of two pathways. Taken from Kelley et al. [10]

In this study, *prior* knowledge or known metabolic pathways associated with aerobic respiration and dark fermentative hydrogen producing phenotypes were selected and compared across multiple species to identify conserved sub-networks. KEGG maps corresponding to each target pathway were aligned and analyzed. Selection of these pathways was derived from top pathways predicted by the NIBBS algorithm. Classification of metabolic pathways as phenotype-specific or not phenotype-specific is based on the following assumption that pathways which are conserved primarily in phenotype-expressing organisms are more likely to be related to the expression of that phenotype.

CHAPTER 5: DISCOVERY OF POTENTIAL CROSS-TALKS AND REGULATORY CONTROLS AMONG METABOLIC PATHWAYS AND FUNCTIONAL MODULES ASSOCIATED WITH HYDROGEN PRODUCTION

5.1 Background and Motivation

Application of genomic and systems-biology studies toward environmental engineering (e.g., waste treatment) generally requires understanding of microbial responses and metabolic capabilities at the genome and metabolic levels. This is particularly true for predicting relationships between microbial traits and metabolic pathways and networks. To date, several studies have used different approaches in an attempt to link microbial phenotypes to conserved sub-networks [78, 101], including the ones described in Chapters 3-4 [102]. While knowledge of phenotype-related genes and metabolic pathways is important for designing genetically modified organisms, expression of a phenotype by microorganisms may be due to the result of interactions between multiple pathways rather than individual pathways.

In biological systems, phenotype-related genes encode for a number of functionally associated proteins which may be found across a number of different metabolic, regulatory, and signaling pathways [17, 103]. Together these pathways form a biologically important and potential set of interacting network of proteins (or genes) that are responsible for the expression of a particular phenotype. Figure 5.1 illustrates the ability of proteins involved in metabolic reactions to communicate. Since proteins can be present in a number of biochemical reactions, or pathways, understanding of the role and interactions of proteins within various networks is necessary to identify which metabolic

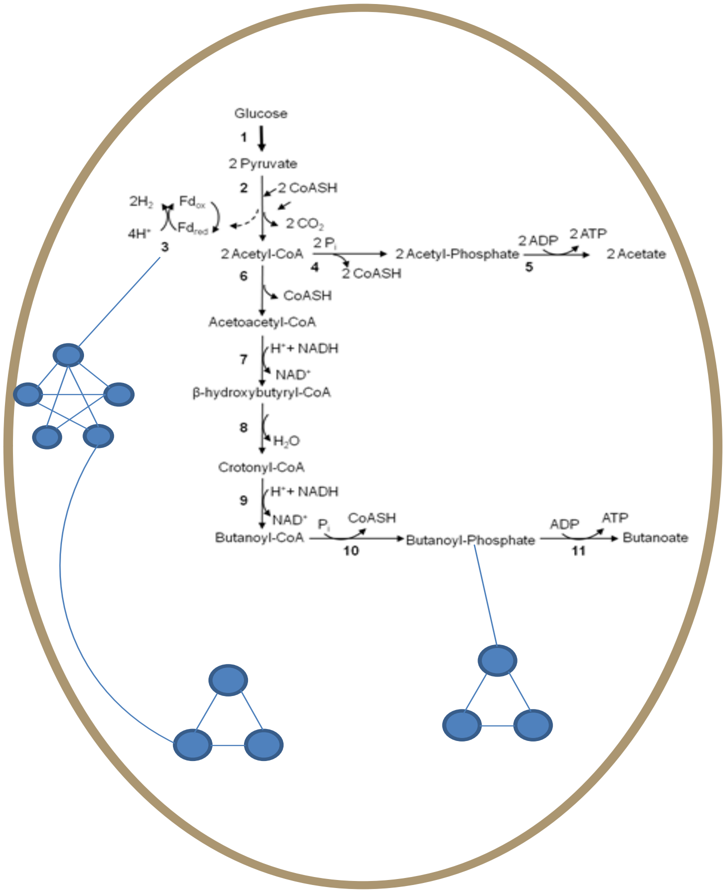


Figure 5.1 Examples of network interactions between pathways in microbial cells. This figure demonstrates that proteins involved in pathways, such as Acetyl-CoA formation, may also be involved in expression of other metabolic pathways.

and cellular networks are important for enhancing or suppressing expression of phenotypic traits. Through analysis of biologically conserved network models, insights into the functional role of phenotype-related genes and interactions between genes in these networks can be obtained. This knowledge can then be used by metabolic engineers to identify which genes are potential candidates for modification studies and to determine how modification of selected genes will impact their desired outcome (e.g. hydrogen production).

5.2 Research Goal and Contributions

The goal of this work is to enable the identification of phenotype-related network modules in gene and protein networks related to expression of microbial phenotypes involved biological hydrogen production. The aims for this study are to (1) identify phenotype-related subsystems or functional modules, (2) enable the identification of functional roles of genes in regulating or signaling pathway expression, (3) enable the identification of clues to functional roles of previously uncharacterized genes related to phenotype expression, and (4) enable the discovery of interplay between phenotype-related subsystems, particularly in high-level functional association problems.

Identification of interplay or "cross-talk" between metabolic networks through computational approaches has not been well studied. As such, this study is one of the first to introduce a potential application for identification of cross-talk between pathways related to bioenergy. The scope of this work is focused on expression of the two phenotypes—dark fermentative, biological hydrogen production and acid-tolerance.

To achieve these aims, three complementary approaches are employed to identify phenotype-related network models and potential functional associations between the networks. A brief overview of the methodologies is provided below, but more detailed descriptions on the computational approach are provided in Sections 5.4 (general overview) and in Chapter 6 (detailed) of this dissertation.

The first approach called the Dense ENriched Subgraph Enumeration (DENSE) algorithm, capitalizes on the availability of partial *"prior* knowledge" about the proteins involved in this process and enriches that knowledge with newly identified sets of functionally associated proteins in individual phenotype-expressing microorganisms.

When applied to a network of functionally associated proteins in the dark fermentative, hydrogen producing bacterium, *Clostridium acetobutylicum*, the algorithm is able to predict known and novel relationships including those with regulatory, signaling, and uncharacterized proteins. The second approach, a bi-clustering algorithm, allows for identification of conserved functional modules by assigning proteins to COG groups. Unlike the previous approach, this method utilizes genes present in metabolic networks rather than those present across entire cellular networks of phenotype expressing organisms. Using this data, pairs of interacting networks of COGs for each organism are identified and matrices consisting of phenotype-related networks constructed. The last approach, called *α,β*-motifs, allows for identification of functional modules that, in addition to metabolic subsystems, could include their regulators, sensors, transporters, and even uncharacterized proteins that are predicted to be related to the target phenotype. By comparing hundreds of genome-scale networks of functionally associated proteins, this method identifies those functional modules that are enriched in at least α networks of phenotype-expressing organisms but may still appear in no more than β networks of organisms that do not exhibit the target phenotype.

5.3 Results

In this section, predicted phenotype-related subsystems and functional relationships between genes (or proteins) present in each subsystem are presented. For each method, a set of experiments was conducted to predict phenotypes-related networks (e.g. functional modules or protein clusters) related to dark fermentative hydrogen production. To further understand which subsystems are involved in acid resistance mechanisms in bacteria, the phenotype acid-tolerance was evaluated using the DENSE algorithm and *α,β*-motif algorithm. Organisms for the acid-tolerant study were predominately representative of the class Bacilli and Clostridia. For each study, genes within predicted subsystems were reviewed to determine their functional role and associations between interacting genes.

5.3.1 Discovery of Phenotype-Related Proteins Using Knowledge Priors

To discover clusters of related to phenotypes and sub-phenotypes associated with hydrogen production from waste materials, the DENSE algorithm was applied to the hydrogen producing bacterium, *Clostridium acetobutylicum* ATCC 824. *C. acetobutylicum* is a widely studied and well-characterized organism for hydrogen production in nutrient-rich systems [104, 105]. In addition to dark fermentative hydrogen production, *C. acetobutylicum* exhibits a number of phenotypes important for growth and production of hydrogen. Such phenotypes include dark fermentative hydrogen production and acid-tolerance down to a pH range of 4.4-6.0 [106]. While Clostridium species are often associated with dark fermentative acidogenesis, they are also known for production of solvents [106, 107]. During solventogenesis, hydrogen produced is consumed and butanol, ethanol, and acetone are generated [106].

The following sections present a description of biological networks identified and the predicted interactions between proteins (and genes) that play a role in uptake and production of hydrogen through regulation, signaling, or synthesis of key enzyme. Specifically, emphasis is placed on key proteins and networks identified in the previous methodologies (e.g, hydrogenase or enzymes for butyrate production). To identify dense, enriched protein-protein interaction networks, three experiments were conducted. In the first experiment proteins directly related to the [FeFe]-hydrogenase (HydA) were identified. In the last two experiments, hydrogen-related and acid-tolerant knowledge priors identified from Chapter 3 T-test and NIBBS experiments were incorporated into the algorithm and the clusters analyzed.

5.3.1.1 Hydrogenase Interactions in *C. acetobutylicum*

In fermentative hydrogen-producing organisms, such as *C. acetobutylicum,* hydrogen yields are dependent on the presence and activation of hydrogen producing enzymes called hydrogenases [47]. Studies evaluating the role of hydrogenase in hydrogen production have shown that organisms can contain more than one type of hydrogenase that can each require sets of accessory proteins for activation. As such, the presence or absence of specific accessory proteins play an important role in regulating the activity of hydrogenase and hydrogen production or uptake in microorganisms. In addition, many hydrogenases are thought to either directly or indirectly regulate other metabolic processes, such as nitrogen metabolism [6]. Therefore, the understanding of phenotype-related proteins required for activation and maturation of hydrogenases is important when metabolically engineering organisms.

When applied to HydA, a hydrogen producing hydrogenase enzyme, the DENSE algorithm was able to identify three maturation proteins that are essential for expression of a [FeFe]- hydrogenase [52]. They are HydE (CAC1631), HydF (CAC1651), and HydG (CAC1356) (Figure 5.2; Table 5.1). When these proteins are present and interact with HydA1, activation of the hydrogen producing [FeFe]-hydrogenase occurs. According to studies on hydrogenases, deletion of one maturation protein will inactivate the [FeFe]-hydrogenase [52].

Table 5.1 Protein-protein interaction network corresponding to Figure 5.1 and description of hydrogenase-related proteins present in *Clostridium acetobutylicum*.

STRING ID	Protein ID	Protein Description
CAC0028	HydA1	Hydrogenase I (Hydrogene dehydrogenase)
CAC0487	-	Uncharacterized protein
CAC1651	HydF	Predicted GTPase with uncharacterized domain
CAC1631	HydE	Biotin synthase family enzyme
CAC1356	HydG	Thiamine biosynthesis enzyme

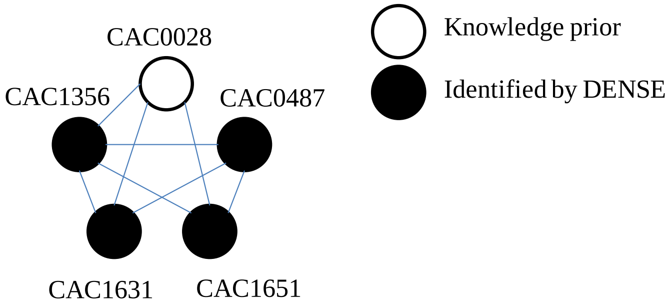


Figure 5.2 Dense and enriched sub-graphs identified by DENSE algorithm**.** White circle represents the knowledge prior protein HydA1. Image borrowed from Hendrix et al. (2010).

In addition to identifying key protein clusters, the algorithm predicted an association between an uncharacterized protein (Figure 5.2; CAC0487) and the three maturation proteins. Identification of uncharacterized proteins may assist scientists and bioengineers in identifying potential functional roles of unknown proteins. For example, in our study [108], functional associations between HydA1 and proteins involved in activation of a hydrogen producing enzyme were identified. Since the uncharacterized protein is highly interconnected with the maturation proteins, it can be predicted that the protein is involved in development of the [FeFe]-hydrogenase (HydA1). Utilizing this information, the role of CAC0487 in relation to the three maturation proteins can be characterized through genetic studies and then applied to bioengineering hydrogen producers.

Application of the clustering algorithm using hydrogen-related enzymes identified with NIBBS resulted in prediction of over 6000 clusters of protein-protein interactions. Of these clusters, a number of interaction networks containing proteins associated with expression of key enzymes related to either hydrogen uptake enzymes were identified. Examples of enzymes include those involved in maturation of hydrogenase (HypE and HypD) and nitrogenase (Nif), and key fermentation pathways for hydrogen production in anaerobic organisms. Within these clusters, both known and new interactions between proteins involved in regulation, synthesis, and signaling of hydrogen producing pathways were identified.

Review of the predicted protein-protein interaction clusters for the phenotype hydrogen producing revealed the presence of only one cluster containing known hydrogenase proteins (Figure 5.3; Table 5.2). Within this cluster are two [NiFe]-maturation hydrogenase proteins (HypE and HypD) and phosphoheptose isomerase (GmhA). Similar to results identified with the α, β- Clique algorithm, HypD (CAC0811) and HypE (CAC0809) proteins are depicted as interacting, further strengthening the importance of [NiFe]-maturation proteins in impacting the overall hydrogen yields in hydrogen-producing organisms. Since Hyp proteins are involved in activation and synthesis of uptake hydrogenase enzymes [47], down-regulation of HypD and HypE in Clostridium species are potential targets for enhancing biological hydrogen production. Together the HypABC proteins, HypD and HypE are functionally important for

Table 5.2 Protein-protein interaction network corresponding to Figure 5.3 and description of hydrogenase-related proteins present in *Clostridium acetotbutylicum*.

STRING ID	Protein ID	Protein Description
CAC3054	GmhA	Phosphoheptose isomerase
CAC0811	HypD	Hydrogenase expression-formation factor
CAC0809	HypE	Hydrogenase formation factor

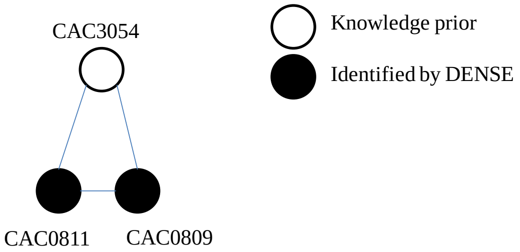


Figure 5.3 Phosphoheptose and interacting proteins identified by DENSE algorithm**.** White circle represents the knowledge prior utilized to identify the dense, enriched cluster.

expression the [NiFe]-hydrogenase and deletion one of the proteins may lead to inactivation [47].

While the interaction between the two Hyp proteins is clearly defined by previous studies [43, 47, 109], their interaction with phosphoheptose isomerase is not well understood. Phosphoheptose isomerase or GmhA (CAC3054) is an enzyme involved in biosynthesis of glycerol-manno-heptose [110]. In *Escherichia coli*, phosphoheptose isomerase is the involved in biosynthesis of ADP-L-glycero-β-D manno-heptose, a compound required in development of lipopolysaccharide (LPS) [110, 111]. Specifically, ADP-L-glycero-β-D manno-heptose utilized in biosynthetic pathways resulting in production of S-layer glycoproteins and production of the inner-core of LPS [111]. While development of lipolysaccharides is typically found in gram negative bacteria, the presence of LPS in Clostridium has been reported [111]. According to the results, all three proteins are shown to interact with one another (Figure 5.3). The knowledge prior GmhA is shown to interact with both HypD and HypE. However, from the diagrams provided in Figure 5.3, it is unclear why and how the two hydrogenase proteins (HypD and HypE) interact with GmhA.

5.3.1.1.1 Pyruvate: Ferredoxin Oxidoreductase and Associated Proteins

Another important enzyme for hydrogen production in *C. acetobutylicum* is pyruvate: ferredoxin oxidoreductase (CAC2229). In anaerobic, hydrogen-producing organisms, pyruvate: ferredoxin oxidoreductase or PFOR is responsible for the conversion of pyruvate to acetyl-CoA [32, 44, 112]. Acetyl-CoA is then utilized by a number of pathways, including acetate and butyrate fermentation routes. During production of acetate and butyrate, hydrogen is also produced as a by-product. In this, the DENSE algorithm was able to predict the interaction of this important enzyme when pyruvate lyase was given as a hydrogen-related enzyme. While pyruvate formate lyase (PFL) is utilized to generate formate and acetyl coenzyme A (Acetyl-CoA) in facultative anaerobic bacteria [32], it is not uncommon to find genes encoding PFL in anaerobic organisms, such as Clostridium [113].

In this study, many clusters containing PFL were identified, but only one that contained PFOR. Figure 5.4 and Table 5.3 demonstrates an example of one cluster containing PFL (CAC0980) identified by the DENSE algorithm. In this cluster, the algorithm identified interactions between the two acetyl-CoA forming enzymes, PFL and PFOR (CAC2229) and a third enzyme involved in the acetyl-CoA pathway—phosphotransacetylase (CAC1742). Phosphotransacetylase (Pta) is involved in the

Table 5.3 Protein-protein interaction network corresponding to Figure 5.4 and description of pyruvate: ferredoxin oxidoreductase and associated proteins present in *Clostridium acetobutylicum*.

STRING ID	Protein ID	Protein Description
CAC0980	-	Pyruvate-formate lyase
CAC2229	-	Pyruvate:ferredoxin oxidoreductase
CAC1742	Pta	Phosphotransacetylase

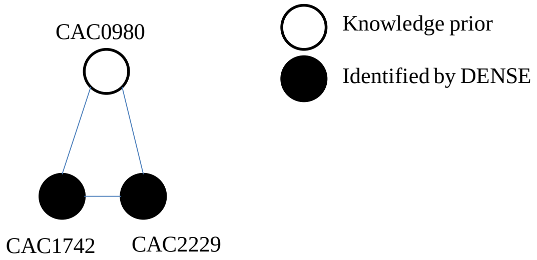


Figure 5.4 DENSE cluster containing pyruvate-ferredoxin oxidoreductase and interacting proteins identified by DENSE algorithm. White circle represents the knowledge prior utilized to identify the dense, enriched cluster.

conversion of acetyl-CoA to acetyl-phosphate [8]. Interactions between phosphotransacetylase and PFOR are consistent with known biochemical data. Although the presence of PFOR and PFL has been described in Clostridium, the direct interaction between the two enzymes is not well known. In *C. acetobutylicum,* PFOR is involved in the pathway for acetyl-CoA and acetogenesis [8]. However PFL, if utilized, may be involved in production of other products, such as solvents, through alternative pathways. Further analysis of the role of PFL in *C. acetobutylicum* is necessary to identify its functional role.

5.3.1.1.2 Butyrate Kinase and Associated Proteins

During dark fermentative hydrogen reactions, such as those that occur in anaerobic wastewater reactors, acetic acid and butyric acid are the two metabolites sought after by scientists and engineers. One reason for this is that through production of these two metabolites hydrogen gas is also co-evolved as a by-product. Therefore, through production or absence of acetate or butyrate by microorganisms, scientists can verify if metabolic fluxes are directed toward hydrogen production rather than hydrogen consumption. As such, understanding the mechanisms involved in production of acetic acid (acetate) or butyric acid (butyrate) is important for enhancing hydrogen production yields.

In this study, application of the DENSE algorithm resulted in identification of a number of clusters including proteins involved in acetate and butyrate formation. From the results, one cluster that contained butyrate kinase, a key enzyme in butyrate formation was identified. Within this cluster, two butyrate kinase proteins (CAC1660 and CAC3075) and one phosphate butyryltransferase (CAC3076) protein are predicted as completely interacting with one another (Figure 5.5; Table 5.4). Such interactions between these two proteins are consistent with known biochemical data regarding butyrate formation [8]. In these studies, both butyrate kinase and phosphate butyryltransferase (Ptb) are described as essential for production of butyric acid [114].

While interactions between the proteins do not appear to be trivial, it is important to note the involvement of Ptb in regulation of metabolic shifts between butyrate and butanol formation. In *C. acetobutylicum*, the switch between acidogenesis and solventogenesis has been shown to occur after formation of butyanol-CoA. In studies

Table 5.4 Protein-protein interaction network corresponding to Figure 5.5 and description of butyrate kinase and associated proteins present in *Clostridium acetobutylicum*.

STRING ID	Protein ID	Protein Description
CAC3076	Ptb	Phosphate butyryltransferase
CAC1660	Buk	Butyrate kinase, BUK
CAC3075	Buk	Butyrate kinase, BUK

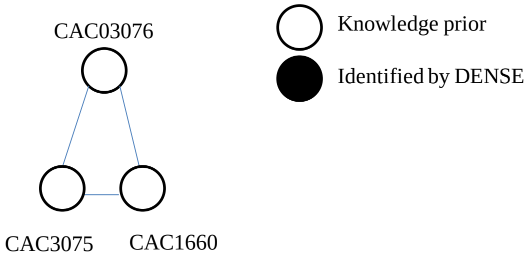


Figure 5.5 DENSE cluster containing butyrate kinase enzymes and phosphate butyryltransferase identified by DENSE algorithm. White circle represents the knowledge prior utilized to identify the dense, enriched cluster.

evaluating activities of the two enzymes, potentially important feedback mechanisms between the activity of Ptb and butyrate formation, and between Ptb and ATP formation were detected [114, 115]. One example of a feedback mechanism is the inhibition of Ptb by ATP during butyrate formation [114]. Based on these flux studies, researchers suggest that Ptb may serve a regulatory role as a signaling protein. When additional interactions between Ptb and other proteins are evaluated, results predicted that Ptb also interacts with two aldehyde dehydrogenases (AdhE2) and acetyl-CoA dehydrogenase. During solvent production, AdhE proteins are responsible for butanol production. Since *C. acetobutylicum* is capable of both solventogenesis and acidogenesis and Ptb is interacting with proteins involved in both butyrate and butanol formation, it can be hypothesized that Ptb is responsible for metabolic shifts involving butyrate fermentation.

5.3.1.2 Acid-Tolerant Interactions in *C. acetobutylicum*

Incorporation of acid-tolerant knowledge priors identified in Chapter 3 to the dark fermentative, acid-tolerant, hydrogen producing bacterium, *Clostridium acetobutylicum* resulted in identification of 889 dense, enriched protein-protein clusters. Due to limitations in identifying a diverse set of completely sequenced organisms, the acid-tolerant proteins incorporated are representative of a small subset of acid-tolerant organisms from the Phylum Firmicutes (9 species) and Proteobacteria (1 species). As such, the clusters identified are based on organisms representative of three classes of bacteria—Bacilli, Clostridia, and α-proteobacteria. Of these clusters, the DENSE algorithm identified 158 as containing proteins involved in a sugar phosphotransferase system (PTS). In organisms, PTS is a system consisting of a number of proteins involved in uptake of sugar (e.g., glucose and fructose) [116]. Each of these proteins are divided into one of two components—E1 and E2. The E1 component consists of two proteins, E1 enzyme and histidine (Hpr), and is responsible for phosphorylation of substrates within the system [116, 117]. The E2 component contains the cytoplasmic proteins, EIIA, EIIB, and EIIC.

In Figure 5.6 and Table 5.5, a densely enriched cluster of PTS proteins identified by DENSE is presented. Proteins involved in this cluster include E1 proteins (CAC0231), EII enzymes (CAC0233 and CAC0234), a transcriptional regulator involved in sugar metabolism (CAC0231), and fructose 1-phosphate kinase (CAC0232). The EII proteins and fructose 1-phosphate kinase are shown to interact with each protein in the cluster. Whereas the transcriptional regulator and EI protein are the only two proteins that do not directly interact. This suggests that the transcriptional regulator is likely

Table 5.5 List of acid-tolerant protein cluster identified by the DENSE algorithm and protein descriptions. Proteins listed correspond to Figure 5.6.

STRING ID	Protein ID	Protein Description
CAC0233	-	PTS system, IIA component
CAC0231	-	Transcriptional regulator of sugar metabolism
CAC3087	-	Phosphoenolpyruvate-protein kinase (PTS system enzyme I)
CAC0232	-	1-phosphofructokinase (fructose 1-phosphate kinase)
CAC0234	-	PTS system fructose-specific IIBC component

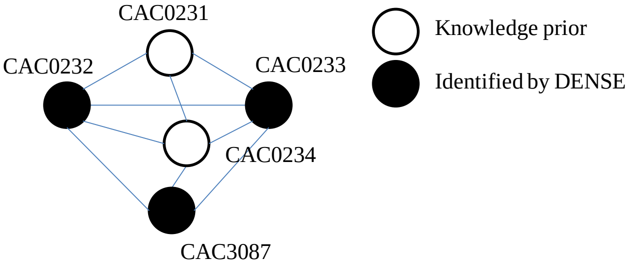


Figure 5.6 DENSE cluster containing phosphotransferase system (PTS) enzymes identified by DENSE algorithm. White circle represents the knowledge prior utilized to identify the dense, enriched cluster.

involved in controlling the interactions between the cytoplasmic proteins in PTS and fructose 1-phosphate kinase. In organisms, fructose 1-phosphate kinase is responsible for conversion of D fructose 1-phsophate to fructose 1,6 biphosphate [116]. Thus, the regulator may play a role in regulating sugar metabolism in *C. acetobutylicum.*

While PTS and sugar metabolism is thought of as involved in acid tolerance, literature reports for acid response mechanisms in *Escherichia coli* and *Streptococcus sobrinus* found that proteins associated with PTS were upregulated during growth at low pH (pH < 6.0) [117, 118]. In a study by Nasciemento et al. [117], PTS activity was shown to be upregulated in *S. sobrinus* when cells were exposed to a pH of 5.0.

However, they found the opposite to be true for *Streptococcus mutans,* with PTS activity decreasing by half when exposed to a pH of 5.0. For *E. coli,* Blankenhorn et al. [118]showed the phosphocarrier protein PtsH and the protein N(pi) phosphohistidine—sugar phosphotransferase (ManX) were induced by *E.coli* during acid stress.

While there is no consistent reaction to acid stress by organisms regarding sugar metabolism and PTS, it does appear that PTS in *C. acetobutylicum* is regulated by a transcriptional factor. Since hydrogen production studies often rely on utilization of glucose (and fructose) as their carbon source, understanding the metabolic response to acid is important. As such, studies evaluating the role of the transcription regulator (CAC0231) on PTS and sugar metabolism in *C. acetobutylicu*, under varying pH conditions are necessary.

5.3.2 The Bi-Clustering Algorithm for Discovery of Cross-Talk

Application of the clustering algorithm resulted in identification of 8 COG clusters associated with organisms capable of light fermentation and bio-photolysis, and 28 COG clusters associated with dark fermenting organisms. Initial review of each light fermentation cluster shows the presence of a set of 13 identical genes found across all 8 COG clusters. These "core" COGs are all observed in Cluster 1 and include genes necessary for synthesis of hydrogenase complex(es). However, for the dark fermentation clusters, we did not observe a large set of COGs present across each cluster. For this set of organisms, only two COGs were identified as present across all clusters. This may be partially due to two reasons. First, the selection of species and their diversity has some impact on the types of clusters generated. Second, dark fermentation organism tend to utilize a greater variety of fermentation pathways than the previous group, such as acetate fermentation and butyrate fermentation pathways [112]. Greater variation in fermentation routes will not produce as large of a “core” set of COGs across each cluster.

Due to limitations in space, the only item discussed is the overall characterization of COGs present in cluster 5 and cluster 14 for light and dark fermentation, respectively. In addition, computational predictions on the cross-talk between the metabolic proteins in the conserved network modules will be discussed in the following sections. Clusters described in this study were selected based on whether they show greater variation in the presence of unique COGs and contained the “core” set of COGs described above. Functional associations between COG groups present in each cluster are validated through literature review and prior knowledge

5.3.2.1 Discovery of Phenotype-Related Proteins Using Knowledge Priors

The overall species composition of clusters containing COGS associated with light fermentation include species representative of light fermentative, hydrogen-producing bacteria, such as *Anabaena* sp. PCC 7120 and *Rhodospeudomonas palustris* [6, 17]. Within the functional modules identified by the bi-clustering algorithm, a number of proteins associated with nitrogen fixation and ferric iron regulation were identified. Proteins associated with nitrogen fixation include ferric iron regulation proteins (*sigK*, *clpB*,and *fur*-related), ammonia ligase (*glnA*), and nitrogenase (*nifH*) [5]. An example of a group of COGs identified is presented in Table 5.6. Clustering of these groups of proteins together suggests iron metabolism and ammonia metabolism may play an important role in regulating nitrogen-fixation.

Nitrogen-fixation is the process which nitrogenase catalyzes the conversion of nitrogen gas to ammonia and inadvertently results in the production of hydrogen gas as a

Table 5.6 The presence (1) or absence (0) of COGs for light fermentation results in one cluster identified by the bi-clustering algorithm. Organisms: *Anabaena variabilis* (ava), *Anabaena* (Nostoc) sp. PCC 7120 (ana), *Rhodobacter sphaeroides* (rsk), *Rhodpseudomonas palustris* (rpa), and *Rhodospirillum rubrum* (rru).

COG_ID	COG_Description	ava	ana	rsk	rpa	rru
COG0068	Hydrogenase maturation factor	1	1	1	1	1
COG0298	Hydrogenase maturation factor	1	1	1	0	1
COG0309	Hydrogenase maturation factor	1	1	1	1	1
COG0374	Ni,Fe-hydrogenase I large subunit	1	1	1	1	1
COG0375	Zn finger protein HypA/HybF (possibly regulating hydrogenase expression)	1	1	1	1	1
COG0378	Ni2+-binding GTPase involved in regulation of expression and maturation of urease and hydrogenase	1	1	0	1	1
COG0409	Hydrogenase maturation factor	1	1	1	1	1
COG0680	Ni,Fe-hydrogenase maturation factor	1	1	1	1	1
COG1740	Ni,Fe-hydrogenase I small subunit	1	1	1	1	1
COG0174	Glutamine synthetase	1	1	1	1	1
COG0535	Predicted Fe-S oxidoreductases	1	1	1	1	1
COG0716	Flavodoxins	1	1	0	1	1
COG1348	Nitrogenase subunit NifH (ATPase)	1	1	1	1	1
COG2082	Precorrin isomerase	1	1	1	1	1
COG2710	Nitrogenase molybdenum-iron protein, alpha and beta chains	1	1	1	1	1
COG2370	Hydrogenase/urease accessory protein	1	1	0	0	0
COG1941	Coenzyme F420-reducing hydrogenase, gamma subunit	1	1	0	0	0
COG3259	Coenzyme F420-reducing hydrogenase, alpha subunit	1	1	0	0	0
COG0735	Fe2+Zn2+ uptake regulation proteins	1	1	1	1	1

byproduct [6, 39]. In this study,two COG groups (COG 2710 and COG 1348), whose functions are associated with expression of two key protein, nitrogenase iron protein (NifH) and molybdenum iron protein [6], were present across all the clusters, including the one in Table 5.6. Although the presence of these two proteins is essential for nitrogen-fixation to be carried out by light fermenting microorganisms, expression of various metabolic genes play an important role in either directly or indirectly regulating the expression of genes encoding NifH proteins.

In previous studies by Lopez-Gollomon [17], the nitrogen regulator protein NtcA was found to work together with the iron-uptake protein, Fur, to co-regulate genes involved in various metabolic functions. Metabolic functions co-regulated include the transcriptional regulation protein and the glutamine synthesis protein [4]. In this study, genes responsible for encoding iron uptake regulator proteins (COG 0735) were clustered together with genes encoding glutamine synthetase. Although nitrogen regulatory genes were not identified in our cluster, the co-appearance of both iron-uptake proteins and glutamine synthetase suggests cross-talk between iron uptake and ammonia assimilation networks may be occurring.

Cross-talk between iron and nitrogen-related metabolic networks is also noted through indirect regulation of nitrogenase by iron uptake proteins. One example of cross-talk is in the nitrogen-fixing cyanobacterium *Anabaena.* In *Anabaena*, iron uptake proteins and some nitrogen proteins (e.g., Ntc) have been shown to regulate genes encoding glutamate synthetase (glnA), an enzyme involved in ammonia assimilation [17]. Review of the role of glutamine synthetase in *Anabaena* indicates that this enzyme is responsible for regulating nitrogenase activity, thus impacting hydrogen production [17].

In this study, close review of genes present in *Anabaena* show glutamate ammonia ligase (*glnA*), a key gene for nitrogenase (*nifH*), and genes encoding proteins for iron uptake, are assembled in the same cluster. Clustering of genes involved in nitrogen and iron metabolism suggest these metabolic networks are necessary for carrying out nitrogen fixation and hence biological hydrogen production.

In addition to nitrogenase, proteins associated with the synthesis of uptake or expression of hydrogenase were identified in 11 of the 19 COGs present in cluster 5. Hydrogen uptake proteins help in removing excess hydrogen to maintain the reducing environment in cells [57]. In this study, we identified a number of proteins (e.g., Hyd and Hyp) involved in formation of [NiFe]-uptake hydrogenases. The presence of maturation hydrogenase factors and accessory proteins for uptake of nickel and expression are consistent with literature reports describing the structure of hydrogenase complexes. Inclusion of hydrogenase proteins in cluster 5 is likely due to the relationship of hydrogenase proteins with iron uptake genes. To function properly, iron is needed to form the NiFe center present in the large hydrogenase subunit (HupL) [47]. As such, hydrogenase maturation is dependent on interactions with iron proteins. In addition, there is indication that hydrogenase proteins, such as HupUV, are involved in regulating the glutamine synthetase gene, *glnAII*, in some organisms [6, 119].

5.3.2.2 Dark Fermentation

Similar to light fermentation, cross-talk between metabolic pathways were identified in COG clusters in dark fermentative bacteria. An example of COG clusters identified is present in Table 5.7. In this cluster, it was identified that 13 different COG groups consisting of proteins that are either directly or indirectly responsible for the

Table 5.7 The presence (1) or absence (0) of COGs for dark fermentation results in one cluster identified by the bi-clustering algorithm**.** Organisms: *Bacillus licheniformis* (bli), *Clostridium acetobutylicum* (cac), *Clostridium beijerinckii* (cbe), *Clostridium perfringens* (cpf), *Caldicellulosiruptor saccharolytics* (csc), *Clostridium thermocellum* (cth), *Escherichia coli* (eco), and *Desulfovibrio vulgaris* subsp. vulgaris Hildenborough (dvu).

COG_ID	COG_Description	bli	cac	cbe	cpf	csc	cth	dvu	eco
COG0298	Hydrogenase maturation factor	0	1	1	0	1	1	0	1
COG0309	Hydrogenase maturation factor	0	1	1	0	1	1	1	0
COG0374	Ni,Fe-hydrogenase I large subunit	0	1	1	0	0	0	1	1
COG0409	Hydrogenase maturation factor	0	1	1	0	1	1	1	0
COG0680	Ni,Fe-hydrogenase maturation factor	0	1	1	0	0	0	1	1
COG1740	Ni,Fe-hydrogenase I small subunit	0	1	1	0	0	0	1	1
COG0535	Predicted Fe-S oxidoreductases	1	1	1	0	1	1	1	0
COG1348	Nitrogenase subunit NifH (ATPase)	0	1	1	0	1	1	1	0
COG2710	Nitrogenase molybdenum-iron protein, alpha and beta chains	0	1	1	0	1	1	1	0
COG0716	Flavodoxins	1	1	1	1	1	1	1	1
COG0735	Fe2+/Zn2+ uptake regulation proteins	1	1	1	1	1	1	1	1
COG2082	Precorrin isomerase	0	1	1	1	0	0	1	0
COG3968	Uncharacterized protein related to glutamine synthetase	0	1	1	0	0	1	1	0

uptake or production of hydrogen, were present. Of these groups, 7 are related to the synthesis or expression of the [NiFe]-hydrogenase, an enzyme involved in either the oxidation or synthesis of hydrogen in microorganisms [119]. The other COG groups present in Cluster 14 are involved in nitrogen and iron metabolic networks. Proteins present in these networks include nitrogenase, iron uptake proteins, ammonia assimilation proteins and proteins involved in electron transfer.

In bacteria, such as *Clostridia acetobutylicum*, capable of dark fermentation, there are three main types of hydrogenase enzymes which may be present. These include the [NiFe]-hydrogenase, [FeFe]-hydrogenase, and non-metal containing hydrogenase enzymes [47]. In this study, analysis of the genes present in COG groups associated with hydrogenase maturation enzymes, indicate [NiFe]-hydrogenase is present in 6 of the 8 species. [NiFe]-hydrogenase is generally associated with oxidation of hydrogen, but in *Escherichia coli*, [NiFe]-hydrogenase (hydrogenase 3) has been shown to evolve hydrogen [119]. From review of the results, the function of hydrogenase is unclear without molecular analysis. However, based on the genes present (e.g., *hybG, hupS*), it can be predicted that the hydrogenase enzymes are associated with hydrogen uptake. This is based on previous findings by Butland et al. [49] that show HybG proteins are not involved in formation of hydrogenase 3 and the presence of proteins typically associated with hydrogen uptake (HypE, HypD, HupS, HupD) [50, 120]. In addition to hydrogenase maturation and expression proteins, Fe-S oxidoreductases were identified. These enzymes form the metal center of hydrogenase and are necessary for synthesis of the protein complex [50].

Similar to the findings for light fermentation, Table 5.7 contains proteins associated with iron uptake and regulation (Fur (COG0735) and glutamine synthetase (COG3968)). As part of the structure of [NiFe]-hydrogenase, Fe-S metal centers are located on the small subunit of the hydrogenase complex [47, 119]. Due to the presence of Fe-S metal centers in the overall structure of an active hydrogenase, it is expected that iron uptake proteins be associated in synthesis of hydrogenase protein complexes and clustered together with hydrogenase-related genes. However, the presence of iron uptake protein may be due to the presence of nitrogenase proteins in some of the dark fermenting organisms [121]. In this cluster, iron uptake proteins may be involved indirectly in nitrogen metabolism through regulation of nitrogenase and maintaining the reducing environment in the cell through hydrogen uptake (hydrogenase) [4,7].

Although iron uptake and regulation has been shown to cross-talk with genes involved in nitrogen metabolism to regulate ammonia assimilation [39], the presence of an uncharacterized glutamine synthetase protein in Table 5.7 is unclear. Regulation of glutamine synthetase by hydrogenase accessory proteins (HupUV) have been demonstrated in light fermenting organisms, such as *R. palustris* [39]. However, to the best of our knowledge, this relationship has not been described in dark fermentation organisms. In our study, the gene encoding the uncharacterized glutamine synthetase proteins was only present in a few species, including *Clostridium acetobutylicum and Clostridium beijerinckii.*

One reason the uncharacterized glutamine synthetase protein may be present in the COG cluster is due to its association with nitrogenase proteins. In our results, nitrogenase was present in 4 of the 8 dark fermenting species. This is due to the presence of nitrogen-fixing, dark fermenting bacteria, such as *C. acetobutylicum* and *C. beijerinckii,* in our analysis. In a study by Chen and Kasap [121], both nitrogenase (Nif) and nitrogen protein (glnB-like gene 1 and 2) were characterized in *C. acetobutylicum and Clostridium pasteurianum.* The gene *glnB*, identified in the study by Chen and Kasap, is thought to be related to a nitrogen regulatory protein identified in *E. coli.* Due to the location of the *glnB* gene in *C. acetobutylicum,* the gln-like protein may indirectly regulate nitrogen-fixation in these organisms. However, molecular analysis is needed to identify the function of the uncharacterized glutamine synthetase protein.

5.3.3 *α,β*-Motifs for Identification of Phenotype-Related Functional Modules

Application of the clustering algorithm resulted in identification of 161 and 71 sets of COGs predicted to be associated with organisms capable of biological hydrogen production and acid tolerance, respectively. Within these set of COGs, enzymes known to be involved in direct production of hydrogen production and acid-tolerance were identified. Enzymes predicted to be involved in light and dark fermentative organisms include nitrogenase, pyruvate: ferredoxin oxidoreductase, and hydrogenase maturation enzymes. Enzymes predicted as acid-tolerant include those responsible for amino acid transport (e.g. argininosuccinate lyase) and nucleotide transport (e.g. orotidine-5'-phosphate decarboxylase).

5.3.3.1 Hydrogenase Related Functional Modules

Four types of COGs for maturation of [NiFe]-hydrogenase noted as present in hydrogen producing organisms and absent in non-hydrogen producing organisms were identified. Proteins associated with these COG groups are HypC (COG0298), HypD (COG0409), HypE (COG0309), and HypF (COG00068) (Figure 5.7; Table 5.8). In model organisms, such as *Escherichia coli*, Hyp*C*DEF proteins are described as regulators for maturation of uptake hydrogenase through participation in development of the active center [47, 49]. Regulation is conducted through the requirement of insertion of Fe, Ni, and diatomic ligands by HypA-F proteins into the hydrogenase center for activation and maturation [43]. In this process, HypE and HypF are responsible for synthesis and insertion of Fe cyanide ligands into the hydrogenase's metal center. However, to carry out this process, HypC and HypD must form a complex for construction of the cyanide ligands to occur [49, 109]. Based on published studies of crystal structures on hydrogenase maturation proteins, we know the presence and coordinated interaction between the proteins is essential for synthesis of [NiFe]-hydrogenase. In this study, we found similar evidence of functional associations between HypCDEF proteins. This is shown in one of the COG-COG networks we identified as associated with hydrogenase. In this network, our algorithm predicted each Hyp protein to be associated with one another (Figure 5.7).

Table 5.8 Genes and COG groups identified by the *α,β*-clique algorithm for sets of hydrogen and non-hydrogen producing organisms corresponding to Figure 5.5.

			Number of organisms	
COG ID	**General Description**	**Gene association**	**Hydrogen-producing**	**Non Hydrogen-producing**
COG0068	Hydrogenase maturation factor	hypF	7	0
COG0298	Hydrogenase maturation factor	hypC	7	0
COG0309	Hydrogenase maturation factor	hypE	7	0
COG0409	Hydrogenase maturation factor	hypD	7	0

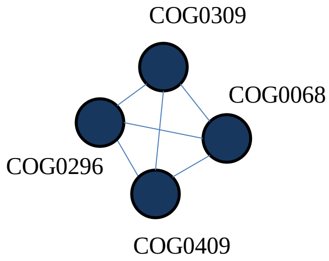


Figure 5.7 Functional module containing hydrogenase enzymes identified by the α,β-clique algorithm.

While associations between maturation proteins have been well characterized in model organisms [49, 122], detailed molecular analysis of [NiFe]-hydrogenase structures and their associated proteins has not been conducted across all phenotype-expressing organisms. Based on the functional association network shown in Figure 5.7 and predicted associations with phenotype and non-phenotype expressing organisms, it can be hypothesized that HypCDEF proteins are related to hydrogen producing organisms and not present in non-phenotype expressing organisms.

In addition to hydrogen maturation proteins, the α,β motifs algorithm was able to identify two COG groups (COG1348 and COG 2710) whose functions were associated with expression of the nitrogen iron protein (NifH) and the molybdenum iron protein (NifD) (Figure 5.8; Table 5.9) [6]. Together these proteins comprise two essential components of nitrogenase, a key enzyme in nitrogen-fixation (N-fixation) [123]. During nitrogen-fixation, nitrogenase catalyzes the conversion of nitrogen gas to ammonia and inadvertently results in the production of hydrogen gas as a byproduct [6, 39]. To carry out this process, NifD serves as the binding site for substrates while NifH assists in biosynthesis of co-factors for NifD [123]. While these proteins are associated with the phenotype nitrogen fixation, results from our algorithm suggest these proteins are highly conserved across various hydrogen producing organisms, thus they may play an indirect role in hydrogen production.

Although the presence of these two proteins is essential for nitrogen-fixation and biological hydrogen production, association of other genes may play an important role in regulating *nif* genes. Examples include proteins such as Cysteine sulfinate desulfinase (COG1104; NifS) and Nitrogen regulatory protein PII (COG0347; GlnK), which are involved in synthesis of the Fe-S cluster and regulation of proteins responsible for nitrogen metabolism [104], respectively. For both GlnK and NifS, the *α,β*-clique algorithm predicted interactions between each COG group and Nif proteins. Specifically, we noted the association of NifH with the regulatory protein PII (GlnK). In nitrogen fixing organisms, GlnK is described as a key signal transducer in NifA in some organisms and regulatory protein in the transcription of the nitrogenase protein NifH in other organisms [124]. In this study, the interaction between COG groups associated

Table 5.9 Set of genes and COG groups associated with nitrogenase formation. COG cluster was identified by the *α,β*-clique algorithm for sets of hydrogen and non-hydrogen producing organisms corresponding to Figure 5.6.

			Number of organisms	
COG ID	**General Description**	**Gene association**	**Hydrogen-producing**	**Non Hydrogen-producing**
COG0388	Predicted amidohydrolase	unknown	9	5
COG0446	Uncharacterized NAD(FAD)-dependent dehydrogenase	HcaD	9	7
COG1063	Threonine dehydrogenase and related Zn-dependent dehdyrogenases	Tdh	9	6
COG1348	Nitrogenase subunit NifH	NifH	7	0
COG2710	Nitrogenase molybdenum-iron protein	NifD	7	0

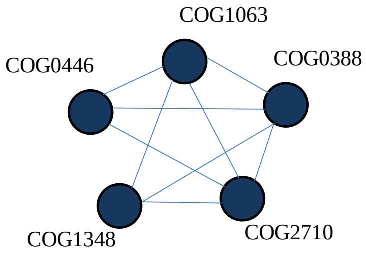


Figure 5.8 Functional module containing nitrogenase enzymes identified by the α,β-clique algorithm.

with NifH and GlnK support experimental evidence that PII proteins are involved in inactivation of nitrogenase across a number of nitrogen-fixing species. In addition, identification of this COG-COG interaction, suggests that PII proteins may play a vital role in hydrogen production via nitrogenase.

5.3.3.2 Hydrogenase Related Functional Modules

When a sub-set of acid-tolerant microorganisms (Phylum Firmicutes and Proteobacteria) were applied to the α,β clique algorithm, the two main mechanisms, lysine/arginine decarboxylase and arginine deaminase, associated with acid-tolerance were not identified. However, eleven types of COGs associated with amino acid transporters were identified, suggesting that amino acid transport is highly related to the set of phenotype-expressing organisms in this study. Within microorganisms, amino acid transporters can participate in a number of metabolic and cellular processes, such as energy metabolism and protein synthesis. In organisms exposed to acid stress, decarboxylation of the two amino acids lysine and arginine is reported as two mechanisms for neutralization of internal pH [66, 74, 75]. During the neutralization process via arginine decarboxylation, antiporters that are responsible for replacing the argmatine generated from arginine with another arginine are brought in from the surrounding environment [75]. In another system, arginine deaminase, ammonia is generated to help protect against acid stress [66]. From our knowledge of these systems, production or uptake of amino acids by microorganisms may play an important role in regulating intracellular pH levels.

In this study, eleven COG groups for amino acid transport were predicted as present across 10 acid-tolerant microorganisms. Proteins associated with these COG groups include argininosuccinate lyase (COB0165; ArgH) and the amino acid transporter LysP (COG0833) (Table 5.10 and Figure 5.9). Argininosuccinate lyase is responsible degrading argininosuccinate to form arginine and fumarate. LysP amino acid transporter is a permease system used by some microorganisms to transport lysine into cells [125].

Table 5.10 Set of genes and COG groups associated with amino acid transport**.** COG cluster was identified by the *α,β*-clique algorithm for sets of acid-tolerant and non-acid-tolerant organisms corresponding to Figure 5.6.

			Number of organisms	
COG ID	**General Description**	**Gene association**	**Hydrogen-producing**	**Non Hydrogen-producing**
COG0165	Argininosuccinate lyase	ArgH	7	5
COG0833	Amino acid transporter	LysP	8	0

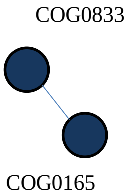


Figure 5.9 Functional module containing enzymes involved in amino acid transport.

Similar to arginine, decarboxylation of lysine has been linked to acid response by some bacteria [126]. While the transport of lysine by the LysP amino acid transporter system is not inhibited by arginine, arginine has been reported to regulate utilization of lysine by the lysine decarboxylation pathways [126]. While the direct interaction between lysine transport and lysine production is not clear, results suggest there is some regulatory control occurring between these two systems.

5.4 Discussion

Application of the three complementary methods resulted in the identification of phenotype-related subsystems. Through co-development and validation of the DENSE algorithm, the functional roles of an uncharacterized protein associated with maturation and activation of hydrogenase accessories proteins in *C. acetobutylicum* were identified. In terms of metabolic reconstructions and engineering, clues toward identifying the functional roles of genes are important for improving the annotation of genomic sequences. Additionally, through analysis of functional modules (α,β motif algorithm) and COG clusters (bi-clustering algorithm), phenotype-related regulators and transporters were identified.

While the main focus of this study was to identify phenotype-related networks, application of the bi-clustering approach resulted in discovery of interactions or "cross-talk" between metabolic pathways and metabolic subsystems. For example, in 5.3.2, the bi-clustering algorithm was capable of identifying COG clusters containing groups of functionally associated proteins in sets of organisms capable of expressing light fermentation. One interesting observation is the presence of genes known to be involved in regulatory networks of individual organisms. This is evident in the presence of genes directly and indirectly involved in regulating nitrogen fixation in light fermenting bacteria. Based on the known function of each gene clustered together, one can observe potential cross-talk between metabolic networks. However, in the dark fermentation results, the exact measure genes involved in nitrogen metabolism are interplaying with iron-uptake genes and/or genes encoding hydrogenase is unclear.

In all organisms, metabolic networks must communicate with each other to carry out necessary functions for survival. Communication in microbial cells is often conducted through responses by regulatory or signaling proteins present in a number of metabolic pathways. Based on the light fermentation results and prior knowledge from literature, the algorithm is capable of identifying functional modules and predicting cross-talk across metabolic pathways associated with these modules. In this study, potential cross-talk between genes involved in nitrogen metabolism (e.g., uncharacterized glutamine synthetase and formation of nitrogenase) and genes involved in hydrogen uptake or electron transfer were predicted. As such, further experimental studies are necessary to identify how the pathways are communicating and how cross-talk could ultimately impact expression of targeted phenotypes.

5.4.1 Diversity of Organisms

One potential limitation for the methodologies used in this study is the selection and diversity of the phenotype-related organisms. For example, the DENSE algorithm utilizes knowledge priors or phenotype-related genes identified in Chapter 3. Therefore, the knowledge priors correspond to the set of phenotype-related organisms. In the case of acid-tolerance, the acid-tolerant knowledge priors were representative of 10 microorganisms consisting predominately of Firmicutes. To identify protein interactions and subsystems related to acid-tolerance, a more diverse group of organisms is necessary. Similarly, data related to acid-tolerance for the α,β-clique algorithm and biclustering algorithm are not truly representative of the phenotype, but rather a small set of organisms capable of acid-tolerance.

5.5 Approach

5.5.1 DENSE: Dense and Enriched Sub-Graph Enumeration Algorithm

Given a phenotype-expressing organism, the DENSE algorithm tackles the problem of identifying genes that are functionally associated to a set of known phenotype-related proteins by enumerating the “dense and enriched” subgraphs in genome-scale networks of functionally associated or interacting proteins (Hendrix et al. 2010). A “dense” subgraph is defined as one in which every vertex is adjacent to at least some γ percentage of the other vertices in the subgraph for some value γ above 50%, which correspond to a set of genes with many strong functional associations between them. In order to incorporate researchers' *prior* knowledge, though, this definition is extended by introducing an "enriched" dense subgraph in which at least μ percentage of the vertices are contained in the knowledge prior query set, for some parameter value μ (Figure 5.10). Genes contained in such dense and enriched subgraphs, or "μ-enriched, γ-dense quasi-cliques," have strong functional relationships with the previously identified genes, and are likely to perform a related task. The assumption behind this is that genes encode for proteins that belong to different metabolic and regulatory pathways and those pathways may result in the representation of a phenotypic trait. If a pathway or network motif is functionally associated for a given phenotype, then it is enriched by the phenotype-related genes.

DENSE uses an agglomerative approach to discover these μ-enriched, γ-dense quasi-cliques, starting with a single query vertex v_0, adding one vertex to the quasi-clique at a time, and backtracking as it finds maximal enriched quasi-cliques or subgraphs that cannot be contained in an enriched quasi-clique. Based on our theoretical results, vertices that can be added to the subgraph in order to reduce the necessary search space are limited. A number of heuristic values are maintained, and when any of the heuristic values is reduced to zero, the related vertex becomes a "dead end" that is not added to the subgraph.

5.5.2 Bi-Clustering Network Motifs

To identify network motifs conserved across sets of phenotype-expressing organisms (e.g., dark fermentative, hydrogen producing), a bi-clustering based algorithm was co-developed and implemented. In this approach, three steps are required. During

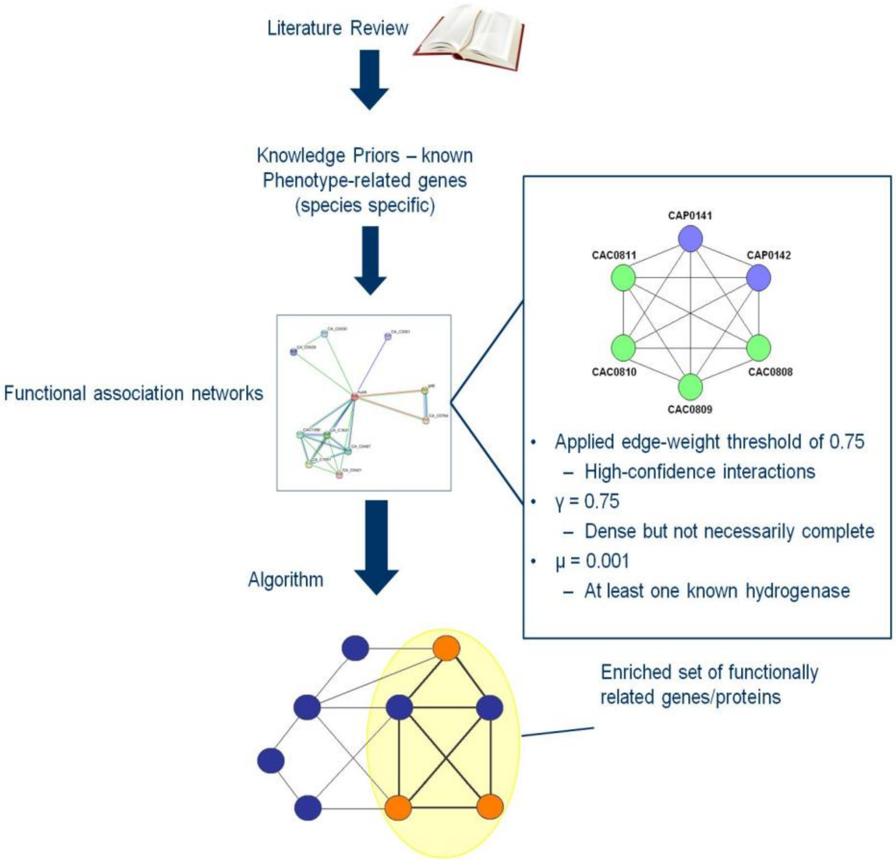

Literature Review
Knowledge Priors – known Phenotype-related genes (species specific)
Functional association networks
Algorithm
CAP0141
CAC0811
CAP0142
CAC0810
CAC0808
CAC0809
• Applied edge-weight threshold of 0.75
– High-confidence interactions
• γ = 0.75
– Dense but not necessarily complete
• μ = 0.001
– At least one known hydrogenase
Enriched set of functionally related genes/proteins

Figure 5.10 Schematic overview of network-based, small-scale computational approach based on utilization of knowledge priors.

the first step, proteins from common orthologous groups (COGs) were identified using entire sets of microorganisms exhibiting the target phenotype. In this study two phenotypes, light fermentation and dark fermentation, were evaluated to provide comparison on the biological methodologies. Complete lists of metabolic genes for each organism were obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathway database and interacting networks of COGs for each organism constructed. Edges in the network represented known functional associations according to the Search Tool for Retrieval of Interacting Genes/Proteins (STRING) database [127].

For each COG pair, the STRING similarity threshold was set to 0.85 to generate a list of COG-COG edges. Next, a matrix was composed where the rows represent organisms, the columns represent COG-COG edges, and the entries in the matrix represent the presence or absence of an edge within the given organism, based on the species similarity threshold. The matrix was pre-processed to remove COG edges that occur in only one organism. This matrix can be viewed as a bipartite graph, where one set contains the COG-COG interactions (or edges in the original organism graphs) and the other contains only organisms. Information obtained was validated based on known literature values and review. A more detailed description of this approach is provided in Chapter 6 of this dissertation.

5.5.3 α,β-Clique Algorithm

Utilizing a set of phenotype expressing organisms, the α,β-clique algorithm aims at identifying functional modules through comparison of hundreds of genome-scale networks of functionally associated proteins. In addition to identification of these subsystems, α,β-clique algorithm also allows for identification of potential regulators, signaling proteins, transporters, and uncharacterized proteins related to organisms expressing the target phenotype. Identification of phenotype-related modules is based on the assumption that a functional module in a single organism will form a maximal clique in the organism's function association networks (Figure 5.11). Therefore, a conserved functional module will form a maximal clique in the phylogenetic functional association network. Because of this relationship, it is assumed that a conserved functional module is a conserved functional module within a set of organisms.

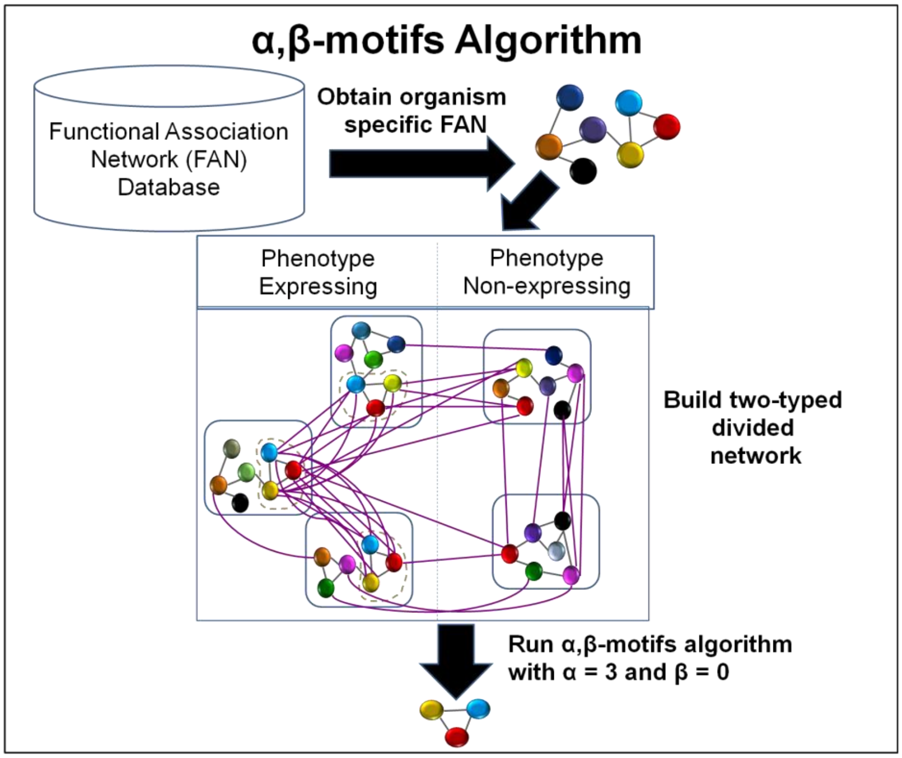

α,β-motifs Algorithm
Functional Association Network (FAN) Database
Obtain organism specific FAN
Phenotype Expressing
Phenotype Non-expressing
Build two-typed divided network
Run α,β-motifs algorithm with α = 3 and β = 0

Figure 5.11 Overview of the α,β-motifs algorithm.

In this approach, the α,β-motifs algorithm identifies conserved functional modules by identifying functional modules that are enriched in at least *α* networks of phenotype-expressing organisms but may still appear in no more than *β* networks of organisms that do not exhibit the target phenotype. Figure 5.11 illustrates how application of the α,β criteria can be used to divide functional association networks a two-typed, divided network. Enumerated α,β-cliques related to phenotype-expressing organisms are then analyzed using sets of known phenotype-related genes and modules. A detailed description of this approach is presented in Chapter 6 of the dissertation.

CHAPTER 6: COMPUTATIONAL METHODOLOGIES

6.1 Overview of Chapter

In this Chapter detailed descriptions of each computational methodology previously described in Chapter 3-5 are presented in order of appearance. In addition, a description of methods for statistical analysis using the Student's T-test and for selection of organisms is included. The descriptions and equations for co-developed computational methodologies are extracted from either published papers or papers in preparation. The submitted computational papers include the NIBBS algorithm[128], multiple alignment algorithm [129], and DENSE algorithm [130].

6.2 Summary of Organism Selection for Methodologies

Since no one microorganism can utilize all potential biomass feedstock sources and organic matter in wastewaters, a combination of anaerobic, fermentative bacteria are often used to carry out dark fermentation processes. While hydrogen producing organisms encompass a wide range of genomes, especially within wastewater and waste materials, the most commonly identified and studied microbial species includes those from the genus Clostridium (e.g. *Clostridium acetobutylicum*) [11, 105]. As such, results generated for each study were evaluated in respect to the model organism *Clostridium acetobutylicum.* In computational studies that evaluate biochemical processes within individual organisms, such as the DENSE and α, β- clique experiments, *Clostridium acetobutylicum* was selected to represent the two phenotypes dark fermentative, hydrogen producing and acid tolerance. This was due to the ability of *C. acetobutylicum* to express both hydrogen production and acid tolerance phenotypes simultaneously.

In studies where multiple phenotype expressing and non-phenotype expressing microorganisms were selected, the organisms selected were identified through extensive literature reviews and from microbial databases, such as the Department of Energy's (DOE's) Joint Genome Institute (JGI) and the National Center for Biotechnology Information (NCBI) database. To ensure all metabolic and cellular components were identified for each organism, all organisms used in the datasets were completely sequenced. Verification of complete genomes was conducted using the NCBI database. In addition, organisms selected consisted of a diverse group to ensure metabolic enzymes and pathways identified were representative of the phenotype and not a specific genera (Figure 6.1).

6.3 Student's T-Test for Identification of Phenotype-Related Genes

To identify input seeds for the algorithm, a statistical comparison of genes present in a group of phenotype expressing organisms (e.g., aerobic, anaerobic) with a group of non-phenotype expressing organisms was conducted. Identification of phenotype-related genes was based on the biological assumption that when comparing genomes and biological networks of organisms, biologically related elements are evolutionarily conserved. In this study, if a metabolic gene is evolutionarily conserved across a group of related microorganisms exhibiting the same phenotypic trait and not conserved in a set of non-phenotype expressing microorganisms, then the gene is likely related to the phenotype.

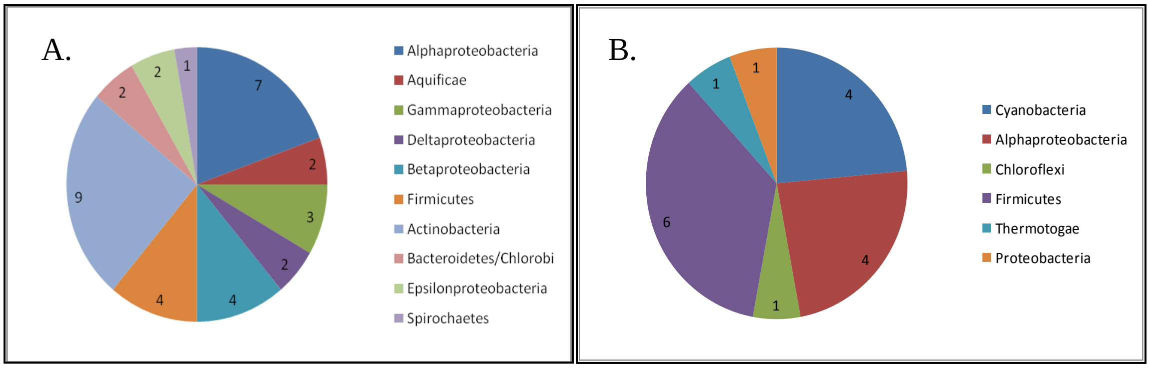


Figure 6.1 Taxomonic diversity of aerobic organisms assessed using the NIBBS algorithm and statistical approach for aerobic respiration (A) and hydrogen producing (B) microorganisms.

Gene phylogenetic profiles (GPPs) were created by obtaining enzymes (E.C. numbers) species maps for each organism from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. GPPs were organized into an enzyme-species distribution table containing species and the total number of enzymes present for each species. Using this approach, one can correlate sets of orthologous genes present in phylogenetic profiles with phenotype-expressing organisms. Based on the ortholog distribution and unequal variance in our dataset, we used a 2-sample T-test to determine which orthologs are related to the target phenotype. The null hypothesis for our study was H_0: $\mu_{aerobic} = \mu_{anaerobic.}$ Because we want to reject the null hypothesis if the p-value is either sufficiently large or small, we calculated the p-value using two-tails.

Based on the probability of a set of organisms, either phenotype expressing and set of non-phenotype expressing organisms, we determined which set were enriched with the gene. If the p-value obtained equals 0.05, as in this study, then there would be a 1 in 20 chance of similarity (5% chance) the ortholog would show preference for one set of organisms over another. Whereas if the p-value is close to 1, then it is likely that the orthologs from both sets of organisms are similar—show no preference for one phenotype over the other. Using this approach we can use the p-values obtained for each gene to identify a set of genes important for the phenotypic trait and a set of genes not important for the phenotypic trait.

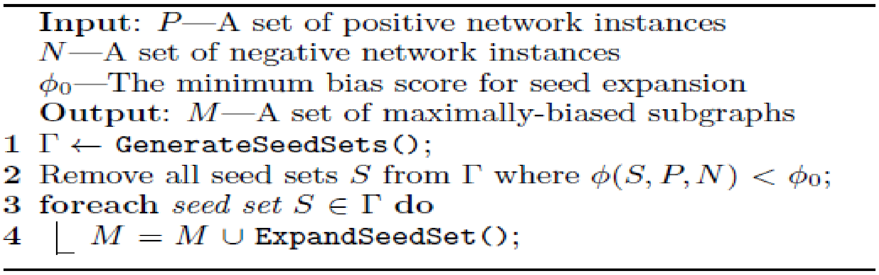
6.4 NIBBS Algorithm

To identify phenotype-related metabolic enzymes and pathways, the NIBBS algorithm was co-developed and applied to the three phenotypes—anaerobic, hydrogen production, and acid tolerant. A general overview of the NIBBS-Search algorithm is given in Algorithm 6.1 below [128]. It is a two-step process that first identifies small seed sets of edges and then expands those sets into the maximally-biased subgraphs.

6.4.1 Seed Set Generation

Informally, seed sets correspond to significant subsets of edges from the network map; they differentiate between common subgraphs that model phenotype-related systems and those that model non-phenotype-related systems, and they improve the NIBBS-Search efficiency by determining the subset of organisms that are predicted to contain the entire phenotype-related system. The motivation behind seed set generation stems from the following observation. The phylogenetic profile of a phenotype-related metabolic system, such as the tricarboxylic acid (TCA) and the reverse TCA (rTCA) cycle is often the same as the phylogenetic profile of a small subset of its constituent enzymes. In other words, this subset defines the set of target organisms that will contain the entire system, and thus reduces the set of network instances that need to be aligned during the expansion process. In addition, it provides hints to the algorithm, that of the many possible common subgraphs that are found when the instances are aligned, only

Algorithm 6.1: NIBBS-Search: Network Instance Based Biased Subgraph Search

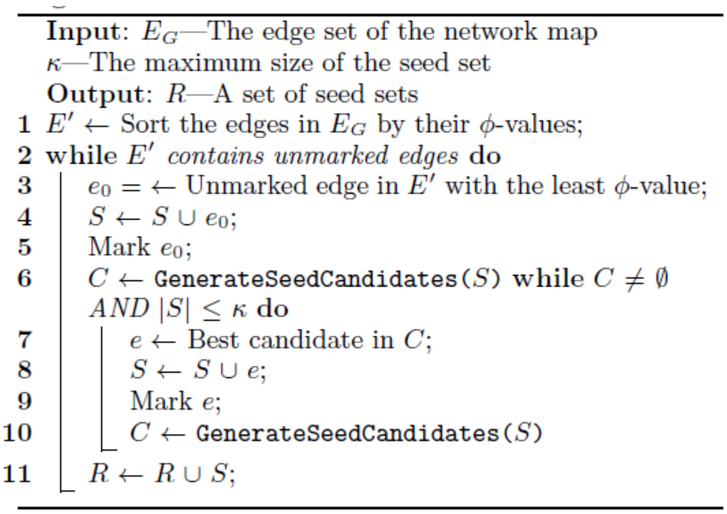


```
Input: $P$—A set of positive network instances
$N$—A set of negative network instances
$\phi_0$—The minimum bias score for seed expansion
Output: $M$—A set of maximally-biased subgraphs
$\Gamma \leftarrow$ GenerateSeedSets();
Remove all seed sets $S$ from $\Gamma$ where $\phi(S, P, N) < \phi_0$;
foreach seed set $S \in \Gamma$ do
    $M = M \cup$ ExpandSeedSet();
```

those that contain the seed set should be predicted to represent phenotype-related systems.

The procedure implemented in the NIBBS-Search algorithm for growing seed sets is provided in Algorithm 6.2. It begins by sorting the set of edges in the network map by their φ-value (Line 1). Then the edge with the least φ-value is used to create a seed set containing only that edge (Line 3). To avoid redundant seed sets, that edge is marked, so it cannot be added to any other seed set (Lines 5 and 9). The algorithm then identifies a set of candidate edges (Line 6), which are the edges whose addition to the seed set decreases its φ-value. Only unmarked edges are considered as possible candidate edges. The algorithm follows a greedy approach by adding the candidate edge that produces the greatest decrease in the seed set's φ -value (Line 7). After an edge is added to the seed set, the set of candidate edges is updated (Line 10). This process continues until the φ -value of the seed set cannot be decreased by adding any candidate edge, or until the seed set reaches a user-defined maximum size (Line 6). The seed set is then added to the set of seed sets, and a new seed set is generated from the unmarked edge that has the least φ -value. This process continues until every edge in the network map is part of a seed set.

Algorithm 6.2: Generate Seed Sets



```
Input: $E_G$—The edge set of the network map
$\kappa$—The maximum size of the seed set
Output: $R$—A set of seed sets
$E' \leftarrow$ Sort the edges in $E_G$ by their $\phi$-values;
while $E'$ contains unmarked edges do
    $e_0 = \leftarrow$ Unmarked edge in $E'$ with the least $\phi$-value;
    $S \leftarrow S \cup e_0$;
    Mark $e_0$;
    $C \leftarrow$ GenerateSeedCandidates($S$) while $C \neq \emptyset$ AND $|S| \leq \kappa$ do
        $e \leftarrow$ Best candidate in $C$;
        $S \leftarrow S \cup e$;
        Mark $e$;
        $C \leftarrow$ GenerateSeedCandidates($S$)
    $R \leftarrow R \cup S$;
```

Two methods of selecting the candidate edges are defined. The first ensures that the seed set forms a connected subgraph. The second does not require that the seed set be connected, but ensures that the seed set be part of a connected to the subgraph after the expansion process. The first method is achieved by only considering edges that are coincident with one of the edges currently in the seed set. The second method considers any edge in the network map as a candidate edge as long as the two edges are connected after the expansion process. To ensure that the two edges are connected, the method determines if a path exists between the edge and one of the edges in the seed set that is present in every positive network instance that the new seed set would be present in.

The user chooses a threshold φ_0 such that only seed sets whose φ -value is less than φ_0 will be expanded into full subgraphs. This allows the user to reduce the number of insignificant subgraphs that are output by the algorithm. Due to the method by which the seed sets are constructed, every edge in the network map will be part of at least one seed set. Some of these seed sets will not have a very low φ-value. Thus, the set of network instances in which these seed sets are present are unlikely to contain a subgraph that models a phenotype-related metabolic system.

6.4.2 Seed Set Expansion

The seed set of edges is unlikely to represent the entire phenotype-related metabolic system. Seed sets are typically small, containing between one and five edges and, depending on the method used to construct them, may form a disconnected subgraph. A metabolic system is likely to form a connected subgraph in a metabolic network containing many more edges [70]. In order to predict the entire set of enzymes belonging to the metabolic system, the NIBBS-Search algorithm expands the seed sets. To ensure that the expansion edges belong to the same metabolic system as the seed edges, the expansion process requires that the expansion edges be present in most if not all of the metabolic networks of phenotype-expressing organisms that also contain the seed edges. The addition of expansion edges to a seed set to form the subgraphs output by the algorithm is called the seed expansion process. During the process, an expansion edge is selected from a set of candidate edges. These candidate edges are determined by two criteria:

- They are coincident with a seed edge or an expansion edge is already in the edge set.
- If added to the current edge set, the resulting edge set will be present in at least α percentage of the positive network instances that the seed set was present in.

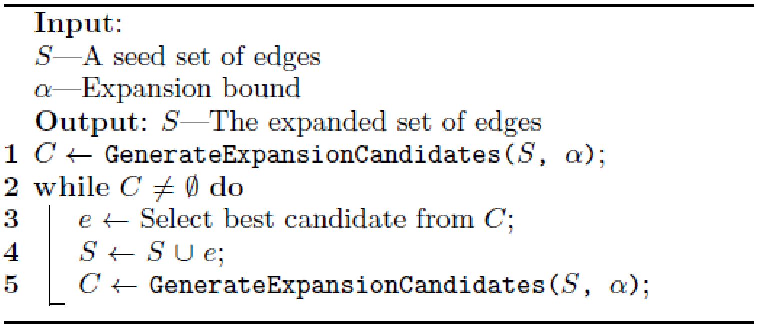
The first criterion ensures that the final edge set will form a connected subgraph. The second criteria allows for noise in the data, while requiring that the final edge set still be present in most if not all of the same positive network instances as the seed set.

The algorithm for expanding the seed set to form the final edge sets is given in Algorithm 6.3. Expansion edges are selected from the set of candidate edges, added to the current edge set, and the set of candidate edges is updated until no candidate edges can be found. The resulting edge set is then output. The order in which candidate edges are added to the edge set will determine the make-up of the output edge set unless $\alpha = 1$. The expansion process determines which candidate edge to add to the edge set by first considering the number of positive network instances that the resulting edge set would be present in. It tries to select the candidate edge that would produce the greatest such number. However, multiple candidate edges may exist that would result in edge sets present in the same number of phenotype-expressing organisms. In this case, the expansion process selects from this set the candidate edge that would produce the greatest decrease in the φ=-value of the edge set. If more than one of these candidates produces the same decrease in the φ-value, then a candidate edge is selected at random from these remaining candidates and added to the edge set.

6.5 Multiple Alignment Algorithm

The multiple alignment algorithm is a heuristic algorithm for the maximum-weighted k-partite matching problem [129]. Note that each matched node set in the solution M produced by the Algorithm 6.4 contains at most one node from the first part U1 and at most r nodes from other parts, but each matched node set of a k-partite matching can contain up to r nodes from each part. In order to allow each matched node

Algorithm 6.3: Expand Seed Sets



```
Input:
$S$—A seed set of edges
$\alpha$—Expansion bound
Output: $S$—The expanded set of edges
$C \leftarrow$ GenerateExpansionCandidates($S$, $\alpha$);
while $C \neq \emptyset$ do
    $e \leftarrow$ Select best candidate from $C$;
    $S \leftarrow S \cup e$;
    $C \leftarrow$ GenerateExpansionCandidates($S$, $\alpha$);
```

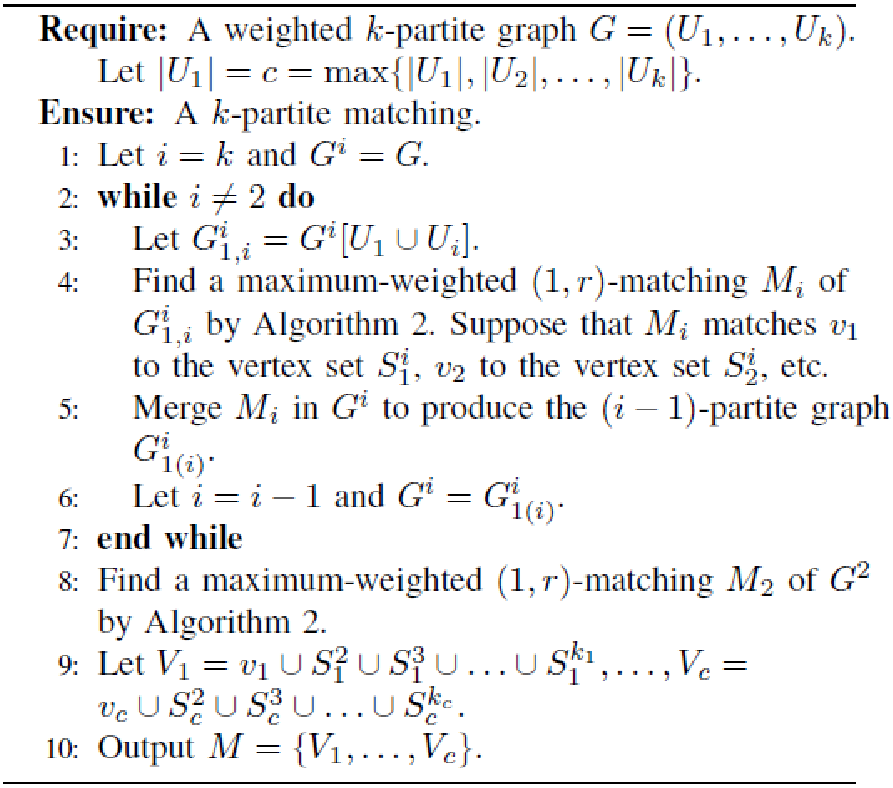
set of the approximation to also contain up to r nodes from U1, a novel heuristic algorithm was co-developed.

Our heuristic algorithm idea is as follows. We first find an (r, r)-matching M for the subgraph of two parts by the greedy algorithm in [131], then perform a merge based on M to get a (k−1)-partite graph G0. Since the definition of merging a (r, r) matching is very similar to the definition of merging a (1, r) matching, we omit it here. Next, we find a matching M0 for the (k − 1)-partite graph G0 using our approximation algorithm, and the matching M and M0 are combined into an approximate solution. The heuristic algorithm for the maximum-weighted k-partite matching problem appears in Algorithm 6.5. In order to explain our algorithm clearly, we reproduced some formatting of the approximation algorithm for the maximum disjoint k-clique problem given in [132].

6.6 DENSE Algorithm

Using the Dense ENriched Subgraph Enumeration (DENSE) algorithm, genes that are functionally associated to a set of known phenotype-related proteins are identified. In this algorithm, a "dense" subgraph is defined as one in which every vertex is adjacent to at least some γ percentage of the other vertices in the subgraph for some value γ above 50%, which correspond to a set of genes with many strong functional associations

Algorithm 6.4: Approximation Algorithm for the Maximum-Weighted *k*-Partite Matching Problem.



```
Require: A weighted $k$-partite graph $G = (U_1, \ldots, U_k)$.
    Let $|U_1| = c = \max\{|U_1|, |U_2|, \ldots, |U_k|\}$.
Ensure: A $k$-partite matching.
Let $i = k$ and $G^i = G$.
while $i \neq 2$ do
    Let $G^i_{1,i} = G^i[U_1 \cup U_i]$.
    Find a maximum-weighted $(1, r)$-matching $M_i$ of
    $G^i_{1,i}$ by Algorithm 2. Suppose that $M_i$ matches $v_1$
    to the vertex set $S^i_1$, $v_2$ to the vertex set $S^i_2$, etc.
    Merge $M_i$ in $G^i$ to produce the $(i-1)$-partite graph
    $G^i_{1(i)}$.
    Let $i = i - 1$ and $G^i = G^i_{1(i)}$.
end while
Find a maximum-weighted $(1, r)$-matching $M_2$ of $G^2$
by Algorithm 2.
Let $V_1 = v_1 \cup S^2_1 \cup S^3_1 \cup \ldots \cup S^{k_1}_1, \ldots, V_c =$
$v_c \cup S^2_c \cup S^3_c \cup \ldots \cup S^{k_c}_c$.
Output $M = \{V_1, \ldots, V_c\}$.
```

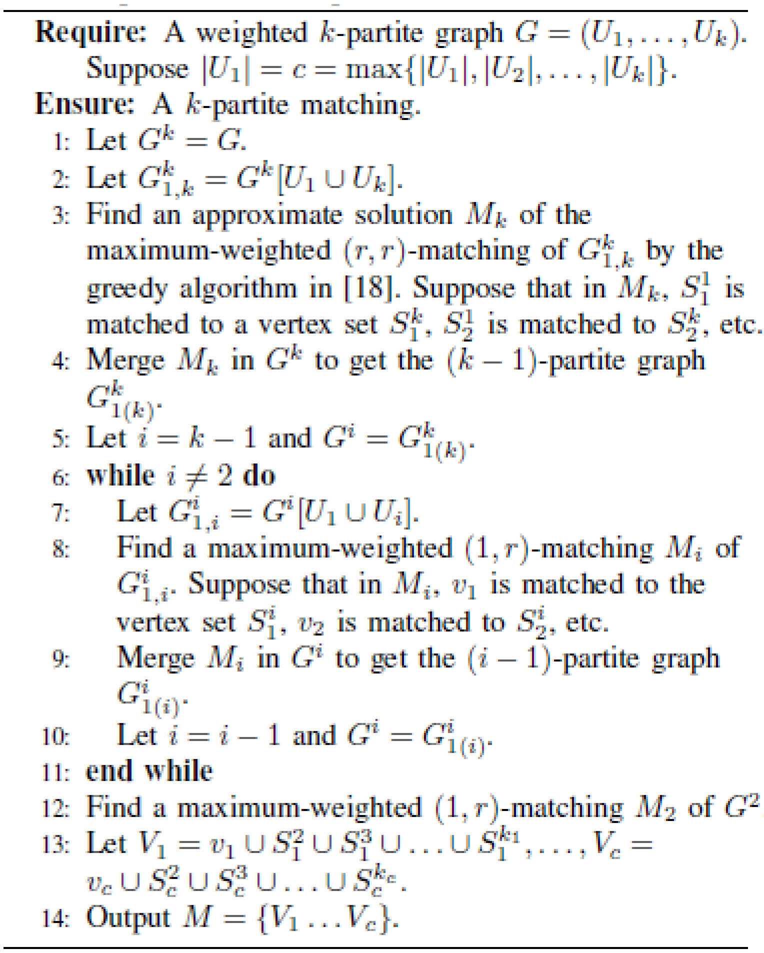
between them [130]. Formal definitions of what it means for a subgraph to be μ-enriched and a γ -dense quasi-clique are provided below.

Definition 1: Given a labeled graph G and a real value γ ∈ (0:5; 1], a subgraph S of G is a γ -dense quasi-clique if and only if every vertex of S is adjacent to at least γ ($|S|$-1) of the other vertices of S.

Definition 2: Given a labeled graph G, a “query” set of vertices Q, a real value γ ∈ [0:5; 1], and a real value μ ∈[0; 1], a γ -dense quasi-clique S is μ -enriched with respect to Q if and only if at least γ$|S|$vertices of S are contained in Q.

As several previous dense subgraph enumeration algorithms have successfully employed an agglomerative (bottom-up) approach, we focused our efforts on developing

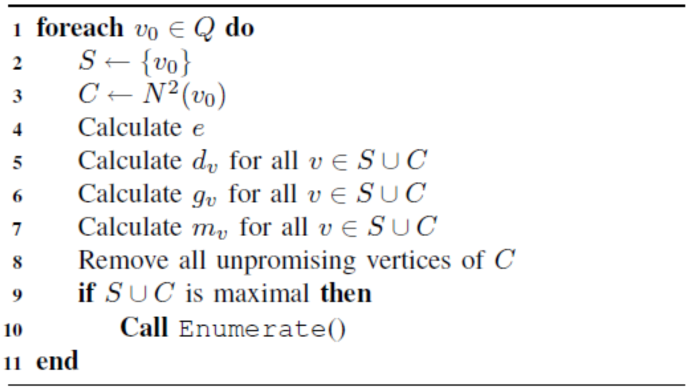
Algorithm 6.5: The Heuristic algorithm A3 for the Maximum *k*-Partite Matching Problem.



```
Require: A weighted $k$-partite graph $G = (U_1, \ldots, U_k)$.
    Suppose $|U_1| = c = \max\{|U_1|, |U_2|, \ldots, |U_k|\}$.
Ensure: A $k$-partite matching.
Let $G^k = G$.
Let $G^k_{1,k} = G^k[U_1 \cup U_k]$.
Find an approximate solution $M_k$ of the maximum-weighted $(r, r)$-matching of $G^k_{1,k}$ by the greedy algorithm in [18]. Suppose that in $M_k$, $S^1_1$ is matched to a vertex set $S^k_1$, $S^1_2$ is matched to $S^k_2$, etc.
Merge $M_k$ in $G^k$ to get the $(k-1)$-partite graph $G^k_{1(k)}$.
Let $i = k - 1$ and $G^i = G^k_{1(k)}$.
while $i \neq 2$ do
    Let $G^i_{1,i} = G^i[U_1 \cup U_i]$.
    Find a maximum-weighted $(1, r)$-matching $M_i$ of $G^i_{1,i}$. Suppose that in $M_i$, $v_1$ is matched to the vertex set $S^i_1$, $v_2$ is matched to $S^i_2$, etc.
    Merge $M_i$ in $G^i$ to get the $(i-1)$-partite graph $G^i_{1(i)}$.
    Let $i = i - 1$ and $G^i = G^i_{1(i)}$.
end while
Find a maximum-weighted $(1, r)$-matching $M_2$ of $G^2$.
Let $V_1 = v_1 \cup S^2_1 \cup S^3_1 \cup \ldots \cup S^{k_1}_1, \ldots, V_c = v_c \cup S^2_c \cup S^3_c \cup \ldots \cup S^{k_c}_c$.
Output $M = \{V_1 \ldots V_c\}$.
```

an algorithm that builds the (μ, γ)-quasi-cliques one vertex at a time, starting with a single query vertex v_0, backtracking as it finds maximal (μ, γ)-quasi-cliques or subgraphs that cannot be contained in a (μ, γ)-quasi-clique. We adopted the convention that S will represent the current subgraph under consideration, and C represents the set of vertices that could extend S to produce a (μ, γ)-quasi-clique (the set of "candidate" vertices). With these sets defined, we provided a pseudocode outline for our enumeration algorithm in Algorithms 6.6 and 6.7. In the theoretical results that follow, we established the

Algorithm 6.6: Pseudocode Outline of the (μ, γ)-Quasiclique Algorithm, Continued in Algorithm 6.7.

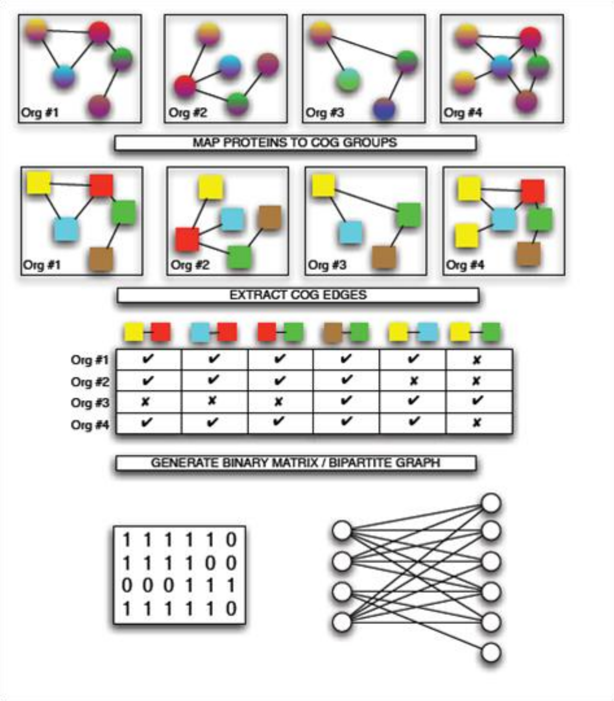


```
foreach $v_0 \in Q$ do
    $S \leftarrow \{v_0\}$
    $C \leftarrow N^2(v_0)$
    Calculate $e$
    Calculate $d_v$ for all $v \in S \cup C$
    Calculate $g_v$ for all $v \in S \cup C$
    Calculate $m_v$ for all $v \in S \cup C$
    Remove all unpromising vertices of $C$
    if $S \cup C$ is maximal then
        Call Enumerate()
end
```

meaning of the values $N^2(v)$, *e*, *d*v, *g*v, and *m*v, and we describe how these values help bind our search for the (μ, γ)-quasi-cliques of the graph.

6.7 Bi-Clustering Algorithm

For each organism, a complete list of gene id's from the KEGG metabolic pathway database and map, each gene to a COG group was generated. Using the STRING database the functional association between COGS for each organism was determined. For each COG pair the STRING similarity threshold was set to 0.85 to generate a list of COG-COG edges. Finally, a matrix in which the rows represent organisms was composed where the columns represent COG-COG edges. The entries in the matrix represent the presence or absence of an edge within the given organisms, based on the specified similarity threshold. The matrix was preprocessed to remove COG edges that occur in only one organism. The resultant matrix, referred to as the COG-COG association matrix, was used as the input to the biclustering algorithm. This matrix can be viewed as a bipartite graph, where one set contains the COG-COG interactions (or edges in the original organism graphs) and the other contains only organisms. An overview of this method is provided in Figure 6.2.


Org #1
Org #2
Org #3
Org #4
MAP PROTEINS TO COG GROUPS
Org #1
Org #2
Org #3
Org #4
EXTRACT COG EDGES
GENERATE BINARY MATRIX / BIPARTITE GRAPH

	🟨–🟥	🟦–🟥	🟥–🟩	🟫–🟩	🟨–🟦	🟨–🟩
Org #1	✔	✔	✔	✔	✔	✘
Org #2	✔	✔	✔	✔	✘	✘
Org #3	✘	✘	✘	✔	✔	✔
Org #4	✔	✔	✔	✔	✔	✘

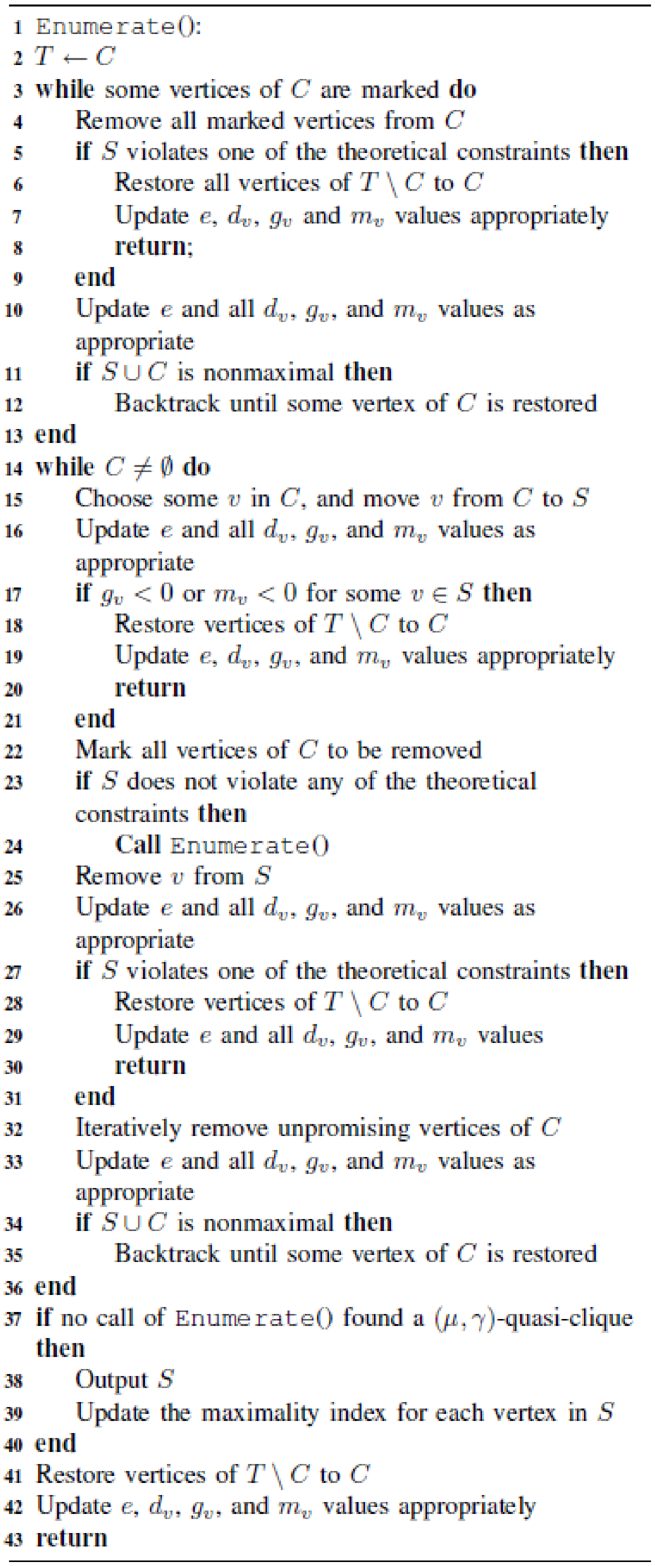
$$\begin{matrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{matrix}$$

Figure 6.2 Overview of matrix generation; individual proteins are first mapped into COG groups and then COG edges are extracted to form a binary matrix**.** The matrix represents a bipartite graph, where one set contains the organisms, the other contains COG-COG edges, and the edges represent the presence of the COG edges in the organism.

With the COG-COG association matrix, Prelic et al.'s [133] Bimax biclustering algorithm was applied to obtain the organism/COG edge clustering. The biclustering algorithm iteratively permutes the rows and columns of the COG-COG association matrix using a divide and conquers strategy. The running time of the algorithm is $O(n{\cdot}m{\cdot}B{\cdot}logB)$, where n and m are the dimensions of the matrix and B is the number of maximal biclusters. The space complexity is $\Omega\ (n{\cdot}m{\cdot}B)$. This effectively re-organizes the matrix and returns a series of maximal bi-cliques.

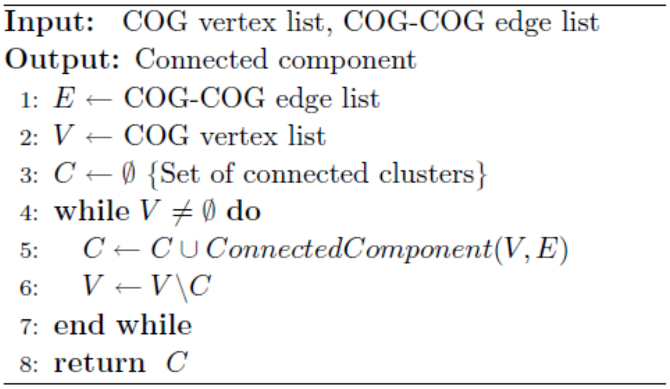
When applying this algorithm, the constraint that a returned cluster must be present in at least two organisms and consist of at least two COG edge (i.e., at least three

Algorithm 6.7: Pseudocode for Enumerate Function



```
Enumerate():
T ← C
while some vertices of C are marked do
    Remove all marked vertices from C
    if S violates one of the theoretical constraints then
        Restore all vertices of $T \setminus C$ to C
        Update $e$, $d_v$, $g_v$ and $m_v$ values appropriately
        return;
    end
    Update $e$ and all $d_v$, $g_v$, and $m_v$ values as
    appropriate
    if $S \cup C$ is nonmaximal then
        Backtrack until some vertex of C is restored
end
while $C \neq \emptyset$ do
    Choose some $v$ in C, and move $v$ from C to S
    Update $e$ and all $d_v$, $g_v$, and $m_v$ values as
    appropriate
    if $g_v < 0$ or $m_v < 0$ for some $v \in S$ then
        Restore vertices of $T \setminus C$ to C
        Update $e$, $d_v$, $g_v$, and $m_v$ values appropriately
        return
    end
    Mark all vertices of C to be removed
    if S does not violate any of the theoretical
    constraints then
        Call Enumerate()
    Remove $v$ from S
    Update $e$ and all $d_v$, $g_v$, and $m_v$ values as
    appropriate
    if S violates one of the theoretical constraints then
        Restore vertices of $T \setminus C$ to C
        Update $e$ and all $d_v$, $g_v$, and $m_v$ values
        return
    end
    Iteratively remove unpromising vertices of C
    Update $e$ and all $d_v$, $g_v$, and $m_v$ values as
    appropriate
    if $S \cup C$ is nonmaximal then
        Backtrack until some vertex of C is restored
end
if no call of Enumerate() found a $(\mu, \gamma)$-quasi-clique
then
    Output S
    Update the maximality index for each vertex in S
end
Restore vertices of $T \setminus C$ to C
Update $e$, $d_v$, $g_v$, and $m_v$ values appropriately
return
```

Algorithm 6.8: Enumerate Connected Components in a Subgraph



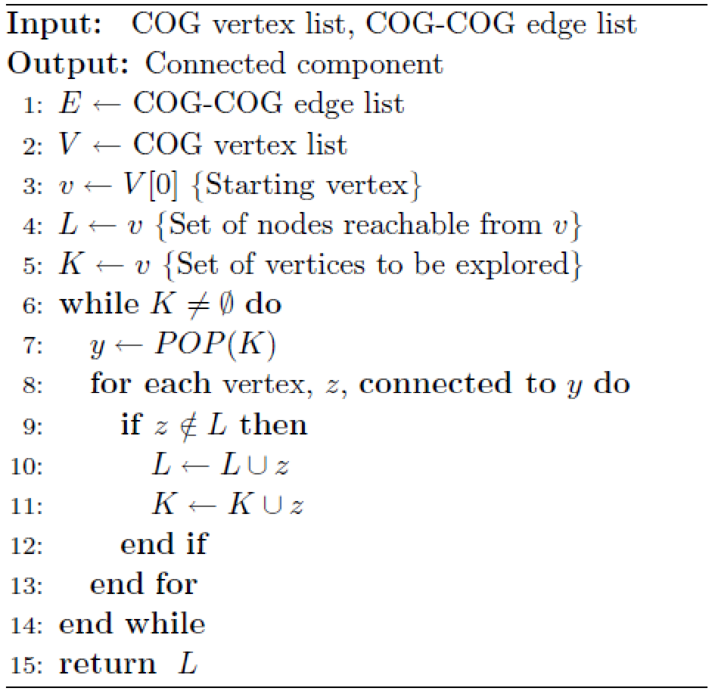
```
Input:   COG vertex list, COG-COG edge list
Output: Connected component
1: $E \leftarrow$ COG-COG edge list
2: $V \leftarrow$ COG vertex list
3: $C \leftarrow \emptyset$ {Set of connected clusters}
4: while $V \neq \emptyset$ do
5:    $C \leftarrow C \cup ConnectedComponent(V, E)$
6:    $V \leftarrow V \backslash C$
7: end while
8: return $C$
```

COGs) must be enforced. For each returned cluster, it must be determined whether it forms a connected subgraph, as we are interested in functional associations. Algorithm 6.8 is applied to each cluster to enumerate the connected components within a subgraph. For each connected component, the requirement for it to consist of at least two COG-edges is enforced.

A cluster may contain zero or more connected components. For a given cluster all connected components that meet the initial criteria of having two organisms and two edges are enumerated. Figure 6.3 shows the connected components for the bi-clique. Given a cluster of edges, Algorithm 6.8 returns a subset of edges that form a connected component, using a depth-first search. During this process, edges in the cluster are partitioned into two sets. The first is a connected component and the second is its complement. Because the complement of the connected component may itself contain a connected component, we iteratively applied Algorithm 6.8 until there were no remaining edges. This was accomplished by Algorithm 6.9.

The inputs to Algorithm 6.8 are the COG-COG edge list from the returned bi-clique and a list of individual COGS present in the edges. Lines 1-3 create variables to

Algorithm 6.9: Enumerate Connected Components



```
Input:  COG vertex list, COG-COG edge list
Output: Connected component
 1: $E \leftarrow$ COG-COG edge list
 2: $V \leftarrow$ COG vertex list
 3: $v \leftarrow V[0]$ {Starting vertex}
 4: $L \leftarrow v$ {Set of nodes reachable from $v$}
 5: $K \leftarrow v$ {Set of vertices to be explored}
 6: while $K \neq \emptyset$ do
 7:    $y \leftarrow POP(K)$
 8:    for each vertex, $z$, connected to $y$ do
 9:       if $z \notin L$ then
10:          $L \leftarrow L \cup z$
11:          $K \leftarrow K \cup z$
12:       end if
13:    end for
14: end while
15: return $L$
```

store the COG edges, COG vertices, and an empty set of clusters. Lines 4 through 7 repeat until the list of vertices is empty. With each iteration, line 5 calls the connected component algorithm, passing in the current vertex and edge lists. The algorithm returns a connected component (if they exist), given the input parameters. This component is added to the set of clusters (line 5) and its vertices are removed from the set (line 6). This repeats until the set of vertices is empty and thus we have enumerated all connected components.

Algorithm 6.9 identifies individual connected components given a set of vertices and edges as input. The algorithm selects an initial vertex (line 3) from the supplied set and performs a depth-first search in order to identify all vertices reachable from the initial vertex. Two sets are defined, in lines 4 and 5, to store the set of vertices reachable from the initial vertex and the set of vertices to be explored, respectively. Lines 6 through 14

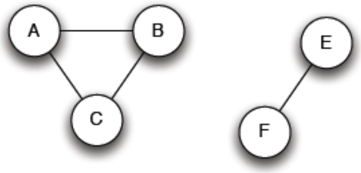


Figure 6.3 Connected components of a bi-clique.

repeat until all vertices have been explored. For each vertex to be visited, starting with the initial vertex, the algorithm tests to see which other vertices are reachable. If a vertex is not reachable, it is added to both the visited set and the “to be explored" set. This results in two sets of vertices. If these sets are the same size then the whole graph is connected. Thus, the reachable set contains a connected subgraph, which is then returned.

6.8 α,β-Clique Algorithm

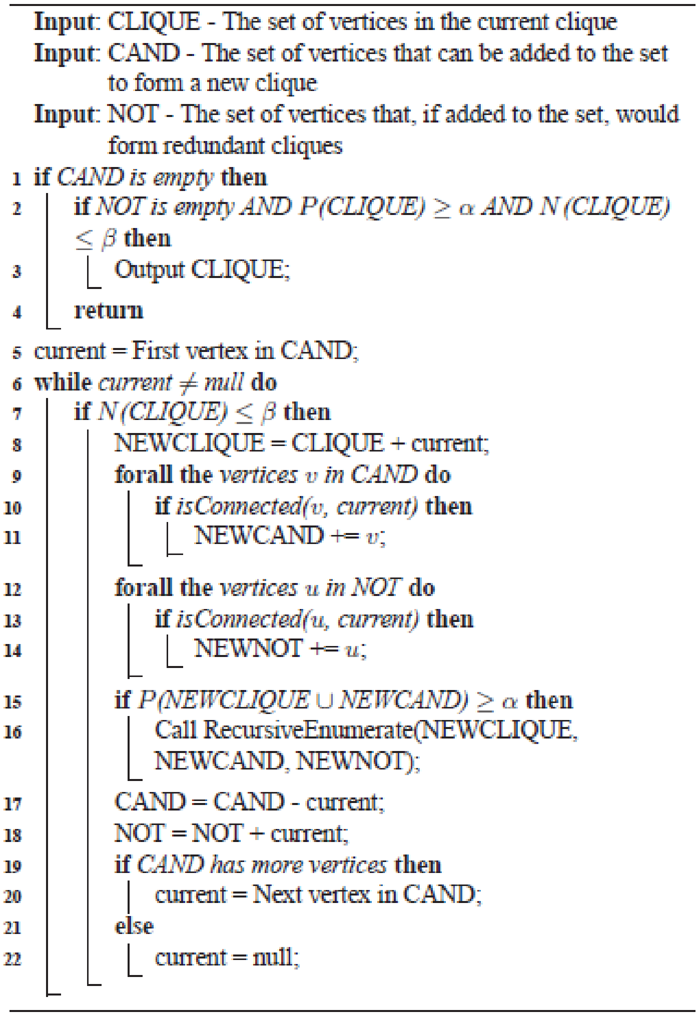
In this section, a description of the α,β-clique algorithm is provided in detail. In Chapter 5, the α,β-clique algorithm was introduced and described as an algorithm for identification of conserved functional modules. To identify these conserved modules, the algorithm identifies functional modules that are enriched in at least α networks of phenotype-expressing organisms, but may still appear in no more than β networks of organisms that do not exhibit the target phenotype.

Since α,β-cliques are maximal cliques that satisfy the α,β-criteria, a straightforward approach to enumerating them would be to enumerate all maximal cliques in the network and only output those that satisfy the α,β- criteria. However, for many values of α and β, this would result in many maximal cliques being enumerated but not output because they do not satisfy the α,β-criteria. Since the enumeration of large numbers of maximal cliques is a process that can require an intractable amount of time, it would be less time-intensive if the search process only enumerated those maximal cliques that could be also α,β -cliques.

The pseudocode of a recursive enumeration algorithm for α,β -cliques is given in Algorithm 6.10. The algorithm is a modification of the maximal clique enumeration algorithm of Bron and Kerbosch [134], which will be referred to as the BK algorithm. Key to this modification is the introduction of the functions P(S) and N(S). The function P(S) takes a set of vertices and returns the number of positive divisions that contain at least one vertex in S. The N(S) function does the same for negative divisions. Introducing the P(S) and N(S) functions to the BK algorithm enables the enumeration of α,β cliques. Line 2 would restrict the output of the algorithm to α,β-cliques. However, this would require the search algorithm to identify all maximal cliques in order to run the test on Line 2.

Algorithm 6.10 introduces bounds on the search process to reduce the search space and becomes more efficient. As the BK algorithm traverses the search tree, it keeps track of three arrays. First is the CLIQUE array, which contains all of the vertices already in the clique. Second is the CAND array that keeps track of all of the vertices that could be added to CLIQUE to form a new larger clique. The third is the NOT array, which contains vertices that, if added to CLIQUE would only identify cliques that have previously been enumerated. The first bound on Line 7 reduces the search space using the value of N(S). If N(CLIQUE) is ever greater than β, it is impossible to add vertices to the CLIQUE set and have N(CLIQUE) be less than or equal to β. Thus, if N(CLIQUE) ever becomes greater than β, the algorithm will bind the search process. The second

Algorithm 6.10: The Recursive Enumerate Algorithm



```
   Input: CLIQUE - The set of vertices in the current clique
   Input: CAND - The set of vertices that can be added to the set
          to form a new clique
   Input: NOT - The set of vertices that, if added to the set, would
          form redundant cliques
1  if CAND is empty then
2      if NOT is empty AND P(CLIQUE) $\geq \alpha$ AND N(CLIQUE)
       $\leq \beta$ then
3          Output CLIQUE;
4      return
5  current = First vertex in CAND;
6  while current $\neq$ null do
7      if N(CLIQUE) $\leq \beta$ then
8          NEWCLIQUE = CLIQUE + current;
9          forall the vertices $v$ in CAND do
10             if isConnected($v$, current) then
11                 NEWCAND += $v$;
12         forall the vertices $u$ in NOT do
13             if isConnected($u$, current) then
14                 NEWNOT += $u$;
15         if P(NEWCLIQUE $\cup$ NEWCAND) $\geq \alpha$ then
16             Call RecursiveEnumerate(NEWCLIQUE,
               NEWCAND, NEWNOT);
17         CAND = CAND - current;
18         NOT = NOT + current;
19         if CAND has more vertices then
20             current = Next vertex in CAND;
21         else
22             current = null;
```

bound on Line 15 reduces the search space using the value of P(S). For any given search node, the maximum P(CLIQUE) value that could exist for any child of the search node is P(CLIQUE [CAND). If this value is less than β, then there is no reason to continue expanding the subtree of the current search node.

CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

7.1 Conclusions

The research presented in this dissertation focused on co-development and application of computationally driven approaches to: (1) identify phenotype-related biochemical processes and (2) demonstrate predictability of these techniques, to hydrogen production by dark fermentative and acid-tolerant bacteria. To predict biochemical processes, a systems biology driven approach was taken, thus allowing for identification of *in silico* phenotype-related metabolic and cellular networks that are unlikely to be found by current *in silico* methodologies.

Information obtained through application of these methodologies can be used as a knowledge base by scientists and engineers to decipher which specific biochemical processes or networks of microorganisms to genetically modify for enhanced hydrogen production. In addition, phenotype-related components that were predicted in this dissertation allow for the development of new research questions and hypotheses regarding expression of target phenotypes. Examples of such questions were provided in Chapter 3-5 and include: “what is the role of previously uncharacterized proteins associated with hydrogenase accessory proteins”, or “is phosphate butyryltransferase associated with metabolic shifts between acidogenesis and solventogenesis?”

Other advantages of the methodologies applied in this research are the ability of the Network Instance-Based Biased Subgraph (NIBBS) search algorithm to identify metabolic subsystems across hundreds of genome scale networks and at orders of magnitude faster than an exact algorithm for enumerating all maximally-based subgraphs. Metabolic subsystems identified by NIBBS not only included those specific to a subset of phenotype-expressing organisms, but also identified sub-phenotypes (e.g., dark fermentation) of the target phenotype (e.g., hydrogen production). Identification of sub-phenotypes provides the first step toward deciphering between sets of organisms capable of expressing more than one phenotype.

The multiple alignment algorithm provided for closer examination of key metabolic pathways in smaller sets of organisms. This algorithm is the first that allows alignment of multiple metabolic pathways based on the similarity of enzymes, compounds, reactions and topological structure. This information is particularly valuable when trying to identify conserved components in specific sets of organisms and in determining which metabolic capabilities are beneficial for biologically engineering microorganisms.

Finally, computational methodologies presented in Chapter 5 were able to successfully identify phenotype-related networks in individual organisms (DENSE) and across multiple organisms (α,β-clique and bi-clustering). Through the utilization of three complementary approaches, important interactions between known and previously uncharacterized hydrogenase and other hydrogen-related enzymes were identified. Through the utilization of knowledge priors, previously uncharacterized proteins were detected and their functional role predicted. In addition, analysis and review of the results from all three methodologies demonstrated our ability to incorporate data beyond metabolic networks. This is particularly important when trying to identify phenotypes, such as acid-tolerance, often regulated by non-metabolic components (e.g., enzymes associated with cell membrane or repair). In this study, results focused on metabolic processes involved in acid-tolerance; however, through utilization of entire microbial networks, information regarding potential phenotype-related transporters and regulatory enzymes was also obtained.

7.2 Future Work

While results for computational methodologies were able to successfully identify conserved and phenotype-related biochemical processes, the methodologies presented are still in the initial stages in the development of evolutionary genome-scale phenotype-centric comparative network analysis tools. One area of improvement would be in the systematic development of methodologies for identification of cross-talk. In this study, cross-talk or communication between metabolic pathways and networks was detected by both the NIBBS algorithm and bi-clustering algorithm. For each methodology, cross-talk was a by-product of the algorithm rather than the targeted output. Due to the importance of identifying cross-talk between biochemical networks, new methodologies targeting systematic characterization of metabolic pathway cross-talk mechanisms is needed. As noted from the dissertation, understanding of cross-talk and underlying biochemical mechanisms will provide scientists and engineers with a better foundation for bioengineering microorganisms.

Another area of improvement is in the detection of phenotype-related components within organisms. In order to survive changes and maintain growth under varying environmental and chemical parameters, organisms often exhibit multiple phenotypes. For example, in this study *Clostridium acetobutylicum* was used as the model organism for analysis of computational results. Selection of *Clostridium acetobutylium* as a model organism was partially due to of the vast amounts of literature regarding its genomic sequence, as well as to its ability to exhibit the three main phenotypes targeted in this research (i.e., anaerobic, dark fermentative hydrogen producing, and acid-tolerant). Similarly, comparison of the list of anaerobic organisms with acid-tolerant and hydrogen producing phenotypes would reveal that many of the organisms share similar phenotypes. Here, only paired phenotypes (e.g., aerobic versus anaerobic) were used in experiments.

One way to improve upon computational methodologies, such as NIBBS, which focuses on sets of organisms, is to evaluate a collection of phenotypes rather than a binary set of phenotypes. In measurement of this parameter, improvement is also necessary in linking predicted phenotype-related components to the actual phenotype. Currently, results presented in Chapters 3-5 can only hypothesize or predict the relation of identified biochemical components to phenotype-expressing organisms rather than the phenotype itself. In addition to evaluating sets of phenotypes, an area of future inquiry would be to see if the knowledgebase of genes, pathways, and networks identified in this study can be used to predict phenotype profiles of unknown organisms. In mixed microbial communities, such information would provide valuable clues to the discovery of potential new species and an understanding of the communities, such as those in anaerobic digesters that are considered as being in a "black box."

The results presented in this study are mostly consistent with literature values. However, in the case of conserved hydrogen-producing enzymes identified by the multiple alignment algorithm, a known enzyme responsible for conversion of pyruvate to Acetyl-CoA was predicted as missing in *Clostridium acetobutylicum*. This mistake is not a result of the algorithm itself, but rather due to the mis-information in published network databases (e.g., the Kyoto Encyclopedia of genes and genomes (KEGG)). Due to the reliance of the NIBBS and multiple alignment algorithm on the KEGG database, some important clues regarding phenotype-related proteins may be missed. To overcome this problem, computational methodologies for NIBBS and the multiple alignment algorithm should consider incorporating or utilizing other databases. In terms of NIBBS, incorporation of multiple databases and perhaps experimental data may improve overall predictions for phenotype-related biochemical processes.

Lastly, the knowledgebase of genes, pathways, and networks generated by this study provides for further understanding and identification of a number of clues related to biochemical processes involved in hydrogen production. To demonstrate the ability of the knowledge obtained to transfer to experimental studies, molecular and then ultimately metabolic engineering studies focused on specific networks identified in this study is necessary. Successful application of metabolic engineering as a result of *in silico* studies would demonstrate the necessity of systems-wide computational approaches for design of engineered organisms important for practical application.

7.3 Application to Engineering Studies

The work presented in this dissertation sets the stage for the role of computational biology in further defining the roles of metabolic and regulatory pathways involved in hydrogen production. Specifically, utilization of the computational tools allows for prediction of potentially key genes, pathways, or metabolic sub-systems directly or indirectly involved in regulating production of hydrogen. Such information is important in further understanding biochemical processes for hydrogen production and for improving the overall yield for biohydrogen production by microbial communities.

In order to improve hydrogen yields, there have been a number of metabolic and genetic engineering studies aimed at either increasing the expression of target genes or removing genes involved hydrogen uptake [39]. Other engineering studies, such as those by Rey et al. [6], focus on the redirection of metabolic pathways. While these studies show the potential for increasing hydrogen yields through bioengineering of organisms, this is only a small piece to a large problem—application of bioengineering towards hydrogen production.

Part of the slow development of bioengineering technologies and application of these technologies is due to the time commitment needed to identify key metabolic or cellular components across entire organisms and genomes. In this study, predictions generated by each algorithm provide scientists with clues toward the metabolic capabilities of organisms. In addition, information predicted can also be used as the underlying foundation for bioengineering microorganisms for enhanced biohydrogen production. Instead of spending years on specific metabolic pathways or individual organisms, computational tools can lead to identification of important phenotype-related biochemical processes on shorter time-scales, across numerous genomes, and across entire metabolic systems.

In this dissertation, computational predictions and hypotheses developed were not experimentally tested. For future work, scientists and engineers would first need to evaluate predicted results. If the results from the validation studies are accurate, then engineers can use their knowledge of hydrogen producing systems to design hydrogen producing organisms for experimental bench top studies. From bench top studies, information regarding the environmental and operational parameters can be obtained for engineered systems, such as small closed systems and hydrogen production from fuel cells. Due to the complexity of large organic-rich systems, such as wastewater facilities (shown in Chapter 2), utilization of engineered organisms for hydrogen production will mostly likely be limited in the short term to research on fuel cells. In order to generate high yields of biological hydrogen in wastewater systems, operational parameters and improved design parameters to maintain any introduced microbial stock is necessary. As such, application to large-scale engineering studies may be limited at this time.

REFERENCES CITED

1. Ueno, Y., S. Otsuka, and M. Morimoto, *Hydrogen production from industrial wastewater by anaerobic microflora in chemostat culture.* J. Ferment. Bioeng., 1996. 82(2): p. 194-197.

2. Li, C. and H.H.P. Fang, *Fermentative Hydrogen Production From Wastewater and Solid Wastes by Mixed Cultures.* Critical Reviews in Environmental Science and Technology, 2007. 37(1): p. 1 - 39.

3. Claassen, P.A., T.d. Vrije, and M.A.W. Budde. *Biological hydrogen production from sweet sorghum by thermophilic bacteria*. in *Proceedings 2nd World Conference on Biomass for Energy*. 2004. Rome.

4. Miyake, J., *Biohydrogen*, ed. O.R. Zaborsky. 1998, USA: Plenum Press.

5. Lathe, W.C., J.M. Williams, M.E. Mangan, and D. Karolchik, *Genomic Data Resources: Challenges and Promises.* Nature Education, 2008. 1(3).

6. Rey, F.E., E.K. Heiniger, and C.S. Harwood, *Redirection of metabolism for biological hydrogen production.* Applied and Environmental Microbiology, 2007. 73(5): p. 1665-1671.

7. Yu, J. and P. Takahashi, *Biophotolysis-based hydrogen production by cyanobacteria and green microalgae*, in *Communicating Current Research and Educational Topics and Trends in Applied Microbiology*, A. Mendez-Vilas, Editor. 2007. p. 79-89.

8. White, D., *The physiology and biochemistry of prokaryotes*, ed. D. White. 2007, New York: Oxford University Press.

9. Classen, P.A.M., J.B.v. Lier, A.M.L. Contreras, E.W.J.v. Niel, L. Sijtsma, A.J.M. Stams, S.S.d. Vries, and R.A. Weushuis, *Utilization of biomass for the supply of energy carriers.* Appl. Microbiol. Biotechnol., 1999. 52: p. 741-755.

10. Kelley, B.P., R. Sharan, R.M. Karp, T. Sittler, D.E. Root, B.R. Stockwell, and T. Ideker, *Conserved pathways within bacteria and yeast as revealed by global protein network alignment.* Proceedings of the National Academy of Sciences of the United States of America, 2003. 100(20): p. 11394-11399.

11. Kapdan, I.K. and F. Kargi, *Bio-hydrogen production from waste materials.* Enzyme and Microbial Technology, 2006. 38(5): p. 569-582.

12. Zhu, H., S. Ueda, Y. Asada, and J. Miyake, *Hydrogen production as a novel process of wastewater treatment--studies on tofu wastewater with entrapped R. sphaeroides and mutagenesis.* International Journal of Hydrogen Energy. 27(11-12): p. 1349-1357.

13. Okamoto, M., T. Miyahara, O. Mizuno, and T. Noike, *Biolgoical hydrogen potential of material characteristics of the organic fraction of municipal solid wastes.* Water Sci. Tech., 2000. 41(3): p. 25-32.

14. Lay, J.J., K.S. Fan, J.I. Chang, and C.H. Ku, *Influence of chemical nature of organic wastes on their conversion to hydrogen by heat-shock digested sludge.* International Journal of Hydrogen Energy, 2003. 28(12): p. 1361-1376.

15. Wagner, R.C., J.M. Regan, S.-E. Oh, Y. Zuo, and B.E. Logan, *Hydrogen and methane production from swine wastewater using microbial electrolysis cells.* Water Research, 2009. 43: p. 1480-1488.

16. Hawkes, F.R., R. Dinsdale, D.L. Hawkes, and I. Hussy, *Sustainable fermentative hydrogen production: challenges for process optimisation.* International Journal of Hydrogen Energy, 2002. 27: p. 1339-1347.

17. López-Gomollón, S., J.A. Hernández, S. Pellicer, V.E. Angarica, M.L. Peleato, and M.F. Fillat, *Cross-talk Between Iron and Nitrogen Regulatory Networks in Anabaena (Nostoc) sp. PCC 7120: Identification of Overlapping Genes in FurA and NtcA Regulons.* Journal of Molecular Biology, 2007. 374(1): p. 267-281.

18. Lin, C.Y. and C.H. Lay, *A nutrient formulation for fermentative hydrogen production using anaerobic sewage sludge microflora.* International Journal of Hydrogen Energy, 2005. 30(3): p. 285-292.

19. Dolfing, J., B. Jiang, A.M. Henstra, A.J.M. Stams, and C.M. Plugge, *Syntrophic Growth on Formate: a New Microbial Niche in Anoxic Environments.* Appl. Environ. Microbiol., 2008. 74(19): p. 6126-6131.

20. Shin, H.-S., J.-H. Youn, and S.-H. Kim, *Hydrogen production from food waste in anaerobic mesophilic and thermophilic acidogenesis.* International Journal of Hydrogen Energy, 2004. 29(13): p. 1355-1363.

21. Thiele, J.H. and G. Zeikus, *Control of Interspecies Electron Flow During Anaerobic Digestion: Significance of Formate Transfer versus Hydrogen Transfer During Synthrophic Methanogenesis in Flocs.* Appl Environ Microbiol, 1988. 54(1): p. 20-29.

22. Nath, K. and D. Das, *Improvement of fermentative hydrogen production: Various approaches.* Appl. Microbiol. Biotechnol., 2004. 65: p. 520-529.

23. Boyles, D., *Bioenergy technology- Thermodynamics and costs.* 1984, New York: Wiley

24. Lin, C.Y. and C.H. Lay, *Effects of carbonate and phosphate concentrations on hydrogen production using anaerobic sewage sludge microflora.* International Journal of Hydrogen Energy, 2004. 29(3): p. 275-281.

25. Yu, H., Z. Zhu, W. Hu, and H. Zhang, *Hydrogen production from rice winery wastewater in an upflow anaerobic reactor by using mixed anaerobic cultures.* International Journal of Hydrogen Energy, 2002. 27(11-12): p. 1359-1365.

26. Li, R.Y. and H.H.P. Fang, *Heterotrophic photofermentative hydrogen production.* Critical Reviews in Environmental Science and Technology, 2009. 39: p. 1081-1108.

27. Maeda, T., V. Sanchez-Torres, and T.K. Wood, *Metabolic engineering to enhance bacterial hydrogen production.* Microbial Biotechnology, 2008. 1(1): p. 30-39.

28. Bansal, A.K., *Bioinformatics in microbial biotechnology--a mini review.* Microb Cell Fact, 2005. 4(1): p. 19.

29. Kastenmuller, G., M.E. Schenk, J. Gasteiger, and H.W. Mewes, *Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes.* Genome Biology, 2009. 10: p. R28.

30. Koyutürk, M., *Algorithmic and analytical methods in network biology.* Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2010. 2(3): p. 277-292.

31. Bailey, J.E., A. Sburlati, V. Hatzimanikatis, K. Lee, W.A. Renner, and P.S. Tsai, *Inverse metabolic engineering: a strategy for directed genetic engineering of useful phenotypes.* Biotechnol Bioeng., 1996. 52: p. 109-121.

32. Mathews, J. and G. Wang, *Metabolic pathway engineering for enhanced biohydrogen production.* International Journal of Hydrogen Energy, 2009. 34(17): p. 7404-7416.

33. Brentner, L.B., J. Peccia, and J.B. Zimmerman, *Challenges in Developing Biohydrogen as a Sustainable Energy Source: Implications for a Research Agenda.* Environmental Science & Technology, 2010. 44: p. 2243-2254.

34. Guest, J.S., S.J. Skerlos, J.L. Barnard, M.B. Beck, G.T. Daigger, H. Hilger, S.J. Jackson, K. Karvazy, L. Kelly, L. Macpherson, J.R. Mihelcic, A. Pramanik, L. Raskin, M.C.M. Van Loosdrecht, D. Yeh, and N.G. Love*, *A New Planning and Design Paradigm to Achieve Sustainable Resource Recovery from Wastewater1.* Environmental Science & Technology, 2009. 43(16): p. 6126-6130.

35. Logan, B.E., *Microbial Fuel Cells*, ed. B.E. Logan. 2008, New York: Wiley & Sons Inc.

36. Hallenbeck, P.C. and J.R. Benemann, *Biological hydrogen production; fundamentals and limiting processes.* International Journal of Hydrogen Energy. 27(11-12): p. 1185-1193.

37. Khanal, S., *Bioenergy generation from residues of biofuel industries*, in *Anaerobic biotechnology for bioenergy production: Principles and applications*, S. Khanal, Editor. 2008, Wiley-Blackwell: USA. p. 161-187.

38. Khanal, S., *Overview of anaerobic biotechnology*, in *Anaerobic biotechnology for bioenergy production. Principles and applications*, S. Khanal, Editor. 2008, Wiley-Blackwell: USA. p. 1-25.

39. Rey, F.E., Y. Oda, and C.S. Harwood, *Regulation of uptake hydrogenase and effects of hydrogen utilization on gene expression in Rhodopseudomonas palustris.* j=Journal of bacteriology, 2006. 188(17): p. 6143-6152.

40. Jones, P.R., *Improving Fermentative Biomass-derived H_2 Production by Engineering Microbial Metabolism.* International Journal of Hydrogen Energy, 2008. 33: p. 5122-5130.

41. Amador-Noguez, D., X.-J. Feng, J. Fan, N. Roquet, H. Rabitz, and J.D. Rabinowitz, *Systems-level metabolic flux profiling elucidates a complete, bifurcated TCA cycle in Clostridium acetobutylicum.* J. Bacteriol., 2010: p. JB.00490-10.

42. Nicolet, Y., J.C. Fontecilla-Camps, and M. Fontecave, *Maturation of [FeFe]-hydrogenases: Structures and mechanisms.* International Journal of Hydrogen Energy, 2010. 35(19): p. 10750-10760.

43. Shomura, Y., H. Komori, N. Miyabe, M. Tomiyama, N. Shibata, and Y. Higuchi, *Crystal Structures of Hydrogenase Maturation Protein HypE in the Apo and ATP-bound Forms.* Journal of Molecular Biology, 2007. 372(4): p. 1045-1054.

44. Lee, J., H. Yun, A. Feist, B. Palsson, and S. Lee, *Genome-scale reconstruction and in silico analysis of the Clostridium acetobutylicum ATCC 824 metabolic network.* Applied Microbiology and Biotechnology, 2008. 80(5): p. 849-862.

45. McKinlay, J.B. and C.S. Harwood, *Carbon dioxide fixation as a central redox cofactor recycling mechanism in bacteria.* Proceedings of the National Academy of Sciences.

46. Peters, J.W., K. Fisher, and D.R. Dean, *Nitrogenase structure and function: A biochemical-genetic perspective.* Annu. Rev. Microbiol. , 1995. 49: p. 335-366.

47. Vignais, P.M., B. Billoud, and J. Meyer, *Classification and phylogeny of hydrogenases* FEMS Microbiology Reviews, 2001. 25(4): p. 455-501.

48. King, P.W., M.C. Posewitz, M.L. Ghirardi, and M. Seibert, *Functional Studies of [FeFe] Hydrogenase Maturation in an Escherichia coli Biosynthetic System.* J. Bacteriol., 2006. 188(6): p. 2163-2172.

49. Butland, G., J.w. Zhang, W. Yang, A. Sheung, P. Wong, J.F. Greenbalt, A. Emili, and D.B. Zamble, *Interactions of the Escherichia coli hydrogenase biosynthetic proteins: HybG complex formation.* FEBS Letters, 2006. 580: p. 677-681.

50. Vignais, P.M., B. Billoud, and J. Meyer, *Classification and phylogeny of hydrogenases<sup>1</sup>.* FEMS Microbiology Reviews, 2001. 25(4): p. 455-501.

51. Velt, A., M.K. Akhtar, T. Mizutani, and P.R. Jones, *Contructing and Testing the Thermodynamic Limits of Synthetic NAD(P)H: H_2 pathways.* Microbial Biotechnology, 2008. 1(5): p. 382-294.

52. Akhtar, M.K. and P.R. Jones, *Engineering of a Synthetic hydF=hydE=hydG=hydA operon for biohydrogen production.* Analytical Biochemistry, 2008. 373: p. 170-172.

53. Antoni, D., V.V. Zverlov, and W.H. Schwarz, *Biofuels from microbes.* Applied Microbiology and Biotechnology, 2007. 77(1): p. 23-35.

54. Khanal, S., *Biohydrogen production: fundamentals, challenges, and operation strategies for enhanced yield*, in *Anaerobic biotechnology for bioenergy production: principles and applications*, S. Khanals, Editor. 2008, Wiley-Blackwell: USA.

55. Wu, S.-Y., C.-N. Lin, J.-S. Chang, and J.-S. Chang, *Biohydrogen production with anaerobic sludge immobilized by ethylene-vinyl acetate copolymer.* International Journal of Hydrogen Energy, 2005. 30(13-14): p. 1375-1381.

56. Lui, H. and H.H.P. Fang, *Hydrogen production from wastewater by acidogenic granular sludge.* Water Sci. Tech., 2002. 147(1): p. 153-158.

57. Kovacs, K.L., Z. Bagi, B. Balint, B.d. Fodor, G. Scnadi, R. Csaki, T. Hanczar, A.T. Kovacs, G. Maroti, K. Perei, A. Toth, and G. Rakhely, *Novel approaches to exploit microbial hydrogen metabolism*, in *Biohydrogen III*, J. Miyake, Y. Igarashi, and M. Rogner, Editors. 2004, Elsevier Inc: USA.

58. Czech, I., A. Silakov, W. Lubitz, and T. Happe, *The [FeFe]-hydrogenase maturase HydF from Clostridium acetobutylicum contains a CO and CN- ligated iron cofactor.* FEBS Letters, 2010. 584(3): p. 638-642.

59. Hoekman, S.K., *Biofuels in the U.S. - Challenges and opportunities.* Renewable Energy, 2009. 34: p. 14-22.

60. Fang, H.H.P. and Y. Lui, *Anaerobic wastewater treatment in (sub-) tropical regions.* , in *Advances in water and wastewater treatment technology*, T. Matsuo, et al., Editors. 2001, Elsevier Science. p. 285-294.

61. Hart, D., *Hydrogen power: The commerical future of the ultimate fuel*. 1997, London: Financial Times Energy Publishing

62. Sparling, R., D. Risbey, and H.M. Poggi-Varaldo, *Hydrogen production from inhibited anaerobic composters.* International Journal of Hydrogen Energy, 2007. 22(6): p. 563-566.

63. Lay, J.J., Y.J. Lee, and T. Noike, *Feasibility of biological hydrogen production from organic fraction of municipal solid waste.* Water Research, 1999. 33(11): p. 2579-2586.

64. Lee, Y.J., T. Miyahara, and T. Noike, *Effect of iron concentratin of hydrogen fermentation* Bioresource Technology, 2001. 80(3): p. 227-231.

65. Dabrock, B., H. Bahl, and G. Gottschalk, *Parameters affecting solvent production by Clostridium pasterurianum.* Appl. Environ. Microbiol, 1992. 58(4): p. 1233-1239.

66. Foster, J.W., *Microbial Response to Acid Stress*, in *Bacterial Stress Responses*, G. Storz and R. Hengge-Aronis, Editors. 2000, ASM Press: Washington, D.C. p. 99-116.

67. Gardner, T.S., D. di Bernardo, D. Lorenz, and J.J. Collins, *Inferring genetic networks and identifying compound mode of action via expression profiling.* Science, 2003. 301(5629): p. 102-5.

68. Krieger, C.J., P. Zhang, L.A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S.Y. Rhee, and P.D. Karp, *MetaCyc: a multiorganism database of metabolic pathways and enzymes.* Nucleic Acids Res, 2004. 32(Database issue): p. D438-42.

69. Korbel, J.O., T. Doerks, L.J.Jensen, C. Perez-Iratxeta, S. Kaczanowski, S.D. Hooper, M.A. Andrade, and P. Bork, *Systematic association of genes to phenotypes by genome and literature mining.* PLOS Biology, 2005. 3(5): p. e134.

70. Kastenmuller, G., M. Schenk, J. Gasteiger, and H.-W. Mewes, *Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes.* Genome Biology, 2009. 10(3): p. R28.

71. Hallenbeck, P.C. and D. Ghosh, *Improvements in fermentative biological hydrogen production through metabolic engineering.* Journal of Environmental Management. In Press, Corrected Proof.

72. Chen, J.-S., J. Toth, and M. Kasap, *Nitrogen-fixation genes and nitrogenase activity in Clostridium acetobutylicum and Clsotridium beijerinckii.* Journal of Industrial Microbiology and Biotechnology, 2001. 27: p. 281-286.

73. Huang, L., C.W. Forsberg, and L.N. Gibbins, *Influence of external pH and fermentation products on Clostridium acetobutylicum Intracellular pH and cellular distribution of fermentation products.* Appl. Environ. Microbiol, 1986. 51(6): p. 1230-1234.

74. Foster, J.W., *Escherichia coli acid resistance: tales of an amateur acidophile.* Nat. Rev. Microbiol., 2004. 2: p. 898-907.

75. Borden, J.R., S.W. Jones, D. Indurthi, Y. Chen, and E. Terry Papoutsakis, *A genomic-library based discovery of a novel, possibly synthetic, acid-tolerance mechanism in Clostridium acetobutylicum involving non-coding RNAs and ribosomal RNA processing.* Metabolic Engineering, 2010. 12(3): p. 268-281.

76. Murray, A.W., *The Biological Significance of Purine Salvage.* Annual Review of Biochemistry, 1971. 40(1): p. 811-826.

77. Anders, M.W. and W. Dekant, *Aminoacylases*, in *Advances in Pharmacology*, M.W. Anders and D. Wolfgang, Editors. 1994, Academic Press. p. 431-448.

78. Jim, K., K. Parmar, M. Singh, and S. Tavazoie, *A cross-genome approach for systematic mapping of phenotypic traits to genes.* Genome Research, 2004. 14: p. 109-115.

79. M.Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates, *Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles.* PNAS, 1999. 96(8): p. 42854-4288.

80. Veit, A., M.K. Akhtar, T. Mizutani, and P.R. Jones, *Contructing and Testing the Thermodynamic Limits of Synthetic NAD(P)H: H_2 pathways.* Microbial Biotechnology, 2008. 1(5): p. 382-294.

81. Madigan, M. and J. Martinko, *Brock Biology of Microorganisms* 11th ed. 2005: Prentice Hall.

82. Vogels, G.D. and C. Van der Drift, *Degradation of purines and pyrimidines by microorganisms.* Microbiol. Mol. Biol. Rev., 1976. 40(2): p. 403-468.

83. Köpke, M., C. Held, S. Hujer, H. Liesegang, A. Wiezer, A. Wollherr, A. Ehrenreich, W. Liebl, G. Gottschalk, and P. Dürre, *Clostridium ljungdahlii represents a microbial production platform based on syngas.* Proceedings of the National Academy of Sciences, 2010. 107(29): p. 13087-13092.

84. Durre, P. and J.R. Andreesen, *Purine and glycine metabolism by purinolytic clostridia.* J. Bacteriol., 1983. 154(1): p. 192-199.

85. Lu, C.-D., *Pathways and regulation of bacterial arginine metabolism and perspectives for obtaining arginine overproducing strains.* Applied Microbiology and Biotechnology, 2006. 70(3): p. 261-272.

86. Pérez-Arellano, I., F. Carmona-Álvarez, J. Gallego, and J. Cervera, *Molecular Mechanisms Modulating Glutamate Kinase Activity. Identification of the Proline Feedback Inhibitor Binding Site.* Journal of Molecular Biology, 2010. 404(5): p. 890-901.

87. Gamper, H. and V. Moses, *Enzyme organization in the proline biosynthetic pathway of Escherichia coli.* Biochimica et Biophysica Acta (BBA) - General Subjects, 1974. 354(1): p. 75-87.

88. Chen, L.M. and S. Maloy, *Regulation of proline utilization in enteric bacteria: cloning and characterization of the Klebsiella put control region.* J. Bacteriol., 1991. 173(2): p. 783-790.

89. Milner, J.L., D.J. McClellan, and J.M. Wood, *Factors Reducing and Promoting the Effectiveness of Proline as an Osmoprotectant in Escherichia coli K12.* J Gen Microbiol, 1987. 133(7): p. 1851-1860.

90. Rodionov, D.A., A.G. Vitreschak, A.A. Mironov, and M.S. Gelfand, *Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems.* Nucleic Acids Research, 2004. 32(11): p. 3340-3353.

91. Andre, G., E. Haudecoeur, M. Monot, K. Ohtani, T. Shimizu, B. Dupuy, and I. Martin-Verstraete *Global regulation of gene expression in response to cysteine availability in Clostridium perfringens.* BMC Microbiology, 2010. 10(1): p. 234.

92. Masip, L., K. Veeravalli, and G. Georgiou, *The many faces of glutathione in bacteria.* Antioxid Redox Signal, 2006. 8(5-6): p. 753-762.

93. Peters, J.W., K. Fisher, and D.R. Dean, *Nitrogenase Structure and Function: A Biochemical-Genetic Perspective.* Annu Rev Microbiol, 1995. 49: p. 335-366.

94. Gaudu, P., Y. Yamamoto, P.R. Jensen, K. Hammer, and A. Gruss, *Genetics of Lactococci*, in *Gram-Positive Pathogens*, V.A. Fischetti, et al., Editors. 2006, ASM Press: Washington D.C. p. 361-362.

95. Koyutürk, M., A. Grama, and W. Szpankowski, *An efficient algorithm for detecting frequent subgraphs in biological networks.* Bioinformatics, 2004. 20(suppl 1): p. i200-i207.

96. Hoffman, P., A. Goodwin, J. Johnsen, K. Magee, and S. Veldhuyzen van Zanten, *Metabolic activities of metronidazole-sensitive and -resistant strains of Helicobacter pylori: repression of pyruvate oxidoreductase and expression of isocitrate lyase activity correlate with resistance.* J. Bacteriol., 1996. 178(16): p. 4822-4829.

97. Huang, K.-x., F.B. Rudolph, and G.N. Bennett, *Characterization of Methylglyoxal Synthase from Clostridium acetobutylicum ATCC 824 and Its Use in the Formation of 1,2-Propanediol.* Appl. Environ. Microbiol., 1999. 65(7): p. 3244-3247.

98. Sawers, R.G., *Formate and its role in hydrogen production in Escherichia coli.* Biochem Soc Trans, 2005. 33: p. 42-6.

99. Tamura, M. and P. D'haeseleer, *Microbial genotype-phenotype mapping by class association rule mining.* Bioinformatics, 2008. 24(13): p. 1523-1529.

100. Schmidt, M.C. and N. F.Samatova. *An Algorithm for Discovery of Phenotype Related Metabolic Pathways*. in *IEEE International Conference on Bioinformatics & Biomedicine* 2009. Washington D.C. .

101. Korbel, J.O., T. Doerks, L.J. Jensen, C. Perez-Iratxeta, S. Kaczanowski, S.D. Hooper, M.A. Andrade, and P. Bork, *Systematic association of genes to phenotypes by genome and literature mining.* PLos Biol, 2005. 3(5): p. e134.

102. Hendrix, W., A.M. Rocha, M. Elmore, J. Trien, and N.F. Samatova, *Discovery of Enriched Biological Network Motifs using Knowledge Priors with Application to Biohydrogen Production.* Accepted. BIOCOMP conference proceedings.

103. Yebra, M.J. and G. Perez-Martinez, *Cross-talk between the L-sorbose and D-sorbitol (D-glucitol) metabolic pathways in Lactobacillus casei.* Microbiology 2002. 148: p. 2351-2359.

104. Zhang, H., M.A. Bruns, and B.E. Logan, *Biological hydrogen production by Clostridium acetobutylicum in an unsaturated flow reactor.* Water Research, 2006. 40(4): p. 728-734.

105. Huang, Y., W. Zong, X. Yang, R. Wang, C.L. Hemme, J. Zhou, and Z. Zhou, *Succesion of the bacterial community and dynamics of hydrogen producers in a hydrogen-producing bioreactor.* Appl. Environ. Microbiol, 2010. 76(10): p. 3387-3390.

106. Alsaker, K.V., C. Paredes, and E.T. Papoutsakis, *Metabolite stress and tolerance in the production of biofuels and chemicals: Gene-expression-based systems analysis of butanol, butyrate, and acetate stresses in the anaerobe Clostridium acetobutylicum.* Biotechnology and Bioengineering, 2010. 105(6): p. 1131-1147.

107. Bahl, H., M. Gottwald, A. Kuhn, V. Rale, W. Andersch, and G. Gottschalk, *Nutritional Factors Affecting the Ratio of Solvents Produced by Clostridium acetobutylicum.* Appl Environ Microbiol, 1986. 52(1): p. 169 - 172.

108. Hendrix, W., A.M. Rocha, M. Elmore, J. Trien, and N.F. Samatova. *Discovery of Enriched Biological Motifs with Knowledge Priors*. in *BIOCOMP'10 - 11th International Conference on Bioinformatics and Computational Biology*. 2010.

109. Blokesch, M., S.P.J. Albracht, B.F. Matzanke, N.M. Drapal, A. Jacobi, and A. Böck, *The Complex Between Hydrogenase-maturation Proteins HypC and HypD is an Intermediate in the Supply of Cyanide to the Active Site Iron of [NiFe]-Hydrogenases.* Journal of Molecular Biology, 2004. 344(1): p. 155-167.

110. Eidels, L. and M.J. Osborn, *Phosphoheptose Isomerase, First Enzyme in the Biosynthesis of Aldoheptose in Salmonella typhimurium.* Journal of Biological Chemistry, 1974. 249(17): p. 5642-5648.

111. Valvano, M.A., P. Messner, and P. Kosma, *Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides.* Microbiology, 2002. 148(7): p. 1979-1989.

112. White, D., *The physiology and biochemistry of prokaryotes*. 2nd ed. 2000, Oxford: Oxford University Press, Inc. 549.

113. Weidner, G. and G. Sawers, *Molecular characterization of the genes encoding pyruvate formate-lyase and its activating enzyme of Clostridium pasteurianum.* J. Bacteriol., 1996. 178(8): p. 2440-2444.

114. Wiesenborn, D., F. Rudolph, and E. Papoutsakis, *Phosphotransbutyrylase from Clostridium acetobutylicum ATCC 824 and its role in acidogenesis.* Appl Environ Microbiol, 1989. 55(2): p. 317-322.

115. Hartmanis, M.G.N. and S. Gatenbeck, *Intermediary Metabolism in Clostridium acetobutylicum: Levels of Enzymes Involved in the Formation of Acetate and Butyrate.* Appl. Environ. Microbiol., 1984. 47(6): p. 1277-1283.

116. Rathi, J., *Mcirobial Physiology Genetics and Ecology*. 2009, Delhi Manglam Publications.

117. Nascimento, M.M., J.A.C. Lemos, J. Abranches, R.B. Goncalves, and R.A. Burne, *Adaptive Acid Tolerance Response of Streptococcus sobrinus.* J. Bacteriol., 2004. 186(19): p. 6383-6390.

118. Blankenhorn, D., J. Phillips, and J.L. Slonczewski, *Acid- and Base-Induced Proteins during Aerobic and Anaerobic Growth of Escherichia coli Revealed by Two-Dimensional Gel Electrophoresis.* J. Bacteriol., 1999. 181(7): p. 2209-2216.

119. Shima, S. and R. K.Thauer, *A third type of hydrogenase catalyzing H2 activation.* The chemical record, 2007. 7: p. 37-46.

120. Guerrini, O., B. Burlat, C. Leger, B. Guigliarelli, P. Soucaille, and L. Girbal, *Chracterization of Two 2[4Fe4S] Ferredoxins from Clostridium acetobutylicum.* Curr Microbiol, 2008. 56:261-267.

121. Chen, J.-S., J. Toth, and M. Kasap, *Nitrogen-fixation genes and nitrogenase activity in Clostridium acetobutylicum and Clostridium beijerinckii.* Journal of Industrial Microbiology & Biotechnology, 2001 27: p. 281 –286.

122. Butland, G., J.M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili, *Interaction network containing conserved and essential protein complexes in Escherichia coli.* Nature, 2005. 433(7025): p. 531-537.

123. Fani, R., R. Gallo, and P. Liò, *Molecular evolution of nitrogen fixation: the evolutionary history of the nifD, nifK, nifE, and nifN genes.* Journal of Molecular Evolution, 2000. 51(1): p. 1-11.

124. Atkinson, M.R., T.A. Blauwkamp, and A.J. Ninfa, *Context-Dependent Functions of the PII and GlnK Signal Transduction Proteins in Escherichia coli.* J. Bacteriol., 2002. 184(19): p. 5364-5375.

125. Steffes, C., J. Ellis, J. Wu, and B.P. Rosen, *The lysP gene encodes the lysine-specific permease.* J. Bacteriol., 1992. 174(10): p. 3242-3249.

126. Chou, H.T., M. Hegazy, and C.-D. Lu, *L-Lysine Catabolism Is Controlled by L-Arginine and ArgR in Pseudomonas aeruginosa PAO1.* J. Bacteriol., 2010. 192(22): p. 5874-5880.

127. Jensen, L.J., M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C.v. Mering, *String 8-a global veiw on protiens and their functional interactions in 630 organisms.* Nucleic Acids Research, 2009. 37: p. D412-654.

128. Schmidt, M.C., A.M. Rocha, K. Padmanabhan, J.D. Young, J.R. Mihelcic, and N.F. Samatova. *NIBBS-Search for Fast and Accurate Prediction of Phenotype-Biased Metabolic Systems*. in *ACM SIGKDD Conference*. submitted.

129. Chen, W., A.M. Rocha, W. Hendrix, M. Schmidt, and N.F. Samatova. *The Multiple Alignment Algorithm for Metabolic Pathways without Abstraction*. in *2010 IEEE International Conference on Data Mining Workshops*. 2010: ICDMW.

130. Hendrix, W., A.M. Rocha, M. Elmore, J. Trien, and N.F. Samatova, *Discovery of Enriched Biological Motifs Using Knowledge Priors with Application to Biohydrogen Production*, in *2010 International Conference on Bioinformatics & Computational Biology, BIOCOMP 2010*. 2010: Las Vegas, Nevada. p. 17-23.

131. Singh, R., J. Xu, and B. Berger, *Global alignment of multiple protein interaction networks with application to functional orthology detection.* Proceedings of the National Academy of Sciences, 2008. 105(35): p. 12763-12768.

132. He, G., J. Liu, and C. Zhao, *Approximation algorithms for some graph partitioning problems.* Journal of Graph Algorithms and Applicaitons, 2000. 4(2): p. 1-11.

133. Prelić, A., S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, *A systematic comparison and evaluation of biclustering methods for gene expression data.* Bioinformatics, 2006. 22(9): p. 1122-1129.

134. Bron, C. and J. Kerbosch, *Algoirthm 457: Finding All Cliques of an Undirected Graph.* Communications of the ACM, 1973. 16(9): p. 757-577.

135. Sims, R.E.H. and N.E. Bassam, *Biomass and Resources*, in *Bioenergy options for a cleaner environment*, R.E.H. Sims, Editor. 2003, Elsevier Ltd.: USA. p. 1-28.

136. Lewis, N.S. and D.G. Nocera, *Powering the planet: Chemical challenges in solar energy utilization.* Proceedings of the National Academy of Sciences, 2006. 103(43): p. 15729-15735.

137. Townsend, A., B. Broas, C. Jenkins, and K. Ray, *Exploring sustainable biodiesel*, ed. A. Townsend, B. Broas, and K. Ray. 2008: Schiffer Publishing Ltd. 10-17.

138. Accounts, T.C.o.P., *The Energy Report*. 2008.

139. Greene, D.L. and A. Schafer. *Reducing greenhouse gas emission from U.S. transportation. PEW Center on Global Climate Changes (2003)*. 2007 [cited 2009 December]; Available from: http://www.environlink.org/resource.html?itemid=200305291231210.23591&catid=6.

140. IEA, *Renewables information, OECD and International Energy Agency Joing Publication*. 2002: Paris.

141. Meng, X., J. Yang, X. Xu, L. Zhang, Q. Nie, and M. Xian, *Biodiesel production from oleaginous microorganisms.* Renewable Energy, 2009. 34(1): p. 1-5.

142. Davidson, S., *Sustainable bioenergy: Genomics and biofuels development.* Nature Education, 2008. 1: p. 1.

143. Yadvika, Santosh, T.R. Sreekrishnan, S. Kohli, and V. Rana, *Enhancement of biogas production from solid substrates using different techniques--a review.* Bioresource Technology, 2004. 95(1): p. 1-10.

144. Tilche, A. and M. Galatola, *The potential of biomethane as bio-fuel/bio-energy for reducing greenhouse gas emissions: A qualitative assessment for Eurpoe in a life cycle perspective.* Water Sci. Tech., 2008. 57(11): p. 1883-1892.

145. Harikishan, S., *Biogas processing and utilization as an energy source,* in *Anaerobic biotechnology for bioenergy production, principles and applications.* , S.K. Khanal, Editor. 2008, Wiley-Blackwell. p. 267-291.

146. de Vrije, T., G.G. de Haas, G.B. Tan, E.R.P. Keijsers, and P.A.M. Claassen, *Pretreatment of Miscanthus for hydrogen production by Thermotoga elfii.* International Journal of Hydrogen Energy. 27(11-12): p. 1381-1390.

147. Fang, H.H.P., H. Liu, and T. Zhang, *Characterization of a hydrogen-producing granular sludge.* Biotechnology and Bioengineering, 2002. 78(1): p. 44-52.

APPENDICES

Appendix A: Brief Overview of Approaches in Energy and Fuel Production

A.1 Introduction

Increase in demand of fossil fuel resources to support the world's growing population has led to rising concern regarding sustainability of natural resources, increase in global warming due to greenhouse gas emissions, energy security, and the potential impacts of limited resources on local and national economies [35, 135-137]. Although the current supply of fossil fuels necessary for energy supply is still abundant, continued consumption of these resources at high rates cannot continue without long-term environmental, social, and economic impacts.

In the U.S. alone, approximately 25% of the world's natural resources are used for heating, electricity, transportation, and industrial use resulting in consumption of over 100 quadrillion British thermal units (BTU's) of energy each year [35, 53]. In 2006, the U.S. used approximately 28.5% of its energy to support transportation [138] and since the transportation sector relies heavily on petroleum based products, this sector is one of the largest contributors to the release of greenhouse emissions [137, 139]. The impact of these greenhouse gas emissions on global climate is of great concern globally and has become one of the leading driving factors for development of cleaner, alternative resources. Accordingly, in the last decade, U.S. government agencies and industries have begun endeavors to develop and improve upon current alternative technologies for production of clean, alternative sources for energy and fuel [35, 38, 135].

Currently, there are a number of promising technologies for production of bioenergy and biofuels, including efforts to harness wind, water, and solar energy [35]. Although utilization of these methods provides clean energy, many of these technologies are limited by supply of the resource they are dependent on. For example, energy power harvested by wind or solar energy is restricted on calm and cloudy days, respectively. Also, these technologies are often regionally applicable, and to incorporate them widely on a national and/or global level would be costly and is unlikely in the near future.

One promising technology currently being assessed is the utilization of biomass as a renewable energy source for the production of bioenergy and biofuels [135, 137]. The term biomass refers to biologically derived organic matter, including plant derivatives, crop and forest residues, municipal solid waste material (e.g., wastewater sludge), animal wastes, and agricultural crops which can be used as substrates in the production of biofuels [135].

Efforts to harness biomass as a renewable energy resource are currently underway in developed and developing countries. According to reports by the International Energy Agency (IEA), bioenergy and biofuel resources can contribute more than 80 tera-Watts of electricity and more than 200 peta-Joules of heat, respectively [135, 140]. This would provide a significant contribution towards meeting energy and fuel daily demands in the U.S. [135]. In addition to providing heat, electricity, and fuels for transportation, use of biomass and production of biofuels is considered to be more carbon neutral, thus providing a cleaner alternative renewable source [38, 135].

To date, a number of potentially important renewable energy resources derived from various biomass resources have been identified. These include: bioethanol [37], biomethane [53], biodiesel [137], and biohydrogen [4, 54]. In each process, the bioavailability of the biomass and selection of microorganisms are key factors in production of biofuels. If microorganisms used in these systems, such as those used for biohydrogen production, are not able to use the available biomass or utilize a small portion of the biomass, biofuels will not be produced or produced in small amounts. As such, when designing systems for biofuel production, it is important to understand what types of microorganisms are commonly used to generate biofuels. To help provide a general idea, a list of microorganisms and the renewable resource they produce is provided in Table A.1. A brief description of each biofuel and biomass resource is provided in the following sections.

A.1.1 Biodiesel Production from Oils and Fatty Acids

Similar to bioethanol, biodiesel is another well-established, renewable energy resource currently produced to help meet industrial and transportation needs [37, 137]. When combined with gasoline, biodiesel can help provide a cleaner energy source by reducing the amount of greenhouse emissions. Production of diesel using biomass resources is considered a green technology since carbon dioxide is not emitted into the atmosphere [141]. Another important feature of biodiesel production is biodiesel generated from industrial plants can be operated at standard temperature and pressures, thus reducing the amount of energy required for trans-esterification of the biomass feedstock [137].

Production of biodiesel occurs through transesterification of triacylgylcerols─a chemical process using alcohol, methanol or other catalyst, and oil or fat-based biomass. Biodiesel feed stocks include animal fats, fat waste products (i.e. grease), soybean oil,

Table A.1 List of renewable resources produced biologically and key microorganisms involved in the process.

Renewable Resource	Method Description	Microorganisms involved	References
Ethanol	Microbial fermentation of corn, glucose, or other biomass; enzymatic hydrolysis of biomass feed stock	*Saccharomyces* yeast, *Zymomonas mobilis, Clostridium*	[6, 13]
Diesel	Biologically via microbial lipid production	Microalgae, some yeast and fungi; Bacteria (*Bacillus alcalophilus, Acinetobacter caloaceticus*)	[35]
Methane	Anaerobic digestion of organic material	Methanogens	[6, 135, 136]
Hydrogen	Light and dark fermentation; Bio-photolysis	*Anabaena vaiabilis, Rhodobacter sphaeroides, Clostridium acetobutylicum, Clostridium thermocellum, Escherichia coli, Enterobacter aerogenes*	[6, 20, 22, 137]

palm, coconut, peanut, and sunflower oils [37, 141]. Chemical processing of oil-based and fat feed stocks is currently the most common technique for production of biodiesel. However, new studies are being conducted to evaluate microbial processes for production of oil by “oleaginous” microorganisms [141]. Oleaginous microorganisms are organisms whose cells contain lipids in excess of 20%. Examples of these organisms include some fungi, yeast, bacillus, and photosynthesizing microorganisms, such as microalgae and cyanobacteria [141, 142]. Unlike transesterification of fats, the use of biological processes to produce diesel may require large energy inputs to support growth of photosynthetic microorganisms [53]. Addition of a light source for microbial growth may result in increase costs compared to chemical processing methods. Although biological production of biodiesel is a promising technique much research is needed to understand the biological metabolic processes for production of diesel.

A.1.2 Biomethane/Biogas as an Alternative Renewable Source

Biogas (e.g., biomethane) produced during anaerobic digestion of organic matter in wastewater and sanitary landfills is another potentially important source for production of bioenergy. During anaerobic digestion of organic wastes, a mixture of known greenhouse gases, including methane, is often produced [143, 144]. Depending on the type of organic materials digested and microorganisms digesting the material, the mixture of biogas may contain up to 65% to 75% methane gas [144, 145]. If released into the environment, methane will contribute to increase global warming because it is a greenhouse gas. However, methane and other biogases recovered from anaerobic digesters could be a potential source of bioenergy [145]. Although production of methane in anaerobic digesters is not new, utilization of biomethane and biogas as alternative fuel and energy sources is still a fairly new concept, and technology for capturing and maximizing production of biomethane is currently being developed.

A.1.3 Biohydrogen as an Alternative Renewable Energy Source

Biological hydrogen production is slowly becoming a popular alternative as a renewable energy source for biofuels and bioenergy [53, 54]. Reasons for this include: (1) the ability of microorganisms to generate biohydrogen using three different biological processes; (2) the availability of some microbial species to produce hydrogen through different metabolic pathways, and (3) the large quantities of biomass available from other renewable resources, such as wastewater, for production of hydrogen [55].

The availability of multiple methods to produce hydrogen allows for development of different designs to suit current treatment plants and systems. The three types of biological production methods for hydrogen production are photo-fermentation of organic matter, dark fermentation of organic matter, and decomposition of water by photosynthesizing microorganisms [36, 37, 53]. An overview of the three biological methods is presented in Table A.2. In these approaches, production of biohydrogen using photosynthesizing microorganisms or photo-fermentation requires a large amount of light energy. Similar to the previous method, artificial light energy would result in increased production costs. If light energy from the sun were used, the production of hydrogen would be reliant upon sunshine and would still require a secondary light source (artificial) to supplement the system when sunlight is not available.

Unlike the previous two methods, dark fermentation processes do not require excessive energy from light since organisms involved in this process use organic materials, such as those found in the wastewater, as their energy source. Utilization of waste materials (e.g., wastewater sludge) and wastewaters as biomass in fermentation is a potential green technology currently being evaluated for simultaneous production of two renewable resources─energy and water. Use of organic materials from wastewater can result in the production of reusable water as a result of microbial degradation of organic materials from wastewater [54].

Table A.2 Overview of the three types of biological hydrogen production

Method	Description	Microorganisms Involved	References
Bio-photolysis	Decomposition of water by photosynthesizing organisms. H_2 (gas) is produced from water and sunlight	Chlamydomonas reinhardtii, Chlorococcum littorale, Chlorella fusca, Anabaena cylindrical, Anabaena variabilis	[15, 24]
Light Fermentation	H_2 (gas) is produced using organic materials, such as waste material. The organic matter is used by heterotrophs as a carbon source. Light is used to supply the energy source.	*Rhodobacter sphaeroides, Rhodobacter capsulatus, Rhodopseudomonas palustris, Cloroflexus aurantiacus*	[15, 24, 138]
Dark Fermentation	H_2 (gas) is produced from organic matter. The organic material is used as the carbon and energy source.	*Escherichia coli, Clostridium thermocellum, Clostridium acetobutylicum, Enterobacter aerogenes, Thermotoga neapolitana*	[6, 15, 25, 138]

Depending on the organic substrate and the microbial community present, it is possible to remove almost all organic material from wastewater, thus producing clean water for reuse or discharge to aquatic ecological systems. In addition, the ability to use high volumes of wastewater derived organic matter allows for development of large fermentation plants and increased hydrogen production. In light and dark fermentation batch reactors, specific hydrogen production rates ranging from 0.7 mL/g VSS h to 2,389 mL/g VSS h have been reported [11]. Variation in hydrogen production rates is dependent on the amount and type of carbohydrates and proteins present in the waste material. An overall estimate of energy production from biohydrogen has been reported to be approximately140 Joules of energy per gram of hydrogen gas. This estimate is more than half of the potential energy produced by biomethane and biogas [54].

To date, several naturally occurring hydrogen-producing organisms have been identified. Types of microorganisms identified range in their phenotypes including thermophilic bacteria (e.g., *Thermotoga neapolitana* DSM 4359), photosynthetic cyanobacteria (e.g., *Chloroflexus aurantiacus*), mesophilic bacteria (e.g., *Escherichia coli*), facultative anaerobes (e.g., *Escherichia coli*), and strict anaerobes (e.g., *Clostridium acetobutylicum*) [3, 11, 12, 22, 53, 54, 146]. Of the hydrogen-producing organisms identified, a majority of these species appear to utilize dark fermentation metabolic processes to produce hydrogen. Since no one microorganism can utilize all potential biomass feedstock sources and organic matter in wastewaters, a combination of anaerobic, fermentative bacteria are often used to carry out dark fermentation processes. Among these species, several *Clostridium* species are capable of utilizing both simple sugars (e.g., glucose) and cellulosic materials [53, 147]. Although the full maximum potential for biohydrogen produced in anaerobic fermentation plants has not been reached, many genomics and bioengineering studies are currently being conducted to identify environmental, metabolic, and chemical parameters necessary to increase biohydrogen production by hydrogen–producing microorganisms.

A.1.4 Concerns Regarding Production of Biofuels and Bioenergy

Production of bioenergy and biofuels from biologically derived materials is not a new concept in the U.S. [53]. In fact, many industrial plants are already established to produce ethanol for industrial use. However, when it comes to production of bioenergy or biofuels on a national scale, there are still many questions and concerns that need to be addressed. One question in particular is “Where will the biomass resource come from?” For all of the renewable resources described above, production is dependent on availability of biologically derived feed stocks for mass production of biofuels and bioenergy [37, 54].

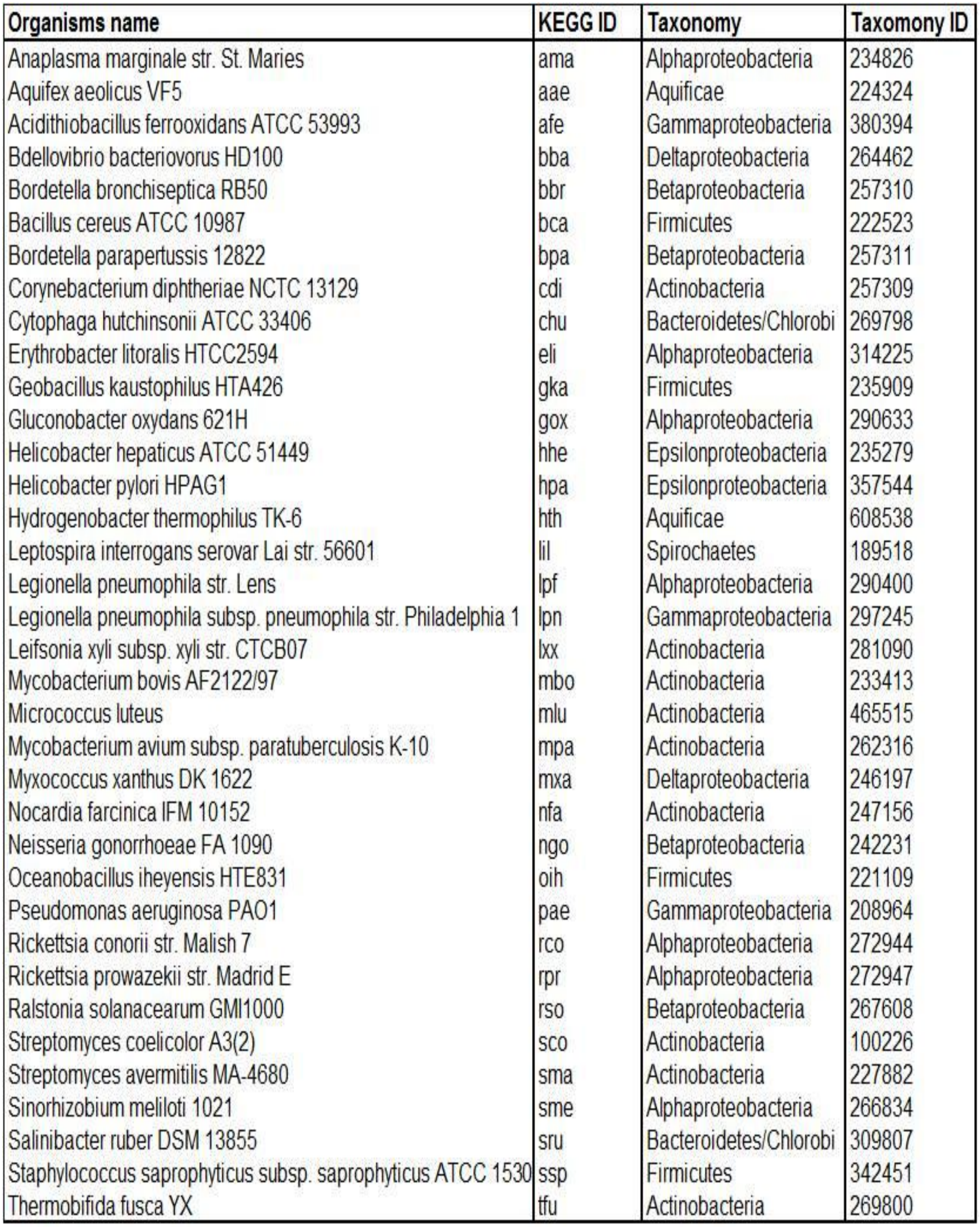
For bioethanol, primary sources for biomass feed stocks are plant residues, sugarcane, and corn. In order to produce large enough quantities to support both ethanol and biofuel for industrial, commercial, and transportation needs, mass production of corn and agricultural crops is necessary. However, mass production of these crops for biofuels creates direct competition with the production of food crops and land needed for agriculture [135]. Similarly, production of biodiesel requires accumulation of biomass feed stocks from sources including agricultural plants.

Biomass feedstock resources for biohydrogen and biomethane production consist primarily of organic matter found in wastewaters, animal wastes and plant residues [54, 143, 144]. However, utilization of biomethane and biogas is still fairly new in terms of technology, and issues regarding the separation of methane from other biogas produced in the anaerobic digester, is still in question. Unlike biomethane, direct utilization of biohydrogen at wastewater treatment plants is possible. Recent studies evaluating the use of hydrogen as energy demonstrate that hydrogen produced by microorganisms can be captured and used directly, like a battery, in systems called microbial fuel cells [35]. If hydrogen-producing microorganisms in these fuel cells can produce enough energy, this energy can possibly be used for transportation and maintenance of wastewater treatment plants [35]. The ability of microorganisms involved in biohydrogen production to aid in treatment wastewater through removal of organic substrates while producing biohydrogen to support biofuel and bioenergy makes biohydrogen an ideal candidate as a renewable energy resource.

Unfortunately, one major limitation to production of biohydrogen in wastewater treatment systems is the ability of anaerobic fermentative bacteria to utilize a wide variety of complex organic materials and carbon sources. Many anaerobic organisms have a variety of fermentation metabolic pathways that are used to degrade different carbon sources, such as glucose. However, wastewater and waste materials contain a mixture of different carbohydrates, proteins, and other biomass materials. Depending on the microbial community present in anaerobic treatment of wastewater, the amount of hydrogen produced will vary. As such, selection of microorganisms capable of utilizing primary carbons sources may have to be determined individually at each wastewater treatment plant [54].

Appendix B: List of Organisms Used in the Student’s T-Test and NIBBS Experiments

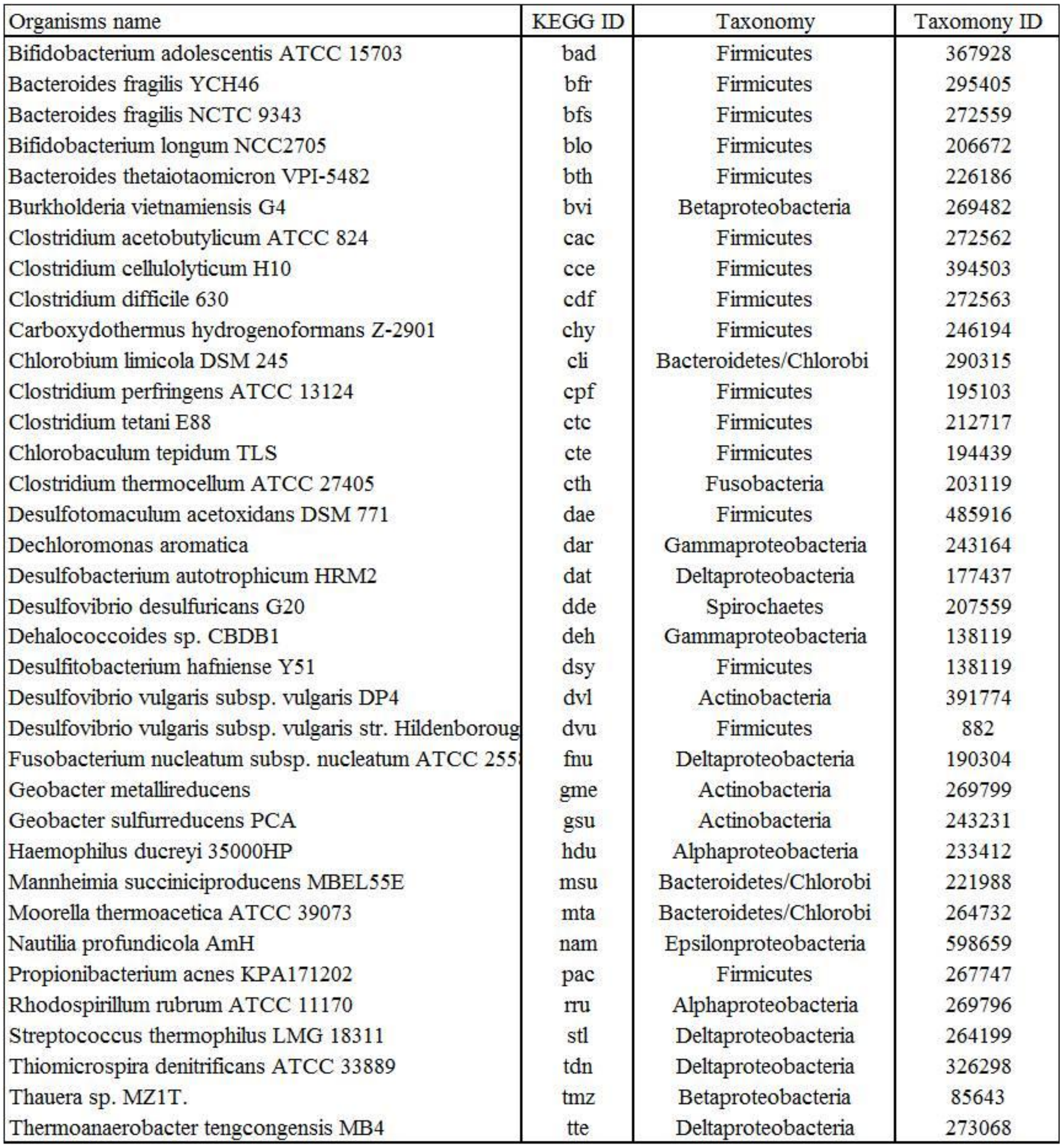
Table B.1 List of aerobic organisms used in the T-test and NIBBS experiment.



Organisms name	KEGG ID	Taxonomy	Taxomony ID
Anaplasma marginale str. St. Maries	ama	Alphaproteobacteria	234826
Aquifex aeolicus VF5	aae	Aquificae	224324
Acidithiobacillus ferrooxidans ATCC 53993	afe	Gammaproteobacteria	380394
Bdellovibrio bacteriovorus HD100	bba	Deltaproteobacteria	264462
Bordetella bronchiseptica RB50	bbr	Betaproteobacteria	257310
Bacillus cereus ATCC 10987	bca	Firmicutes	222523
Bordetella parapertussis 12822	bpa	Betaproteobacteria	257311
Corynebacterium diphtheriae NCTC 13129	cdi	Actinobacteria	257309
Cytophaga hutchinsonii ATCC 33406	chu	Bacteroidetes/Chlorobi	269798
Erythrobacter litoralis HTCC2594	eli	Alphaproteobacteria	314225
Geobacillus kaustophilus HTA426	gka	Firmicutes	235909
Gluconobacter oxydans 621H	gox	Alphaproteobacteria	290633
Helicobacter hepaticus ATCC 51449	hhe	Epsilonproteobacteria	235279
Helicobacter pylori HPAG1	hpa	Epsilonproteobacteria	357544
Hydrogenobacter thermophilus TK-6	hth	Aquificae	608538
Leptospira interrogans serovar Lai str. 56601	lil	Spirochaetes	189518
Legionella pneumophila str. Lens	lpf	Alphaproteobacteria	290400
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	lpn	Gammaproteobacteria	297245
Leifsonia xyli subsp. xyli str. CTCB07	lxx	Actinobacteria	281090
Mycobacterium bovis AF2122/97	mbo	Actinobacteria	233413
Micrococcus luteus	mlu	Actinobacteria	465515
Mycobacterium avium subsp. paratuberculosis K-10	mpa	Actinobacteria	262316
Myxococcus xanthus DK 1622	mxa	Deltaproteobacteria	246197
Nocardia farcinica IFM 10152	nfa	Actinobacteria	247156
Neisseria gonorrhoeae FA 1090	ngo	Betaproteobacteria	242231
Oceanobacillus iheyensis HTE831	oih	Firmicutes	221109
Pseudomonas aeruginosa PAO1	pae	Gammaproteobacteria	208964
Rickettsia conorii str. Malish 7	rco	Alphaproteobacteria	272944
Rickettsia prowazekii str. Madrid E	rpr	Alphaproteobacteria	272947
Ralstonia solanacearum GMI1000	rso	Betaproteobacteria	267608
Streptomyces coelicolor A3(2)	sco	Actinobacteria	100226
Streptomyces avermitilis MA-4680	sma	Actinobacteria	227882
Sinorhizobium meliloti 1021	sme	Alphaproteobacteria	266834
Salinibacter ruber DSM 13855	sru	Bacteroidetes/Chlorobi	309807
Staphylococcus saprophyticus subsp. saprophyticus ATCC 1530	ssp	Firmicutes	342451
Thermobifida fusca YX	tfu	Actinobacteria	269800

Appendix B (Continued)

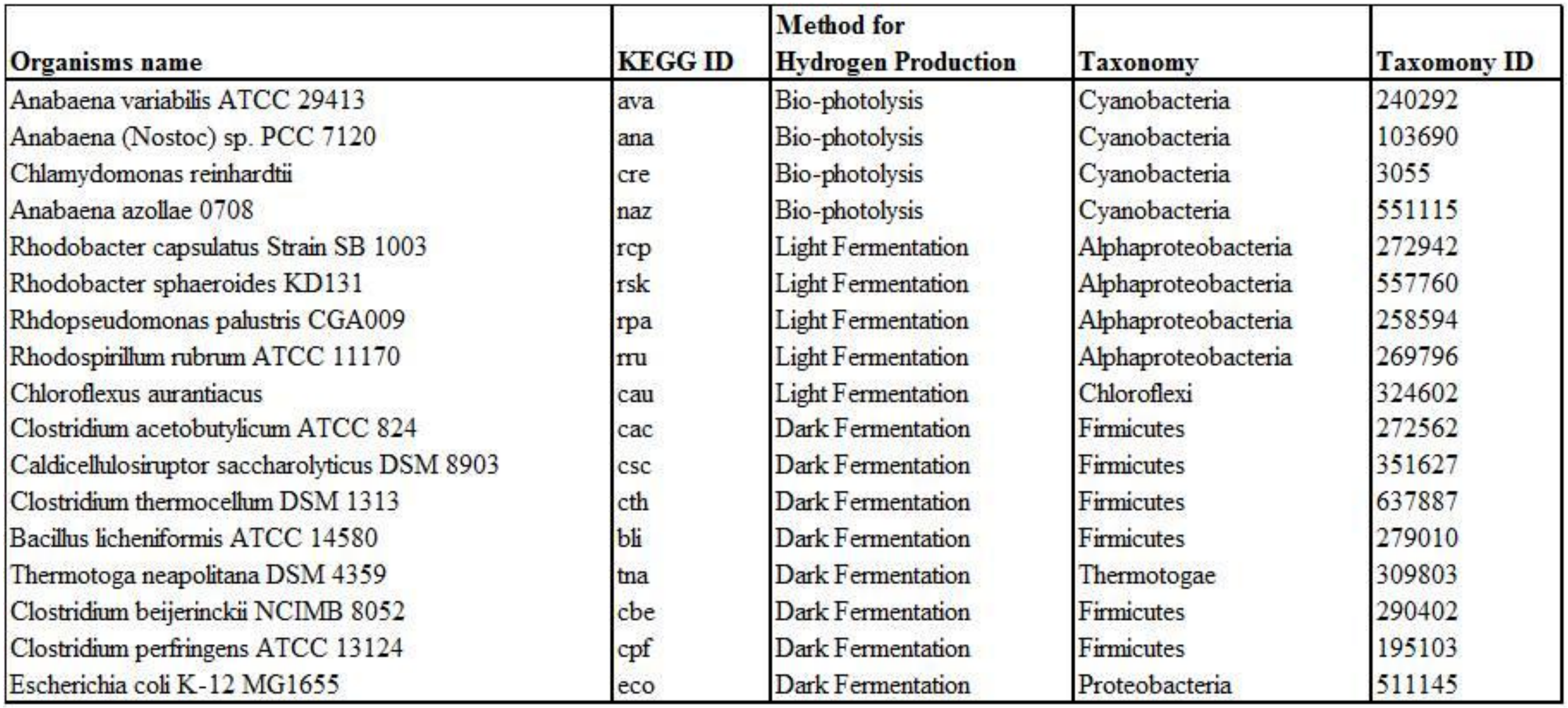
Table B.2 List of anaerobic organisms used in the T-test and NIBBS experiment.



Organisms name	KEGG ID	Taxonomy	Taxomony ID
Bifidobacterium adolescentis ATCC 15703	bad	Firmicutes	367928
Bacteroides fragilis YCH46	bfr	Firmicutes	295405
Bacteroides fragilis NCTC 9343	bfs	Firmicutes	272559
Bifidobacterium longum NCC2705	blo	Firmicutes	206672
Bacteroides thetaiotaomicron VPI-5482	bth	Firmicutes	226186
Burkholderia vietnamiensis G4	bvi	Betaproteobacteria	269482
Clostridium acetobutylicum ATCC 824	cac	Firmicutes	272562
Clostridium cellulolyticum H10	cce	Firmicutes	394503
Clostridium difficile 630	cdf	Firmicutes	272563
Carboxydothermus hydrogenoformans Z-2901	chy	Firmicutes	246194
Chlorobium limicola DSM 245	cli	Bacteroidetes/Chlorobi	290315
Clostridium perfringens ATCC 13124	cpf	Firmicutes	195103
Clostridium tetani E88	ctc	Firmicutes	212717
Chlorobaculum tepidum TLS	cte	Firmicutes	194439
Clostridium thermocellum ATCC 27405	cth	Fusobacteria	203119
Desulfotomaculum acetoxidans DSM 771	dae	Firmicutes	485916
Dechloromonas aromatica	dar	Gammaproteobacteria	243164
Desulfobacterium autotrophicum HRM2	dat	Deltaproteobacteria	177437
Desulfovibrio desulfuricans G20	dde	Spirochaetes	207559
Dehalococcoides sp. CBDB1	deh	Gammaproteobacteria	138119
Desulfitobacterium hafniense Y51	dsy	Firmicutes	138119
Desulfovibrio vulgaris subsp. vulgaris DP4	dvl	Actinobacteria	391774
Desulfovibrio vulgaris subsp. vulgaris str. Hildenboroug	dvu	Firmicutes	882
Fusobacterium nucleatum subsp. nucleatum ATCC 255	fnu	Deltaproteobacteria	190304
Geobacter metallireducens	gme	Actinobacteria	269799
Geobacter sulfurreducens PCA	gsu	Actinobacteria	243231
Haemophilus ducreyi 35000HP	hdu	Alphaproteobacteria	233412
Mannheimia succiniciproducens MBEL55E	msu	Bacteroidetes/Chlorobi	221988
Moorella thermoacetica ATCC 39073	mta	Bacteroidetes/Chlorobi	264732
Nautilia profundicola AmH	nam	Epsilonproteobacteria	598659
Propionibacterium acnes KPA171202	pac	Firmicutes	267747
Rhodospirillum rubrum ATCC 11170	rru	Alphaproteobacteria	269796
Streptococcus thermophilus LMG 18311	stl	Deltaproteobacteria	264199
Thiomicrospira denitrificans ATCC 33889	tdn	Deltaproteobacteria	326298
Thauera sp. MZ1T.	tmz	Betaproteobacteria	85643
Thermoanaerobacter tengcongensis MB4	tte	Deltaproteobacteria	273068

Appendix B (Continued)

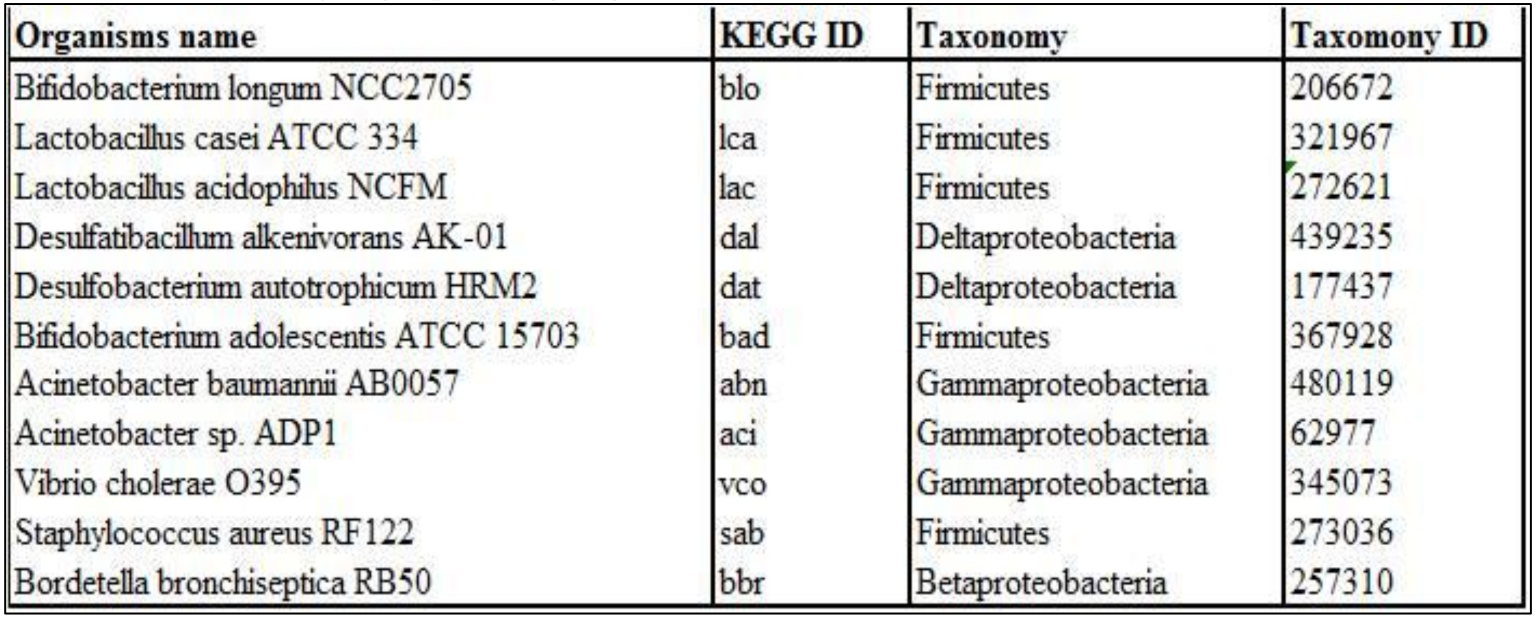
Table B.3 List of hydrogen producing organisms used in the T-test and NIBBS experiment.



Organisms name	KEGG ID	Method for Hydrogen Production	Taxonomy	Taxomony ID
Anabaena variabilis ATCC 29413	ava	Bio-photolysis	Cyanobacteria	240292
Anabaena (Nostoc) sp. PCC 7120	ana	Bio-photolysis	Cyanobacteria	103690
Chlamydomonas reinhardtii	cre	Bio-photolysis	Cyanobacteria	3055
Anabaena azollae 0708	naz	Bio-photolysis	Cyanobacteria	551115
Rhodobacter capsulatus Strain SB 1003	rcp	Light Fermentation	Alphaproteobacteria	272942
Rhodobacter sphaeroides KD131	rsk	Light Fermentation	Alphaproteobacteria	557760
Rhdopseudomonas palustris CGA009	rpa	Light Fermentation	Alphaproteobacteria	258594
Rhodospirillum rubrum ATCC 11170	rru	Light Fermentation	Alphaproteobacteria	269796
Chloroflexus aurantiacus	cau	Light Fermentation	Chloroflexi	324602
Clostridium acetobutylicum ATCC 824	cac	Dark Fermentation	Firmicutes	272562
Caldicellulosiruptor saccharolyticus DSM 8903	csc	Dark Fermentation	Firmicutes	351627
Clostridium thermocellum DSM 1313	cth	Dark Fermentation	Firmicutes	637887
Bacillus licheniformis ATCC 14580	bli	Dark Fermentation	Firmicutes	279010
Thermotoga neapolitana DSM 4359	tna	Dark Fermentation	Thermotogae	309803
Clostridium beijerinckii NCIMB 8052	cbe	Dark Fermentation	Firmicutes	290402
Clostridium perfringens ATCC 13124	cpf	Dark Fermentation	Firmicutes	195103
Escherichia coli K-12 MG1655	eco	Dark Fermentation	Proteobacteria	511145

Appendix B (Continued)

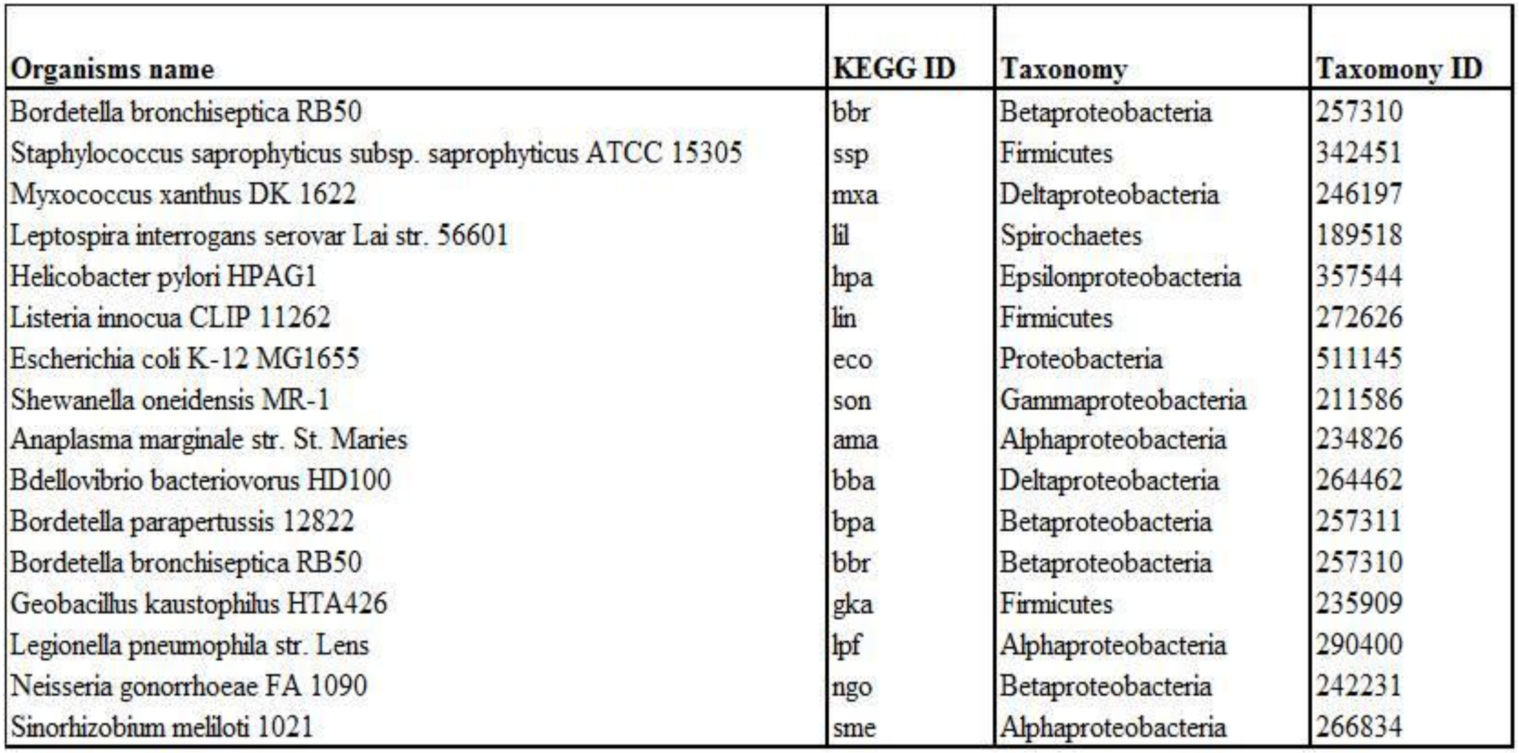
Table B.4 List of 'non-hydrogen producing' organisms used in the T-test and NIBBS experiment.



Organisms name	KEGG ID	Taxonomy	Taxomony ID
Bifidobacterium longum NCC2705	blo	Firmicutes	206672
Lactobacillus casei ATCC 334	lca	Firmicutes	321967
Lactobacillus acidophilus NCFM	lac	Firmicutes	272621
Desulfatibacillum alkenivorans AK-01	dal	Deltaproteobacteria	439235
Desulfobacterium autotrophicum HRM2	dat	Deltaproteobacteria	177437
Bifidobacterium adolescentis ATCC 15703	bad	Firmicutes	367928
Acinetobacter baumannii AB0057	abn	Gammaproteobacteria	480119
Acinetobacter sp. ADP1	aci	Gammaproteobacteria	62977
Vibrio cholerae O395	vco	Gammaproteobacteria	345073
Staphylococcus aureus RF122	sab	Firmicutes	273036
Bordetella bronchiseptica RB50	bbr	Betaproteobacteria	257310

Appendix B (Continued)

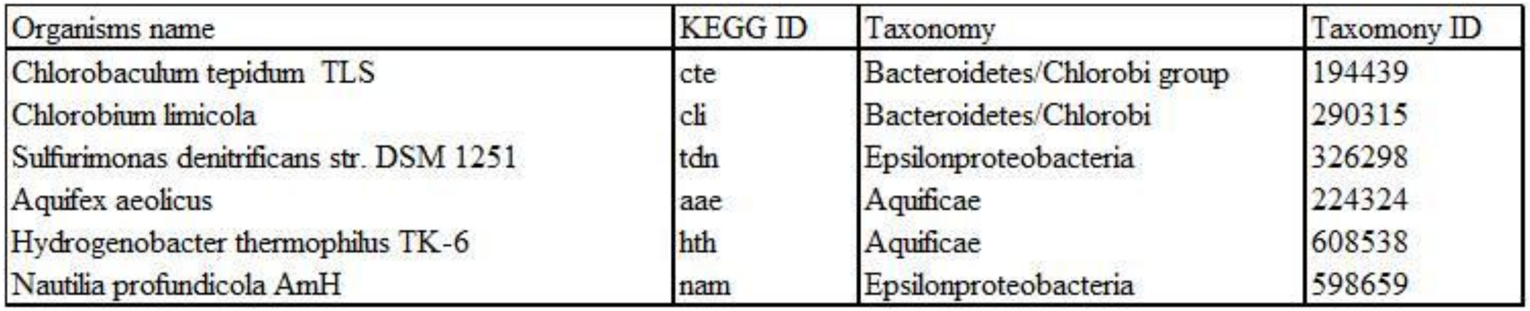
Table B.5 List of TCA expressing organisms used in the T-test and NIBBS experiment.



Organisms name	KEGG ID	Taxonomy	Taxomony ID
Bordetella bronchiseptica RB50	bbr	Betaproteobacteria	257310
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	ssp	Firmicutes	342451
Myxococcus xanthus DK 1622	mxa	Deltaproteobacteria	246197
Leptospira interrogans serovar Lai str. 56601	lil	Spirochaetes	189518
Helicobacter pylori HPAG1	hpa	Epsilonproteobacteria	357544
Listeria innocua CLIP 11262	lin	Firmicutes	272626
Escherichia coli K-12 MG1655	eco	Proteobacteria	511145
Shewanella oneidensis MR-1	son	Gammaproteobacteria	211586
Anaplasma marginale str. St. Maries	ama	Alphaproteobacteria	234826
Bdellovibrio bacteriovorus HD100	bba	Deltaproteobacteria	264462
Bordetella parapertussis 12822	bpa	Betaproteobacteria	257311
Bordetella bronchiseptica RB50	bbr	Betaproteobacteria	257310
Geobacillus kaustophilus HTA426	gka	Firmicutes	235909
Legionella pneumophila str. Lens	lpf	Alphaproteobacteria	290400
Neisseria gonorrhoeae FA 1090	ngo	Betaproteobacteria	242231
Sinorhizobium meliloti 1021	sme	Alphaproteobacteria	266834

Appendix B (Continued)

Table B.6 List of rTCA expressing organisms used in the T-test and NIBBS experiment.



Organisms name	KEGG ID	Taxonomy	Taxomony ID
Chlorobaculum tepidum TLS	cte	Bacteroidetes/Chlorobi group	194439
Chlorobium limicola	cli	Bacteroidetes/Chlorobi	290315
Sulfurimonas denitrificans str. DSM 1251	tdn	Epsilonproteobacteria	326298
Aquifex aeolicus	aae	Aquificae	224324
Hydrogenobacter thermophilus TK-6	hth	Aquificae	608538
Nautilia profundicola AmH	nam	Epsilonproteobacteria	598659

Appendix C: Summary of Enzymes in the Dataset Identified by the NIBBS Algorithm and Student's T-Test for Each Phenotype

Table C.1 Summary of enzymes in the dataset identified (True) by the NIBBS algorithm and Student's T-test for the phenotype anaerobic respiration.

EC Number	Enzyme Name	In T-Test	NIBBS
3.5.4.12	dCMP deaminase;	TRUE	TRUE
3.6.1.23	dUTP diphosphatase;		TRUE
2.3.1.54	formate C-acetyltransferase;	TRUE	TRUE
1.17.4.2	ribonucleoside-triphosphate reductase;	TRUE	TRUE
1.2.7.3	2-oxoglutarate synthase;	TRUE	TRUE
1.4.1.13	glutamate synthase (NADPH);	TRUE	TRUE
4.1.1.19	arginine decarboxylase;		TRUE
2.3.3.14	homocitrate synthase;	TRUE	TRUE
1.18.6.1	nitrogenase	TRUE	TRUE
1.2.7.5	aldehyde ferredoxin oxidoreductase;	TRUE	TRUE
1.2.1.41	glutamate-5-semialdehyde dehydrogenase;		TRUE
1.4.1.14	glutamate synthase (NADH);	TRUE	TRUE
6.3.4.3	formate---tetrahydrofolate ligase;	TRUE	TRUE
4.1.1.3	oxaloacetate decarboxylase;	TRUE	TRUE
2.7.9.3	selenide, water dikinase;	TRUE	TRUE
2.3.1.1	amino-acid N-acetyltransferase;		TRUE
2.7.2.11	glutamate 5-kinase;		TRUE
5.1.1.3	glutamate racemase		TRUE
6.1.1.18	glutamine---tRNA ligase;	TRUE	TRUE
4.99.1.3	sirohydrochlorin cobaltochelatase;		TRUE
1.2.1.10	acetaldehyde dehydrogenase (acetylating);	TRUE	TRUE
2.1.1.151	cobalt-factor II C20-methyltransferase;		TRUE
2.4.2.22	xanthine phosphoribosyltransferase;	TRUE	TRUE
3.5.4.10	IMP cyclohydrolase;		TRUE
2.3.1.51	1-acylglycerol-3-phosphate O-acyltransferase;		TRUE
2.7.1.107	diacylglycerol kinase;	TRUE	TRUE
2.7.8.26	adenosylcobinamide-GDP ribazoletransferase;	TRUE	TRUE
6.3.1.10	adenosylcobinamide-phosphate synthase;	TRUE	TRUE
2.4.2.21	nicotinate-nucleotide---dimethylbenzimidazole	TRUE	TRUE
2.9.1.1	L-seryl-tRNASec selenium transferase;	TRUE	TRUE
2.7.1.156	adenosylcobinamide kinase;	TRUE	TRUE
6.1.1.17	glutamate---tRNA ligase;		TRUE
6.3.1.1	aspartate---ammonia ligase;	TRUE	TRUE
3.4.13.3	Xaa-His dipeptidase;	TRUE	TRUE
2.2.1.7	1-deoxy-D-xylulose-5-phosphate synthase;	TRUE	TRUE
2.7.7.60	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase;	TRUE	TRUE
1.2.7.1	pyruvate synthase;		TRUE

Appendix C (Continued)

Table C.1 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.7.1.11	6-phosphofructokinase;	TRUE	TRUE
3.1.3.15	histidinol-phosphatase;	TRUE	TRUE
4.3.1.1	aspartate ammonia-lyase;	TRUE	TRUE
6.3.4.5	argininosuccinate synthase;		TRUE
6.3.5.4	asparagine synthase (glutamine-hydrolysing);		TRUE
2.4.1.21	starch synthase;	TRUE	TRUE
5.1.3.14	UDP-N-acetylglucosamine 2-epimerase;	TRUE	TRUE
1.1.1.267	1-deoxy-D-xylulose-5-phosphate reductoisomerase;	TRUE	TRUE
2.5.1.17	cob(I)yrinic acid a,c-diamide adenosyltransferase;	TRUE	TRUE
2.3.1.30	serine O-acetyltransferase;		TRUE
5.4.99.5	chorismate mutase;	TRUE	TRUE
2.7.1.50	hydroxyethylthiazole kinase;		TRUE
1.1.1.29	glycerate dehydrogenase;	TRUE	TRUE
2.7.1.45	2-dehydro-3-deoxygluconokinase;		TRUE
4.3.1.17	L-serine ammonia-lyase;		TRUE
6.3.1.-	<Ligases; Forming carbon-nitrogen bonds; Acid--ammonia (or amine) ligases		TRUE
6.3.5.10	adenosylcobyric acid synthase (glutamine-hydrolysing);	TRUE	TRUE
6.3.5.9	hydrogenobyrinic acid a,c-diamide synthase (glutamine-hydrolysing);	TRUE	TRUE
3.1.3.73	alpha-ribazole phosphatase;	TRUE	TRUE
1.17.1.2	4-hydroxy-3-methylbut-2-enyl diphosphate reductase;		TRUE
1.8.99.2	adenylyl-sulfate reductase;	TRUE	TRUE
2.4.2.3	uridine phosphorylase;		TRUE
2.4.2.9	uracil phosphoribosyltransferase;		TRUE
2.7.7.13	mannose-1-phosphate guanylyltransferase;		TRUE
2.7.7.22	mannose-1-phosphate guanylyltransferase (GDP);		TRUE
4.2.1.47	GDP-mannose 4,6-dehydratase;		TRUE
4.2.1.8	mannonate dehydratase;		TRUE
2.3.1.19	phosphate butyryltransferase;	TRUE	TRUE
2.7.2.7	butyrate kinase	TRUE	TRUE
2.7.1.40	pyruvate kinase;		TRUE
4.6.1.12	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase;		TRUE
3.5.4.5	cytidine deaminase;		TRUE
3.5.99.6	glucosamine-6-phosphate deaminase;		TRUE
5.3.1.8	mannose-6-phosphate isomerase;		TRUE
2.1.2.5	glutamate formimidoyltransferase;	TRUE	TRUE
5.3.1.23	S-methyl-5-thioribose-1-phosphate isomerase;		TRUE
2.1.1.132	precorrin-6Y C5,15-methyltransferase (decarboxylating);		TRUE
1.17.7.1	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase;		TRUE
4.1.1.12	aspartate 4-decarboxylase;	TRUE	TRUE
4.2.1.10	3-dehydroquinate dehydratase;		TRUE
2.1.1.130	precorrin-2 C20-methyltransferase	TRUE	TRUE
2.7.9.1	pyruvate, phosphate dikinase;		TRUE

Appendix C (Continued)

Table C.1 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
3.5.4.2	adenine deaminase;		TRUE
1.1.1.271	GDP-L-fucose synthase;		TRUE
2.4.2.8	hypoxanthine phosphoribosyltransferase;		TRUE
1.1.1.25	shikimate dehydrogenase;		TRUE
2.7.1.33	pantothenate kinase;		TRUE
2.7.1.71	shikimate kinase;	TRUE	TRUE
6.3.2.5	phosphopantothenate---cysteine ligase;		TRUE
1.2.1.9	glyceraldehyde-3-phosphate dehydrogenase (NADP+);		TRUE
1.2.99.5	formylmethanofuran dehydrogenase;	TRUE	TRUE
5.4.99.1	methylaspartate mutase;	TRUE	TRUE
2.7.1.35	pyridoxal kinase;	TRUE	TRUE
6.3.3.2	5-formyltetrahydrofolate cyclo-ligase;		TRUE
2.7.7.24	glucose-1-phosphate thymidylyltransferase;	TRUE	TRUE
2.7.7.33	glucose-1-phosphate cytidylyltransferase;		TRUE
4.1.1.50	adenosylmethionine decarboxylase;		TRUE
1.1.1.27	L-lactate dehydrogenase;		TRUE
1.4.1.4	glutamate dehydrogenase (NADP+);		TRUE
1.4.3.16	L-aspartate oxidase;		TRUE
2.6.1.21	D-amino-acid transaminase;		TRUE
3.2.1.26	beta-fructofuranosidase;		TRUE
3.2.1.35	hyaluronoglucosaminidase;		TRUE
3.5.1.14	aminoacylase;		TRUE
3.5.4.1	cytosine deaminase;		TRUE
5.4.2.2	phosphoglucomutase;		TRUE
5.4.2.9	phosphoenolpyruvate mutase;	TRUE	TRUE
2.1.2.3	phosphoribosylaminoimidazolecarboxamide formyltransferase;	TRUE	TRUE
2.7.1.26	riboflavin kinase;		TRUE
4.2.3.5	chorismate synthase;		TRUE
5.3.1.9	glucose-6-phosphate isomerase;		TRUE
6.3.3.1	phosphoribosylformylglycinamidine cyclo-ligase;		TRUE
2.7.8.-	Transferases; Transferring phosphorus-containing groups		TRUE
3.1.3.18	phosphoglycolate phosphatase;		TRUE
2.-.-.-	<Transferases>		TRUE
2.6.1.1	aspartate transaminase;		TRUE
2.7.7.12	UDP-glucose---hexose-1-phosphate uridylyltransferase;	TRUE	TRUE

Table C.1 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.1.1.13	methionine synthase;	TRUE	TRUE
2.1.1.133	precorrin-4 C11-methyltransferase;	TRUE	TRUE
5.4.1.2	precorrin-8X methylmutase;	TRUE	TRUE
5.-.-.-	<Isomerases>		TRUE
2.7.1.49	hydroxymethylpyrimidine kinase;		TRUE
2.4.2.1	purine-nucleoside phosphorylase;		TRUE
2.4.1.25	4-alpha-glucanotransferase;		TRUE
2.7.2.1	acetate kinase;	TRUE	TRUE
1.1.1.44	phosphogluconate dehydrogenase (decarboxylating);		TRUE
1.12.98.1	coenzyme F420 hydrogenase;		TRUE
1.18.-.-	<Oxidoreductases; Acting on iron-sulfur proteins as donors>		TRUE
1.2.1.70	glutamyl-tRNA reductase		TRUE
1.2.1.72	erythrose-4-phosphate dehydrogenase;		TRUE
2.1.1.10	homocysteine S-methyltransferase;		TRUE
2.1.1.45	thymidylate synthase;		TRUE
2.4.1.119	dolichyl-diphosphooligosaccharide---protein glycotransferase;	TRUE	TRUE
2.7.1.21	thymidine kinase;		TRUE
2.8.3.1	propionate CoA-transferase;	TRUE	TRUE
3.1.1.31	6-phosphogluconolactonase;		TRUE
3.1.2.14	oleoyl-[acyl-carrier-protein] hydrolase;	TRUE	TRUE
3.4.11.2	membrane alanyl aminopeptidase;		TRUE
3.6.1.19	nucleoside-triphosphate diphosphatase;		TRUE
4.1.1.11	aspartate 1-decarboxylase;		TRUE
4.1.1.22	histidine decarboxylase;		TRUE
4.1.1.32	phosphoenolpyruvate carboxykinase (GTP);		TRUE
5.1.3.4	L-ribulose-5-phosphate 4-epimerase;	TRUE	TRUE
5.4.4.2	isochorismate synthase;		TRUE
3.6.1.15	nucleoside-triphosphatase;		TRUE
2.3.1.46	homoserine O-succinyltransferase;		TRUE
2.7.1.23	NAD+ kinase;		TRUE
2.7.7.18	nicotinate-nucleotide adenylyltransferase;		TRUE
6.3.5.1	NAD+ synthase (glutamine-hydrolysing);		TRUE
3.1.3.5	5'-nucleotidase;		TRUE
5.1.3.2	UDP-glucose 4-epimerase;		TRUE
1.3.1.-	<Oxidoreductases; Acting on the CH-CH group of donors;		TRUE
2.7.6.2	thiamine diphosphokinase;		TRUE
2.7.1.48	uridine kinase;		TRUE
1.-.-.-	<Oxidoreductases>		TRUE

Appendix C (Continued)

Table C.1 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
5.3.1.12	glucuronate isomerase;		TRUE
1.2.1.12	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating);		TRUE
1.3.3.1	dihydroorotate oxidase;	TRUE	TRUE
2.5.1.6	methionine adenosyltransferase;		TRUE
2.7.2.3	phosphoglycerate kinase;		TRUE
4.1.1.23	orotidine-5'-phosphate decarboxylase;		TRUE
5.1.3.1	ribulose-phosphate 3-epimerase;		TRUE
6.3.5.5	carbamoyl-phosphate synthase (glutamine-hydrolysing);		TRUE
4.-.-.-	<Lyases>		TRUE
1.1.1.169	2-dehydropantoate 2-reductase;		TRUE
1.14.13.81	magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase;		TRUE
1.3.1.83	geranylgeranyl diphosphate reductase;		TRUE
1.4.1.2	glutamate dehydrogenase;		TRUE
1.5.1.34	6,7-dihydropteridine reductase;		TRUE
2.2.1.9	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid		TRUE
2.3.1.117	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase;		TRUE
2.3.3.8	ATP citrate synthase;	TRUE	TRUE
2.6.1.17	succinyldiaminopimelate transaminase;		TRUE
2.7.1.113	deoxyguanosine kinase;		TRUE
2.7.1.20	adenosine kinase;		TRUE
2.7.1.29	glycerone kinase;		TRUE
2.7.7.63	lipoate---protein ligase;		TRUE
2.8.1.2	3-mercaptopyruvate sulfurtransferase;		TRUE
3.5.4.3	guanine deaminase;		TRUE
4.1.1.18	lysine decarboxylase;		TRUE
4.1.2.14	2-dehydro-3-deoxy-phosphogluconate aldolase;		TRUE
4.1.3.40	chorismate lyase;		TRUE
4.4.1.8	cystathionine beta-lyase;		TRUE
5.1.3.7	UDP-N-acetylglucosamine 4-epimerase;		TRUE
2.7.6.3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase;		TRUE
2.8.1.6	biotin synthase	TRUE	TRUE
2.5.1.15	dihydropteroate synthase;		TRUE
2.5.1.19	3-phosphoshikimate 1-carboxyvinyltransferase;		TRUE
4.1.1.20	diaminopimelate decarboxylase;		TRUE
6.3.4.4	adenylosuccinate synthase;		TRUE
6.3.5.3	phosphoribosylformylglycinamidine synthase;		TRUE
2.1.3.3	ornithine carbamoyltransferase;		TRUE
4.2.1.-	<Lyases; Carbon-oxygen lyases; Hydro-lyases>		TRUE

Table C.1 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
3.2.1.23	beta-galactosidase;		TRUE
1.2.1.2	formate dehydrogenase;		TRUE
2.8.3.8	acetate CoA-transferase;		TRUE
3.2.2.9	adenosylhomocysteine nucleosidase;		TRUE
1.1.1.262	4-hydroxythreonine-4-phosphate dehydrogenase;		TRUE
1.1.1.41	isocitrate dehydrogenase (NAD+);		TRUE
2.3.1.35	glutamate N-acetyltransferase;		TRUE
2.4.1.18	1,4-alpha-glucan branching enzyme;		TRUE
3.2.1.1	alpha-amylase;		TRUE
2.4.1.182	lipid-A-disaccharide synthase		TRUE
3.1.3.45	3-deoxy-manno-octulosonate-8-phosphatase		TRUE
1.1.1.86	ketol-acid reductoisomerase;		TRUE
2.2.1.6	acetolactate synthase;		TRUE
2.3.1.15	glycerol-3-phosphate O-acyltransferase;		TRUE
4.2.1.9	dihydroxy-acid dehydratase;		TRUE
4.2.3.12	6-pyruvoyltetrahydropterin synthase;		TRUE
1.1.1.95	phosphoglycerate dehydrogenase;		TRUE
3.5.4.26	diaminohydroxyphosphoribosylaminopyrimidine deaminase		TRUE
4.3.2.1	argininosuccinate lyase;		TRUE
1.1.1.42	isocitrate dehydrogenase (NADP+);		TRUE
1.1.1.58	tagaturonate reductase;		TRUE
1.1.1.6	glycerol dehydrogenase;		TRUE
1.17.99.1	4-cresol dehydrogenase (hydroxylating);		TRUE
1.4.1.3	glutamate dehydrogenase [NAD(P)+];		TRUE
2.3.1.31	homoserine O-acetyltransferase;		TRUE
2.4.1.83	dolichyl-phosphate beta-D-mannosyltransferase;		TRUE
2.4.2.17	ATP phosphoribosyltransferase;		TRUE
2.4.2.28	S-methyl-5'-thioadenosine phosphorylase;		TRUE
2.5.1.62	chlorophyll synthase		TRUE
2.7.1.1	hexokinase;		TRUE
2.7.1.2	glucokinase;		TRUE
2.7.1.30	glycerol kinase;		TRUE
3.1.1.32	phospholipase A1		TRUE
3.2.1.31	beta-glucuronidase;		TRUE
3.4.11.23	PepB aminopeptidase;		TRUE
3.5.1.53	N-carbamoylputrescine amidase;		TRUE
3.8.1.8	atrazine chlorohydrolase;		TRUE
4.1.1.28	aromatic-L-amino-acid decarboxylase;		TRUE

Appendix C (Continued)

Table C.1 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
4.1.1.82	phosphonopyruvate decarboxylase;		TRUE
4.1.1.9	malonyl-CoA decarboxylase;		TRUE
4.2.1.33	3-isopropylmalate dehydratase;		TRUE
4.2.1.7	altronate dehydratase;		TRUE
4.2.99.20	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase;		TRUE
4.3.1.2	methylaspartate ammonia-lyase;		TRUE
4.3.1.4	formimidoyltetrahydrofolate cyclodeaminase;		TRUE
6.1.1.24	glutamate---tRNAGln ligase;		TRUE
6.2.1.25	benzoate---CoA ligase;		TRUE
6.4.1.1	pyruvate carboxylase;		TRUE
1.2.1.38	N-acetyl-gamma-glutamyl-phosphate reductase;		TRUE
6.-.-.-	null		TRUE
1.5.1.2	pyrroline-5-carboxylate reductase;		TRUE
2.5.1.47	cysteine synthase;		TRUE
2.6.1.9	histidinol-phosphate transaminase;		TRUE
2.7.7.41	phosphatidate cytidylyltransferase;		TRUE
3.1.3.27	phosphatidylglycerophosphatase;		TRUE
1.1.1.14	L-iditol 2-dehydrogenase;		TRUE
1.1.1.37	malate dehydrogenase;		TRUE
1.1.1.40	malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+);		TRUE
4.1.1.49	phosphoenolpyruvate carboxykinase (ATP);		TRUE
2.7.1.19	phosphoribulokinase;		TRUE
5.1.99.1	methylmalonyl-CoA epimerase;		TRUE
2.4.1.157	1,2-diacylglycerol 3-glucosyltransferase;		TRUE
4.1.2.5	threonine aldolase;		TRUE
2.1.1.131	precorrin-3B C17-methyltransferase;		TRUE
2.5.1.49	O-acetylhomoserine aminocarboxypropyltransferase;	TRUE	TRUE
1.1.1.3	homoserine dehydrogenase;		TRUE
2.3.3.13	2-isopropylmalate synthase;		TRUE
2.1.1.107	uroporphyrinogen-III C-methyltransferase;		TRUE
4.2.1.75	uroporphyrinogen-III synthase;		TRUE
2.1.2.2	phosphoribosylglycinamide formyltransferase;		TRUE
2.4.2.14	amidophosphoribosyltransferase;		TRUE
2.6.1.16	glutamine---fructose-6-phosphate transaminase (isomerizing);		TRUE
4.3.2.2	adenylosuccinate lyase;		TRUE
5.4.99.18	5-(carboxyamino)imidazole ribonucleotide mutase;		TRUE
6.3.2.6	phosphoribosylaminoimidazolesuccinocarboxamide synthase;		TRUE
1.1.-.-	<Oxidoreductases; Acting on the CH-OH group of donors>		TRUE

Appendix C (Continued)

Table C.1 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.4.2.2	pyrimidine-nucleoside phosphorylase;		TRUE
2.3.1.39	[acyl-carrier-protein] S-malonyltransferase;		TRUE
2.6.1.-	<Transferases; Transferring nitrogenous groups; Transaminases>		TRUE
1.5.1.20	methylenetetrahydrofolate reductase [NAD(P)H];		TRUE
2.3.1.41	beta-ketoacyl-acyl-carrier-protein synthase I;		TRUE
2.2.1.1	transketolase;		TRUE
2.7.7.3	pantetheine-phosphate adenylyltransferase;		TRUE
5.3.1.4	L-arabinose isomerase;		TRUE
4.2.1.46	dTDP-glucose 4,6-dehydratase;		TRUE
2.7.1.31	glycerate kinase;		TRUE
6.3.3.3	dethiobiotin synthase;		TRUE
2.7.7.27	glucose-1-phosphate adenylyltransferase;		TRUE
2.7.7.9	UTP---glucose-1-phosphate uridylyltransferase;		TRUE
2.4.1.8	maltose phosphorylase		TRUE
2.7.7.1	nicotinamide-nucleotide adenylyltransferase;		TRUE
3.2.1.45	glucosylceramidase;		TRUE
4.1.1.31	phosphoenolpyruvate carboxylase;		TRUE
2.3.1.179	beta-ketoacyl-acyl-carrier-protein synthase II;		TRUE
6.4.1.2	acetyl-CoA carboxylase;		TRUE
2.7.6.1	ribose-phosphate diphosphokinase;		TRUE
1.1.1.26	glyoxylate reductase;		TRUE
1.1.1.57	fructuronate reductase;		TRUE
1.1.1.9	D-xylulose reductase;		TRUE
1.2.1.-	Oxidoreductases; Acting on the aldehyde or oxo group of donors		TRUE
1.4.1.9	leucine dehydrogenase;		TRUE
1.5.1.7	saccharopine dehydrogenase (NAD+, L-lysine-forming);		TRUE
2.1.1.14	5-methyltetrahydropteroyltriglutamate---homocysteine		TRUE
2.4.1.14	sucrose-phosphate synthase;		TRUE
2.5.1.1	dimethylallyltranstransferase;		TRUE
2.5.1.16	spermidine synthase;		TRUE
2.5.1.29	farnesyltranstransferase;		TRUE
2.6.1.57	aromatic-amino-acid transaminase;		TRUE
2.7.1.4	fructokinase;		TRUE
2.7.1.60	N-acylmannosamine kinase;		TRUE
2.7.1.76	deoxyadenosine kinase;		TRUE
3.1.1.23	acylglycerol lipase;		TRUE
3.11.1.1	phosphonoacetaldehyde hydrolase;		TRUE

Appendix C (Continued)

Table C.2 Summary of enzymes in the dataset identified (True) by the NIBBS algorithm and Student’s T-test for the phenotype aerobic respiration

EC Number	Enzyme Name	In T-Test	NIBBS
3.5.2.3	dihydroorotase;		TRUE
3.5.4.25	GTP cyclohydrolase II;		TRUE
6.3.4.18	5-(carboxyamino)imidazole ribonucleotide synthase;		TRUE
2.2.1.6	acetolactate synthase;		TRUE
2.7.1.40	pyruvate kinase;		TRUE
2.3.1.9	acetyl-CoA C-acetyltransferase;	TRUE	TRUE
4.99.1.1	ferrochelatase;	TRUE	TRUE
3.5.4.13	dCTP deaminase;	TRUE	TRUE
1.1.1.157	3-hydroxybutyryl-CoA dehydrogenase;		TRUE
6.3.4.14	biotin carboxylase;		TRUE
2.7.4.14	cytidylate kinase;	TRUE	TRUE
6.4.1.2	acetyl-CoA carboxylase;		TRUE
4.1.1.37	uroporphyrinogen decarboxylase;	TRUE	TRUE
2.3.1.61	dihydrolipoyllysine-residue succinyltransferase;	TRUE	TRUE
2.1.1.45	thymidylate synthase;		TRUE
2.3.1.12	dihydrolipoyllysine-residue acetyltransferase;	TRUE	TRUE
6.4.1.1	pyruvate carboxylase;		TRUE
1.1.1.49	glucose-6-phosphate dehydrogenase;	TRUE	TRUE
2.7.1.39	homoserine kinase;		TRUE
4.3.1.19	threonine ammonia-lyase;	TRUE	TRUE
1.2.4.1	pyruvate dehydrogenase (acetyl-transferring);	TRUE	TRUE
2.3.3.1	citrate (Si)-synthase;	TRUE	TRUE
4.2.1.3	aconitate hydratase;		TRUE
2.7.1.2	glucokinase;		TRUE
2.7.1.17	xylulokinase;		TRUE
3.5.3.8	formimidoylglutamase;	TRUE	TRUE
6.3.4.2	CTP synthase;		TRUE
1.2.4.2	oxoglutarate dehydrogenase (succinyl-transferring);	TRUE	TRUE
1.8.1.4	dihydrolipoyl dehydrogenase;	TRUE	TRUE
6.2.1.5	succinate---CoA ligase (ADP-forming);		TRUE
1.2.1.27	methylmalonate-semialdehyde dehydrogenase (acylating);	TRUE	TRUE
6.4.1.3	propionyl-CoA carboxylase;	TRUE	TRUE
1.13.11.27	4-hydroxyphenylpyruvate dioxygenase;	TRUE	TRUE
2.3.1.39	[acyl-carrier-protein] S-malonyltransferase;		TRUE
2.7.1.4	fructokinase;		TRUE
2.3.1.35	glutamate N-acetyltransferase;		TRUE
2.3.3.9	malate synthase;	TRUE	TRUE
2.4.2.7	adenine phosphoribosyltransferase;		TRUE
2.4.2.8	hypoxanthine phosphoribosyltransferase;		TRUE
3.1.3.3	phosphoserine phosphatase		TRUE
5.1.1.3	glutamate racemase		TRUE
5.3.1.8	mannose-6-phosphate isomerase;		TRUE
1.14.13.-	<Oxidoreductases; Acting on paired donors with incorporation of molecular oxygen>		TRUE
1.1.1.2	alcohol dehydrogenase (NADP+);	TRUE	TRUE
2.7.1.21	thymidine kinase;		TRUE
3.5.4.19	phosphoribosyl-AMP cyclohydrolase;		TRUE
3.6.1.31	phosphoribosyl-ATP diphosphatase;		TRUE
1.4.1.13	glutamate synthase (NADPH);		TRUE

Appendix C (Continued)

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
1.4.1.2	glutamate dehydrogenase;		TRUE
1.4.4.2	glycine dehydrogenase (decarboxylating);		TRUE
3.2.1.20	alpha-glucosidase;		TRUE
4.1.1.11	aspartate 1-decarboxylase;	TRUE	TRUE
4.1.3.1	isocitrate lyase;	TRUE	TRUE
1.13.11.11	tryptophan 2,3-dioxygenase;	TRUE	TRUE
1.13.11.5	homogentisate 1,2-dioxygenase;	TRUE	TRUE
1.14.-.-	<Oxidoreductases; Acting on paired donors with incorporation of molecular oxygen>		TRUE
1.1.1.169	2-dehydropantoate 2-reductase;		TRUE
1.1.1.27	L-lactate dehydrogenase;		TRUE
1.1.1.44	phosphogluconate dehydrogenase (decarboxylating);	TRUE	TRUE
1.1.1.85	3-isopropylmalate dehydrogenase;		TRUE
2.3.1.1	amino-acid N-acetyltransferase;		TRUE
2.5.1.48	cystathionine gamma-synthase;		TRUE
1.11.1.6	catalase;	TRUE	TRUE
1.2.1.16	succinate-semialdehyde dehydrogenase [NAD(P)+];	TRUE	TRUE
2.4.2.18	anthranilate phosphoribosyltransferase;		TRUE
2.7.7.23	UDP-N-acetylglucosamine diphosphorylase;		TRUE
4.1.2.14	2-dehydro-3-deoxy-phosphogluconate aldolase;		TRUE
1.3.99.7	glutaryl-CoA dehydrogenase;	TRUE	TRUE
3.7.1.3	kynureninase	TRUE	TRUE
5.4.2.10	phosphoglucosamine mutase		TRUE
2.3.2.2	gamma-glutamyltransferase;	TRUE	TRUE
1.5.1.12	1-pyrroline-5-carboxylate dehydrogenase;	TRUE	TRUE
1.3.1.9	enoyl-[acyl-carrier-protein] reductase (NADH);	TRUE	TRUE
2.1.3.2	aspartate carbamoyltransferase;		TRUE
2.4.2.9	uracil phosphoribosyltransferase;		TRUE
2.6.1.13	ornithine aminotransferase;	TRUE	TRUE
2.6.1.19	4-aminobutyrate transaminase;	TRUE	TRUE
2.6.1.76	diaminobutyrate---2-oxoglutarate transaminase;	TRUE	TRUE
3.5.3.6	arginine deiminase;	TRUE	TRUE
1.2.1.3	aldehyde dehydrogenase (NAD+);	TRUE	TRUE
4.2.1.17	enoyl-CoA hydratase;	TRUE	TRUE
3.1.1.45	carboxymethylenebutenolidase;	TRUE	TRUE
1.1.1.35	3-hydroxyacyl-CoA dehydrogenase;	TRUE	TRUE
1.4.3.5	pyridoxal 5'-phosphate synthase;	TRUE	TRUE
2.3.1.16	acetyl-CoA C-acyltransferase;	TRUE	TRUE
1.3.99.3	acyl-CoA dehydrogenase;	TRUE	TRUE
1.1.1.30	3-hydroxybutyrate dehydrogenase;	TRUE	TRUE
1.14.16.1	phenylalanine 4-monooxygenase;	TRUE	TRUE
3.5.1.2	glutaminase;		TRUE
1.5.99.8	proline dehydrogenase;	TRUE	TRUE
1.4.1.9	leucine dehydrogenase;		TRUE
1.4.3.16	L-aspartate oxidase;		TRUE
2.7.1.63	polyphosphate---glucose phosphotransferase;	TRUE	TRUE
2.8.1.8	lipoyl synthase;	TRUE	TRUE
3.6.1.19	nucleoside-triphosphate diphosphatase;		TRUE
4.1.1.32	phosphoenolpyruvate carboxykinase (GTP);		TRUE

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
4.3.1.17	L-serine ammonia-lyase;		TRUE
6.3.2.2	glutamate---cysteine ligase;		TRUE
1.3.99.1	succinate dehydrogenase;	TRUE	TRUE
4.2.1.2	fumarate hydratase;		TRUE
1.1.1.31	3-hydroxyisobutyrate dehydrogenase;		TRUE
4.1.3.4	hydroxymethylglutaryl-CoA lyase;	TRUE	TRUE
1.1.1.42	isocitrate dehydrogenase (NADP+);		TRUE
1.3.1.76	precorrin-2 dehydrogenase;		TRUE
1.2.4.4	3-methyl-2-oxobutanoate dehydrogenase	TRUE	TRUE
2.7.1.12	gluconokinase;	TRUE	TRUE
2.1.2.11	3-methyl-2-oxobutanoate hydroxymethyltransferase;	TRUE	TRUE
2.1.3.3	ornithine carbamoyltransferase;		TRUE
2.3.1.157	glucosamine-1-phosphate N-acetyltransferase		TRUE
2.3.1.31	homoserine O-acetyltransferase;		TRUE
3.2.2.1	purine nucleosidase;		TRUE
3.7.1.2	fumarylacetoacetase;	TRUE	TRUE
5.2.1.2	maleylacetoacetate isomerase;	TRUE	TRUE
6.3.2.3	glutathione synthase;	TRUE	TRUE
3.5.1.5	urease	TRUE	TRUE
3.4.11.1	leucyl aminopeptidase;	TRUE	TRUE
1.3.3.3	coproporphyrinogen oxidase;	TRUE	TRUE
6.3.2.12	dihydrofolate synthase;		TRUE
2.5.1.54	3-deoxy-7-phosphoheptulonate synthase;		TRUE
5.3.1.6	ribose-5-phosphate isomerase;		TRUE
3.1.1.31	6-phosphogluconolactonase;		TRUE
3.1.3.25	inositol-phosphate phosphatase;	TRUE	TRUE
4.2.1.12	phosphogluconate dehydratase;	TRUE	TRUE
4.2.1.44	myo-inosose-2 dehydratase;		TRUE
2.1.2.10	aminomethyltransferase;	TRUE	TRUE
1.1.2.3	L-lactate dehydrogenase (cytochrome);	TRUE	TRUE
1.3.3.6	acyl-CoA oxidase;	TRUE	TRUE
2.3.1.168	dihydrolipoyllysine-residue (2-methylpropanoyl)transferase;	TRUE	TRUE
2.5.1.19	3-phosphoshikimate 1-carboxyvinyltransferase;		TRUE
2.5.1.26	alkylglycerone-phosphate synthase;	TRUE	TRUE
2.5.1.47	cysteine synthase;		TRUE
3.1.3.77	acireductone synthase;	TRUE	TRUE
3.5.4.16	GTP cyclohydrolase I;		TRUE
4.1.1.19	arginine decarboxylase;		TRUE
4.1.1.48	indole-3-glycerol-phosphate synthase;		TRUE
4.4.1.1	cystathionine gamma-lyase;	TRUE	TRUE
5.5.1.4	inositol-3-phosphate synthase;	TRUE	TRUE
2.3.1.117	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase;	TRUE	TRUE
5.3.3.2	isopentenyl-diphosphate Delta-isomerase;	TRUE	TRUE
1.3.99.-	<Oxidoreductases; Acting on the CH-CH group of donors; With other acceptors>		TRUE
3.5.4.4	adenosine deaminase;	TRUE	TRUE
4.4.1.16	selenocysteine lyase;	TRUE	TRUE
1.6.1.2	NAD(P)+ transhydrogenase (AB-specific);	TRUE	TRUE
1.1.1.37	malate dehydrogenase;	TRUE	TRUE

Appendix C (Continued)

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.7.4.9	dTMP kinase;		TRUE
1.5.3.1	sarcosine oxidase		TRUE
2.3.1.74	naringenin-chalcone synthase;	TRUE	TRUE
5.4.99.5	chorismate mutase;		TRUE
6.2.1.3	long-chain-fatty-acid---CoA ligase;		TRUE
3.1.1.1	carboxylesterase;		TRUE
2.7.4.6	nucleoside-diphosphate kinase;	TRUE	TRUE
2.5.1.32	phytoene synthase;		TRUE
1.1.1.95	phosphoglycerate dehydrogenase;		TRUE
1.14.13.83	precorrin-3B synthase;	TRUE	TRUE
1.14.19.3	linoleoyl-CoA desaturase;	TRUE	TRUE
1.2.1.39	phenylacetaldehyde dehydrogenase	TRUE	TRUE
1.8.1.2	sulfite reductase (NADPH);	TRUE	TRUE
2.2.1.7	1-deoxy-D-xylulose-5-phosphate synthase;		TRUE
2.3.1.178	diaminobutyrate acetyltransferase;	TRUE	TRUE
2.5.1.10	geranyltranstransferase;		TRUE
2.6.1.21	D-amino-acid transaminase;		TRUE
3.5.3.1	arginase;		TRUE
3.5.3.11	agmatinase;		TRUE
4.99.1.4	sirohydrochlorin ferrochelatase;		TRUE
5.4.4.2	isochorismate synthase;	TRUE	TRUE
2.4.2.10	orotate phosphoribosyltransferase;		TRUE
2.4.2.17	ATP phosphoribosyltransferase;		TRUE
4.2.3.4	3-dehydroquinate synthase;		TRUE
3.5.1.18	succinyl-diaminopimelate desuccinylase;	TRUE	TRUE
4.1.1.44	4-carboxymuconolactone decarboxylase;		TRUE
1.2.1.-	<Oxidoreductases; Acting on the aldehyde or oxo group of donors;>		TRUE
3.1.2.-	<Hydrolases; Acting on ester bonds; Thiolester hydrolases>		TRUE
2.3.1.181	lipoyl(octanoyl) transferase;	TRUE	TRUE
1.14.15.3	alkane 1-monooxygenase;	TRUE	TRUE
2.1.1.131	precorrin-3B C17-methyltransferase;		TRUE
2.1.1.152	precorrin-6A synthase (deacetylating);	TRUE	TRUE
4.2.1.41	5-dehydro-4-deoxyglucarate dehydratase;		TRUE
3.7.1.-	<Hydrolases; Acting on carbon-carbon bonds; In ketonic substances>		TRUE
4.2.1.108	ectoine synthase;	TRUE	TRUE
6.3.5.7	glutaminyl-tRNA synthase (glutamine-hydrolysing);	TRUE	TRUE
3.1.1.24	3-oxoadipate enol-lactonase;	TRUE	TRUE
1.14.13.39	nitric-oxide synthase;	TRUE	TRUE
1.14.14.1	unspecific monooxygenase;	TRUE	TRUE
2.7.1.29	glycerone kinase;		TRUE
3.1.2.23	4-hydroxybenzoyl-CoA thioesterase	TRUE	TRUE
3.5.99.6	glucosamine-6-phosphate deaminase;		TRUE
3.8.1.2	(S)-2-haloacid dehalogenase;	TRUE	TRUE
4.1.1.15	glutamate decarboxylase;		TRUE
4.1.2.-	<Lyases; Carbon-carbon lyases; Aldehyde-lyases>		TRUE
4.1.3.39	4-hydroxy-2-oxovalerate aldolase;		TRUE
5.4.2.2	phosphoglucomutase;		TRUE
1.1.1.22	UDP-glucose 6-dehydrogenase;		TRUE

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.7.7.9	UTP---glucose-1-phosphate uridylyltransferase;		TRUE
5.1.3.2	UDP-glucose 4-epimerase;		TRUE
2.4.2.19	nicotinate-nucleotide diphosphorylase (carboxylating);		TRUE
3.5.1.19	nicotinamidase;	TRUE	TRUE
1.1.1.23	histidinol dehydrogenase;		TRUE
4.3.1.3	histidine ammonia-lyase;		TRUE
2.6.1.52	phosphoserine transaminase;	TRUE	TRUE
2.1.1.64	3-demethylubiquinone-9 3-O-methyltransferase;	TRUE	TRUE
1.6.1.1	NAD(P)+ transhydrogenase (B-specific);	TRUE	TRUE
2.3.1.37	5-aminolevulinate synthase;		TRUE
4.1.2.5	threonine aldolase;		TRUE
4.2.1.80	2-oxopent-4-enoate hydratase;		TRUE
5.4.99.18	5-(carboxyamino)imidazole ribonucleotide mutase;		TRUE
5.5.1.2	3-carboxy-cis,cis-muconate cycloisomerase;	TRUE	TRUE
1.13.11.3	protocatechuate 3,4-dioxygenase;		TRUE
2.7.1.36	mevalonate kinase;	TRUE	TRUE
3.3.1.1	adenosylhomocysteinase;		TRUE
3.5.1.1	asparaginase;		TRUE
4.4.1.8	cystathionine beta-lyase;		TRUE
1.4.3.4	monoamine oxidase;	TRUE	TRUE
1.3.3.4	protoporphyrinogen oxidase;	TRUE	TRUE
6.6.1.1	magnesium chelatase;		TRUE
6.6.1.2	cobaltochelatase;		TRUE
3.4.11.2	membrane alanyl aminopeptidase;	TRUE	TRUE
1.13.12.-	<Oxidoreductases; Acting on single donors with incorporation of molecular oxygen (oxygenases)>		TRUE
1.14.13.9	kynurenine 3-monooxygenase;	TRUE	TRUE
1.14.18.1	monophenol monooxygenase;	TRUE	TRUE
1.3.-.-	<Oxidoreductases; Acting on the CH-CH group of donors>		TRUE
1.3.1.12	prephenate dehydrogenase;		TRUE
1.7.3.3	factor-independent urate hydroxylase;	TRUE	TRUE
1.8.7.1	sulfite reductase (ferredoxin);	TRUE	TRUE
2.1.1.10	homocysteine S-methyltransferase;		TRUE
2.2.1.2	transaldolase;		TRUE
2.3.3.10	hydroxymethylglutaryl-CoA synthase;		TRUE
2.4.2.1	purine-nucleoside phosphorylase;		TRUE
2.4.2.2	pyrimidine-nucleoside phosphorylase;		TRUE
2.4.2.4	thymidine phosphorylase;		TRUE
2.5.1.1	dimethylallyltranstransferase;		TRUE
2.6.1.18	beta-alanine---pyruvate transaminase;		TRUE
2.7.1.20	adenosine kinase;		TRUE
3.1.3.1	alkaline phosphatase;		TRUE
3.5.1.16	acetylornithine deacetylase;		TRUE
3.5.3.12	agmatine deiminase;		TRUE
3.5.4.3	guanine deaminase;		TRUE
4.1.1.47	tartronate-semialdehyde synthase;		TRUE
5.3.1.5	xylose isomerase;		TRUE
6.1.1.24	glutamate---tRNAGln ligase;		TRUE
6.2.1.26	o-succinylbenzoate---CoA ligase;		TRUE

Appendix C (Continued)

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.2.1.9	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid		TRUE
4.1.3.27	anthranilate synthase;		TRUE
1.3.99.10	isovaleryl-CoA dehydrogenase;		TRUE
6.3.2.1	pantoate---beta-alanine ligase;	TRUE	TRUE
4.1.1.31	phosphoenolpyruvate carboxylase;	TRUE	TRUE
2.3.1.47	8-amino-7-oxononanoate synthase;		TRUE
1.14.13.2	4-hydroxybenzoate 3-monooxygenase;		TRUE
4.2.1.113	o-succinylbenzoate synthase;	TRUE	TRUE
3.1.1.17	gluconolactonase;	TRUE	TRUE
3.5.3.4	allantoicase		TRUE
3.8.1.5	haloalkane dehalogenase;		TRUE
3.5.2.17	hydroxyisourate hydrolase;		TRUE
4.1.1.33	diphosphomevalonate decarboxylase;	TRUE	TRUE
3.1.3.11	fructose-bisphosphatase;		TRUE
2.7.4.7	phosphomethylpyrimidine kinase;		TRUE
1.5.1.3	dihydrofolate reductase;		TRUE
1.5.1.5	methylenetetrahydrofolate dehydrogenase (NADP+);		TRUE
2.1.2.1	glycine hydroxymethyltransferase;		TRUE
3.1.3.-	<Hydrolases; Acting on ester bonds; Phosphoric monoester hydrolases>		TRUE
2.7.1.30	glycerol kinase;		TRUE
6.2.1.1	acetate---CoA ligase;	TRUE	TRUE
4.1.2.25	dihydroneopterin aldolase;		TRUE
6.3.2.17	tetrahydrofolate synthase;		TRUE
4.1.3.36	1,4-dihydroxy-2-naphthoyl-CoA synthase;		TRUE
2.6.1.62	adenosylmethionine---8-amino-7-oxononanoate transaminase;		TRUE
1.1.1.122	D-threo-aldose 1-dehydrogenase;		TRUE
1.1.1.215	gluconate 2-dehydrogenase;		TRUE
1.1.1.41	isocitrate dehydrogenase (NAD+);		TRUE
1.13.11.6	3-hydroxyanthranilate 3,4-dioxygenase;		TRUE
1.14.13.70	sterol 14-demethylase;		TRUE
1.17.1.4	xanthine dehydrogenase;		TRUE
1.2.3.3	pyruvate oxidase;		TRUE
1.4.1.3	glutamate dehydrogenase [NAD(P)+];		TRUE
1.4.3.21	primary-amine oxidase;		TRUE
2.1.1.107	uroporphyrinogen-III C-methyltransferase;		TRUE
2.1.1.37	DNA (cytosine-5-)-methyltransferase;		TRUE
2.4.2.22	xanthine phosphoribosyltransferase;		TRUE
2.6.1.1	aspartate transaminase;		TRUE
2.7.1.69	protein-Npi-phosphohistidine---sugar phosphotransferase;		TRUE
2.7.7.10	UTP---hexose-1-phosphate uridylyltransferase;		TRUE
3.11.1.1	phosphonoacetaldehyde hydrolase;		TRUE
3.2.1.10	oligo-1,6-glucosidase;		TRUE
3.2.1.3	glucan 1,4-alpha-glucosidase;		TRUE
3.5.1.6	beta-ureidopropionase		TRUE
3.5.3.19	ureidoglycolate hydrolase		TRUE
4.1.1.17	ornithine decarboxylase;		TRUE
5.1.1.17	isopenicillin-N epimerase		TRUE
6.3.5.4	asparagine synthase (glutamine-hydrolysing);		TRUE

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
3.1.1.3	triacylglycerol lipase;		TRUE
1.1.1.53	3alpha(or 20beta)-hydroxysteroid dehydrogenase;		TRUE
1.1.99.3	gluconate 2-dehydrogenase (acceptor);		TRUE
1.14.13.1	salicylate 1-monooxygenase;	TRUE	TRUE
1.14.13.7	phenol 2-monooxygenase;		TRUE
3.5.99.3	hydroxydechloroatrazine ethylaminohydrolase;		TRUE
4.1.1.4	acetoacetate decarboxylase;		TRUE
4.1.1.45	aminocarboxymuconate-semialdehyde decarboxylase;		TRUE
5.3.3.1	steroid Delta-isomerase;		TRUE
5.3.3.4	muconolactone Delta-isomerase;		TRUE
2.2.1.1	transketolase;		TRUE
2.4.2.14	amidophosphoribosyltransferase;		TRUE
4.1.2.13	fructose-bisphosphate aldolase;		TRUE
6.3.4.13	phosphoribosylamine---glycine ligase;		TRUE
2.7.4.16	thiamine-phosphate kinase;		TRUE
3.5.2.7	imidazolonepropionase;		TRUE
1.11.1.15	peroxiredoxin;		TRUE
1.17.4.1	ribonucleoside-diphosphate reductase;	TRUE	TRUE
3.5.2.14	N-methylhydantoinase (ATP-hydrolysing);		TRUE
6.4.1.4	methylcrotonoyl-CoA carboxylase;	TRUE	TRUE
2.4.2.11	nicotinate phosphoribosyltransferase;		TRUE
2.7.7.18	nicotinate-nucleotide adenylyltransferase;		TRUE
6.3.1.5	NAD+ synthase;		TRUE
3.2.1.-	<Hydrolases; Glycosidases; Glycosidases>		TRUE
1.1.1.1	alcohol dehydrogenase;		TRUE
2.7.1.92	5-dehydro-2-deoxygluconokinase;		TRUE
4.1.1.74	indolepyruvate decarboxylase;		TRUE
5.1.1.7	diaminopimelate epimerase		TRUE
1.1.1.79	glyoxylate reductase (NADP+);		TRUE
2.1.1.17	phosphatidylethanolamine N-methyltransferase;		TRUE
1.11.1.7	peroxidase;		TRUE
2.7.1.-	<Transferases; Transferring phosphorus-containing groups>		TRUE
4.2.1.49	urocanate hydratase;		TRUE
1.1.1.34	hydroxymethylglutaryl-CoA reductase (NADPH);		TRUE
1.1.1.50	3alpha-hydroxysteroid dehydrogenase (B-specific);		TRUE
1.1.1.6	glycerol dehydrogenase;		TRUE
1.1.1.60	2-hydroxy-3-oxopropionate reductase;		TRUE
1.1.1.81	hydroxypyruvate reductase;		TRUE
1.1.3.8	L-gulonolactone oxidase;		TRUE
1.1.5.2	quinoprotein glucose dehydrogenase;		TRUE
1.2.7.1	pyruvate synthase;		TRUE
1.3.1.-	<Oxidoreductases; Acting on the CH-CH group of donors>		TRUE
1.3.1.2	dihydropyrimidine dehydrogenase (NADP+);		TRUE
1.3.99.12	2-methylacyl-CoA dehydrogenase;		TRUE
1.4.1.20	phenylalanine dehydrogenase;		TRUE
1.5.1.7	saccharopine dehydrogenase (NAD+, L-lysine-forming);		TRUE
2.1.1.6	catechol O-methyltransferase;		TRUE
2.3.1.57	diamine N-acetyltransferase;		TRUE

Appendix C (Continued)

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.4.1.10	levansucrase;		TRUE
2.4.1.83	dolichyl-phosphate beta-D-mannosyltransferase;		TRUE
2.4.2.28	S-methyl-5'-thioadenosine phosphorylase;		TRUE
2.5.1.21	squalene synthase;		TRUE
2.5.1.29	farnesyltranstransferase;		TRUE
2.5.1.49	O-acetylhomoserine aminocarboxypropyltransferase;		TRUE
2.6.1.36	L-lysine 6-transaminase;		TRUE
2.6.1.66	valine---pyruvate transaminase;		TRUE
2.7.1.100	S-methyl-5-thioribose kinase;		TRUE
2.7.1.48	uridine kinase;		TRUE
2.7.4.2	phosphomevalonate kinase;		TRUE
2.7.7.63	lipoate---protein ligase;		TRUE
2.8.1.2	3-mercaptopyruvate sulfurtransferase;		TRUE
3.1.4.12	sphingomyelin phosphodiesterase;		TRUE
3.2.1.1	alpha-amylase;		TRUE
3.2.1.15	polygalacturonase;		TRUE
3.2.1.23	beta-galactosidase;		TRUE
3.2.1.26	beta-fructofuranosidase;		TRUE
3.2.1.37	xylan 1,4-beta-xylosidase;		TRUE
3.5.2.2	dihydropyrimidinase;		TRUE
3.5.4.2	adenine deaminase;		TRUE
4.1.1.18	lysine decarboxylase;		TRUE
4.1.1.39	ribulose-bisphosphate carboxylase;		TRUE
4.1.2.27	sphinganine-1-phosphate aldolase;		TRUE
4.1.3.40	chorismate lyase;		TRUE
4.2.1.-	<Lyases; Carbon-oxygen lyases; Hydro-lyases>		TRUE
4.2.1.47	GDP-mannose 4,6-dehydratase;		TRUE
4.2.1.7	altronate dehydratase;		TRUE
4.2.1.75	uroporphyrinogen-III synthase;		TRUE
4.2.3.12	6-pyruvoyltetrahydropterin synthase;		TRUE
4.2.99.20	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase;		TRUE
4.3.1.1	aspartate ammonia-lyase;		TRUE
5.1.1.4	proline racemase		TRUE
5.4.99.2	methylmalonyl-CoA mutase;		TRUE
6.-.-.-	null		TRUE
6.2.1.17	propionate---CoA ligase;		TRUE
1.1.99.8	alcohol dehydrogenase (acceptor);		TRUE
1.13.11.39	biphenyl-2,3-diol 1,2-dioxygenase;		TRUE
2.8.3.6	3-oxoadipate CoA-transferase;		TRUE
1.2.1.70	glutamyl-tRNA reductase		TRUE
2.3.1.-	<Transferases; Acyltransferases; Transferring groups other than amino-acyl groups>		TRUE
3.1.4.3	phospholipase C;		TRUE
2.6.1.11	acetylornithine transaminase;		TRUE
6.2.1.-	<Ligases; Forming carbon-sulfur bonds; Acid--thiol ligases>		TRUE
2.6.1.17	succinyldiaminopimelate transaminase;		TRUE
2.7.7.4	sulfate adenylyltransferase;		TRUE
1.14.99.-	<Oxidoreductases; Acting on paired donors with incorporation of molecular oxygen>		TRUE

Appendix C (Continued)

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
3.5.1.-	<Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bondss>		TRUE
2.7.8.5	CDP-diacylglycerol---glycerol-3-phosphate 3-phosphatidyltransferase;	TRUE	TRUE
2.7.8.8	CDP-diacylglycerol---serine O-phosphatidyltransferase;		TRUE
2.7.1.24	dephospho-CoA kinase;	TRUE	TRUE
6.3.2.10	UDP-N-acetylmuramoyl-tripeptide---D-alanyl-D-alanine ligase;		TRUE
2.6.1.57	aromatic-amino-acid transaminase;		TRUE
5.3.2.-	<Isomerases; Intramolecular oxidoreductases; Interconverting keto- and enol-groups>		TRUE
2.4.2.-	<Transferases; Glycosyltransferases; Pentosyltransferases>		TRUE
2.4.1.-	<Transferases; Glycosyltransferases; Hexosyltransferases>		TRUE
5.3.1.-	<Isomerases; Intramolecular oxidoreductases; Interconverting aldoses and ketoses>		TRUE
2.3.1.5	arylamine N-acetyltransferase;		TRUE
2.5.1.16	spermidine synthase;		TRUE
4.1.1.77	4-oxalocrotonate decarboxylase;		TRUE
6.3.1.8	glutathionylspermidine synthase;		TRUE
1.13.11.4	gentisate 1,2-dioxygenase;		TRUE
1.2.1.28	benzaldehyde dehydrogenase (NAD+);		TRUE
2.7.1.16	ribulokinase;		TRUE
2.7.8.20	phosphatidylglycerol---membrane-oligosaccharide		TRUE
4.1.1.8	oxalyl-CoA decarboxylase;		TRUE
5.1.3.12	UDP-glucuronate 5'-epimerase;		TRUE
1.-.-.-	<Oxidoreductases>		TRUE
1.3.1.54	precorrin-6A reductase;		TRUE
2.1.1.132	precorrin-6Y C5,15-methyltransferase (decarboxylating);		TRUE
6.3.1.-	<Ligases; Forming carbon-nitrogen bonds; Acid--ammonia (or amine) ligases>		TRUE
5.1.99.4	alpha-methylacyl-CoA racemase		TRUE
3.5.4.1	cytosine deaminase;		TRUE
1.1.1.220	6-pyruvoyltetrahydropterin 2'-reductase;		TRUE
1.1.99.25	quinate dehydrogenase (pyrroloquinoline-quinone);		TRUE
1.10.3.3	L-ascorbate oxidase;		TRUE
1.14.12.-	<Oxidoreductases; Acting on paired donors with incorporation of molecular oxygen>		TRUE
1.17.1.2	4-hydroxy-3-methylbut-2-enyl diphosphate reductase;		TRUE
1.17.4.2	ribonucleoside-triphosphate reductase;		TRUE
1.2.1.46	formaldehyde dehydrogenase;		TRUE
1.2.1.59	glyceraldehyde-3-phosphate dehydrogenase (NAD(P)+)		TRUE
1.2.1.72	erythrose-4-phosphate dehydrogenase;		TRUE
1.21.3.1	isopenicillin-N synthase;		TRUE
1.3.3.-	null		TRUE
1.3.99.2	butyryl-CoA dehydrogenase;		TRUE
1.4.3.3	D-amino-acid oxidase;		TRUE
1.97.1.-	<Oxidoreductases; Other oxidoreductases>		TRUE
2.1.1.71	phosphatidyl-N-methylethanolamine N-methyltransferase;		TRUE
2.3.1.179	beta-ketoacyl-acyl-carrier-protein synthase II;		TRUE
2.3.1.41	beta-ketoacyl-acyl-carrier-protein synthase I;		TRUE
2.3.1.46	homoserine O-succinyltransferase;		TRUE
2.4.1.144	beta-1,4-mannosyl-glycoprotein		TRUE
2.4.99.-	<Transferases; Glycosyltransferases; Transferring other glycosyl groups>		TRUE
2.6.1.9	histidinol-phosphate transaminase;		TRUE
2.7.1.11	6-phosphofructokinase;		TRUE
2.7.1.113	deoxyguanosine kinase;		TRUE

Appendix C (Continued)

Table C.2 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.7.7.12	UDP-glucose---hexose-1-phosphate uridylyltransferase;		TRUE
2.7.7.33	glucose-1-phosphate cytidylyltransferase;		TRUE
2.8.3.8	acetate CoA-transferase;		TRUE
2.9.1.1	L-seryl-tRNASec selenium transferase;		TRUE
3.1.1.11	pectinesterase;		TRUE
3.1.3.5	5'-nucleotidase;		TRUE
3.1.3.74	pyridoxal phosphatase;		TRUE
3.5.1.14	aminoacylase;		TRUE
3.5.1.53	N-carbamoylputrescine amidase;		TRUE
3.5.1.54	allophanate hydrolase;		TRUE
4.1.1.49	phosphoenolpyruvate carboxykinase (ATP);		TRUE
4.1.1.9	malonyl-CoA decarboxylase;		TRUE
4.2.1.51	prephenate dehydratase;		TRUE
4.3.3.2	strictosidine synthase;		TRUE
5.3.1.23	S-methyl-5-thioribose-1-phosphate isomerase;		TRUE
5.3.1.4	L-arabinose isomerase;		TRUE
5.4.2.9	phosphoenolpyruvate mutase;		TRUE
5.5.1.7	chloromuconate cycloisomerase;		TRUE
6.3.2.7	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate---L-lysine ligase;		TRUE
1.1.1.158	UDP-N-acetylmuramate dehydrogenase;		TRUE
2.5.1.-	<Transferases; Transferring alkyl or aryl groups, other than methyl groups>		TRUE
6.3.1.2	glutamate---ammonia ligase;		TRUE
1.1.1.-	<Oxidoreductases; Acting on the CH-OH group of donors>		TRUE
4.2.1.20	tryptophan synthase;		TRUE
1.14.12.10	benzoate 1,2-dioxygenase;		TRUE
2.6.99.2	pyridoxine 5'-phosphate synthase;		TRUE
3.2.-.-	<Hydrolases; Glycosidases>		TRUE
2.5.1.9	riboflavin synthase;		TRUE
2.7.7.43	N-acylneuraminate cytidylyltransferase;		TRUE
3.1.1.32	phospholipase A1		TRUE
4.1.1.-	<Lyases; Carbon-carbon lyases; Carboxy-lyases>		TRUE
1.1.1.133	dTDP-4-dehydrorhamnose reductase;		TRUE
1.1.1.90	aryl-alcohol dehydrogenase;		TRUE
4.1.1.55	4,5-dihydroxyphthalate decarboxylase;		TRUE
1.1.1.100	3-oxoacyl-[acyl-carrier-protein] reductase;	TRUE	
1.1.1.193	5-amino-6-(5-phosphoribosylamino)uracil reductase;	TRUE	
1.1.1.205	IMP dehydrogenase;	TRUE	
1.1.1.38	malate dehydrogenase (oxaloacetate-decarboxylating);	TRUE	
1.13.11.53	acireductone dioxygenase (Ni2+-requiring);	TRUE	
1.13.11.54	acireductone dioxygenase [iron(II)-requiring];	TRUE	
1.4.3.19	glycine oxidase	TRUE	
2.5.1.15	dihydropteroate synthase;	TRUE	
2.5.1.61	hydroxymethylbilane synthase;	TRUE	
4.2.1.109	methylthioribulose 1-phosphate dehydratase;	TRUE	
4.2.1.24	porphobilinogen synthase;	TRUE	
4.2.1.52	dihydrodipicolinate synthase;	TRUE	
5.4.3.8	glutamate-1-semialdehyde 2,1-aminomutase;	TRUE	
6.3.5.3	phosphoribosylformylglycinamidine synthase;	TRUE	

Table C.3 Summary of enzymes in the dataset identified (True) by the NIBBS algorithm and Student’s T-test for the phenotype hydrogen production.

EC Number	Enzyme Name	In T-Test	NIBBS
1.1.1.1	alcohol dehydrogenase;		TRUE
1.1.1.133	dTDP-4-dehydrorhamnose reductase;		TRUE
1.1.1.14	L-iditol 2-dehydrogenase;		TRUE
1.1.1.169	2-dehydropantoate 2-reductase;		TRUE
1.1.1.18	inositol 2-dehydrogenase;	TRUE	TRUE
1.1.1.193	5-amino-6-(5-phosphoribosylamino)uracil reductase;		TRUE
1.1.1.206	tropinone reductase I;		TRUE
1.1.1.219	dihydrokaempferol 4-reductase;		TRUE
1.1.1.22	UDP-glucose 6-dehydrogenase;		TRUE
1.1.1.23	histidinol dehydrogenase;		TRUE
1.1.1.25	shikimate dehydrogenase;		TRUE
1.1.1.26	glyoxylate reductase;		TRUE
1.1.1.267	1-deoxy-D-xylulose-5-phosphate reductoisomerase;		TRUE
1.1.1.27	L-lactate dehydrogenase;		TRUE
1.1.1.271	GDP-L-fucose synthase;	TRUE	TRUE
1.1.1.35	3-hydroxyacyl-CoA dehydrogenase;		TRUE
1.1.1.37	malate dehydrogenase;		TRUE
1.1.1.41	isocitrate dehydrogenase (NAD+);		TRUE
1.1.1.42	isocitrate dehydrogenase (NADP+);		TRUE
1.1.1.44	phosphogluconate dehydrogenase (decarboxylating);		TRUE
1.1.1.49	glucose-6-phosphate dehydrogenase;		TRUE
1.1.1.57	fructuronate reductase;	TRUE	TRUE
1.1.1.58	tagaturonate reductase;	TRUE	
1.1.1.6	glycerol dehydrogenase;		TRUE
1.1.1.60	2-hydroxy-3-oxopropionate reductase;		TRUE
1.1.1.65	pyridoxine 4-dehydrogenase;		TRUE
1.1.1.79	glyoxylate reductase (NADP+);		TRUE
1.1.1.81	hydroxypyruvate reductase;		TRUE
1.1.1.86	ketol-acid reductoisomerase;		TRUE
1.1.3.15	(S)-2-hydroxy-acid oxidase;		TRUE
1.12.98.1	coenzyme F420 hydrogenase;		TRUE
1.13.11.2	catechol 2,3-dioxygenase;		TRUE
1.13.11.27	4-hydroxyphenylpyruvate dioxygenase;		TRUE
1.13.11.53	acireductone dioxygenase (Ni2+-requiring);	TRUE	TRUE
1.13.11.54	acireductone dioxygenase [iron(II)-requiring];	TRUE	
1.14.-.-	<Oxidoreductases; Acting on paired donors with incorporation of molecular oxygen>		TRUE
1.14.13.-	<Oxidoreductases; Acting on paired donors with incorporation of molecular oxygen>		TRUE
1.14.13.81	magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase;	TRUE	TRUE

Appendix C (Continued)

Table C.3 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
1.14.14.1	unspecific monooxygenase;		TRUE
1.14.99.-	<Oxidoreductases; Acting on paired donors with incorporation of molecular oxygen>		TRUE
1.14.99.30	carotene 7,8-desaturase;		TRUE
1.17.1.2	4-hydroxy-3-methylbut-2-enyl diphosphate reductase;		TRUE
1.17.1.4	xanthine dehydrogenase;		TRUE
1.17.4.2	ribonucleoside-triphosphate reductase;		TRUE
1.17.7.1	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase;		TRUE
1.18.6.1	nitrogenase	TRUE	TRUE
1.2.1.-	<Oxidoreductases; Acting on the aldehyde or oxo group of donors; >		TRUE
1.2.1.16	succinate-semialdehyde dehydrogenase [NAD(P)+];		TRUE
1.2.1.2	formate dehydrogenase;		TRUE
1.2.1.27	methylmalonate-semialdehyde dehydrogenase (acylating);		TRUE
1.2.1.3	aldehyde dehydrogenase (NAD+);		TRUE
1.2.1.38	N-acetyl-gamma-glutamyl-phosphate reductase;		TRUE
1.2.1.41	glutamate-5-semialdehyde dehydrogenase;		TRUE
1.2.1.59	glyceraldehyde-3-phosphate dehydrogenase (NAD(P)+)		TRUE
1.2.1.70	glutamyl-tRNA reductase		TRUE
1.2.1.9	glyceraldehyde-3-phosphate dehydrogenase (NADP+);		TRUE
1.2.4.1	pyruvate dehydrogenase (acetyl-transferring);		TRUE
1.2.4.2	oxoglutarate dehydrogenase (succinyl-transferring);		TRUE
1.2.4.4	3-methyl-2-oxobutanoate dehydrogenase		TRUE
1.2.7.1	pyruvate synthase;		TRUE
1.2.7.3	2-oxoglutarate synthase;		TRUE
1.2.7.5	aldehyde ferredoxin oxidoreductase;		TRUE
1.3.1.-	<Oxidoreductases; Acting on the CH-CH group of donors>		TRUE
1.3.1.33	protochlorophyllide reductase;		TRUE
1.3.1.54	precorrin-6A reductase;	TRUE	TRUE
1.3.1.76	precorrin-2 dehydrogenase;		TRUE
1.3.1.9	enoyl-[acyl-carrier-protein] reductase (NADH);		TRUE
1.3.3.3	coproporphyrinogen oxidase;		TRUE
1.3.3.4	protoporphyrinogen oxidase;		TRUE
1.3.99.2	butyryl-CoA dehydrogenase;		TRUE
1.3.99.7	glutaryl-CoA dehydrogenase;		TRUE
1.4.1.3	glutamate dehydrogenase [NAD(P)+];		TRUE
1.4.1.4	glutamate dehydrogenase (NADP+);		TRUE
1.4.1.9	leucine dehydrogenase;		TRUE
1.4.3.-	<Oxidoreductases; Acting on the CH-NH2 group of donors>		TRUE
1.4.3.16	L-aspartate oxidase;	TRUE	TRUE

Table C.3 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
1.4.3.21	primary-amine oxidase;		TRUE
1.4.3.4	monoamine oxidase;		TRUE
1.4.3.5	pyridoxal 5'-phosphate synthase;		TRUE
1.4.4.2	glycine dehydrogenase (decarboxylating);		TRUE
1.5.1.12	1-pyrroline-5-carboxylate dehydrogenase;		TRUE
1.5.3.1	sarcosine oxidase		TRUE
1.5.99.8	proline dehydrogenase;		TRUE
1.6.1.2	NAD(P)+ transhydrogenase (AB-specific);		TRUE
1.8.1.4	dihydrolipoyl dehydrogenase;		TRUE
2.-.-.-	<Transferases>		TRUE
2.1.1.10	homocysteine S-methyltransferase;		TRUE
2.1.1.107	uroporphyrinogen-III C-methyltransferase;		TRUE
2.1.1.11	magnesium protoporphyrin IX methyltransferase	TRUE	TRUE
2.1.1.13	methionine synthase;		TRUE
2.1.1.130	precorrin-2 C20-methyltransferase	TRUE	TRUE
2.1.1.131	precorrin-3B C17-methyltransferase;	TRUE	TRUE
2.1.1.132	precorrin-6Y C5,15-methyltransferase (decarboxylating);	TRUE	TRUE
2.1.1.133	precorrin-4 C11-methyltransferase;	TRUE	
2.1.1.151	cobalt-factor II C20-methyltransferase;		TRUE
2.1.1.152	precorrin-6A synthase (deacetylating);	TRUE	TRUE
2.1.1.17	phosphatidylethanolamine N-methyltransferase;		TRUE
2.1.1.37	DNA (cytosine-5-)-methyltransferase;		TRUE
2.1.1.95	tocopherol O-methyltransferase;		TRUE
2.1.2.10	aminomethyltransferase;		TRUE
2.1.2.11	3-methyl-2-oxobutanoate hydroxymethyltransferase;		TRUE
2.1.3.3	ornithine carbamoyltransferase;		TRUE
2.2.1.1	transketolase;	TRUE	TRUE
2.2.1.2	transaldolase;		TRUE
2.2.1.7	1-deoxy-D-xylulose-5-phosphate synthase;		TRUE
2.2.1.9	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid		TRUE
2.3.1.1	amino-acid N-acetyltransferase;		TRUE
2.3.1.12	dihydrolipoyllysine-residue acetyltransferase;		TRUE
2.3.1.15	glycerol-3-phosphate O-acyltransferase;		TRUE
2.3.1.157	glucosamine-1-phosphate N-acetyltransferase		TRUE
2.3.1.179	beta-ketoacyl-acyl-carrier-protein synthase II;	TRUE	TRUE
2.3.1.180	beta-ketoacyl-acyl-carrier-protein synthase III;		TRUE
2.3.1.181	lipoyl(octanoyl) transferase;		TRUE
2.3.1.19	phosphate butyryltransferase;		TRUE

Appendix C (Continued)

Table C.3 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.3.1.30	serine O-acetyltransferase;		TRUE
2.3.1.35	glutamate N-acetyltransferase;		TRUE
2.3.1.37	5-aminolevulinate synthase;		TRUE
2.3.1.46	homoserine O-succinyltransferase;		TRUE
2.3.1.54	formate C-acetyltransferase;		TRUE
2.3.1.57	diamine N-acetyltransferase;		TRUE
2.3.1.61	dihydrolipoyllysine-residue succinyltransferase;		TRUE
2.3.1.9	acetyl-CoA C-acetyltransferase;		TRUE
2.3.3.14	homocitrate synthase;	TRUE	TRUE
2.3.3.9	malate synthase;		TRUE
2.4.1.-	<Transferases; Glycosyltransferases; Hexosyltransferases>		TRUE
2.4.1.13	sucrose synthase;		TRUE
2.4.1.157	1,2-diacylglycerol 3-glucosyltransferase;		TRUE
2.4.1.21	starch synthase;		TRUE
2.4.1.25	4-alpha-glucanotransferase;		TRUE
2.4.1.80	ceramide glucosyltransferase;		TRUE
2.4.1.83	dolichyl-phosphate beta-D-mannosyltransferase;	TRUE	TRUE
2.4.2.1	purine-nucleoside phosphorylase;		TRUE
2.4.2.19	nicotinate-nucleotide diphosphorylase (carboxylating);		TRUE
2.4.2.2	pyrimidine-nucleoside phosphorylase;		TRUE
2.4.2.21	nicotinate-nucleotide---dimethylbenzimidazole		TRUE
2.4.2.22	xanthine phosphoribosyltransferase;		TRUE
2.4.2.28	S-methyl-5'-thioadenosine phosphorylase;	TRUE	TRUE
2.4.2.4	thymidine phosphorylase;		TRUE
2.4.2.7	adenine phosphoribosyltransferase;		TRUE
2.4.2.9	uracil phosphoribosyltransferase;		TRUE
2.5.1.-	<Transferases; Transferring alkyl or aryl groups, other than methyl groups>		TRUE
2.5.1.1	dimethylallyltranstransferase;		TRUE
2.5.1.10	geranyltranstransferase;		TRUE
2.5.1.16	spermidine synthase;		TRUE
2.5.1.17	cob(I)yrinic acid a,c-diamide adenosyltransferase;		TRUE
2.5.1.19	3-phosphoshikimate 1-carboxyvinyltransferase;		TRUE
2.5.1.21	squalene synthase;		TRUE
2.5.1.29	farnesyltranstransferase;		TRUE
2.5.1.3	thiamine-phosphate diphosphorylase;		TRUE
2.5.1.32	phytoene synthase;		TRUE
2.5.1.47	cysteine synthase;		TRUE
2.5.1.48	cystathionine gamma-synthase;		TRUE

Appendix C (Continued)

Table C.3 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.5.1.49	O-acetylhomoserine aminocarboxypropyltransferase;		TRUE
2.5.1.54	3-deoxy-7-phosphoheptulonate synthase;		TRUE
2.5.1.62	chlorophyll synthase	TRUE	TRUE
2.5.1.72	quinolinate synthase;	TRUE	
2.5.1.9	riboflavin synthase;		TRUE
2.6.1.1	aspartate transaminase;		TRUE
2.6.1.11	acetylornithine transaminase;		TRUE
2.6.1.13	ornithine aminotransferase;		TRUE
2.6.1.17	succinyldiaminopimelate transaminase;		TRUE
2.6.1.19	4-aminobutyrate transaminase;		TRUE
2.6.1.21	D-amino-acid transaminase;		TRUE
2.6.1.45	serine---glyoxylate transaminase		TRUE
2.6.1.52	phosphoserine transaminase;		TRUE
2.6.1.9	histidinol-phosphate transaminase;		TRUE
2.7.-.-	<Transferases; Transferring phosphorus-containing groups>		TRUE
2.7.1.107	diacylglycerol kinase;		TRUE
2.7.1.11	6-phosphofructokinase;	TRUE	TRUE
2.7.1.148	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase;		TRUE
2.7.1.156	adenosylcobinamide kinase;	TRUE	TRUE
2.7.1.16	ribulokinase;		TRUE
2.7.1.17	xylulokinase;		TRUE
2.7.1.19	phosphoribulokinase;	TRUE	TRUE
2.7.1.2	glucokinase;	TRUE	TRUE
2.7.1.21	thymidine kinase;		TRUE
2.7.1.24	dephospho-CoA kinase;		TRUE
2.7.1.31	glycerate kinase;		TRUE
2.7.1.4	fructokinase;		TRUE
2.7.1.40	pyruvate kinase;		TRUE
2.7.1.45	2-dehydro-3-deoxygluconokinase;	TRUE	TRUE
2.7.1.48	uridine kinase;		TRUE
2.7.1.63	polyphosphate---glucose phosphotransferase;		TRUE
2.7.1.71	shikimate kinase;		TRUE
2.7.1.92	5-dehydro-2-deoxygluconokinase;	TRUE	TRUE
2.7.2.1	acetate kinase;		TRUE
2.7.2.11	glutamate 5-kinase;		TRUE
2.7.2.7	butyrate kinase		TRUE
2.7.2.8	acetylglutamate kinase;		TRUE
2.7.4.14	cytidylate kinase;		TRUE

Appendix C (Continued)

Table C.3 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.7.4.6	nucleoside-diphosphate kinase;		TRUE
2.7.6.2	thiamine diphosphokinase;		TRUE
2.7.6.3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase;		TRUE
2.7.7.12	UDP-glucose---hexose-1-phosphate uridylyltransferase;		TRUE
2.7.7.13	mannose-1-phosphate guanylyltransferase;	TRUE	TRUE
2.7.7.18	nicotinate-nucleotide adenylyltransferase;		TRUE
2.7.7.22	mannose-1-phosphate guanylyltransferase (GDP);		TRUE
2.7.7.23	UDP-N-acetylglucosamine diphosphorylase;		TRUE
2.7.7.27	glucose-1-phosphate adenylyltransferase;		TRUE
2.7.7.3	pantetheine-phosphate adenylyltransferase;		TRUE
2.7.7.33	glucose-1-phosphate cytidylyltransferase;	TRUE	TRUE
2.7.7.4	sulfate adenylyltransferase;		TRUE
2.7.7.62	adenosylcobinamide-phosphate guanylyltransferase;	TRUE	
2.7.7.63	lipoate---protein ligase;		TRUE
2.7.8.26	adenosylcobinamide-GDP ribazoletransferase;	TRUE	
2.7.9.1	pyruvate, phosphate dikinase;		TRUE
2.7.9.3	selenide, water dikinase;		TRUE
2.8.1.2	3-mercaptopyruvate sulfurtransferase;		TRUE
2.8.1.6	biotin synthase		TRUE
2.8.1.8	lipoyl synthase;		TRUE
2.8.3.8	acetate CoA-transferase;		TRUE
3.1.1.1	carboxylesterase;		TRUE
3.1.1.23	acylglycerol lipase;		TRUE
3.1.2.14	oleoyl-[acyl-carrier-protein] hydrolase;		TRUE
3.1.3.1	alkaline phosphatase;		TRUE
3.1.3.11	fructose-bisphosphatase;		TRUE
3.1.3.15	histidinol-phosphatase;		TRUE
3.1.3.25	inositol-phosphate phosphatase;		TRUE
3.1.3.3	phosphoserine phosphatase		TRUE
3.1.3.37	sedoheptulose-bisphosphatase;	TRUE	TRUE
3.1.3.5	5'-nucleotidase;		TRUE
3.1.3.73	alpha-ribazole phosphatase;		TRUE
3.2.-.-	<Hydrolases; Glycosidases>		TRUE
3.2.1.1	alpha-amylase;		TRUE
3.2.1.20	alpha-glucosidase;		TRUE
3.2.1.23	beta-galactosidase;		TRUE
3.2.1.26	beta-fructofuranosidase;		TRUE
3.2.1.3	glucan 1,4-alpha-glucosidase;		TRUE

Appendix C (Continued)

Table C.3 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
3.2.1.31	beta-glucuronidase;		TRUE
3.2.1.45	glucosylceramidase;		TRUE
3.2.1.52	beta-N-acetylhexosaminidase;		TRUE
3.2.2.1	purine nucleosidase;		TRUE
3.3.1.1	adenosylhomocysteinase;		TRUE
3.4.11.1	leucyl aminopeptidase;		TRUE
3.5.1.14	aminoacylase;		TRUE
3.5.1.16	acetylornithine deacetylase;		TRUE
3.5.1.2	glutaminase;		TRUE
3.5.1.53	N-carbamoylputrescine amidase;		TRUE
3.5.2.14	N-methylhydantoinase (ATP-hydrolysing);		TRUE
3.5.2.17	hydroxyisourate hydrolase;		TRUE
3.5.2.2	dihydropyrimidinase;	TRUE	TRUE
3.5.2.5	allantoinase		TRUE
3.5.3.1	arginase;		TRUE
3.5.3.11	agmatinase;		TRUE
3.5.3.12	agmatine deiminase;		TRUE
3.5.3.19	ureidoglycolate hydrolase		TRUE
3.5.4.1	cytosine deaminase;		TRUE
3.5.4.12	dCMP deaminase;		TRUE
3.5.4.13	dCTP deaminase;		TRUE
3.5.4.16	GTP cyclohydrolase I;	TRUE	TRUE
3.5.4.2	adenine deaminase;		TRUE
3.5.4.25	GTP cyclohydrolase II;		TRUE
3.5.4.26	diaminohydroxyphosphoribosylaminopyrimidine deaminase		TRUE
3.5.4.3	guanine deaminase;		TRUE
3.5.4.4	adenosine deaminase;		TRUE
3.5.4.5	cytidine deaminase;	TRUE	TRUE
3.5.99.6	glucosamine-6-phosphate deaminase;		TRUE
3.6.1.19	nucleoside-triphosphate diphosphatase;		TRUE
3.6.1.23	dUTP diphosphatase;		TRUE
3.6.1.31	phosphoribosyl-ATP diphosphatase;		TRUE
3.7.1.-	<Hydrolases; Acting on carbon-carbon bonds; In ketonic substances>		TRUE
3.7.1.3	kynureninase		TRUE
3.8.1.2	(S)-2-haloacid dehalogenase;		TRUE
3.8.1.5	haloalkane dehalogenase;		TRUE
4.-.-.-	<Lyases>		TRUE
4.1.1.-	<Lyases; Carbon-carbon lyases; Carboxy-lyases>		TRUE

Appendix C (Continued)

Table C.3 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
4.1.1.11	aspartate 1-decarboxylase;		TRUE
4.1.1.12	aspartate 4-decarboxylase;		TRUE
4.1.1.17	ornithine decarboxylase;		TRUE
4.1.1.18	lysine decarboxylase;		TRUE
4.1.1.19	arginine decarboxylase;		TRUE
4.1.1.22	histidine decarboxylase;		TRUE
4.1.1.28	aromatic-L-amino-acid decarboxylase;		TRUE
4.1.1.3	oxaloacetate decarboxylase;		TRUE
4.1.1.31	phosphoenolpyruvate carboxylase;		TRUE
4.1.1.36	phosphopantothenoylcysteine decarboxylase;		TRUE
4.1.1.37	uroporphyrinogen decarboxylase;		TRUE
4.1.1.39	ribulose-bisphosphate carboxylase;	TRUE	TRUE
4.1.1.48	indole-3-glycerol-phosphate synthase;		TRUE
4.1.1.49	phosphoenolpyruvate carboxykinase (ATP);		TRUE
4.1.1.50	adenosylmethionine decarboxylase;	TRUE	TRUE
4.1.1.74	indolepyruvate decarboxylase;		TRUE
4.1.1.9	malonyl-CoA decarboxylase;		TRUE
4.1.2.-	<Lyases; Carbon-carbon lyases; Aldehyde-lyases>		TRUE
4.1.2.14	2-dehydro-3-deoxy-phosphogluconate aldolase;	TRUE	TRUE
4.1.2.25	dihydroneopterin aldolase;		TRUE
4.1.2.5	threonine aldolase;		TRUE
4.1.2.9	phosphoketolase;		TRUE
4.1.3.16	4-hydroxy-2-oxoglutarate aldolase;	TRUE	
4.1.3.36	1,4-dihydroxy-2-naphthoyl-CoA synthase;		TRUE
4.1.3.39	4-hydroxy-2-oxovalerate aldolase;		TRUE
4.2.1.-	<Lyases; Carbon-oxygen lyases; Hydro-lyases>		TRUE
4.2.1.17	enoyl-CoA hydratase;		TRUE
4.2.1.18	methylglutaconyl-CoA hydratase;		TRUE
4.2.1.2	fumarate hydratase;		TRUE
4.2.1.24	porphobilinogen synthase;		TRUE
4.2.1.3	aconitate hydratase;		TRUE
4.2.1.33	3-isopropylmalate dehydratase;		TRUE
4.2.1.44	myo-inosose-2 dehydratase;		TRUE
4.2.1.46	dTDP-glucose 4,6-dehydratase;		TRUE
4.2.1.47	GDP-mannose 4,6-dehydratase;	TRUE	TRUE
4.2.1.60	3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase;		TRUE
4.2.1.7	altronate dehydratase;		TRUE
4.2.1.75	uroporphyrinogen-III synthase;		TRUE

Table C.4 Summary of enzymes in the dataset identified (True) by the NIBBS algorithm and Student's T-test for the phenotype dark fermentative hydrogen production.

EC Number	Enzyme Name	In T-Test	NIBBS
1.17.7.1	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase;		TRUE
2.7.1.11	6-phosphofructokinase;	TRUE	TRUE
4.1.1.50	adenosylmethionine decarboxylase;	TRUE	TRUE
5.3.1.8	mannose-6-phosphate isomerase;		TRUE
2.7.1.2	glucokinase;	TRUE	TRUE
1.1.1.27	L-lactate dehydrogenase;		TRUE
1.4.3.16	L-aspartate oxidase;	TRUE	TRUE
2.1.3.3	ornithine carbamoyltransferase;	TRUE	TRUE
4.1.1.11	aspartate 1-decarboxylase;		TRUE
4.1.2.14	2-dehydro-3-deoxy-phosphogluconate aldolase;	TRUE	TRUE
5.1.3.4	L-ribulose-5-phosphate 4-epimerase;		TRUE
5.3.1.12	glucuronate isomerase;	TRUE	TRUE
6.3.5.4	asparagine synthase (glutamine-hydrolysing);		TRUE
2.4.2.1	purine-nucleoside phosphorylase;	TRUE	TRUE
2.4.2.7	adenine phosphoribosyltransferase;		TRUE
2.4.2.9	uracil phosphoribosyltransferase;		TRUE
1.1.1.18	inositol 2-dehydrogenase;	TRUE	TRUE
1.2.1.10	acetaldehyde dehydrogenase (acetylating);		TRUE
2.2.1.7	1-deoxy-D-xylulose-5-phosphate synthase;		TRUE
2.3.1.46	homoserine O-succinyltransferase;		TRUE
3.2.1.20	alpha-glucosidase;		TRUE
3.5.4.2	adenine deaminase;	TRUE	TRUE
4.2.1.7	altronate dehydratase;		TRUE
5.1.3.14	UDP-N-acetylglucosamine 2-epimerase;		TRUE
5.3.1.5	xylose isomerase;		TRUE
3.5.4.25	GTP cyclohydrolase II;		TRUE
1.17.4.2	ribonucleoside-triphosphate reductase;		TRUE
2.4.2.19	nicotinate-nucleotide diphosphorylase (carboxylating);		TRUE
2.7.7.18	nicotinate-nucleotide adenylyltransferase;		TRUE
3.1.3.15	histidinol-phosphatase;	TRUE	TRUE
3.5.4.26	diaminohydroxyphosphoribosylaminopyrimidine deaminase		TRUE
2.7.6.2	thiamine diphosphokinase;	TRUE	TRUE
1.1.1.133	dTDP-4-dehydrorhamnose reductase;		TRUE
1.1.1.6	glycerol dehydrogenase;		TRUE
1.17.1.2	4-hydroxy-3-methylbut-2-enyl diphosphate reductase;		TRUE
1.2.7.3	2-oxoglutarate synthase;		TRUE
1.4.1.4	glutamate dehydrogenase (NADP+);		TRUE
2.3.1.30	serine O-acetyltransferase;		TRUE
2.3.1.35	glutamate N-acetyltransferase;		TRUE

Appendix C (Continued)

Table C.4 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.4.2.2	pyrimidine-nucleoside phosphorylase;		TRUE
2.6.1.19	4-aminobutyrate transaminase;		TRUE
2.7.1.107	diacylglycerol kinase;		TRUE
2.7.7.12	UDP-glucose---hexose-1-phosphate uridylyltransferase;		TRUE
3.2.1.26	beta-fructofuranosidase;		TRUE
3.5.99.6	glucosamine-6-phosphate deaminase;		TRUE
3.6.1.31	phosphoribosyl-ATP diphosphatase;		TRUE
4.3.1.1	aspartate ammonia-lyase;	TRUE	TRUE
5.3.1.-	<Isomerases; Intramolecular oxidoreductases; Interconverting aldoses and ketoses>		TRUE
2.7.2.7	butyrate kinase	TRUE	TRUE
3.1.3.73	alpha-ribazole phosphatase;	TRUE	TRUE
3.2.1.37	xylan 1,4-beta-xylosidase;	TRUE	TRUE
4.1.2.25	dihydroneopterin aldolase;		TRUE
6.3.3.2	5-formyltetrahydrofolate cyclo-ligase;		TRUE
2.4.1.21	starch synthase;		TRUE
2.7.7.27	glucose-1-phosphate adenylyltransferase;		TRUE
2.3.1.179	beta-ketoacyl-acyl-carrier-protein synthase II;		TRUE
2.4.1.-	<Transferases; Glycosyltransferases; Hexosyltransferases>		TRUE
2.7.1.21	thymidine kinase;	TRUE	TRUE
4.-.-.-	<Lyases>		TRUE
1.1.1.267	1-deoxy-D-xylulose-5-phosphate reductoisomerase;		TRUE
1.1.1.49	glucose-6-phosphate dehydrogenase;		TRUE
1.1.1.57	fructuronate reductase;	TRUE	TRUE
1.2.1.38	N-acetyl-gamma-glutamyl-phosphate reductase;		TRUE
1.2.7.1	pyruvate synthase;		TRUE
1.3.1.54	precorrin-6A reductase;		TRUE
1.8.1.4	dihydrolipoyl dehydrogenase;		TRUE
2.1.1.10	homocysteine S-methyltransferase;		TRUE
2.3.1.54	formate C-acetyltransferase;		TRUE
2.5.1.1	dimethylallyltranstransferase;		TRUE
2.5.1.19	3-phosphoshikimate 1-carboxyvinyltransferase;		TRUE
2.7.1.16	ribulokinase;		TRUE
2.7.1.92	5-dehydro-2-deoxygluconokinase;		TRUE
2.7.4.6	nucleoside-diphosphate kinase;		TRUE
3.5.3.11	agmatinase;		TRUE
3.5.4.12	dCMP deaminase;	TRUE	TRUE
3.5.4.16	GTP cyclohydrolase I;		TRUE
4.1.1.48	indole-3-glycerol-phosphate synthase;		TRUE

Appendix C (Continued)

Table C.4 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
4.1.2.-	<Lyases; Carbon-carbon lyases; Aldehyde-lyases>		TRUE
4.2.1.44	myo-inosose-2 dehydratase;		TRUE
6.1.1.24	glutamate---tRNAGln ligase;		TRUE
6.3.1.1	aspartate---ammonia ligase;		TRUE
2.7.1.6	galactokinase;	TRUE	TRUE
3.2.1.52	beta-N-acetylhexosaminidase;		TRUE
5.3.1.4	L-arabinose isomerase;	TRUE	TRUE
6.-.-.-	null		TRUE
1.1.3.15	(S)-2-hydroxy-acid oxidase;		TRUE
2.3.1.19	phosphate butyryltransferase;		TRUE
3.2.1.1	alpha-amylase;		TRUE
3.7.1.-	<Hydrolases; Acting on carbon-carbon bonds; In ketonic substances>		TRUE
6.3.1.10	adenosylcobinamide-phosphate synthase;		TRUE
2.5.1.47	cysteine synthase;		TRUE
3.2.1.23	beta-galactosidase;		TRUE
4.2.1.51	prephenate dehydratase;		TRUE
2.5.1.9	riboflavin synthase;		TRUE
2.8.1.6	biotin synthase	TRUE	TRUE
4.2.1.-	<Lyases; Carbon-oxygen lyases; Hydro-lyases>		TRUE
2.7.9.1	pyruvate, phosphate dikinase;		TRUE
1.1.1.14	L-iditol 2-dehydrogenase;		TRUE
1.1.1.271	GDP-L-fucose synthase;		TRUE
1.1.1.42	isocitrate dehydrogenase (NADP+);		TRUE
1.2.1.9	glyceraldehyde-3-phosphate dehydrogenase (NADP+);		TRUE
1.2.4.4	3-methyl-2-oxobutanoate dehydrogenase		TRUE
1.2.7.5	aldehyde ferredoxin oxidoreductase;		TRUE
1.3.1.76	precorrin-2 dehydrogenase;		TRUE
1.3.1.9	enoyl-[acyl-carrier-protein] reductase (NADH);		TRUE
2.1.1.107	uroporphyrinogen-III C-methyltransferase;		TRUE
2.1.1.130	precorrin-2 C20-methyltransferase		TRUE
2.1.1.37	DNA (cytosine-5-)-methyltransferase;		TRUE
2.1.2.11	3-methyl-2-oxobutanoate hydroxymethyltransferase;		TRUE
2.2.1.1	transketolase;	TRUE	TRUE
2.3.1.180	beta-ketoacyl-acyl-carrier-protein synthase III;		TRUE
2.4.1.18	1,4-alpha-glucan branching enzyme;		TRUE
2.4.1.8	maltose phosphorylase		TRUE
2.4.2.28	S-methyl-5'-thioadenosine phosphorylase;		TRUE
2.4.2.3	uridine phosphorylase;		TRUE

Appendix C (Continued)

Table C.4 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.5.1.48	cystathionine gamma-synthase;		TRUE
2.5.1.49	O-acetylhomoserine aminocarboxypropyltransferase;		TRUE
2.6.1.21	D-amino-acid transaminase;		TRUE
2.7.1.40	pyruvate kinase;		TRUE
2.7.1.71	shikimate kinase;		TRUE
2.7.7.13	mannose-1-phosphate guanylyltransferase;		TRUE
3.2.1.31	beta-glucuronidase;		TRUE
3.2.1.45	glucosylceramidase;		TRUE
3.2.2.1	purine nucleosidase;		TRUE
3.4.13.3	Xaa-His dipeptidase;		TRUE
3.5.1.14	aminoacylase;		TRUE
3.5.1.2	glutaminase;		TRUE
3.5.2.2	dihydropyrimidinase;		TRUE
3.5.4.1	cytosine deaminase;		TRUE
3.5.4.13	dCTP deaminase;		TRUE
4.1.1.19	arginine decarboxylase;		TRUE
4.1.1.49	phosphoenolpyruvate carboxykinase (ATP);		TRUE
4.4.1.16	selenocysteine lyase;		TRUE
6.3.2.2	glutamate---cysteine ligase;		TRUE
6.3.4.18	5-(carboxyamino)imidazole ribonucleotide synthase;		TRUE
6.3.5.7	glutaminyl-tRNA synthase (glutamine-hydrolysing);		TRUE
1.18.6.1	nitrogenase		TRUE
5.3.1.23	S-methyl-5-thioribose-1-phosphate isomerase;		TRUE
2.6.1.1	aspartate transaminase;		TRUE
2.7.6.3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase;		TRUE
2.7.7.3	pantetheine-phosphate adenylyltransferase;		TRUE
4.1.1.36	phosphopantothenoylcysteine decarboxylase;		TRUE
1.3.99.2	butyryl-CoA dehydrogenase;		TRUE
2.1.1.131	precorrin-3B C17-methyltransferase;		TRUE
2.4.2.4	thymidine phosphorylase;		TRUE
4.1.1.18	lysine decarboxylase;		TRUE
5.1.1.7	diaminopimelate epimerase		TRUE
6.3.3.3	dethiobiotin synthase;		TRUE
1.3.1.-	<Oxidoreductases; Acting on the CH-CH group of donors>		TRUE
2.6.1.62	adenosylmethionine---8-amino-7-oxononanoate transaminase;		TRUE
2.7.1.31	glycerate kinase;		TRUE
4.2.1.75	uroporphyrinogen-III synthase;		TRUE
1.1.-.-	<Oxidoreductases; Acting on the CH-OH group of donors>		TRUE

Table C.4 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
6.4.1.3	propionyl-CoA carboxylase;		TRUE
2.-.-.-	<Transferases>		TRUE
2.3.1.15	glycerol-3-phosphate O-acyltransferase;		TRUE
2.5.1.10	geranyltranstransferase;		TRUE
2.7.4.14	cytidylate kinase;		TRUE
5.-.-.-	<Isomerases>		TRUE
1.8.99.2	adenylyl-sulfate reductase;		TRUE
2.1.1.151	cobalt-factor II C20-methyltransferase;		TRUE
2.7.7.4	sulfate adenylyltransferase;		TRUE
2.4.1.83	dolichyl-phosphate beta-D-mannosyltransferase;		TRUE
2.8.3.8	acetate CoA-transferase;		TRUE
3.5.2.5	allantoinase		TRUE
3.8.1.2	(S)-2-haloacid dehalogenase;		TRUE
4.1.1.22	histidine decarboxylase;		TRUE
4.1.1.4	acetoacetate decarboxylase;		TRUE
2.7.8.-	<Transferases; Transferring phosphorus-containing groups>		TRUE
2.7.8.8	CDP-diacylglycerol---serine O-phosphatidyltransferase;		TRUE
1.1.1.1	alcohol dehydrogenase;		TRUE
1.14.14.1	unspecific monooxygenase;		TRUE
1.3.1.12	prephenate dehydrogenase;		TRUE
1.3.99.-	<Oxidoreductases; Acting on the CH-CH group of donors; With other acceptors>		TRUE
1.4.1.2	glutamate dehydrogenase;		TRUE
1.4.3.4	monoamine oxidase;		TRUE
2.4.1.46	monogalactosyldiacylglycerol synthase;		TRUE
2.6.1.37	2-aminoethylphosphonate---pyruvate transaminase;		TRUE
2.7.1.113	deoxyguanosine kinase;		TRUE
2.7.1.12	gluconokinase;		TRUE
2.7.1.89	thiamine kinase;		TRUE
2.7.8.20	phosphatidylglycerol---membrane-oligosaccharide		TRUE
3.1.1.11	pectinesterase;		TRUE
3.1.3.25	inositol-phosphate phosphatase;		TRUE
3.2.1.35	hyaluronoglucosaminidase;		TRUE
3.2.1.67	galacturan 1,4-alpha-galacturonidase;		TRUE
3.5.2.17	hydroxyisourate hydrolase;		TRUE
3.5.3.1	arginase;		TRUE
3.5.3.19	ureidoglycolate hydrolase		TRUE
4.1.1.-	<Lyases; Carbon-carbon lyases; Carboxy-lyases>		TRUE
4.1.1.1	pyruvate decarboxylase;		TRUE

Appendix C (Continued)

Table C.4 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
4.1.1.15	glutamate decarboxylase;		TRUE
4.1.1.2	oxalate decarboxylase;		TRUE
4.1.1.3	oxaloacetate decarboxylase;		TRUE
4.2.1.80	2-oxopent-4-enoate hydratase;		TRUE
5.1.3.7	UDP-N-acetylglucosamine 4-epimerase;		TRUE
5.1.99.1	methylmalonyl-CoA epimerase;		TRUE
5.3.1.22	hydroxypyruvate isomerase		TRUE
5.4.2.9	phosphoenolpyruvate mutase;		TRUE
5.5.1.4	inositol-3-phosphate synthase;		TRUE
6.1.1.18	glutamine---tRNA ligase;		TRUE
6.3.1.5	NAD+ synthase;		TRUE
6.3.5.1	NAD+ synthase (glutamine-hydrolysing);		TRUE
2.5.1.3	thiamine-phosphate diphosphorylase;		TRUE
6.3.4.3	formate---tetrahydrofolate ligase;		TRUE
1.1.1.58	tagaturonate reductase;	TRUE	
2.4.2.8	hypoxanthine phosphoribosyltransferase;	TRUE	
2.5.1.16	spermidine synthase;	TRUE	
2.5.1.72	quinolinate synthase;	TRUE	
2.7.1.4	fructokinase;	TRUE	
2.7.1.45	2-dehydro-3-deoxygluconokinase;	TRUE	
2.7.1.48	uridine kinase;	TRUE	
3.5.4.5	cytidine deaminase;	TRUE	
4.1.3.16	4-hydroxy-2-oxoglutarate aldolase;	TRUE	
4.2.1.8	mannonate dehydratase;	TRUE	

Appendix C (Continued)

Table C.5 Summary of enzymes in the dataset identified (True) by the NIBBS algorithm and Student’s T-test for the phenotype acid-tolerant.

EC Number	Enzyme Name	In T-Test	NIBBS
1.1.1.14	L-iditol 2-dehydrogenase;		TRUE
1.1.1.169	2-dehydropantoate 2-reductase;		TRUE
1.1.1.18	inositol 2-dehydrogenase;		TRUE
1.1.1.219	dihydrokaempferol 4-reductase;		TRUE
1.1.1.39	malate dehydrogenase (decarboxylating);		TRUE
1.1.1.40	malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+);		TRUE
1.1.1.42	isocitrate dehydrogenase (NADP+);		TRUE
1.1.1.49	glucose-6-phosphate dehydrogenase;		TRUE
1.1.1.56	ribitol 2-dehydrogenase;		TRUE
1.1.1.60	2-hydroxy-3-oxopropionate reductase;		TRUE
1.1.1.79	glyoxylate reductase (NADP+);		TRUE
1.1.1.90	aryl-alcohol dehydrogenase;		TRUE
1.1.5.2	quinoprotein glucose dehydrogenase;		TRUE
1.17.4.1	ribonucleoside-diphosphate reductase;		TRUE
1.17.4.2	ribonucleoside-triphosphate reductase;		TRUE
1.2.1.27	methylmalonate-semialdehyde dehydrogenase (acylating);		TRUE
1.2.3.3	pyruvate oxidase;		TRUE
1.2.7.5	aldehyde ferredoxin oxidoreductase;		TRUE
1.3.1.-	<Oxidoreductases; Acting on the CH-CH group of donors>		TRUE
1.3.1.9	enoyl-[acyl-carrier-protein] reductase (NADH);		TRUE
1.4.3.4	monoamine oxidase;		TRUE
1.5.1.-	<Oxidoreductases; Acting on the CH-NH group of donors>		TRUE
1.5.1.3	dihydrofolate reductase;		TRUE
2.1.1.10	homocysteine S-methyltransferase;		TRUE
2.1.1.14	5-methyltetrahydropteroyltriglutamate---homocysteine		TRUE
2.1.1.37	DNA (cytosine-5-)-methyltransferase;		TRUE
2.1.1.45	thymidylate synthase;		TRUE
2.2.1.9	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid		TRUE
2.3.1.-	<Transferases; Acyltransferases; Transferring groups other than amino-acyl groups>		TRUE
2.3.1.117	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase;		TRUE
2.3.1.19	phosphate butyryltransferase;		TRUE
2.3.1.46	homoserine O-succinyltransferase;		TRUE
2.3.1.54	formate C-acetyltransferase;		TRUE
2.3.2.-	<Transferases; Acyltransferases; Aminoacyltransferases>		TRUE
2.3.3.10	hydroxymethylglutaryl-CoA synthase;		TRUE
2.3.3.14	homocitrate synthase;		TRUE
2.4.1.10	levansucrase;		TRUE
2.4.1.157	1,2-diacylglycerol 3-glucosyltransferase;	TRUE	TRUE
2.4.1.18	1,4-alpha-glucan branching enzyme;		TRUE

Appendix C (Continued)

Table C.5 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.4.1.25	4-alpha-glucanotransferase;		TRUE
2.4.1.8	maltose phosphorylase		TRUE
2.4.1.80	ceramide glucosyltransferase;		TRUE
2.4.1.83	dolichyl-phosphate beta-D-mannosyltransferase;		TRUE
2.4.2.11	nicotinate phosphoribosyltransferase;		TRUE
2.4.2.21	nicotinate-nucleotide---dimethylbenzimidazole		TRUE
2.4.2.22	xanthine phosphoribosyltransferase;		TRUE
2.4.2.3	uridine phosphorylase;		TRUE
2.4.2.7	adenine phosphoribosyltransferase;		TRUE
2.5.1.17	cob(I)yrinic acid a,c-diamide adenosyltransferase;		TRUE
2.6.1.-	<Transferases; Transferring nitrogenous groups; Transaminases>		TRUE
2.6.1.1	aspartate transaminase;		TRUE
2.6.1.17	succinyldiaminopimelate transaminase;		TRUE
2.6.1.2	alanine transaminase;		TRUE
2.6.1.37	2-aminoethylphosphonate---pyruvate transaminase;		TRUE
2.6.1.57	aromatic-amino-acid transaminase;		TRUE
2.6.1.66	valine---pyruvate transaminase;		TRUE
2.7.1.-	<Transferases; Transferring phosphorus-containing groups>		TRUE
2.7.1.100	S-methyl-5-thioribose kinase;		TRUE
2.7.1.107	diacylglycerol kinase;		TRUE
2.7.1.113	deoxyguanosine kinase;		TRUE
2.7.1.12	gluconokinase;		TRUE
2.7.1.16	ribulokinase;		TRUE
2.7.1.17	xylulokinase;		TRUE
2.7.1.29	glycerone kinase;		TRUE
2.7.1.31	glycerate kinase;		TRUE
2.7.1.35	pyridoxal kinase;	TRUE	TRUE
2.7.1.4	fructokinase;	TRUE	TRUE
2.7.1.48	uridine kinase;		TRUE
2.7.1.6	galactokinase;	TRUE	TRUE
2.7.1.69	protein-Npi-phosphohistidine---sugar phosphotransferase;	TRUE	TRUE
2.7.2.1	acetate kinase;		TRUE
2.7.2.7	butyrate kinase		TRUE
2.7.4.2	phosphomevalonate kinase;		TRUE
2.7.4.6	nucleoside-diphosphate kinase;		TRUE
2.7.6.2	thiamine diphosphokinase;		TRUE
2.7.7.15	choline-phosphate cytidylyltransferase;		TRUE
2.7.8.-	<Transferases; Transferring phosphorus-containing groupss>		TRUE

Appendix C (Continued)

Table C.5 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
2.7.8.8	CDP-diacylglycerol---serine O-phosphatidyltransferase;		TRUE
2.7.9.1	pyruvate, phosphate dikinase;		TRUE
2.8.3.1	propionate CoA-transferase;		TRUE
3.1.1.3	triacylglycerol lipase;		TRUE
3.1.1.31	6-phosphogluconolactonase;		TRUE
3.1.1.45	carboxymethylenebutenolidase;		TRUE
3.1.2.14	oleoyl-[acyl-carrier-protein] hydrolase;	TRUE	TRUE
3.1.3.18	phosphoglycolate phosphatase;		TRUE
3.1.3.73	alpha-ribazole phosphatase;		TRUE
3.1.4.3	phospholipase C;		TRUE
3.11.1.1	phosphonoacetaldehyde hydrolase;		TRUE
3.2.1.-	<Hydrolases; Glycosidases; Glycosidases>		TRUE
3.2.1.1	alpha-amylase;		TRUE
3.2.1.10	oligo-1,6-glucosidase;		TRUE
3.2.1.20	alpha-glucosidase;		TRUE
3.2.1.20	alpha-glucosidase;		TRUE
3.2.1.23	beta-galactosidase;		TRUE
3.2.1.26	beta-fructofuranosidase;	TRUE	TRUE
3.2.1.35	hyaluronoglucosaminidase;		TRUE
3.2.1.45	glucosylceramidase;		TRUE
3.2.2.1	purine nucleosidase;		TRUE
3.2.2.9	adenosylhomocysteine nucleosidase;		TRUE
3.4.11.2	membrane alanyl aminopeptidase;		TRUE
3.4.13.3	Xaa-His dipeptidase;		TRUE
3.4.13.3	Xaa-His dipeptidase;		TRUE
3.5.1.-	<Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bonds>		TRUE
3.5.1.1	asparaginase;		TRUE
3.5.1.14	aminoacylase;		TRUE
3.5.1.18	succinyl-diaminopimelate desuccinylase;		TRUE
3.5.1.24	choloylglycine hydrolase;	TRUE	TRUE
3.5.1.53	N-carbamoylputrescine amidase;		TRUE
3.5.2.14	N-methylhydantoinase (ATP-hydrolysing);		TRUE
3.5.3.12	agmatine deiminase;		TRUE
3.5.4.1	cytosine deaminase;		TRUE
3.5.4.12	dCMP deaminase;		TRUE
3.5.4.13	dCTP deaminase;		TRUE
3.5.4.16	GTP cyclohydrolase I;		TRUE
3.5.4.3	guanine deaminase;		TRUE

Appendix C (Continued)

Table C.5 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
3.5.4.4	adenosine deaminase;	TRUE	TRUE
3.5.4.5	cytidine deaminase;		TRUE
3.5.99.3	hydroxydechloroatrazine ethylaminohydrolase;		TRUE
3.5.99.6	glucosamine-6-phosphate deaminase;		TRUE
3.6.1.23	dUTP diphosphatase;	TRUE	TRUE
3.8.1.2	(S)-2-haloacid dehalogenase;		TRUE
4.1.1.12	aspartate 4-decarboxylase;		TRUE
4.1.1.15	glutamate decarboxylase;		TRUE
4.1.1.17	ornithine decarboxylase;		TRUE
4.1.1.18	lysine decarboxylase;		TRUE
4.1.1.20	diaminopimelate decarboxylase;		TRUE
4.1.1.22	histidine decarboxylase;		TRUE
4.1.1.3	oxaloacetate decarboxylase;		TRUE
4.1.1.31	phosphoenolpyruvate carboxylase;		TRUE
4.1.1.4	acetoacetate decarboxylase;		TRUE
4.1.1.44	4-carboxymuconolactone decarboxylase;	TRUE	TRUE
4.1.1.74	indolepyruvate decarboxylase;		TRUE
4.1.2.-	<Lyases; Carbon-carbon lyases; Aldehyde-lyases>		TRUE
4.1.2.13	fructose-bisphosphate aldolase;		TRUE
4.1.2.14	2-dehydro-3-deoxy-phosphogluconate aldolase;		TRUE
4.1.2.9	phosphoketolase;		TRUE
4.2.1.46	dTDP-glucose 4,6-dehydratase;		TRUE
4.3.1.1	aspartate ammonia-lyase;		TRUE
4.4.1.16	selenocysteine lyase;		TRUE
4.4.1.8	cystathionine beta-lyase;		TRUE
5.1.3.14	UDP-N-acetylglucosamine 2-epimerase;		TRUE
5.1.3.2	UDP-glucose 4-epimerase;		TRUE
5.1.3.4	L-ribulose-5-phosphate 4-epimerase;	TRUE	TRUE
5.3.1.8	mannose-6-phosphate isomerase;	TRUE	TRUE
5.3.1.9	glucose-6-phosphate isomerase;		TRUE
5.3.2.-	<Isomerases; Intramolecular oxidoreductases; Interconverting keto- and enol-groups>		TRUE
5.3.3.2	isopentenyl-diphosphate Delta-isomerase;		TRUE
5.4.2.2	phosphoglucomutase;		TRUE
5.5.1.2	3-carboxy-cis,cis-muconate cycloisomerase;		TRUE
5.5.1.4	inositol-3-phosphate synthase;		TRUE
6.1.1.17	glutamate---tRNA ligase;		TRUE
6.1.1.18	glutamine---tRNA ligase;		TRUE
6.2.1.17	propionate---CoA ligase;		TRUE

Appendix C (Continued)

Table C.5 Continued

EC Number	Enzyme Name	In T-Test	NIBBS
6.3.1.1	aspartate---ammonia ligase;		TRUE
6.3.2.2	glutamate---cysteine ligase;		TRUE
6.3.2.3	glutathione synthase;		TRUE
6.3.4.14	biotin carboxylase;		TRUE
6.3.4.3	formate---tetrahydrofolate ligase;		TRUE
6.3.4.3	formate---tetrahydrofolate ligase;		TRUE
6.3.4.4	adenylosuccinate synthase;		TRUE
6.3.5.1	NAD+ synthase (glutamine-hydrolysing);		TRUE
6.3.5.5	carbamoyl-phosphate synthase (glutamine-hydrolysing);		TRUE
6.4.1.1	pyruvate carboxylase;		TRUE
2.7.6.1	ribose-phosphate diphosphokinase; ribose-phosphate pyrophosphokinase; PRPP synthetase; phosphoribosylpyrophosphate synthetase; PPRibP synthetase; PP-ribose P synthetase; 5-phosphoribosyl-1-pyrophosphate synthetase; 5-phosphoribose pyrophosphorylase; 5-phosphoribosyl-alpha-1-pyrophosphate synthetase; phosphoribosyl-diphosphate synthetase; phosphoribosylpyrophosphate synthase; pyrophosphoribosylphosphate synthetase; ribophosphate pyrophosphokinase; ribose-5-phosphate pyrophosphokinase	TRUE	
5.4.2.1	phosphoglycerate mutase; phosphoglycerate phosphomutase; phosphoglyceromutase; glycerate phosphomutase (diphosphoglycerate cofactor); monophosphoglycerate mutase; monophosphoglyceromutase; diphosphoglycomutase; diphosphoglycerate mutase; bisphosphoglyceromutase; GriP mutase; MPGM; PGA mutase; PGAM-i; PGAM; PGAM-d; PGM	TRUE	
1.1.1.27	L-lactate dehydrogenase; lactic acid dehydrogenase; L(+)-nLDH; L-(+)-lactate dehydrogenase; L-lactic dehydrogenase; L-lactic acid dehydrogenase; lactate dehydrogenase; lactate dehydrogenase NAD+-dependent; lactic dehydrogenase; NAD+-lactate dehydrogenase	TRUE	
4.1.2.9	phosphoketolase; D-xylulose-5-phosphate D-glyceraldehyde-3-phosphate-lyase (phosphate-acetylating)	TRUE	