Transcription Regulation of the *Clostridium thermocellum* Cellulase System

by

Michael James Newcomb

Submitted in Partial Fulfillment
of the
Requirements for the Degree
Doctor of Philosophy

Supervised by
Professor J. H. David Wu

Department of Chemical Engineering
The College
School of Engineering and Applied Sciences

University of Rochester
Rochester, New York

2007
Curriculum Vitae

The author was born in Dansville, New York on December 30th, 1979. He attended the University of Rochester from 1998 to 2002, and graduated with a Bachelor of Science degree in Biomedical Engineering. He began his graduate work at the University of Rochester in the Fall of 2002. He pursued his research in Chemical Engineering under the direction of Professor David Wu.
Acknowledgements

During the course of my graduate career I have learned a lot and I owe much of that to the people that I have encountered. I have many people to thank, who have all assisted me in my research and in life. First, I would like to thank my advisor, Dr. J.H. David Wu, for his advice and support throughout the last five years. I appreciate that he gave me the opportunity to lead a project from my very first day in the lab and stuck with me when things didn’t go as planned.

I want to thank Matt Heckman, Chun-Yu Chen, and Melissa Krebs for all of their technical advice, especially when I first started on my project. I’d especially like to thank my current and former labmates, Sara Ladd, Tushar Dabral, Sunny Wu, and Melissa Krebs, who have joined me in working on the cellulase project. I’d like to thank Eric Newcomb and Kelly White for editing this thesis. Finally, I’d like to thank my family for their love and support throughout my graduate years.
Abstract

Biomass conversion to ethanol as a liquid fuel may potentially offset some of the world’s dependence on petroleum for energy. *Clostridium thermocellum* is an anaerobic, thermophilic, cellulolytic, and ethanogenic bacterium that holds much promise as an industrial ethanol producer. The organism is able to degrade cellulose into fermentable sugars by utilizing a complex cellulase system. Very little is known about how the bacterium regulates the expression of its cellulase system. We have identified *glyR3* as the first transcriptional regulator of cellulase genes discovered in the organism. Northern blot showed that *glyR3* is co-transcribed with cellulase/hemicellulase genes *celC* and *licA*. Using Real-Time quantitative Polymerase Chain Reaction (RT-qPCR), it was shown that when the bacterium is grown on β-1,3 linked glucans, the expression of *celC*, *glyR3*, and *licA* is increased when compared to growth on β-1,4 linked glucans. Gel-shift assays revealed that recombinant GlyR3 bound specifically to the *celC* promoter region. GlyR3 was identified from cell lysate on lichenan-grown cells but could not be identified from cellobiose-grown cells. DNase I footprinting and competitive gel-shift proved the binding site was an 18-bp near perfect palindromic sequence. Gel-shift and *in vitro* transcription assays showed that the DNA-binding activity was specifically inhibited by laminaribiose, a β-1,3 linked glucose dimer. The results prove that GlyR3 is a negative regulator of the *celC* operon with laminaribiose as an inducer. Also described in this work is an
uncharacterized gene cluster of five cellulase system components, named Cluster X. Cluster X consists of five genes that encode proteins containing dockerin domains, meaning they are components of the cellulosome. The cellulosome is an extracellular multiprotein complex that efficiently degrades cellulose. Northern blot revealed that all five genes in Cluster X are co-transcribed and primer extension revealed two initiation sites for the transcript. RT-qPCR discovered that all of the genes have transcripts that are more highly expressed when the bacterium is grown on cellulose, as compared to cellobiose. Additionally, the highest point of expression, regardless of substrate, was found in the stationary growth phase. The results are a first step in understanding the regulation of the cellulase system in *C. thermocellum*. 
Table of Contents

Chapter 1  Introduction 1

1.1 The Need for an Alternative Energy Source to Replace Fossil Fuels 1

1.2 Overview of the Cellulosic Biomass to Ethanol Process 3

1.3 Clostridium thermocellum as an Industrial Ethanol Producer 4

1.4 The Cellulase System of Clostridium thermocellum 5

1.5 Carbon Source Regulation of Cellulase Synthesis in C. thermocellum 7

1.6 Clostridia Gene Clusters and Negative Regulation of Cellulase Genes 10

1.7 Molecular Mechanisms of Cellulase Regulation in Other Organisms 13

1.8 The Need for Knowledge of Molecular Mechanisms for Cellulase Regulation in C. thermocellum 14

1.9 Synopsis of Results 15

1.10 Chapter 1 References 19

Chapter 2  The celC Gene Cluster and its Expression Profile 27

2.1 Introduction 27

2.2 Results 28

2.3 Discussion 32

2.4 Materials and Methods 34

2.5 Chapter 2 References 49
<table>
<thead>
<tr>
<th>Chapter 3</th>
<th>Characterization of GlyR3</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>50</td>
</tr>
<tr>
<td>3.2</td>
<td>Results</td>
<td>51</td>
</tr>
<tr>
<td>3.3</td>
<td>Discussion</td>
<td>57</td>
</tr>
<tr>
<td>3.4</td>
<td>Materials and Methods</td>
<td>60</td>
</tr>
<tr>
<td>3.5</td>
<td>Chapter 3 References</td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>Bioinformatic Analysis of Cluster X and Study of its Transcript Expression Profile</th>
<th>77</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>77</td>
</tr>
<tr>
<td>4.2</td>
<td>Results</td>
<td>78</td>
</tr>
<tr>
<td>4.3</td>
<td>Discussion</td>
<td>80</td>
</tr>
<tr>
<td>4.4</td>
<td>Materials and Methods</td>
<td>82</td>
</tr>
<tr>
<td>4.5</td>
<td>Chapter 4 References</td>
<td>99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 5</th>
<th>Conclusions and Future Work</th>
<th>101</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Conclusions</td>
<td>101</td>
</tr>
<tr>
<td>5.2</td>
<td>Future Work</td>
<td>103</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Primers and their function for Chapter 2 experiments</td>
<td>48</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Primers and their functions for Chapter 3 experiments</td>
<td>66</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>The DNA binding half-sites of various LacI/GalR family regulators</td>
<td>72</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Primers and their function for Chapter 4 experiments</td>
<td>98</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>An illustration of the cellulosome of <em>Clostridium thermocellum</em></td>
<td>16</td>
</tr>
<tr>
<td>1.2</td>
<td>Cellulase system gene clusters in <em>C. thermocellum</em></td>
<td>17</td>
</tr>
<tr>
<td>1.3</td>
<td>Gene clusters in Mesophilic Clostridia related to the cellulase system of each organism</td>
<td>18</td>
</tr>
<tr>
<td>2.1</td>
<td>A schematic of the <em>celC</em> gene cluster</td>
<td>39</td>
</tr>
<tr>
<td>2.2</td>
<td>Northern blot results of <em>ceC</em>, <em>glyR3</em>, and <em>licA</em></td>
<td>40</td>
</tr>
<tr>
<td>2.3</td>
<td>Reverse transcriptase mapping of <em>orf4</em>, <em>manB</em>, and <em>celT</em></td>
<td>41</td>
</tr>
<tr>
<td>2.4</td>
<td>Transcription initiation site as revealed by primer extension using a primer corresponding to the <em>celC</em> sequence</td>
<td>42</td>
</tr>
<tr>
<td>2.5</td>
<td>Transcription initiation site as revealed by primer extension using a primer corresponding to the <em>orf4</em> and <em>manB</em> sequence</td>
<td>43</td>
</tr>
<tr>
<td>2.6</td>
<td>A growth curve showing intracellular protein for cultures grown using celllobiose, cellulose, and laminarin</td>
<td>44</td>
</tr>
<tr>
<td>2.7</td>
<td>The expression patterns for <em>celC</em>, <em>glyR3</em> and <em>licA</em> at different growth phases using different substrates</td>
<td>45</td>
</tr>
<tr>
<td>2.8</td>
<td>The expression pattern for <em>orf4</em> at different growth phases using different substrates</td>
<td>46</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>2.9</td>
<td>The expression patterns for <em>manB</em> and <em>celT</em> at different growth phases using different substrates</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Alignment of GlyR1, GlyR2, GlyR3, and LacI</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Schematic drawing of the <em>celC</em> operon</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Binding of rGlyR3 to the <em>celC</em> promoter region as revealed by EMSA</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>GlyR3 DNA-binding site as determined by DNase I footprinting analysis</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>Competitive EMSA confirming the rGlyR3 DNA-binding site</td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td>Inhibition of GlyR3 DNA-binding activity by laminaribiose as analyzed by EMSA</td>
<td></td>
</tr>
<tr>
<td>3.7</td>
<td>Laminaribiose induction of <em>celC</em> by inactivating GlyR3 as revealed by <em>in vitro</em> transcription assay</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>A schematic of the genes located in Cluster X</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>The protein domain structures for all members of Cluster X</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>Northern blot analysis of the genes from Cluster X</td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>Primer extension results for <em>gene1</em> of Cluster X</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Growth curves of <em>C. thermocellum</em> using cellobiose or cellulose as carbon sources</td>
<td></td>
</tr>
<tr>
<td>4.6</td>
<td>RT-qPCR results for <em>gene1</em> of Cluster X</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>RT-qPCR results for <em>gene2</em> of Cluster X</td>
<td></td>
</tr>
<tr>
<td>4.8</td>
<td>RT-qPCR results for <em>gene3</em> of Cluster X</td>
<td></td>
</tr>
</tbody>
</table>
4.9  RT-qPCR results for gene4 of Cluster X 94
4.10 RT-qPCR results for gene5 of Cluster X 95
4.11 RT-qPCR analysis of celS transcript level 96
4.12 RT-qPCR analysis of xynC transcript level 97
Chapter 1

Introduction

1.1) The Need for an Alternative Energy Source to Replace Fossil Fuels

Rising oil prices have led to a renewed effort to find an economically competitive, environmentally friendly source of renewable energy to replace fossil fuels. As the world’s biggest consumer of fossil fuels, various political and economic factors make it worthwhile for the United States to lead the effort in researching alternative energy. Over half of all gasoline used in the United States for transportation and for industrial purposes must be imported [1]. This is in addition to the crude oil that is imported to the United States to be refined into gasoline within the country. The importing of crude oil and refined gasoline is a major contributor to the federal trade deficit facing the United States at this time. In the first quarter of 2007, the importing of petroleum accounted for over 30% of the overall trade deficit [2]. These facts make the search for a “homegrown” energy source very appealing, as the advent of a new energy source that does not depend so heavily on imports would lead to a more energy independent United States.
One possibility for a homegrown energy source would be biomass feedstock. In 2005, Oak Ridge National Laboratory conducted a study which found that the United States can produce 1.3 billion tons of biomass feedstock annually with only modest changes made in farming practices and land usage [3]. One billion tons of feedstock could be produced without affecting the food, feed, and exporting demands placed on the agriculture industry and would produce enough fuel to offset roughly 30% of the current transportation fuel demands in the United States. As technology improves, it may be possible to increase the amount of feedstock available, and as conversion technology increases, it may be possible to offset an ever higher percent of the energy demand (40-60%). However, the eventual percentage of our total energy consumption that is met by biomass fuels will largely be based on the efficiency of the conversion process. A more efficient process will give a more abundant yield and be more economical.

The most abundant renewable natural resource that is available for conversion into fuel is Lignocellulose. The liquid fuel that is produced from this biomass substrate is ethanol. Some aspects of ethanol make it a superior fuel when compared to gasoline [4, 5]. For example, unblended ethanol burns more cleanly than gasoline, which leads to greater burn efficiency and to a reduction of greenhouse gases released. Additionally, low levels of smog producing gasses are produced by burning ethanol [6]. Ethanol is also much less toxic to humans and animals than gasoline. Current technology is able to produce ethanol from biomass (mostly using corn based
substrates) but is unable to efficiently use more cellulosic substrates. The current process has also been the source of controversy in regards to the net energy produced. One study has estimated that there is a net energy loss when converting biomass to ethanol [7]. All other studies done have estimated a 1.2-1.7:1 net energy gain for the current process [8-11]. Using a cellulosic feedstock, such as switchgrass, for biomass conversion greatly increases the net energy gain from the process. The return on energy invested for a cellulosic ethanol process is estimated to be ~4.75:1. The underlying problem with cellulosic ethanol is that it has never been demonstrated, on a large scale, to be economically competitive with other forms of energy. Cellulosic ethanol technology has to be improved to a point where it is possible to utilize the vast amount of raw cellulose material present in the world today to produce ethanol at a competitive price.

1.2) Overview of the Cellulosic Biomass to Ethanol Process

The most expensive part of the biomass conversion to ethanol process is the degradation of cellulose into fermentable sugars. Microorganisms are known to be very efficient in degrading cellulose into fermentable sugars using several enzymes termed cellulases and hemicellulases [12, 13]. Several properties of Lignocellulose make it difficult to hydrolyze. Lignocellulose is associated with hemicellulose, therefore to degrade cellulose you need to break it free of the hemicellulose. For the most efficient process it would be ideal if the hemicellulose could be utilized to produce ethanol along with the cellulose. Lignocellulose is also surrounded with a lignin seal which needs to be degraded. Finally, most cellulose has a crystalline
structure which can consist of six hydrogen bonds [14]. This leads to a highly ordered, tightly packed structure that is difficult to hydrolyze.

The current cellulosic biomass conversion process utilizes pretreatment of the Lignocellulose to remove the lignin seal, solubolize the hemicellulose, weaken the crystalline structure, and increase the surface area of cellulose available. The next step is to use enzymes (mostly from fungi, *Trichoderma reesei*) to hydrolyze the cellulose into fermentable sugars. Finally the sugars are converted to ethanol by yeast (*Saccharomyces cerevisiae*). Recently, it has become clear that by using microorganisms that have unique properties for cellulose degradation it is possible to convert cellulose into ethanol using less steps. Using the special talents of these organisms may greatly increase the efficiency of the process. One such organism that is a possible candidate is the bacterium *Clostridium thermocellum* [15-19].

### 1.3) *Clostridium thermocellum* as an Industrial Ethanol Producer

The cellulase system of *C. thermocellum* is the most widely studied in the bacterial world. As such, much is known about its properties that make it an ideal organism to utilize for biomass conversion to ethanol. *C. thermocellum* has several abilities that make it an ideal organism for fermentation of cellulose into ethanol. *C. thermocellum* is anaerobic, thermophilic, cellulolytic, and ethanogenic. All of these properties are advantages for *C. thermocellum* in an industrial setting. The bacterium grows at an optimum temperature of 60 °C and needs an oxygen free environment. The high temperature reduces cooling costs during the process, and also makes contamination
less of a problem. The higher temperature also makes ethanol removal and recovery easier than organisms grown at lower temperatures. The thermophilic nature also allows the enzymes produced by *C. thermocellum* to be very heat stable. The anaerobic nature makes expensive oxygen transfer unnecessary to grow the bacterium. While the thermophilic and anaerobic properties of *C. thermocellum* are good for an industrial setting, its most important aspects are its ability to be cellulolytic and ethanogenic. This allows for saccharification of cellulose and fermentation of the simple sugars produced into ethanol in a single step. The combination of what are usually two steps in the cellulose to ethanol process allows for a more efficient procedure. *C. thermocellum* breaks down cellulose into cellobiose and cellodextrins using cellulases. Cellobiose is a disaccharide consisting of two glucose molecules with a β-1,4 linkage. Once inside the cell, the cellobiose is further utilized and the final end products are ethanol, acetic acid, lactic acid, hydrogen, and carbon dioxide [20]. The cellodextrins produced can also be up-taken by the cell, further degraded, and then converted into the end products listed above [21]. The capacity for *C. thermocellum* to degrade such a tough substrate like crystalline cellulose stems from the complex system of enzymes it employs to accomplish the task. All of the characteristics of *C. thermocellum* listed above make it a great candidate to be used in the biomass to ethanol conversion process.

1.4) **The Cellulase System of *Clostridium thermocellum***

The cellulase system of *Clostridium thermocellum* consists of a vast number of genes and proteins. A bioinformatic search of the genome shows that well over 100 genes
are likely to be involved in the cellulase system to some degree [22]. The main component of the cellulase system is an extracellular, multi-protein complex termed the cellulosome (Fig. 1.1) [23]. The cellulosome consists of a large number (over 70) of proteins, which may include many different types of cellulases and hemicellulases. The main types of enzymes that are responsible for most of the cellulase activity are endoglucanases and exoglucanases. Endoglucanases break the internal bonds of cellulose creating new cellulose polysaccharide chains. Exoglucanases cleave from the end of these exposed chains and can progressively work down the cellulose fiber. Other enzymes that can be included in the cellulosome are xylanases, lichenases, and mannanases. Most studies to this point have focused on the proteins that are important to the cellulosome. At the center of the cellulosome is a 250 kDa non-catalytic polypeptide, CipA, which binds to cellulose and also serves as a scaffold for the catalytic subunits [24-26]. CipA contains a series of receptor domains for the catalytic subunits to bind. Nine highly homologous domains, termed cohesin, anchor the catalytic subunits to CipA. Each catalytic subunit that is part of the cellulosome also contains a highly conserved domain that interacts with cohesion to bind the subunits to the scaffolding. This domain is called the dockerin and is almost exclusively found at the C-terminus of each cellulosomal subunit.

The most important cellulosomal subunit found to date is CelS [27]. CelS is the major catalytic component of the cellulosome. It is an exoglucanase [28-32] with a molecular weight of 80,670 Daltons. When CelS was first discovered it became the first member of a new cellulase family [33]. It is classified as a family 48 glycosyl
The family 48 GH’s have been shown to be important components in other cellulase systems after its discovery in *C. thermocellum*. The cellulosome has been shown to account for 95% of the endoglucanase activity associated with the entire cellulase system [34]. Some of the cloned endoglucanase genes that are cellulosomal include: *celA* [35], *celB* [36], *celD* [37], *celF* [38], *celG* [39], *celN* [40], *celR* [22], *celQ* [41], and *celT* [42]. There are several more predicted cellulosomal endoglucanases in the genome sequence. Cloned and characterized cellulosomal xylanases include *xynD*, *xynC* [43], *xynA* [44], and *xynB* [44]. There are putative glycosidases, carbohydrate esterases, pectinases, proteases, protease inhibitors, and a xyloglucan hydrolase in the genome sequence that are all putative cellulosomal components.

In addition to the cellulosome subunits, the organism secretes catalytic enzymes that are not part of the cellulosome. These enzymes usually contain Carbohydrate Binding Modules (CBM) of their own and some of them have Surface Layer Homology (SLH) domains that facilitate the protein binding to the cell surface. Examples of these enzymes, which have been cloned and characterized, include *licA*, *celC* [45], *celI* [40, 46, 47], and *celX* [48]. Other putative free cellulases and hemicellulases are found in the genome sequence but remain uncharacterized.

1.5) **Carbon Source Regulation of Cellulase Synthesis in *C. thermocellum***

The different strains of *C. thermocellum* have been shown to utilize several different sugars for growth. The literature contains several contradictions on the exact carbon
sources that can be used as growth substrates for specific strains. One group reports that strain 651 can grow on cellulose, hemicellulose, cellobiose, and xylose but does not grow using fructose, glucose, sorbitol, or mannitol [49]. However, other groups found that strain 651 did grow on mannitol, glucose, xylose, and fructose [50, 51]. Strain LQ8 did not grow on glucose or xylose [52]. Strain YS reportedly only grows on cellulose and cellobiose [53]. *C. thermocellum* ATCC 27405, the subject of all research in this dissertation, grows well on cellobiose and cellulose. Growth of this strain using glucose, fructose, and sorbitol can be seen but only after a long lag time of greater than 100 hours [54]. The strain does not seem to grow on xylose or xylan [55], despite the existence of several xylanase genes in the genome. Growth on laminarin, lichenan, and laminaribiose has been shown to be very robust [56, 57].

Despite the many studies that have been conducted on the cellulase system of the organism and its growth substrates, very little is known about how the bacterium controls the expression of the different components of the system. One phenomenon of regulation that has been widely recognized is that growth conditions affect the cellulase activity of *C. thermocellum* [54, 58-60]. The profile of cellulosomal components is also affected by growth conditions [61]. When the organism is grown using cellulose as a carbon source, the cellulase activity is much higher compared to the bacterium grown using cellobiose [59, 61-63]. Total endoglucanase activity seems to be constitutive and is not subjected to carbon source regulation. The true cellulase activity is dependent upon the carbon source available [54]. Cellulase activity was detected when *C. thermocellum* ATCC 27405 was grown with cellulose, cellobiose,
glucose, fructose, or sorbitol. The specific amount of cellulase activity measured varied over 200-fold between the different substrates. As mentioned above, glucose, fructose, and sorbitol have a lag time of over 100 hours before growth is measurable. Interestingly, fructose and sorbitol produced high amounts of cellulase activity after the lag time, while glucose did not. The high cellulase activity produced by growth on fructose was quickly repressed by adding cellobiose into the culture. As the literature shows, the global phenomenon associated with carbon source regulating production of cellulases is well established.

The expression of specific cellulase genes under different growth conditions has been the topic of recent studies. Transcripts of endoglucanase genes celA, celD, and celF were subjected to Northern blot analysis with the bacterium grown on cellobiose [64]. Transcripts could not be detected until the late exponential phase of growth. It is postulated that cellobiose concentration in the cell drops later in the growth cycle, which relieves repression of the three genes. Current studies have shown that the expression of celS, cipA, olpB, orf2A, celB, celG, and celD are regulated by growth rate as revealed by chemostat experiments under cellobiose or nitrogen limitation [53, 65, 66]. Growth rate does not affect the synthesis of sdbA or xynC.

Immunoassays have shown that the two most abundant components of the cellulosome, CipA and CelS, are subject to carbon source regulation [67, 68]. Utilizing enzyme-linked immunosorbent assay (ELISA) with an antibody raised against a segment of CipA, it was reported that in batch cultures CipA production was
nine-fold higher in cellulose cultures than in cellobiose cultures. Using the same technique, but with continuous cultures, it was shown that CipA production was 1.3- to 2.4-fold greater on cellulose at the same dilution rates as on cellobiose.

1.6) Clostridia Gene Clusters and Negative Regulation of Cellulase Genes

This section is based on the following publication:

Arnold L. Demain, Michael Newcomb, and J. H. David Wu
Cellulase, Clostridia, and Ethanol

Despite the fact that cellulase and xylanase genes of *C. thermocellum* are mostly scattered over the chromosome [69], several gene clusters have been found, suggesting the existence of operons as units of gene regulation (Fig. 1.2). The first two operons found are in the *cipA* cluster as two tandem operons including *cipA* + *olpB* and *orf2* + *olpA*. These are all structure genes, as *cipA* is the cellulosomal scaffolding, *olpB*, *orf2*, and *olpA* all help cellulosomes or enzymes bind to the bacterial cell. More recently, a non-cellulosomal endoglucanase, *Cell*, was shown to be clustered with cellulosomal cellulase *CelN* and a possible cellulosomal structural component *CseP* (cellulosomal element protein) [40]. *cseP* codes for a type I dockerin-containing protein and is in the opposite strand and likely transcribed independently from *cell* and *celN*. The recombinant CseP displayed no detectable enzymatic activities. Although the exact function of CseP is unknown, western blot analysis revealed that the cellulosome contains a substantial amount of this protein. Another gene cluster consists of the genes encoding the non-cellulosomal endoglucanase *CelC*, a LacI-like protein, the lichenan degrading enzyme LicA, and a
gene encoding a potential membrane protein with eight putative transmembrane regions [56]. We [70] reported that this gene cluster is immediately upstream from the genes encoding ManB [71] and CelT [42]. The cluster therefore has a total of six genes [70]. The gene cluster is preceded by a putative transposase gene and orf2. All of these genes are on the same strand. As will be described below, our data indicates that the LacI-like protein regulates the expression of the gene cluster by binding to the promoter region. A third gene cluster consists of celA, chiA and orfZ [22]. CelA is a non-cellulosomal endoglucanase. ChiA (family 18) is the only chitinase in the cellulose. OrfZ has a domain of unknown function and a dockerin. Finally, a five-gene cluster consisting of genes encoding five dockerin-containing proteins was found in analyzing the C. thermocellum genome [72]. Interestingly, none of these five genes have been previously cloned. The first gene product is a putative family 2 glycosyl hydrolase. The second encoded protein is a putative family 43 glycosyl hydrolase, which contains a family 4 CBD. The third encoded protein contains a domain of unknown function and a family 6 CBD. The fourth encoded protein is a putative esterase with a family 6 CBD. The last gene product is a putative family 5 cellulase with a family 4 CBD and a fibronectin-like domain. Gene clusters presumably resulted from gene duplication include the gene pairs of celK-cbhA [73] and xynB-xynA [44]. As the genome sequencing of C. thermocellum is being completed, more cellulase and hemicellulase gene clusters may be detected.

In mesophilic clostridia, large cellulase gene clusters have been found (Fig. 1.3). Interestingly, these gene clusters all start with a scaffoldin gene, followed by a gene
encoding a family 48 enzyme (exoglucanase working from the reducing end), one to two endoglucanase genes, and another gene encoding an exoglucanase working from the non-reducing end [74]. Similarities among these gene clusters indicate that these clostridia are closely related. In *C. cellulovorans*, the nine-gene cluster contains the genes encoding the scaffolding protein CbpA, the exoglucanase ExgS, several family 9 endoglucanases, the mannanase ManA, and the hydrophobic protein HbpA containing a surface layer homology domain and a hydrophobic (or cohesin) domain [75, 76]. The sequence of the nine-gene cluster is *cbpA-exgS-engH-engK-hbpA-engL-manA-engM-engN* and spans about 22 kb in length. Several possible transcription terminators have been found between some of the genes. Northern hybridization revealed that the cellulosomal *cbpA* gene cluster is transcribed as polycistronic mRNAs of 8 and 12 kb, respectively. The 8-kb mRNA coded for CbpA and ExgS, and the 12-kb mRNA coded for CbpA, ExgS, EngH, and EngK. [77]. Conversely, Northern hybridization revealed that *manA* is transcribed as a monocistronic messenger.

In *C. cellulolyticum*, a similar gene cluster has been reported. The sequence of the 10-gene, 20 kb cluster is *cipC-celF-celC-celG-celE-orfXp-celH-celJ-manK-celM*. Results of northern blot analysis revealed that *celC* and *celG* are expressed into a polycistronic transcriptional unit which possibly includes a third gene [78]. The genome sequence of *C. acetobutylicum* revealed at least eleven genes encoding dockerin- or cohesin-containing proteins [79]. Ten of these genes form a large cluster, similar to the *C. cellulovorans* and *C. cellulolyticum* gene clusters, with the
sequence of \textit{cipA-celA-celB-celC-orfXp-celD-celE-celF-manG-celH} [80]. Existence of these sequences suggests that the bacterium is capable of producing a cellulosome-like structure and such a structure has been identified as mentioned above. However, the bacterial strain is incapable of hydrolyzing amorphous or crystalline cellulose although it hydrolyzes CMC. In addition, most of the xylanase genes in this bacterium are located in a predicted operon [79]. In \textit{C. josui}, a gene cluster containing at least four genes has been reported [81]. The scaffolding gene \textit{cipA} is the first member of this cluster, followed by \textit{celD}, \textit{celB}, and \textit{celE}.

\textbf{1.7) Molecular Mechanisms of Cellulase Regulation in Other Organisms}

There are no known molecular mechanisms that control cellulase biosynthesis in \textit{C. thermocellum} found in the literature (outside of our GlyR3 studies, [57, 70]). However, in the fungus \textit{Trichoderma reesei}, a series of activators and repressors have been found to control the level of cellulase and xylanase. These include the activator ACE II, and negative elements, ACEI and CRE1 [82, 83]. The ruminant anaerobe \textit{Prevotella bryantii} uses a multidomain protein, XynR, to regulate xylanase gene expression [84]. This protein consists of a histidine kinase domain and a helix-turn-helix AraC-like DNA binding motif and acts as a positive regulator for the \textit{xynABD} operon. The soil bacterium \textit{Thermobifida fusca} has six known cellulase genes, \textit{celA}, \textit{celB}, \textit{celC}, \textit{celD}, \textit{celE}, and \textit{celF}. A protein, CelR, that binds to a 14 bp inverted repeat found in the promoter region of these genes functions as a repressor [85, 86]. The DNA-binding activity of CelR is deactivated by cellobiose but not other sugars. Recently, it has been shown that laminaribiose may be involved in the regulation of
the cellulase genes in this organism [87]. In *Clostridium celloborans*, Northern blot and Reverse Transcriptase Polymerase Chain Reaction mapping studies have revealed coordinated expression of the cellulase and hemicellulase genes for different carbon sources [88].

1.8) The Need for Knowledge of Molecular Mechanisms for Cellulase Regulation in *C. thermocellum*

In order to engineer *C. thermocellum* to more efficiently degrade cellulose and convert it to ethanol, the biological processes and pathways that the organism utilizes to accomplish this task must be well understood. The bottleneck for the conversion process is the degradation of cellulose into fermentable sugars. By understanding which genes and proteins are active under certain culture conditions it will be possible to “turn on” or “turn off” specific enzymes to tailor the bacterium to degrade certain substrates more efficiently. Furthermore, identifying and characterizing transcription factors that are associated with the cellulase system will enhance the ability to manipulate these regulators in the future. Once a genetic system of transformation becomes available it will be possible to knock-out negative regulators and over-express positive regulators, leading to an over-active, enhanced cellulase system. The bulk of this thesis is devoted to determining the expression and regulation of specific cellulase system components in *C. thermocellum*.
1.9) Synopsis of Results

The following work uses RT-qPCR to show that the \textit{celC} gene cluster is up-regulated when the organism is grown using a $\beta$-1,3 linked sugar. In addition the \textit{celC} gene cluster functions as an operon, with \textit{celC-licA-glyR3} all co-transcribed and regulated together. The very first transcriptional regulator of cellulase genes in \textit{C. thermocellum}, GlyR3, is identified and characterized. Laminaribiose is shown to be the inducer associated with GlyR3. Finally, an uncharacterized cellulase gene cluster is described and its mRNA expression pattern determined when the organism is grown on cellulose and cellobiose. The newly discovered cluster is the largest set of putative cellulosomal genes in the genome and the results indicate they may be important for growth on cellulose.
Fig. 1.1) An illustration of the cellulosome of *Clostridium thermocellum*.
Fig 1.2) Cellulase system gene clusters in *C. thermocellum*. Figure taken from Demain, Newcomb and Wu (2005), reference [72].
**Fig 1.3**) Gene clusters in Mesophilic Clostridia related to the cellulase system of each organism. Figure taken from Demain, Newcomb and Wu (2005), reference [72].
1.10) Chapter 1 References


Chapter 2

The celC Gene Cluster and its Transcript Expression Profile

The text of this chapter is based on the following publication.

Michael Newcomb and J. H. David Wu
Co-transcription and Expression of the celC Gene Cluster in Clostridium thermocellum
Journal of Bacteriology (submitted).

2.1) Introduction

As described in Chapter One, gene clusters are utilized in the control of gene expression in many different bacteria. It would stand to reason that C. thermocellum may use gene clusters to at least partially regulate the expression of its cellulase system. The genome of the organism was searched for gene clusters that contained known or putative members of the cellulase system. One such gene cluster (the celC gene cluster) was discovered that contained characterized members of the cellulase system. The celC gene cluster consisted of: celC [1, 2], an uncharacterized gene (now known as glyR3), licA [3], an uncharacterized gene, manB [4], and celT [5]. In order to understand whether this gene cluster functions as an operon, several experiments were conducted to examine the transcription of the genes in the cluster.
Although, two pairs of co-transcribed genes exist in the *cipA* gene cluster *cipA-olpB-orf2p-olpA*, including *cipA-olpB* and *orf2p-olpA* [6], no co-transcription of glycosyl hydrolase genes in *C. thermocellum* has been reported. In this work the *celC* gene cluster transcripts were mapped using Northern blot and the *celC* transcription initiation site was determined by primer extension. Using Real Time Quantitative Polymerase Chain Reaction (RT-qPCR), we are able to show the transcript expression pattern of all the genes in the cluster when the bacterium is grown on three separate substrates. The results showed that *celC-glyR3-licA* are co-transcribed with one major transcription initiation site 8 bp upstream from the putative ribosome binding site. Transcript mapping further revealed that *manB* [4] and *celT* [5], two cellulosomal genes immediately downstream from the *celC-glyR3-licA* cluster, are also co-transcribed. Primer extension showed that *orf4* has two transcription initiation sites 130 bp and 138 bp upstream of the start codon and *manB* has one initiation site, 233 bp upstream of its start codon. Finally, qPCR proves that when the cells are grown on laminarin, *celC*, *glyR3*, and *licA* show an expression peak at the late exponential phase of growth that is at least 2.5 fold greater than when the cells are grown using cellobiose or cellulose. Growth on cellulose or on cellobiose did not produce such high expression of the genes.

2.2) Results

The *celC* gene cluster consists of six genes.

We searched the genome sequence of *C. thermocellum* for potential genes encoding glycosyl hydrolases and found the *celC* cluster is larger than reported [3]. The cluster
**Transcript Mapping of the celC gene cluster.**

To determine if the members of the celC gene cluster are co-transcribed, we mapped the mRNA species transcribed from the celC gene cluster using RT- (reverse-transcriptase-) PCR and Northern Blot analysis. Probes for celC, glyR3, and licA all hybridized to the Northern blot at a size just over 6 kb (Fig. 2.2). The expected size of celC, glyR3, and licA, if they were all co-transcribed, would be 6,261 bp. When grown on laminarin, the Northern blot indicates that celC, glyR3, and licA are all transcribed as one unit. Northern blot was attempted using probes for orf4, manB, and celT but was unsuccessful due to low expression levels of the transcript(s) (data not shown). In order to determine the transcript pattern that is used by the bacterium for these genes, we mapped the mRNA species transcribed from the celC gene cluster using Reverse-Transcriptase PCR. In this analysis, the PCR primers, each representing a sequence from one of the two adjacent genes, were used as previously described [9]. To determine if a pair of adjacent genes are co-transcribed, the primers were designed to amplify the 3’ end of the upstream gene, the intergenic sequence, and 5’ end of the downstream gene. If a pair of genes is co-transcribed, the PCR product with the expected size would be obtained. Positive controls using C. thermocellum genomic DNA as the template were included to verify that the primers
would correctly amplify the DNA segment (Fig. 2.3, Lanes 2, 5, and 8). Negative controls using the total RNA as the template but omitting reverse transcriptase (RT) in the RT reaction were also included to determine the background level (Fig. 2.3, Lanes 4, 7, and 10). orf4 appears to be transcribed alone, as it was not possible to amplify the upstream bridging fragment (290 bp; Fig. 2.3, Lane 3) and downstream bridging fragment (445 bp; Fig. 2.3, Lane 6) above background levels (lanes 4 and 7, respectively) despite repeated attempts. For positive controls, PCR products of correct sizes were obtained (lanes 2 and 5, respectively). Finally, it was possible to amplify the bridging region between manB and celT (450 bp; Fig. 2.3, Lane 9), indicating that these two genes are co-transcribed. Therefore, under the growth conditions used in this work, the celC gene cluster, with its six members, is transcribed into a celC-glyR3-licA poly-cistronic mRNA, an orf4 mono-cistronic mRNA, and a manB-celT bi-cistronic mRNA.

**Primer Extension of celC, orf4, and manB.**

We developed a non-radioisotope technique for primer extension analysis to determine the transcription initiation site for the celC-glyR3-licA genes. In this method, total RNA was isolated at the exponential growth phase as mentioned above. A reverse primer corresponding to the nucleotide positions 14 to 33 of celC was end-labeled with fluorescein (Table 2.1) and used for primer extension employing MMLV reverse transcriptase (Invitrogen) and the manufacturer’s protocol. The resulting cDNA was analyzed by capillary electrophoresis using a Model 3100 Genetic Analyzer (Applied Biosystems). As shown in Fig. 2.4A, a 53 bp extended cDNA fragment was obtained, corresponding to the transcription initiation site located at the
-20 nucleotide position of celC (Fig. 2.4B). The transcription initiation site is 8 bp upstream of the putative Shine-Dalgarno sequence. The technique was applied (Fig. 2.5A, B) to discover the transcription initiation sites for orf1 and manB. orf4 was shown to have two starts sites that were 130 bp and 138 bp upstream from the start codon. The single transcription initiation start for manB was found to be 233 bp upstream of the start codon sequence.

**Expression of the celC gene cluster using different carbon sources.**

*C. thermocellum* was grown on three different carbon sources and growth curves were constructed for each set of cultures using intracellular protein (Fig. 2.6). At different growth stages the RNA was harvested from the cultures and subjected to RT-qPCR. A comparison was made between the expression of all the genes in the celC gene cluster at each growth phase for all three carbon sources (Fig. 2.7A, B, C). The results show that at late exponential growth, celC, glyR3, and licA all have a spike in expression when the cultures were grown on laminarin (Fig. 2.7A, B, C). The spike in expression is not seen when the cells are grown on cellobiose or on cellulose. Once the cells enter early stationary phase, the spike disappears. As will be discussed later, these results fit nicely with the mechanism we have proposed for GlyR3 regulation in the celC gene cluster (Chapter 3 of this thesis) [10]. The expression pattern of orf4, manB, and celT do not have a spike when grown on any of the substrates (Fig. 2.8, Fig. 2.9A, B). It is possible that a different carbon source is needed in order to get maximum expression of orf4, manB, and celT.
2.3) Discussion

*C.thermocellum* produces a highly active cellulase system that has over 100 different components. At this time, only one molecular mechanism has been proven for the regulation of cellulase system components. GlyR3 has been shown to bind to the *celC* promoter region, inhibiting transcription of *celC* until an inducer (laminaribiose) is present [10]. One of the goals of this study was to determine what other genes in the *celC* gene cluster may be regulated by GlyR3. By measuring the transcript expression level of the genes in the *celC* gene cluster we can determine if growth on cellulose or cellobiose has any effect on the regulation of the cellulase genes in this cluster.

Co-transcription of *celC-glyR3-licA* and *manA-celT*, respectively, as demonstrated in this work, represents the first two cases of co-transcription of glycosyl hydrolase genes discovered in *C. thermocellum*. Co-transcription of genes is a method of coordinated gene regulation in prokaryotes. In this regard, we have previously found that GlyR3 serves as a negative regulator that binds to a palindromic sequence in the *celC* promoter region [10]. The binding is inhibited by laminaribiose, a β-1,3 glucose dimer, leading to the induction of *celC-glyR3-licA*. Since CelC [1, 2] and LicA [3] are both active on β-1,3 glucans, the results indicate that *C. thermocellum* may coordinate the expression of glycosyl hydrolases in response to the availability of a particular biomass substrate. Despite the insolubility of biomass substrate, this regulation can be mediated by a soluble sugar.
celC, glyR3, and licA are all co-transcribed (Fig. 2.2), leading to their co-regulation. A basal level of expression is present for celC and for licA, as shown by the transcript present at the exponential stage of growth on all substrates (Fig. 2.7A, B, C). When the bacterium encounters a substrate that contains β-1,3 linkages, some degree of degradation occurs using at least the enzyme products of these genes. The degradation produces laminaribiose, which is able to enter the cell. GlyR3, that is inhibiting transcription of the celC-glyR3-licA transcript binds to the laminaribiose, undergoes a conformation change, and is released from binding to the celC promoter region. Transcription of celC-glyR3-licA is now allowed to proceed uninhibited, as evidenced by the expression level measured at late exponential growth for all genes in the transcript when grown on laminarin (Fig. 2.7A, B, C). By having GlyR3 regulate its own expression, a feedback loop is created. Once the amount of GlyR3 produced overwhelms the amount of laminaribiose present in the cell, transcription of celC-glyR3-licA is severely down-regulated, as shown by the transcript levels of celC-glyR3-licA at early stationary phase (Fig. 2.7A, B, C). The spike in expression of celC-glyR3-licA only occurs when the organism is grown with a substrate containing β-1,3 linkages. The same expression pattern has been observed when lichenan (β-1,3 and β-1,4 linkages) or laminaribiose have been used as the carbon source (data not shown).

It is noteworthy that, in this cluster, the cellulosomal genes (celC and licA) and the non-cellulosomal genes (manB and celT) belong to two different transcription units. The RT-qPCR analysis for celC-glyR3-licA supports the Northern blot data as each
gene from the unit shows the same expression pattern throughout the growth phases. It would be interesting to see if the transcripts change in configuration when different carbon sources are used as growth substrates. The discovery of a regulation mechanism for manB-celT would also prove to be useful in helping to determine if cellulosomal and non-cellulosomal transcripts have similar molecular control mechanisms. The unknown transmembrane orf4 does not seem to fit in with the other genes in the cluster. It is transcribed alone and has a low level of expression on all of the substrates studied. It will be interesting to see if orf4 is related to cellulose degradation, either as a part of the transport system or in some other capacity.

Our results confirm that $\beta$-1,3 linked glucan is able to up-regulate the expression of the RNA transcripts for celC, glyR3, and licA. Since GlyR3 has been shown to control transcription of celC, it is likely that all three genes in this transcript are controlled by the regulatory protein. The celC operon is the first regulatory unit with a proven molecular mechanism. The bacterium tailors the expression of the celC operon to fit the substrates present in its environment. As the study of C. thermocellum continues, other co-transcription of cellulase genes in the organism will definitely be discovered along with molecular mechanisms of cellulase system regulation.

2.4) Materials and Methods

**Bacterial Strains.** The source of all genomic DNA and RNA was *C. thermocellum* ATCC 27405.

**Culture Conditions.** Hungate tubes were used to culture batches of *C. thermocellum* in chemically-defined MJ medium [7]. The medium contained 0.5% of the
appropriate carbon source (cellobiose, cotton, or laminarin). Seed cultures were all grown for 12 hours on cellobiose. All cultures were incubated at 60° C.

**Measuring Pellet Protein for Growth Curve.** To measure the growth curve for cellobiose, laminarin, and cotton cells the pelleted protein per milliliter of culture was estimated. The cells were centrifuged at 12,000 x g for 15 minutes (cotton cultures were filtered through cheesecloth, the remaining cotton fibers were washed with 0.85% NaCl solution). The supernatant was removed from the sample and the pellet was washed and spun down twice with 0.85% NaCl solution. This was done to minimize extracellular protein in the sample. Before the last wash was centrifuged, 1 ml of the resuspended pellet was taken for cellular protein analysis, the rest of the resuspended pellet was spun down for RNA isolation. The removed 1 ml was centrifuged and the final pellet was resuspended with 1 ml of 0.2 N NaOH and heated in boiling water for 15 minutes. After boiling, the sample was centrifuged again at 12,000 x g for 15 minutes. Finally, 1 ml of 0.2 N HCl was added to neutralize the supernatant. The amount of protein in the supernatant was now measured using the Bradford [8] reagent (Bio-Rad). The supernatant was diluted so that spectrophotometer readings fell within the linear range of a bovine serum albumin (Sigma) standardized curve.

**RNA Isolation.** *C. thermocellum* cell pellets were obtained as outlined above. The Trizol (Invitrogen) method of RNA isolation was used to obtain total RNA. The RNA was digested with DNase I (ABgene) and quantified using a spectrophotometer at 260/280 nm.
**Reverse Transcription.** A standard reverse transcription reaction protocol was followed using MMLV reverse transcriptase (Invitrogen).

**Preparation of Northern Blot Probes.** Northern blot probes were created using PCR with the primers listed in table I. A standard protocol was followed for ABgene Thermostart Taq Polymerase. The annealing temperature was set to 58 °C and an extension time of 20 seconds were used for each probe. After PCR, the reaction was purified using the Promega Wizard SV Gel and PCR Clean-Up System.

**Northern Blot.** *C. thermocellum* was grown using cotton as a carbon source for 60 hours. Total RNA was isolated as described above. The Ambion NorthernMax-Gly kit was used to perform the Northern blot. The manufacturers’ protocol was followed with the subsequent conditions. 20 μg of RNA was loaded into a 0.8% agarose gel. An RNA molecular weight marker (Novagen) was also run on the gel to estimate transcript size. A downward capillary blot was used to transfer RNA to Biodyne B (Pall) positively charged membrane. The RNA marker was stained using 0.2% Methylene Blue to verify the success of the blotting. The blot was mixed with ULTRAhyb buffer from the NorthernMax kit and allowed to prehybridize for 45 minutes at 42° C. The probes were prepared by using a PCR reaction with primers from table I. dNTPs were used in the reaction that had the dATP labeled with biotin in a 5.25:1 ratio (New England Biolabs). Probes were heated to 95° C for 10 minutes then added to the prehybridization solution to a final concentration of 2 ng/ml. The hybridization was allowed to continue for 24 hours. After hybridization, the membrane was washed with low stringency buffer for 10 minutes at room temperature and washed twice with high stringency buffer at 42° C for 15 minutes.
each. To develop the blot, the standard manufacturers’ protocol for the Chemiluminescent Nucleic Acid Detection Module (Pierce) was followed.

**Reverse Transcription Mapping.** To determine if *orf1*, *manB*, and *celT* of the *celC* gene cluster are co-transcribed, we mapped the mRNA species transcribed from the *celC* gene cluster using RT- (reverse-transcriptase-) PCR. In this analysis, the PCR primers, each representing a sequence from one of the two adjacent genes, were used as previously described [9].

**Primer Extension.** A primer was created (Table 2.1) and end-labeled with fluorescein (Invitrogen). A standard reverse transcription reaction was carried out for a gene specific primer using MMLV Reverse Transcriptase (Invitrogen). The University of Rochester Functional Genomics Center used capillary electrophoresis to quantify the size of the DNA fragment created by the reaction.

**Quantitative Real-Time PCR.** Gene expression analysis was carried out using iQ Syber Green Supermix (Bio-Rad) on a Bio-Rad iCycler with gene specific primers. All primer sets were tested to make sure data falls within the linear range of quantification, and that amplification efficiency was greater than 90%. The two-step PCR protocol used consisted of activation of the polymerase followed by 40 cycles of 95 °C for 15 seconds and 60 °C for 30 seconds. All sample/primer combinations were analyzed in triplicate. The 16S ribosomal RNA was used to normalize the experiment for sample-to-sample variation; expression data for genes of interest was reported relative to 16S expression.

**Bioinformatic Analysis.** Most of the bioinformatic data was accumulated from the Oak Ridge National Laboratory (ORNL) website devoted to *C. thermocellum*
Tab-delimited files were imported into Microsoft Excel and sorted according to their top BLAST hit. BLAST results that contained similarity to cellulases were then sorted according to their place in the genome. Cellulase gene clusters were found in this manner. To confirm the results, the entire set of gene BLAST results from the ORNL website were reviewed in sequential order based on gene number.
Fig. 2.1) A schematic of the celC gene cluster.
Fig. 2.2) Northern blot results of celC, glyR3, and licA. Probes for all three genes hybridized to the same size transcript, suggesting they are all co-transcribed. The expected size of celC, glyR3 and licA if they were all co-transcribed is 6,261 bp. Northern blot of orf4, manB, and celT was unsuccessful due to low expression levels (data not shown).
Fig. 2.3) Reverse transcriptase mapping of orf4, manB, and celT. Lane 1, Molecular Weight Marker; Lanes 2, 5, and 8, positive controls of PCR using genomic DNA as the template; lanes 4, 7, 10, negative controls omitting the reverse transcriptase (RT) in the RT reaction; Lanes 3 and 6 the negative results indicating the lack of co-transcription of licA-orf1 and orf1-manB, respectively; Lane 9 the 450 bp RT-PCR product indicating co-transcription of manB-celT.
Fig. 2.4) Transcription initiation site as revealed by primer extension using a primer corresponding to the celC sequence. (A) Capillary electrophoresis of the extended cDNA fragment showing a 53 bp fragment, matching up to the transcription initiation site; (B) DNA sequence of the celC promoter region. The transcription initiation site matching the 53 bp extended cDNA fragment is indicated with an arrow. The GlyR3 binding site [10] is bolded.
Fig. 2.5) Transcription initiation site as revealed by primer extension using a primer corresponding to the orf4 and manB sequence. (A) Capillary electrophoresis of the extended cDNA fragment showing 170 bp and 178 bp fragments, matching up to the transcription initiation sites for orf4, the first 40 bp of each fragment were the start of the coding region for orf4; (B) Capillary electrophoresis of the extended cDNA fragment showing a 278 bp fragment, matching up to the transcription initiation sites for manB. The first 45 bp of the fragment correspond to the first part of the coding region for manB.
Fig. 2.6) A growth curve showing intracellular protein for cultures grown using cellulose, celllobiose, and laminarin. The numbers above points represent the growth stage for that particular carbon source. 1, Exponential Growth; 2, Late Exponential Growth; 3, Early Stationary Phase. Errors bars represent the standard deviation of triplicate experiments.
Fig. 2.7) The expression patterns for celC, glyR3 and licA at different growth phases using different substrates. Transcript level is shown relative to 16S expression and error bars represent standard deviation between triplicate samples. (A) celC (B) glyR3 (C) licA
Fig. 2.8) The expression pattern for orf4 at different growth phases using different substrates. Transcript level is shown relative to 16S expression and error bars represent standard deviation between triplicate samples.
Fig. 2.9) The expression patterns for *manB* and *celT* at different growth phases using different substrates. Transcript level is shown relative to 16S expression and error bars represent standard deviation between triplicate samples. (A) *manB* (B) *celT*
<table>
<thead>
<tr>
<th>No.</th>
<th>Primer Sequence</th>
<th>Use of Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F: <em>celC</em>-ATGGTGAGTTTTAAAAGGAGGTA</td>
<td>Northern Blot Probe, Biotin Labeled</td>
</tr>
<tr>
<td>2</td>
<td>R: <em>celC</em>-AATATAAGAAGGCCATCTTCTT</td>
<td>Biotin Labeled</td>
</tr>
<tr>
<td>3</td>
<td>F: <em>glyR3</em>-ATGACCAGTGGAAGAAATAGCAA</td>
<td>Northern Blot Probe, Biotin Labeled</td>
</tr>
<tr>
<td>4</td>
<td>R: <em>glyR3</em>-TCAGATATGTCATAATCCAA ACC</td>
<td>Biotin Labeled</td>
</tr>
<tr>
<td>5</td>
<td>F: <em>licA</em>-ATGTATAAAAAGATTATTGTCGTC</td>
<td>Northern Blot Probe</td>
</tr>
<tr>
<td>6</td>
<td>R: <em>licA</em>-AAATTCGGCTCTTGTATATTA</td>
<td>Biotin Labeled</td>
</tr>
<tr>
<td>7</td>
<td>F: <em>licA</em>-orf1-F-CAACTACTGTCTATCTTGTGTA</td>
<td>RT-PCR mapping</td>
</tr>
<tr>
<td>8</td>
<td>R: <em>licA</em>-orf1-R-AAGAGCCATCAGGCAAGGCC</td>
<td>RT-PCR mapping</td>
</tr>
<tr>
<td>9</td>
<td>F: <em>orf1</em>-manB-F-ATGAGTATTTCCGACAAATATGC</td>
<td>RT-PCR mapping</td>
</tr>
<tr>
<td>10</td>
<td>R: <em>orf1</em>-manB-R-GCAAGCCTCAATGCCAGGAAAA</td>
<td>RT-PCR mapping</td>
</tr>
<tr>
<td>11</td>
<td>F: <em>manB</em>-celT-F-CGCATAATAACGGAGTTTCCC</td>
<td>RT-PCR mapping</td>
</tr>
<tr>
<td>12</td>
<td>R: <em>manB</em>-celT-R-CAATAACCAGCAGCAGAAA</td>
<td>Primer extension, fluoroscein labeled</td>
</tr>
<tr>
<td>13</td>
<td>Primer extension <em>celC</em>-GCCTAAATTTATACCTGCTT</td>
<td>Primer extension, fluoroscein labeled</td>
</tr>
<tr>
<td>14</td>
<td>Primer Extension <em>orf1</em>-AACAGCCGCTATAAGCT</td>
<td>Primer extension, fluoroscein labeled</td>
</tr>
<tr>
<td>15</td>
<td>Primer Extension <em>manB</em>-AAGCATCAATGCGCCAGGAAA</td>
<td>Primer extension, fluoroscein labeled</td>
</tr>
<tr>
<td>16</td>
<td>F: qPCR 16S- AATTCGAAGCAACGCGAAGAAC</td>
<td>qPCR</td>
</tr>
<tr>
<td>17</td>
<td>R: qPCR 16S- GCGGGACTTAACCCAACATCTC</td>
<td>qPCR</td>
</tr>
<tr>
<td>18</td>
<td>F: qPCR <em>celC</em>-CGGGAACATATTGCCTTTGACAC</td>
<td>qPCR</td>
</tr>
<tr>
<td>19</td>
<td>R: qPCR <em>celC</em>-GGTGGGAATCATTGCCCTTGCATTG</td>
<td>qPCR</td>
</tr>
<tr>
<td>20</td>
<td>F: qPCR <em>glyR3</em>-GGGCATAGAACGCTATGAGGA</td>
<td>qPCR</td>
</tr>
<tr>
<td>21</td>
<td>R: qPCR <em>glyR3</em>-TATAGCCGCTGTCATCAAGGAA</td>
<td>qPCR</td>
</tr>
<tr>
<td>22</td>
<td>F: qPCR <em>licA</em>-TTGACCAAGGTCCGAACAGAA</td>
<td>qPCR</td>
</tr>
<tr>
<td>23</td>
<td>R: qPCR <em>licA</em>-TTGACCAAGGTCCGAACAGAA</td>
<td>qPCR</td>
</tr>
<tr>
<td>24</td>
<td>F: qPCR <em>orf1</em>-TCACTGCTTGTACCTCCTGTTTGT</td>
<td>qPCR</td>
</tr>
<tr>
<td>25</td>
<td>R: qPCR <em>orf1</em>-ACTGCTTGTACCTCCTGTTTGT</td>
<td>qPCR</td>
</tr>
<tr>
<td>26</td>
<td>F: qPCR <em>manB</em>-GGTATCCATAAAGGTGCCCAGA</td>
<td>qPCR</td>
</tr>
<tr>
<td>27</td>
<td>R: qPCR <em>manB</em>-ATTCAGGAGTGCCTTGTACC</td>
<td>qPCR</td>
</tr>
<tr>
<td>28</td>
<td>F: qPCR <em>celT</em>- TCTGGATTCCCAGAACCACCAAC</td>
<td>qPCR</td>
</tr>
<tr>
<td>29</td>
<td>R: qPCR <em>celT</em>-CTCAGGGAAACCACCAACTTCAC</td>
<td>qPCR</td>
</tr>
</tbody>
</table>

Table 2.1) Primers and their uses for Chapter 2 experiments.
2.5) Chapter 2 References


Chapter 3

Characterization of GlyR3

The text of this chapter is based on the following publication.

Michael Newcomb, Chun-Yu Chen, and J. H. David Wu
Induction of the celC operon of Clostridium thermocellum by laminaribiose

3.1) Introduction

The discovery of the celC gene cluster and the subsequent work outlined in Chapter 2 left many unanswered questions about the molecular mechanism of regulation for the operon. The putative transcriptional regulator located in the gene cluster was intriguing. Does it control the expression of celC, glyR3, and licA? The following chapter outlines the discovery and characterization of the regulation mechanism for the celC operon, answering many of the questions opened by the studies completed in
Chapter 2. First a brief summary as to where this important work fits into the literature.

*C. thermocellum* is an anaerobic, thermophilic, cellulolytic, and ethanogenic bacterium. It produces a cellulase system highly active on crystalline cellulose [1]. The extracellular cellulase components form an ordered protein complex termed the cellulosome [2]. In addition, many free glycosyl hydrolases are produced. The core of the cellulosome is CipA, a 250-kDa non-catalytic, scaffold protein [2-5]. CipA contains nine cohesin domains. Binding to the cohesin is mediated by the dockerin domain borne on the catalytic subunit [6-9]. CipA further contains a cellulose-binding module (CBM), which anchors the array of catalytic components to the cellulose surface [4, 10, 11].

Searching the genome sequence of *C. thermocellum* revealed more than 70 genes encoding dockerin-containing proteins, which are presumed to be the cellulosome components [12, 13]. Thus, including the genes encoding the cellulosome components, the scaffold proteins, and the free enzymes but without counting the regulatory and sugar-transport genes, there are likely more than 100 genes involved in biomass degradation by this bacterium. How the organism regulates the expression of such a large number of genes and proteins for biomass degradation is an intriguing question, yet so little is known. The issue is further complicated by the fact that
biomass is typically a solid substrate incapable of diffusing into the cell to regulate gene expression.

It has been demonstrated that production of the overall cellulase activity by \( C. \) \textit{thermocellum} is influenced by the carbon source [14-18]. But it is not clear how many individual genes are subject to carbon source regulation. Recent studies focus on a few specific cellulase components. The most abundant catalytic component of the cellulosome is an exoglucanase called CelS [10, 11, 19-24]. At the protein level, CelS [25, 26] and CipA [26] are upregulated by growth on cellulose as compared to cellobiose. In addition, growth rate has been shown to affect the expression of several cellulase genes. The expression of \textit{celS} is growth rate-dependent as revealed by chemostat experiments [25, 27]. Similarly, the transcript levels of \textit{cipA}, \textit{olpB}, \textit{orf2p}, \textit{celB}, \textit{celG}, and \textit{celD} are dependent on growth rate [28, 29]. In contrast, the expression of \textit{sdbA} and \textit{xynC} are independent from growth rate.

Despite these studies, molecular mechanisms governing the carbon-source regulation of the cellulase biosynthesis in this bacterium remain unidentified. Here we report the first cellulase gene transcriptional regulatory protein, GlyR3, of \( C. \) \textit{thermocellum}. GlyR3 specifically binds to an 18-bp near perfect palindrome in the promoter region of the non-cellulosomal cellulase gene \textit{celC}. GlyR3 is shown to repress \textit{celC} in an \textit{in vitro} transcription assay. The repression is reversed by laminaribiose, a \( \beta \)-1-3 linked glucose dimer, which inhibits GlyR3’s DNA-binding activity. The negative
regulation is the first cellulase regulation mechanism found in *C. thermocellum*. Since *celC*, *glyR3*, and *licA* are co-transcribed into a polycistronic mRNA (Chapter 2, Fig. 2.2), these three genes form a cellulase operon, the first demonstrated in *C. thermocellum*.

### 3.2) Results

**GlyR3 Structure.** GlyR3 (353 amino acids) is homologous to LacI (360 amino acids) of *Escherichia coli* (27% identical and 49% similar; Fig. 3.1). BLAST search [30] revealed two other *C. thermocellum* proteins homologous to LacI, GlyR1 (342 amino acids, 22% identical and 43% similar) and GlyR2 (345 amino acids, 29% identical and 49% similar). GlyR3 was particularly interesting because its gene is a member of the *celC* gene cluster and is co-transcribed with *celC* and *licA*, two cellulase or hemicellulase genes (Fig. 3.2A; Figure 2.2). GlyR3 contains two distinct domains, as GlyR1 and GlyR2[31, 32]: a helix-turn-helix DNA-binding motif at the N-terminal end and a sugar-binding domain at the C-terminal end (Fig 3.2B), suggesting that it is a regulatory protein controlled by a sugar. The location of *glyR3* suggests that GlyR3 controls the expression of the *celC* gene cluster by binding to its promoter region.

**rGlyR3 binds to the *celC* promoter region.** To study the function of GlyR3, we cloned its gene into *E. coli* with a chitin-binding domain (CBD) fused to the C-terminus of the recombinant protein. Fusion with the CBD facilitated purification by affinity chromatography using chitin beads as the affinity ligand. rGlyR3 was cleaved
off from the CBD, which bound to the chitin bead, by dithiothreitol (DTT) treatment and appeared as the predominant protein species with the expected size (39,330 daltons) on an SDS-gel (data not shown). The ability of rGlyR3 to bind to the promoter region of the celC gene cluster was examined by EMSA (electrophoretic mobility shift assay). The EMSA probe, prepared by PCR using biotin-labeled primers 3 and 5 (Table 3.1), represented the DNA sequence 100 to 200 bp upstream from the start codon of the celC gene, considered as the promoter region. In EMSA, adding rGlyR3 to the reaction resulted in gel-shift of the probe (lane 2, Fig. 3.3), indicating that rGlyR3 binds to the celC promoter region. Alternatively, under the same condition, rGlyR3 did not bind to the probe representing the CipA promoter region (data not shown), indicating that the binding of rGlyR3 is specific. The apparent dissociation constant ($K_D$), estimated as the concentration of rGlyR3 needed to shift 50% of the probe, was $4 \times 10^{-14}$ M.

To determine that GlyR3 is indeed expressed in vivo and the protein thus expressed binds to the same sequence, the EMSA was carried out using the cell lysate of C. thermocellum as the source of the DNA-binding protein. Although the lysate of the cellobiose-grown cells failed to bind to the celC promoter probe in two different concentrations (lanes 3-4, Fig. 3.3), the lysate of the lichenan-grown cells retarded the probe’s gel mobility to the same level as rGlyR3 (lane 5, Fig. 3.3). To verify that the lysate protein responsible for this shift is indeed GlyR3, we eluted the shifted
band from the EMSA gel and subjected it to SDS-PAGE analysis. The silver-stained protein, which was the only protein detected, had an apparent molecular weight of 39 kD as expected for GlyR3 (data not shown). The 39 kD protein was further eluted from the SDS-gel. MALDI-TOF (matrix assisted laser desorption/ionization time of flight) analysis demonstrated that the eluted protein was GlyR3 (33% sequence coverage; data not shown). These results indicate that GlyR3 is induced by lichenan and binds specifically to the \textit{celC} promoter region.

**Determination of the GlyR3 binding site by DNase I Footprinting.** To determine the GlyR3 binding site, we developed a non-isotope DNase I footprinting technique. In this method, a fluorescein-labeled DNA fragment corresponding to the 200 bp region immediately upstream of the start codon of \textit{celC} was partially digested by DNase I in the presence and absence of rGlyR3. The digested products were resolved by capillary electrophoresis and detected by using a fluorescence detector. As shown in Fig. 3.4, the fluorescence signals of a stretch of 18 bp were suppressed by rGlyR3 (comparing panels A and B). The protected region corresponds to an 18 bp palindromic sequence, typical for a DNA-binding site, with only one mismatch: AATGAACGC GCGTACATT (Fig. 3.4C). The ability of rGlyR3 to bind to this 18 bp sequence was verified by competitive EMSA, in which an excessive amount of unlabeled, double-stranded 18 bp sequence was used to compete for binding to rGlyR3 with the biotin-labeled 100 bp \textit{celC} promoter probe previously mentioned (Fig. 3.3). As shown in Fig. 3.5, the unlabeled 18 bp sequence at 100-fold concentration completely inhibited the binding of rGlyR3 to the 100 bp \textit{celC}
promoter probe (**lane 3**). In contrast, an unrelated 18 bp sequence from another site of the *celC* promoter region (probes 8 and 9, Table 3.1) failed to compete in the EMSA at the same concentration (**lane 4; Fig. 3.5**). These results indicate that rGlyR3 binds specifically to the 18 bp palindromic sequence.

**Laminaribiose inhibits GlyR3 binding to the *celC* promoter region.** The existence of a sugar-binding domain suggests that the DNA-binding activity of GlyR3 is regulated by a sugar. Various sugars were examined for their effects on the GlyR3’s DNA-binding activity using EMSA. Among all the sugars tested, only laminaribiose, a β-1,3 linked glucose disaccharide, was found to inhibit rGlyR3’s ability to bind the 100 bp *celC* promoter probe at the concentration of 15 mM (**lane 3, Fig. 3.6A**). In contrast, cellobiose at the same concentration had no effect (**lane 4, Fig. 3.6A**). Other sugars, including cellotriose, cellotetraose, cellopentose, glucose, sucrose, lactose, maltose and gentibiose, as cellobiose, showed little effect on the binding reaction (data not shown). Laminaribiose similarly inhibited the formation of the DNA-protein complex when the 18 bp binding site was used as the probe (**Fig. 3.6B**). The inhibition was dose dependent with an observable inhibitory effect at 0.5 mM laminaribiose (**lane 2**).

**rGlyR3 is a negative regulator subject to inactivation by laminaribiose as revealed by *in vitro* transcription assay.** To determine if GlyR3 serves as a transcription regulator for the expression of *celC*, we examined its ability to modulate the transcription of *celC* in an *in vitro* transcription assay. The assay utilized a DNA template consisting of the *celC* promoter region and the 5’ end of the *celC* gene. The
resulting \textit{celC} transcript was quantified by using quantitative reverse transcriptase-(RT-) mediated, Real-Time PCR. As shown in \textbf{Fig. 3.7A}, transcription of \textit{celC} was repressed by rGlyR3 in a dose-dependent manner. Furthermore, laminaribiose reversed the repressive effect of rGlyR3, also in a dose-dependent manner (\textbf{columns 1-4, Fig. 3.7B}). The rGlyR3-repressed transcription was completely restored at 10 mM laminaribiose (\textbf{column 4}). In contrast, cellobiose did not reverse the adverse effect of rGlyR3 (\textbf{column 6}). Laminaribiose alone at 10 mM had little effect on transcription (\textbf{column 5}). These results indicate that rGlyR3 serves as a negative regulator for the \textit{celC} gene in these experiments, presumably by binding to the promoter region. The gene is induced by laminaribiose, which inactivates the binding.

\textbf{3.3) Discussion}

\textit{C. thermocellum} produces a highly complicated biomass-degrading enzyme system, including the cellullosome that contains more than 70 subunits and many free enzymes. Despite intensive studies, how the organism coordinates the expression of such a large number of enzymes to degrade a particular biomass substrate or a mixture of substrates remains elusive.

GlyR3 is the first transcriptional regulator of glycosyl hydrolase genes identified in \textit{C. thermocellum}. It binds specifically to a near perfect 18-bp palindrome in the \textit{celC} promoter region. Its binding site notably bears similarity to many previously reported binding sites for transcriptional regulators that are homologous to LacI and control carbon metabolism in a variety of microorganisms (Table 3.2). The dissociation
constant ($K_D$) for GlyR3 is estimated to be $4 \times 10^{-14}$ M. This is near the same order of magnitude as the value for LacI ($K_D = 10^{-13}$ M) [33]. At this time, we cannot rule out the possibility of the existence of a second binding site with a lower affinity as has been reported for LacI.

The role of GlyR3 as a negative regulator is evidenced by the results of the in vitro transcription assay, in which the transcription of celC was repressed by GlyR3 in a dose-dependent manner. The repression is presumed to be due to the binding of GlyR3 to the 18 bp binding site (the operator) in the promoter region. Laminaribiose serves as an inducer, presumably by binding to the sugar-binding domain of GlyR3 and inhibiting its DNA-binding activity. Since we demonstrated that celC-glyR3-licA are co-transcribed (Fig 2.2), the three genes therefore form an operon repressible by GlyR3 and inducible by laminaribiose. The celC operon thus is similar to the lac operon, both operating in a negative mode. On the other hand, since glyR3 is part of the celC operon, induction of the operon would increase the level of the repressor and create a feedback loop. A continuous supply of the inducer, laminaribiose, would be needed to keep the operon in the induced state. In this regard, the celC operon functions like the E. coli hut operon, in which the repressor is part of the operon. In the absence of a continuous supply of the inducer, we expect the induction of the operon to be transient. In the soil bacterium Thermobifida fusca, a similar regulator, CelR, has been reported [34]. CelR binds to a 14 bp inverted repeat in the promoter region of each of the six cellulase genes. The binding is inactivated by cellobiose, the
presumed inducer. Recently, laminaribiose has been shown to induce cellulase production in *T. fusca* [35]. It has been theorized that CelR may bind to laminaribiose to induce expression of the cellulase genes in the bacterium.

Both CelC [36, 37] and LicA [38] are active on polysaccharides containing β-1,3 glucan such as lichenan and laminarin. A natural substrate found in plants would be callose. Callose is a plant cell wall polysaccharide that consists of β-1,3 linked glucose. Constitutive low-level expression of the *celC* operon likely generates low levels of CelC and LicA. When a substrate containing β-1,3 glucan becomes available, these two enzymes would generate the inducer, laminaribiose, as the hydrolysis product. Laminaribiose diffused or transported into the cell would turn on the operon for the biosynthesis of more enzymes. This regulation scheme is corroborated by our observation that GlyR3 was detected in the cell lysate only when the bacterium was grown on lichenan. This regulation scheme further implies that CelC and LicA are the major β-1,3 glucan-degrading enzymes in this bacterium. LicA has indeed been reported to be the major enzyme that degrades β-1,3 glucan [38]. LicA was characterized as an endo-1,3(4)-β-glucanase active on barley-β-glucan and laminarin. It was shown to be upregulated when growing on laminarin or barley-β-glucan as opposed to cellobiose or cellulose. We independently found that *C. thermocellum* grows on laminaribiose as the sole carbon source (data not shown). These results are consistent with the proposed regulation mechanism of the *celC* operon presented above. It is noteworthy that both CelC and LicA are non-
cellulosomal enzymes, suggesting that degradation of β-1,3 glucan does not benefit from the enzymes serving as the cellulosomal components in *C. thermocellum*.

Our results indicate that, despite the water insolubility of the biomass substrates, coordination of the expression of biomass-degrading enzymes can be accomplished through soluble sugars. The *celC* operon as a unit of gene regulation provides the first clue to the puzzle of how the bacterium coordinates the biosynthesis of such a large number of glycosyl hydrolases. GlyR3 is the first transcription regulator found in *C. thermocellum*. It is also the first time laminaribiose is found to serve as an inducer. Foresee ably, more transcription factors and inducers will be found, which will further illuminate how the bacterium commands a myriad of enzymes to attack the complicated biomass substrate containing many different forms of glycans. The results will be particular illuminating in understanding if a particular set of the cellulosome components are selected by the bacterium to optimize its activity on a particular biomass substrate.

3.4) Materials and Methods

**Bacterial Strains and Plasmids.** *C. thermocellum* ATCC 27405 was used as the source for genomic DNA, RNA, and cell lysates. *E. coli* Top10 (Invitrogen) was used as the cloning host for plasmid PTXB1 (New England Biolabs). *E. coli* strain BL21(DE3) (Stratagene) was used for expressing recombinant GlyR3.
Culture Conditions. *C. thermocellum* was grown in Hungate tubes or anaerobic flasks in chemically-defined MJ medium [39] containing 0.5% carbon source (cellobiose, lichenan, or laminaribiose). Seed cultures were grown on cellobiose. The cultures were incubated at 60º C. *E. coli* strains containing recombinant plasmids were grown at 37º C in a shaker or on agar plates containing Luria-Bertani medium [40] supplemented with 0.1 mg/ml ampicillin. Isopropylthiogalactoside (IPTG; 50 mM) was used to induce the expression of cloned *glyR3*.

Cloning of *glyR3*. PCR was employed to clone *glyR3* using *C. thermocellum* genomic DNA as the template, primers 1 and 2, (Table 3.1), which incorporated the *EcoRV* and *XhoI* restriction sites, respectively, and a hi-fidelity DNA polymerase (Extensor; ABgene). The PCR product was digested with *EcoRV* and *XhoI*, cloned into the *NruI* and *XhoI* sites of pTXB1, and transformed by electroporation into *E. coli* TOP10 cells. Restriction digests and DNA sequencing using the dye termination cycle sequencing method and a Model 3100 Genetic Analyzer (Applied Biosystems) were used to verify the cloned gene.

Expression and Purification of rGlyR3. *E. coli* BL21(DE3) harboring pTXB1 containing the clone *glyR3* was induced with 50 mM IPTG in the exponential growth phase for four hours. The cells were harvested by centrifugation and lysed by sonication. rGlyR3 in the lysate was purified by affinity chromatography using chitin beads as the affinity ligand following the IMPACT system protocol (New England Biolabs). The purified protein was concentrated by ultrafiltration using a Microsep
3K column (Pall) and examined for size and purity using an SDS-PAGE on a 12% gel [41].

**Protein Assay.** Protein concentrations were determined using the Bradford [42] reagent (Bio-Rad) and bovine serum albumin (Sigma) as a standard.

**Electrophoresis Mobility Shift Assay (EMSA).** The 100 bp EMSA probe was made by PCR using Taq DNA polymerase (Thermostart; ABgene), primer 3 labeled with biotin, and primer 5 (Table 3.1). The 18 bp probe consisted of complementary DNA fragments annealed by heating to 94º C and slowly cooling to room temperature (probes 6 and 7, Table 3.1). All EMSA experiments were performed on 4% polyacrylamide gels in Tris-Borate-EDTA buffer (45 mM Tris-borate, 1 mM EDTA). Each EMSA reaction mixture contained 500 ng poly (dI-dC), 1X Lightshift EMSA kit binding buffer (Pierce), 1X Lightshift loading dye (Pierce), and appropriate amounts of the DNA probe and protein preparations. Sugars were added in some experiments to test their inhibitory effect as indicated. EMSA gels were electroblotted onto Biodyne B membrane (Pall Corporation). Signal development followed the Lightshift Chemiluminescent EMSA kit protocol (Pierce) using Biomax films (Kodak) for luminescence detection.

**DNase I Footprinting.** PCR was used to amplify the 200 bp celC promoter region using primer 3 labeled with fluorescein and primer 4 (Table 3.1). The reaction mixture contained 400 ng of the amplified DNA fragment, binding buffer (10 mM Tris, 50 mM KCl, 1 mM DTT), 300 ng dI-dC, 1 U DNase I (Invitrogen), and with or without 60 ng rGlyR3. After incubation at 37º C for 7 min, 1 mM EDTA was added
and the mixture was heated to 70° C for 15 min. The DNase I-digested DNA products were resolved and detected using a Model 3100 Genetic Analyzer (Applied Biosystems).

**in vitro Transcription Assay.** In this assay [43, 44], the DNA template was generated using primers 10 and 11 (Table 3.1) to amplify the 200 bp celC promoter region along with the first 650 bp of celC of the *C. thermocellum* genomic DNA. Each assay mixture contained 10 μl *C. thermocellum* cell lysate (cellobiose-grown), 2 μl RNase inhibitor (RNase Out; Invitrogen), 1X RNA polymerase buffer, 1 μg DNA template, 25 nM rNTP’s, different amounts of rGlyR3 and laminaribiose, and DEPC-water to a total volume of 50 μl. The reactions were incubated at 60° C for 50 min. The resulting RNA was isolated using the Trizol method (Invitrogen), subjected to DNase I digestion, reverse transcribed using random primers, and quantified using Real-Time PCR with the primers specific to celC as described below.

**Quatitative Real-Time PCR.** Each reaction mixture contained 1 μl cDNA template, 7.5 μl SYBR Green Supermix (Bio-Rad), 5.75 μl water, and 250 nM of each primer (primers 12 and 13, Table 3.1). Real-Time PCR was carried out using an iCycler IQ.

**Bioinformatic Analysis** The LacI protein sequence (from *E. coli*) was found in NCBI’s genebank and used as a bait to be BLASTed against the entire *C. thermocellum* predicted protein database, kept on the ORNL website. As described in the results section of this chapter (3.2), domain structure of the GlyR proteins was modeled by NCBI Conserved Domain Search. As described in figure 3.1, ClustalW was used to align the amino acid sequences of GlyR1, GlyR2, GlyR3, and LacI.
**Fig. 3.1** Alignment of GlyR1, GlyR2, GlyR3, and LacI. The putative DNA-binding domain of GlyR3 is underlined and the putative sugar-binding domain is bolded. “*”, identical residues; “:”, conserved residues; “.”, semi-conserved residues according to the convention of the European Bioinformatics Institute (http://www.ebi.ac.uk/clustalw/). GenBank accession numbers: GlyR1: ZP_00509723, GlyR2: ZP_00503684, GlyR3: ZP_00504673.
Fig. 3.2) Schematic drawing of the *celC* operon (A) and the domain structure of GlyR3 (B) of *C. thermocellum*. 
<table>
<thead>
<tr>
<th>No.</th>
<th>Sequence&lt;sup&gt;1,2&lt;/sup&gt;</th>
</tr>
</thead>
</table>
| 1   | **F:** glyR3-F-EcoRV-  
|     | GCGCGATATACCCAGTGAGAAATAGCAAAAATTAG | **R:** glyR3-R-XhoI-  
|     | GCGCCTCGAGGAATTCCAAAGCCCTCTTGGTTTTA |
| 2   | **F:** Entire<sub>celC</sub>Prom-F-biotin (or fluorescien)-  
|     | CCGAATTTAAAAACCTGGACAGAG | **R:** Entire<sub>celC</sub>Prom-R-Unlab-  
|     | TCCTCCTGAAATATTGTGTGTTTAA |
| 3   | **R:** celC Prom<sub>1st_100bp</sub>-R-Unlab-  
|     | TGAAACCATTTAACACTGGATTAT |
| 4   | **F:** BS-F-Biotin(or Unlab)-AATGAACGCCTACATT |
| 5   | **R:** BS-R-Unlab-AATGTACGCCGCTTCATT |
| 6   | **F:** Control 18-mer-F-Unlab-  
|     | AACTGGACAGAGAAGAAG |
| 7   | **R:** Control 18-mer-R-Unlab-  
|     | CTTCTTCTCTGCCTGTT |
| 8   | **F:** Invnt-F-CCGAATTTAAACCTGGACAGAG |
| 9   | **R:** Invnt-R-CCAGTGGGCTTTCTGATGC |
| 10  | **F:** celC-F-CGGGAACATATTGCCTTTGAAC |
| 11  | **R:** celC-R-GGTGGAATCTTTCCCTGATT |

<sup>1</sup>F: Forward, R: Reverse  
<sup>2</sup>Restriction sites are underlined.

Table 3.1) Primers and their functions for Chapter 3 experiments.
**Fig. 3.3)** Binding of rGlyR3 to the *celC* promoter region as revealed by EMSA.

All reactions contained 5 ng of a biotin-labeled 100 bp DNA fragment corresponding to the *celC* promoter region. Lanes: 1, no protein; 2, rGlyR3 (1 ng); 3-4, cell lysate from the cellobiose-grown *C. thermocellum* culture (200 ng and 500 ng protein, respectively); 5, cell lysate from the lichenan-grown *C. thermocellum* culture (120 ng protein). The shifted band from lane 5 was excised and subjected to MALDI-TOF analysis, confirming the binding protein to be GlyR3.
Fig. 3.4) GlyR3 DNA-binding site as determined by DNase I footprinting analysis. The flourescein-labeled, 200 bp DNA fragment corresponding to the celC promoter region was subjected to DNase I digestion without (A) and with (B) rGlyR3. The digested products were resolved by capillary electrophoresis and detected by a fluorescence detector. The DNA sequence corresponding to the suppressed peaks (Protected Region) is palindromic with one mismatch (C). The peaks shown in red are the internal size standards.
Fig. 3.5) Competitive EMSA confirming the rGlyR3 DNA-binding site. All reactions contained 5 ng of a biotin-labeled 100 bp DNA fragment corresponding to the celC promoter region. Lanes: 1, no protein; 2, 0.5 ng rGlyR3; 3, 0.5 ng rGlyR3 and 100-fold unlabeled 18 bp binding site; 4, 0.5 ng rGlyR3 and 100-fold unlabeled 18 bp control fragment.
Fig. 3.6) Inhibition of GlyR3 DNA-binding activity by laminaribiose as analyzed by EMSA. (A) 100 bp DNA fragment corresponding to the celC promoter region as the probe. All reactions contained 5 ng biotin-labeled probe. Lanes: 1, probe only; 2, probe and 0.5 ng rGlyR3; 3-4, probe and 0.5 ng rGlyR3 plus 15 mM of laminaribiose and cellobiose, respectively. (B) 18 bp GlyR3 DNA-binding site as the probe. All reactions contained 5 ng biotin-labeled probe and 0.5 ng rGlyR3. Lanes: 1, no laminaribiose, 2-5, 0.5, 1, 5, and 10 mM laminaribiose, respectively.
Fig. 3.7) Laminaribiose induction of celC by inactivating GlyR3 as revealed by *in vitro* transcription assay. (A) Relative transcript level determined by quantitative RT-PCR in the presence of various amounts of rGlyR3. (B) Relative transcript level in the presence of rGlyR3 and cellobiose or various amounts of laminaribiose. The data represent the averages of the results from triplicate experiments. Vertical bars represent standard deviations.
<table>
<thead>
<tr>
<th>Regulator</th>
<th>Sequence$^1$</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlyR3</td>
<td>AATGACGTC</td>
<td><em>C. thermocellum</em></td>
</tr>
<tr>
<td>CelR</td>
<td>TGGGAAGC</td>
<td><em>T. fusca</em></td>
</tr>
<tr>
<td>LacI</td>
<td>TTGTGAGC</td>
<td><em>E. coli</em></td>
</tr>
<tr>
<td>CcpA</td>
<td>TGTAAGC</td>
<td><em>B. subtilis</em></td>
</tr>
<tr>
<td>GalR</td>
<td>GTGAKANC</td>
<td><em>E. coli</em></td>
</tr>
<tr>
<td>GalS</td>
<td>GTGAKANC</td>
<td><em>E. coli</em></td>
</tr>
</tbody>
</table>

CelR-GalS binding half-sites were taken from [34].

$^1$ K = G/T, N = any base; conserved nucleotides are bolded.

**Table 3.2) The DNA binding half-sites of various LacI/GalR family regulators.**
3.5) Chapter 3 References


Chapter 4

Bioinformatic Analysis of Cluster X and Study of its Transcript Expression Profile

4.1) Introduction

While searching the *C.thermocellum* genome for gene clusters, another interesting discovery was made. A gene cluster of five uncharacterized putative cellulosome components was present in the genome. The cluster consisted of five genes that all contained putative dockerin domains (identifying them as potential members of the cellulosome). This was the largest cluster of dockerin containing genes discovered in the *C. thermocellum* genome. Since none of these genes had been cloned and characterized a bioinformatic study was conducted in order to get a better idea of what types of proteins were expressed from these genes. Experiments were also conducted that were similar to what we had undertaken for the *celC* gene cluster (Chapter 2). Transcription of the genes was studied to determine if the cluster functioned as an operon and to tell when the genes were expressed. The results indicated that all of the genes were transcribed into a single ~12 kb transcript. Primer
extension revealed two distinct transcript initiation sites. RT-qPCR determined that the genes all have a similar expression profile (as was expected from the Northern Blot data), and that they have a much higher expression level in cells grown using cellulose rather than cellobiose.

4.2) Results

**Bioinformatic analysis of Cluster X.** Cluster X consists of five genes that encode putative cellulosomal components (Fig. 4.1). This is the largest cluster of cellulosomal components found in the genome. The non-coding regions between the genes are very short, raising the possibility that all of the genes in the cluster are co-transcribed. The proteins expressed from the genes in the cluster all contain Type I dockerin domains, a unique characteristic of cellulosomal components. The first gene (gene1) encodes a putative family 2 glycosyl hydrolase (Fig. 4.2). The second gene (gene2) encodes a potential family 43 glycosyl hydrolase, while the third gene (gene3) contains an unknown domain followed by a family 6 CBM. The fourth gene (gene4) encodes a protein that is a putative esterase. Finally, the fifth gene (gene5) product is a putative family 5 glycosyl hydrolase with a fibronectin-like domain and a family 6 CBM.

**Northern blot analysis of the gene cluster.** Northern blot was conducted to determine if the genes in the cluster were co-transcribed. The results show that probes for each gene hybridize to an RNA transcript that is roughly 12 kb (Fig. 4.3). The Northern blot was conducted using total RNA isolated at stationary phase (Fig. 4.5 at 60 hrs) from cellulose (cotton) grown cultures. Thus, at the conditions tested, the entire gene cluster is transcribed as one large mRNA transcript. It is interesting to
note that Northern Blot was attempted using total RNA isolated from cellobiose grown cultures, but no transcript could be detected (data not shown).

**Primer Extension and promoter analysis of Cluster X.** Primer extension yielded two transcription initiation sites upstream from the coding region of *gene1*. The initiation sites were 74 bp and 151 bp upstream of the putative start codon for *gene1* (Fig. 4.4).

**The effect of carbon source and growth phase on the transcript levels of the genes in Cluster X.** *C. thermocellum* was grown using 0.5% of two different carbon sources, cellobiose and cellulose (cotton). The growth curve for each carbon source was measured using pelleted cell protein (Fig. 4.5). The results show that the growth rate of *C. thermocellum* is faster on cellobiose than on cellulose, which was to be expected [1-7]. RNA was isolated from cultures at different time points throughout the growth curve. The growth curve was used to make sure that cells from cellobiose and cellulose cultures were in the same growth stage when compared by RT-qPCR. As was expected from the Northern Blot results, all of the genes in Cluster X have the same expression profile (Fig. 4.6-4.10). At all growth phases cellulose grown cells have a higher expression level for the Cluster X genes. The expression trend for the Cluster X genes is partially growth rate dependent. The expression level of Cluster X genes for cellulose grown cells drops during late-exponential and early stationary phase. The level then rises back near the level seen at exponential growth once the cells are well into stationary phase. A similar trend is observed for cells grown on cellobiose. Stationary phase consistently gives the maximum amount of transcript measured for each gene in cellobiose grown cultures (Fig. 4.6-4.10 stationary phase).
To insure that our RT-qPCR system was functioning correctly we tested it on genes that have known transcript patterns in the literature. The transcript expression pattern for *celS* (Fig. 4.11) and *xynC* (Fig. 4.12) compared favorable to Northern blot and RNase protection assays previously shown in the literature [8, 9].

### 4.3) Discussion

The discovery of Cluster X illustrates that there are many new components of the cellulosome that remain to be characterized. Cluster X is the largest set of putative cellulosome components found in the genome. The bioinformatic results are interesting when the genes are looked at as a set. Several different Glycosyl hydrolase families are present in the cluster (families 2, 5, and 43). Also, gene 3 has an unknown function and gene 4 is an esterase. Since these genes are expressed concurrently within the same transcript, it is possible they have a synergistic effect in the process of degrading specific substrates. Our data shows that they are more highly expressed when the bacterium is grown on cellulose. *C. thermocellum* is a soil bacterium that is found in compost, its natural surroundings most likely do not contain pure cellulose such as cotton. In nature cellulose is associated with hemicellulose, which in turn is often bound to lignin. In order for *C. thermocellum* to efficiently degrade cellulose to obtain energy, it must be able to get through any hemicellulose or lignin blocking its path. The gene products of Cluster X must be characterized to get a true picture of what use they are to the organism. However, based mostly on bioinformatics, *gene1* encodes a possible β-glucosidase or β-galactosidase, *gene2* and *gene3* encode hemicellulases, *gene4* encodes an esterase,
Cluster X possibly contains a set of proteins that are capable of unlocking cellulose from hemicellulose and lignin, then degrading the cellulose into cellobiose or even glucose. The RT-qPCR results indicate that, at every time point, all of the genes in Cluster X are more highly expressed when the bacterium is grown on cellulose rather than cellobiose (Fig. 4.6-4.10). The higher expression on cellulose signifies that the products of Cluster X are important for growth on cellulose.

The highest expression of Cluster X occurs at the stationary phase of growth (Fig. 4.6-4.10). Other cellulase genes in the bacterium have also shown to exhibit this phenomenon [10]. It has been postulated that cellobiose concentration in the cell drops once the bacterium reaches stationary phase, as most of the cellobiose has been converted into energy. The organism senses the drop in cellobiose, realizes that it must seek out new sources of energy, and produces the Cluster X enzymes to acquire more substrate. Newer studies have cast doubt that cellobiose is the sole contributor to this action by the cell. It has been shown that the cell uptakes other cellodextrins at even faster rates than cellobiose [11]. At mid-exponential growth the cells have just encountered the cellulose, they are still producing Cluster X enzymes (along with other cellulase system components) in order to efficiently degrade the cellulose into usable products. Once the cells get to late exponential and early stationary phases, the amount of usable products from the cellulose are at a maximum concentration in the cell. The bacterium turns down production of the Cluster X enzymes, as the production of these proteins is not needed because of the high concentration of
products from the degradation of cellulose already present in the cell. As the cell starts to get later into its growth cycle, most of the cellulose has been degraded and the concentration of convertible products from the cellulose in the cell begins to drop. Now well into stationary phase, the cell ramps up production of the Cluster X enzymes in order to aid the discovery and degradation of new cellulose sources.

The Northern blot results indicated that Cluster X is transcribed as one single unit. The results lead to the hypothesis that the genes in Cluster X are regulated together in an operon. The RT-qPCR data showing that all of the genes have similar expression profiles backs up this conclusion. Future work with Cluster X will include the cloning and characterization of all of the individual genes and corresponding proteins. The regulation of the cluster will also be studied. Preliminary EMSA data shows a protein that binds to the promoter region of gene1 from the cluster when the organism is grown on cellulose. The protein(s) that binds to the promoter region must be identified in order to determine if it is related to the regulation of the cluster.

4.4) Materials and Methods

**Bacterial Strains.** The source of all genomic DNA and RNA was *C. thermocellum* ATCC 27405.

**Culture Conditions.** Hungate tubes were used to culture batches of *C. thermocellum* in chemically-defined MJ medium [12]. The medium contained 0.5% of the appropriate carbon source (cellobiose or cotton). Seed cultures were all grown for 12 hours on cellobiose. All cultures were incubated at 60° C.
**Measure Pellet Protein for Growth Curve.** To measure the growth curve for cellobiose and cotton cells the pelleted protein per milliliter of culture was estimated. The cells were centrifuged at 12,000 x g for 15 minutes (cotton cultures were filtered through cheesecloth, the remaining cotton fibers were washed with 0.85% NaCl solution). The supernatant was removed from the sample and the pellet was washed and spun down twice with 0.85% NaCl solution. This was done to minimize extracellular protein in the sample. Before the last wash was centrifuged, 1 ml of the resuspended pellet was taken for cellular protein analysis, the rest of the resuspended pellet was spun down for RNA isolation. The removed 1 ml was centrifuged and the final pellet was resuspended with 1 ml of 0.2 N NaOH and heated in boiling water for 15 minutes. After boiling, the sample was centrifuged again at 12,000 x g for 15 minutes. Finally, 1 ml of 0.2 N HCl was added to neutralize the supernatant. The amount of protein in the supernatant was now measured using the Bradford [13] reagent (Bio-Rad). The supernatant was diluted so that spectrophotometer readings fell within the linear range of a bovine serum albumin (Sigma) standardized curve.

**RNA Isolation.** *C. thermocellum* cell pellets were obtained as outlined above. The Trizol (Invitrogen) method of RNA isolation was used to obtain total RNA. The RNA was digested with DNase I (ABgene) and quantified using a spectrophotometer at 260/280 nm.

**Reverse Transcription.** A standard reverse transcription reaction protocol was followed using MMLV reverse transcriptase (Invitrogen).

**Quantitative Real-Time PCR.** Gene expression analysis was carried out using iQ Syber Green Supermix (Bio-Rad) on a Bio-Rad iCycler with gene specific primers
(Table 4.1). All primer sets were tested to make sure data falls within the linear range of quantification, and that amplification efficiency was greater than 90%. The two-step PCR protocol used consisted of activation of the polymerase followed by 40 cycles of 95 °C for 15 seconds and 60 °C for 30 seconds. All sample/primer combinations were analyzed in triplicate. The 16S ribosomal RNA was used to normalize the experiment for sample-to-sample variation; expression data for genes of interest was reported relative to 16S expression.

**Preparation of Northern Blot Probes.** Northern blot probes were created using PCR with the primers listed in Table 4.1. A standard protocol was followed for ABgene Thermostart Taq Polymerase. The annealing temperature was set to 58 °C and an extension time of 20 seconds was used for each probe. After PCR, the reaction was purified using the Promega Wizard SV Gel and PCR Clean-Up System. The PCR was checked for size using a 1.2% agarose gel.

**Northern Blot.** *C. thermocellum* was grown using cotton as a carbon source for 60 hours. Total RNA was isolated as described above. The Ambion NorthernMax-Gly kit was used to perform the northern blot. The manufacturers’ protocol was followed with the subsequent conditions. 20 μg of RNA was loaded into a 0.8% agarose gel. An RNA molecular weight marker (Novagen) was also run on the gel to estimate transcript size. A downward capillary blot was used to transfer RNA to Biodyne B (Pall) positively charged membrane. The RNA marker was stained using 0.2% Methylene Blue to verify the success of the blotting. The blot was mixed with ULTRAdyhyb buffer from the NorthernMax kit and allowed to prehybridize for 45 minutes at 42° C. Probe was prepared by using a PCR reaction with primers from
table I. dNTPs were used in the reaction that had the dATP labeled with biotin in a 5.25:1 ratio (New England Biolabs). Probes were heated to 95°C for 10 minutes then added to the prehybridization solution to a final concentration of 2 ng/ml. The hybridization was allowed to continue for 24 hours. After hybridization, the membrane was washed with low stringency buffer for 10 minutes at room temperature and washed twice with high stringency buffer at 42°C for 15 minutes each. To develop the blot the standard manufacturers’ protocol for the Chemiluminescent Nucleic Acid Detection Module (Pierce) was followed.

**Primer Extension.** A primer was created (Table 4.1) and end-labeled with FAM490 (Invitrogen). A standard reverse transcription reaction was carried out for a gene specific primer using MMLV Reverse Transcriptase (Invitrogen). The University of Rochester Functional Genomics Center used capillary electrophoresis to quantify the size of the DNA fragment created by the reaction.

**Bioinformatic Analysis.** As previously mentioned, the entire genome sequence of *C. thermocellum* was searched for gene clusters using the ORNL website. Cluster X was then explored in depth by utilizing BLAST search found on the NCBI website (http://www.ncbi.nlm.nih.gov/). Conserved Domain search from the NCBI website was also used to map out the structure of all the proteins in Cluster X.
Cluster X

Fig. 4.1) A schematic of the genes located in Cluster X. All five of the genes are uncharacterized and are putative members of the cellulosome.
Fig. 4.2) The protein domain structures for all members of Cluster X. All proteins contain a cellulose-binding module and a dockerin. GH: glycosyl hydrolase family; CBM: cellulose-binding module. Figure adapted from [14].
Fig 4.3: Northern blot analysis of the genes from Cluster X. All probes hybridized to an RNA transcript that is roughly 12 kb in length. These results indicate that all of the genes in Cluster X are co-transcribed under the conditions tested.
Fig. 4.4) Primer extension results for *gene1* of Cluster X. The results show peaks at 107 bp and 184 bp. These correspond to two transcription initiation sites 74 bp and 151 bp upstream of the start codon for *gene1*. 
4.5) **Growth curves of C. thermocellum using cellobiose or cellulose as carbon sources.** Pellet protein was plotted versus time for each culture. Each value on the curve is the average of three independent cultures. The numbers (1,2,3,4) represent the time point for each growth curve at which RNA was isolated for RT-qPCR analysis.
Fig. 4.6) RT-qPCR results for gene1 of Cluster X. The time point taken for each sample is numbered on the growth curve in Fig. 4.5. The results are average of triplicate experiments and error bars represent standard deviation between the individual data.
Fig. 4.7) RT-qPCR results for gene2 of Cluster X. The time point taken for each sample is numbered on the growth curve in Fig. 4.5. The results are average of triplicate experiments and error bars represent standard deviation between the individual data.
Fig. 4.8) RT-qPCR results for gene3 of Cluster X. The time point taken for each sample is numbered on the growth curve in Fig. 4.5. The results are average of triplicate experiments and error bars represent standard deviation between the individual data.
**Fig. 4.9** RT-qPCR results for *gene4* of Cluster X. The time point taken for each sample is numbered on the growth curve in Fig. 4.5. The results are average of triplicate experiments and error bars represent standard deviation between the individual data.
Fig. 4.10) **RT-qPCR results for gene5 of Cluster X.** The time point taken for each sample is numbered on the growth curve in Fig. 4.5. The results are average of triplicate experiments and error bars represent standard deviation between the individual data.
Fig. 4.11) RT-qPCR analysis of *celS* transcript level. The results were compared to known expression levels of *celS* transcript. This confirms that our RT-qPCR system is in agreement with previously detected levels of *celS* transcript [8].
Fig. 4.12) **RT-qPCR analysis of xynC transcript level.** The results were compared to known expression levels of xynC transcript. This confirms that our RT-qPCR system is in agreement with previously detected levels of xynC transcript [9].
<table>
<thead>
<tr>
<th>No.</th>
<th>Primer Sequence</th>
<th>Use of Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F:16S-AATTCGAAGCAACGGAAGAA</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>2</td>
<td>R:16S-GCGGGACTTAACCAACATCTC</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>3</td>
<td>F:Gene 1-AATCAACAAACAGCCAAC</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>4</td>
<td>R:Gene 1-TCCGCAATAATCCACATGAACC</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>5</td>
<td>F:Gene 2-GGGGTCAGTGCGGAAAGAAA</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>6</td>
<td>R:Gene 2-TGTCCGATACTGCAACTCAAT</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>7</td>
<td>F:Gene 3-TGTCCAGAAGACCGGCTCAAC</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>8</td>
<td>R:Gene 3-GCCTGGACACAGGTATCAACACC</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>9</td>
<td>F:Gene 4-GCCTGGGTCTTAGCATGGTTTTG</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>10</td>
<td>R:Gene 4-ATTAACAACCTGCTCTCTG</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>11</td>
<td>F:Gene 5-ATGACTGAGTGATGCGCGGAGGTG</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>12</td>
<td>R:Gene 5-GTTGGCCGGAATGTACTGGAAG</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>13</td>
<td>F:celS-CATGGACCGGACATACACAGG</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>14</td>
<td>R:celS-CAAGTGAAACCTGCAACACCAAG</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>15</td>
<td>F:xynC-TTGACGATTTCCAGCAACACC</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>16</td>
<td>R:xynC-GCAACCACCCACTTGGAAATAACC</td>
<td>RT-qPCR</td>
</tr>
<tr>
<td>17</td>
<td>F:Gene 1-GTGGGGGTATTTTTGCAGG</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>18</td>
<td>R:Gene 1-GAGTCAAGTCATATTCAAAAGA</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>19</td>
<td>F:Gene 2-ATGGGGGGAAGGTTTTATTGTCTC</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>20</td>
<td>R:Gene 2-GGAGTGCTACGTCTCAGTCA</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>21</td>
<td>F:Gene 3-ATGGGTAATGGAATTTAGGA</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>22</td>
<td>R:Gene 3-TTGGCAATGCTCTAAAGGGTG</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>23</td>
<td>F:Gene 4-ATGGTTAAGAAGGTCTTTAGTG</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>24</td>
<td>R:Gene 4-ATATAACTTGGTCCGGGCGTG</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>25</td>
<td>F:Gene 5-ATGGGGGGAATGATTTAAACA</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>26</td>
<td>R:Gene 5-GTCCATTCGGGATGGAAG</td>
<td>Northern Blot</td>
</tr>
<tr>
<td>27</td>
<td>F:Gene 1-AATGTACCATATACCGGCA</td>
<td>Primer Exten.</td>
</tr>
</tbody>
</table>

Table 4.1) Primers and their functions for Chapter 4 experiments.
4.5) Chapter 4 References


Chapter 5

Conclusions and Future Work

5.1) Conclusions

The results from this work have yielded several interesting discoveries that will be built upon in the future to further understand and optimize the biomass to ethanol process. Up to this point, most of the work in the cellulase field that has dealt with *C. thermocellum* has been the discovery and characterization of the enzymes involved in the degradation of cellulose. The literature contains very little about the regulation of expression for these enzymes. While there are still important enzymes to characterize, as evidenced by Cluster X which was discovered in this work, understanding how and why the enzymes are expressed at specific times is the next key in optimizing the biomass to ethanol process. If *Clostridium thermocellum* is to be used in an industrial setting, the knowledge of which enzymes are produced during specific conditions needs to be understood in order to make a more efficient and economical process.

Our discovery that *celC*, *glyR3* and *licA* are all co-transcribed is the first evidence that an operon model of regulation is used by *C. thermocellum* in relation to its cellulase system.
This was the first instance in the organism that a cellulase and a hemicellulase gene were shown to be co-transcribed. The discovery is further confirmation that cellulases and hemicellulases need to work in tandem to efficiently degrade cellulose that is present in the environment. This work is one of the first to use Real-time qPCR to study transcript expression of enzymes over the course of a growth curve. While few papers are available in the literature that focus on the mRNA levels of specific genes, almost all that have been done used traditional protection assays, or Northern blot. The discovery of a spike in expression for celC, glyR3, and licA when the bacterium was grown with β-1,3 linked glucan present in the media was an indication that these sugars can induce the expression of specific genes from the cellulase system. No such spike was detected for manB or for celT; the co-transcribed genes celC, glyR3, and licA were specifically induced.

The main thrust of the work presented in this thesis is the identification and characterization of GlyR3. GlyR3 is the first transcriptional regulator protein that controls expression of cellulase genes characterized in the organism. To our knowledge, GlyR3 is the very first transcriptional regulator of cellulase genes in any organism that is proven to have β-1,3 linked glucans (specifically laminaribiose) as an inducer. This is the first instance where laminaribiose was shown to induce the expression of cellulase and hemicellulase genes. Most work with thermophilic and mesophilic ethanol producers centers on the degradation of β-1,4 linked glucans, as these are the building blocks of cellulose. We have demonstrated that the organism thrives on β-1,3 linked substrates and that genes are activated which are specifically used to help the organism survive utilizing those sugars. GlyR3 is the first regulatory protein in C. thermocellum that has a molecular
mechanism, complete with the sequence of its DNA-binding site and *in vitro* transcription assays, which prove laminaribiose is the inducer of the *celC* operon. The GlyR3 story addresses a key segment of research in the study of thermophilic ethanol producers that has been overlooked in the literature to date.

Another important discovery made in the course of this work is the existence of Cluster X, the largest cluster of cellulosomal genes in the genome of the organism. None of the five genes in Cluster X have been cloned or characterized. The bioinformatic results show that several different enzymes which can help degrade hemicellulose, lignin, and cellulose are present in the cluster. The Northern blot results show that when grown on cotton, all of the genes in Cluster X are co-transcribed. Real time qPCR results show that the genes all are expressed to a higher degree when the bacterium is grown on cellulose, as compared to cellobiose. This indicates that the genes are important for growth on cellulose. The results are also a signal that the genes may all be regulated together by the same control mechanism.

**5.2) Future Work**

There are many avenues of research that may be pursued which would build on the work presented in this thesis. The cellulase system of *C. thermocellum* consists of well over 100 proteins. The molecular mechanisms of regulation for the vast majority of these proteins are unknown. Once the mechanisms are understood, they must be manipulated in a way to increase the efficiency of cellulose degradation.
Specifically, the newly characterized transcription regulator, GlyR3, can be further studied to determine if it controls any other components of the cellulase system. Once a genetic system is widely known, removing the gene in order to determine which members of the cellulase family are more highly expressed over the wild-type would be an interesting experiment. Furthermore, in this work two proteins were identified that are similar to GlyR3 in terms of sequence and domain structure. In preliminary experiments, GlyR1 and GlyR2 have been cloned and expressed as soluble recombinant proteins in *E. coli*. It is likely that GlyR1 and GlyR2 control transcription of other cellulase system genes. Experiments such as whole genome PCR, and bacteria one-hybrid systems may be used to discover the DNA binding sequences of the proteins. Once the DNA binding sequence is known, it will be possible to go into the genome and predict which genes are under the control of each regulator.

The promoter regions of other cellulase system components can be studied using EMSA to determine if proteins exist in different cell lysates that specifically bind to that DNA. Preliminary experiments have detected proteins that bind to the *celS* promoter region, along with the promoter region of *gene1* from Cluster X. The proteins responsible for the band shifts need to be isolated. They can be cut from the EMSA gel, eluted into a buffer, and finally analyzed using SDS-PAGE. The resulting bands on the SDS-PAGE can be cut out and subjected to trypsin digest coupled with Mass Spectroscopy for protein identification. This technique was already used to isolate GlyR3 from crude lysate, as shown in Chapter Three of this thesis. Once the protein(s) is (are) identified the gene responsible for encoding the regulator will cloned and expressed in *E. coli*. 
Characterization of the regulators will then ensue, utilizing DNase I footprinting, in vitro transcription and EMSA experiments.

The complexity and vast quantities of components that are part of the cellulase system in *C. thermocellum* make it necessary for high-throughput methods to be employed to gain a complete understanding of how its’ system is regulated. Only by understanding how the entire system works, will it be possible to engineer the most efficient process for cellulose conversion to ethanol. The completion of the genome sequence during the time of this thesis will be a treasure trove for high-throughput studies. With the genome sequence widely available, the next step is to study the transcriptome, proteome, and secretome of the organism. Microarrays will be applied to determine when genes are turned on and off on a genome wide scale. 2D gel, and mass spectroscopy will be utilized to study the proteins expressed inside the cell, along with the proteins secreted by the organism at different growth phases and when the bacterium is grown using different carbon sources. Using these techniques to identify the differences in genes expression on the mRNA level and then the protein level under specific conditions will speed up the process of identifying and characterizing transcriptional regulators and the targets they control.

Chapter Four of this thesis deals with the existence and transcript expression profile of Cluster X. As mentioned previously, Cluster X is the largest cluster of putative cellulosomal components in the genome. Each of these components must be cloned and characterized in order to understand why these genes are clustered together. Looking
beyond the characterization of the Cluster X enzymes, many more cellulase system components need to be studied. There are roughly forty cellulosomal components that are uncharacterized in the genome. Most of these components can be used in bioinformatics to determine the putative function of the expressed protein; however, around twelve of the predicted proteins have totally unknown domains and functions. These proteins could be enzymes or structural components of the cellulosome that are necessary for efficient biomass conversion to ethanol. Eventually all of the cellulosome (and cellulase system) components will need to be characterized to understand how \textit{C. thermocellum} degrades cellulose so efficiently.

Cellulosic ethanol will become a key alternative energy source as the price of oil based energy products increases. The amount of market share in the energy industry that cellulosic ethanol will eventually control is directly tied to the efficiency of the biomass to ethanol process. The research presented in this thesis is the first step in understanding how the paradigm ethanogenic bacterium, \textit{C. thermocellum}, controls the expression of its cellulase genes. By understanding and eventually controlling the expression of the cellulase genes, it will be possible to overcome a bottleneck (the degradation of cellulose into fermentable sugars) in the biomass to ethanol conversion process. The work presented will be used as a building block by future researchers to engineer a more efficient biomass conversion to ethanol process.