2013

COMPUTATIONAL PREDICTION OF THE SPORULATION NETWORK IN CLOSTRIDIUM THERMOCELLUM

Changyi Jiang
Michigan Technological University

COMPUTATIONAL PREDICTION OF THE SPORULATION NETWORK IN *CLOSTRIDIUM THERMOCELLUM*

By

Changyi Jiang

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Chemical Engineering

MICHIGAN TECHNOLOGICAL UNIVERSITY

2013

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Chemical Engineering.

Department of Chemical Engineering

Thesis Advisor: *Wen Zhou*

Committee Member: *Xiaoqing Tang*

Committee Member: *Timothy C Eisele*

Department Chair: *S. Komar Kawatra*

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my great appreciation to Dr.Wen Zhou, my research advisor, for his valuable and constructive suggestions during the planning and development of my research work.

I would like to thank to Dr.Xizeng Mao and Dr.Qin Ma at the University of Georgia, for their advice and support. I would also like to extend my thanks to my entire research group for their help.

Finally, I wish to thank my parents for their support and encouragement throughout my study

Abstract

Sporulation is a process in which some bacteria divide asymmetrically to form tough protective endospores, which help them to survive in a hazardous environment for a quite long time. The factors which can trigger this process are diverse. Heat, radiation, chemicals and lacking of nutrition can all lead to the formation of endospores. This phenomenon will lead to low productivity during industrial production. However, the sporulation mechanism in a spore-forming bacterium, *Clostridium theromcellum*, is still unclear. Therefore, if a regulation network of sporulation can be built, we may figure out ways to inhibit this process. In this study, a computational method is applied to predict the sporulation network in *Clostridium theromcellum*. A working sporulation network model with 40 new predicted genes and 4 function groups is built by using a network construction program, CINPER. 5 sets of microarray expression data in *Clostridium theromcellum* under different conditions have been collected. The analysis shows the predicted result is reasonable.

Chapter 1 Introduction

1.1 Background

Sporulation is a phenomenon in which some bacteria can form a small, tough, protective and metabolically dormant endospore . This process often takes hours, which is quite different from other adaptive responses in bacteria. Members of Bacilli and Clostridia class can form tiny tough endospores which can help them survive in hostile environment, such as heat, radiation, lacking nutrition or strongly acidic and alkali condition. There are lots of factors that might be contributed to the toughness of these spores, for example, the dehydration of the spore core and compaction of chromosomal DNA. [1] The spores are quite different from the growing cell in shape. Because the division is asymmetric, two sized cells form. The smaller one is called the forespore，and the larger one is called the mother cell. It is believed that there is an information exchange system between the mother cell and the spore, so that they can coordinate with each other and work well.

Bacillus subtilis is a kind of gram-positive spore-bearing bacterium which exists widely in nature. It is rod-shaped and has a strong enzyme activity. There are many reasons for this kind of bacteria to be applied in food, enzyme industry, aquaculture and some other biosynthesis related production. High growth rate , capacity to secrete proteins into the extracellular medium and its safety proved by Food and Drug Administration make it an attractive industrial bacteria.

Clostridium acetobutylicum is another kind of bacteria which is of great commercial value. It can produce butanol, propionic acid and ether by digesting not only sugar but also whey, starch and cellulose. *Clostridium acetobutylicum* needs a anaerobic conditions to grow. It can only survive for several hours in an oxygen-enriched environment. Normally, it will produce spores which can survive for a few years to deal with such situation.

Two different developmental processes, which are temporal change and cellular differentiation, are involved at the same time. In this case, the sporulation process attracted many biologists' interest. As two important industrial bacteria, *Bacillus subtilis* and *Clostridium acetobutylicum* have been paid more attention than other bacteria. This study describes the process of how to build a working network of sporulation in *Clostridium thermocellum* in a computational way. Three key steps are involved during this process, namely⑴ the building of template based on the biological information form known organisms and known information about the target network,⑵ prediction of probably related genome in the target network and⑶ mapping the template model to the target network. In this thesis, the resource biological information of a template is from *Bacillus subtilis sp 168* and *Clostridium acetobutylicm* ATCC 824 and the target genome is from *Clostridium thermocellum* ATCC 27405. Chapter 1 is the introduction part. A literature review is present in Chapter 2. Chapter 3 introduces methods and the result, and the results and discussion are covered in Chapter 4. Chapter 5 is about conclusions and future work.

1.2 Research significance

The sporulation process of *Bacillus subtilis* are of operability and simplicity, which make it a great model for the study of development in prokaryote cell. The research of endospore-forming is not only of great significance for basic science, but also has a full potential for industrial application. However, the mechanism of sporulation in Clostridia is not understood as well as that in *Bacillus subtilis*. Actually, researchers hold different opinions on this issue, which is still under the development. Even the way of phosphorylation of master gene spo0A in clostridia has several hypotheses.

The computational prediction of sporulation regulatory network is based on the known information in *Clostridium acetobutylicum* and *Bacillus subtilis* is a possible way to accurately infer the biological network in target organism. It supplies a new method for studying the complex network in a certain organism which is not deeply engaged. Also, it can provide a new direction for future research.

Besides, the using of CINPER (CSBL INteractive Pathway BuildER) simplfies the prediction process. It handles the P-Map and BLAST mission in background processing , which automates the mapping the initial model section. An easier interface also enables users to learn the operation steps fast and deal with the data in a more systematic manner.

Chapter 2 Literature Review

2.1 Computational prediction

It is an attractive and challenging mission for biologists to predict the regulatory network for some certain organism using multiple template pathways .The appearance of large-scale omic data, and modern calculation devices makes it feasible for people to computationally infer a working model for an organism in a systematic manner.

This idea has been applied to predict a model of the osmoregulation network in response to hyperosmotic stress of *Synechococcus* sp strain WH8102. By using comparative genome analyses and computational prediction, key transporters, synthetases, signal sensor proteins and transcriptional regulator proteins are identified. [2]

This prediction process of osmoregulation network is consist of several steps including the following[9]:

1 Build the template networks.

2 Map the genes from template to build the initial model of target organism.

3 Expand the initial model.

4 Validation and refinement

5 Analyze the result.

2.1.1 Build the template model

Before building the template, it is necessary to indentify the typical components of osmoregulation process. Under hyperosmotic stress, Na^+ inside the cell is released while the K^+ will be taken into the cell. Besides, some compatible osmolytes can also be taken into the cell or synthesized inside the cell to be a substitute for K^+ .[2] By related literature searching ,63 genes are found to be involved with this process in five species. For example, *Aphanothece halophytica* has 3 genes involved with encoding a Na^+/H^+

exchanger. This integral membrane protein can export the Na^+ out of the cell. [3,4]Also, a two-component regulatory system serves as an osmotic stress sensor[5] and a transporter for uptaking betaine[6] in *C.glut*. Actually, it also does exist in other bacteria and perform a similar function. There are also 31 genes involved in this two-component system of *E.coli*.

Such data are used as the templates information and mapped into WH8102.

2.1.2 Map the genes from template to build the initial model of target organism

Two methods are used to map the template network into an initial model, which are P-MAP and BLAST. The existing methods mainly rely on sequence-based orthologous gene mapping[7], which means there would be something missing in the mapping result because only sequence-similarity information is not enough. However, P-MAP method uses not only sequence-similarity to map a template network into a target genome but also operon information to map a template network onto a target genome. The mapping process is finding the orthologous gene of the template in the target genome. When the target sequence-similarity information and genomic structure (operons and regulons) considered, the pathway-mapping accuracy could be greatly improved over the methods that used sequence-similarity information alone.

The other mapping method, PSI-BLAST (Position-Specific Iterated Basic Local Alignment Search Tool) can run at a speed about three times than the original one. This method is introduced for automatically combining statistically significant alignments produced by BLAST into a position-specific score matrix, and searching the database using this matrix.[8] Except for the mapping speed, another advantage is that PSI-BLAST is much more sensitive to weak but biologically relevant sequence similarities. Therefore, it is a greatly improved method for mapping compared with the original BLAST.

Also, there would have been a situation that multiple genes from different organisms in the template network provided are mapped into the same gene in the target genome. If so, the gene with a closer evolutionary relationship would be chosen.

2.1.3 The expanded model

The initial network model needs to be expanded based on co-location, co-regulation, and co-evolution information, so that a final working regulation model can be derived.

The basic idea of the expansion is that if protein A is in the initial model but B is not, we will consider adding B to the model if A and B are related based on the analyses.[7]

It is notorious that genes in the same operon are functionally related. Hence, new genes can be added into the model, if they share the same operon with the initial ones. However, this kind of prediction is not strong enough. Further experimental data are needed to validate the prediction.

Another way to expand the initial model is expanding based on protein-protein interactions information. If one protein can form a protein complex with the one which is already in the initial model, it will be added.

The third method is expanding the model based on regulon information. A global regulator is needed so that related genes can be added based on the orthology mapping from an original organism to the target one.

2.1.4 Validation and refinement

Three methods are applied to validate the prediction. Firstly, related works of literature are checked to confirm the predicted network. This network could be a regulation network. It depends what kind of pathway network researcher would like to focus on. Whole-genome microarray gene expression data is another way to validate the accuracy of the prediction. The third way is using the protein domain architecture information from public databases.[2]

Related literature research work can validate the prediction result if the predicted model is highly consistent with the experimental result. And also, predicted genes can be checked against the microarray dataset. Genes which show different expressions under different conditions are collected .

Conservation information of protein domains is also an important tool to check the genes pairs predicted. The genes pairs here indicate the relevant genes from the original and target genome mapped by the P-MAP algorithm. It is believed that true orthologous genes from two related genome should have the same architecture.[2] Checking the protein domains of genes pairs can demonstrate if the previous mapping work did well or not. Several pairs among predicted result may not have a good protein domain compared result, which may indicate that these genes are not correctly mapped.

2.2 CINPER: an interactive web system for pathway prediction for prokaryotes

CINPER is a web-based network-construction system, which is short for Computational System Biology Laboratory INteractive Pathway BuildER. It can provide a user interface to build a network model for a prokaryotic organism in an intuitive manner. The prediction process follows four steps as in Figure 2.1:

1) Collection of template networks based on known pathways of related organism(s) from the SEED or BioCyc database and the published literature.

2) Construction of the initial network model based on the template networks using the P-Map program.

3) Expansion of the initial model, based on the association information derived from operons, protein-protein interactions, con-expression modules and phylogenetic profiles.

4) Computational validation of the predicted model based on gene expression data.[9]

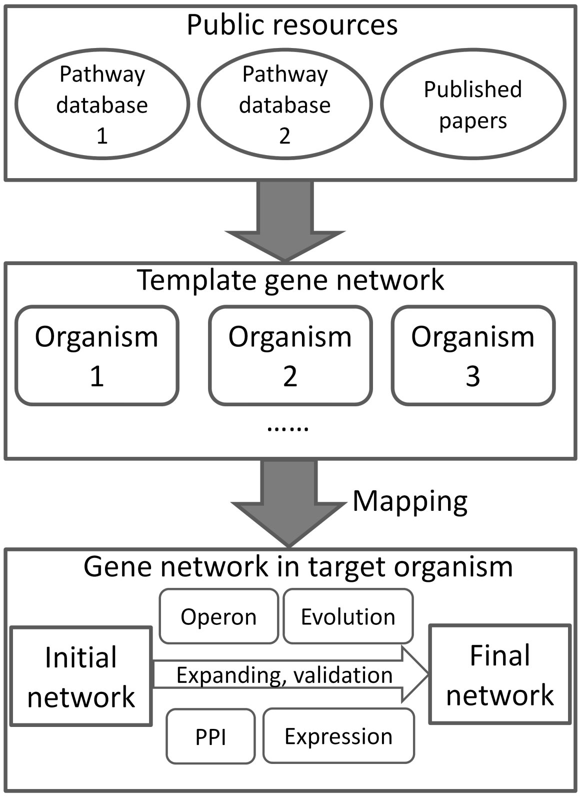

Fig 2.1 A diagram of workflow of CINPER[9]

We might face various problems when using the traditional methods, such as the fragmented result or missing pieces in the mapped model. CINPER has a built -in P-MAP and PSI-BLAST program, which will solve such an information-insufficient problem by considering both location and sequence similarity information. And also, CINPER can help a user finish this process fast and neatly. The website will provide the user a step by step wizard, so that the user can easily create and build a pathway model for prokaryotes. Several types of information resources are considered, including homologous template pathways of related organisms in literature and public database, functional annotation of genes, predicted operons, protein-protein interactions, phylogenetic profiles and gene-expression data for the target genome. [9]All this information is gathered and input by the user and CINPER can automate a manual prediction process. [2,10-12] The expanding and revising process of the model work in an interactive way so that the user can edit the incorrect pathways and add more biological information when desired. Also, CINPER will show the predicted result through its graphical user-interface. The position of each gene can be relocated, and the initial evidence of the predicted result can be easily found from the table displayed in the final modeling result. It is a more intuitive way for a user to search for additional information or revise the model. Gene expression data, for example, microarray dataset of the target genome can also be used by CINPER to assess the consistency between the final working network and gene-expression data, which would give more information for a user to revise their modeling process.

Also, CINPER has various tools for deriving information from the public databases such as RefSeq, KEGG, SEED,DOOR and STRING. With the help of these build-in tools, the gathering functional and interaction information process from the public databases would be automatic, which save much time and makes the modeling work more efficient.

To assess the performance of CINPER, *E.coli* and *Bacillus subtilis* are used as template and target organisms respectively, since these two bacteria are both well studied. Precision and recall are two methods used to quantize the performance, which are defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

TP is the number of template genes that are mapped to the known target genes. FN is the number of template genes that cannot be mapped by P-MAP or have no matched target genes and FN is the number of missed target genes by P-MAP.[9] The results show that the overall precision rate is 90% and recall rate is 76% on 17 well-studied pathways in the MetaCyc Date, which means that the initial model is well built. After template expanded, the results of the final model show that the overall precision rate(87%) raise a little higher and the recall rate declines to 28%. It may due to the pathway models in MetaCyc are far from being complete.

Previous work has been done to predict iron homeostasis network in *synchocystis* PCC6803 by using CINPER. 27 template genes in four organisms, *Sinorhizobium meliloti,Escherichiacloli* K12, *Prochlorococcus marinus* MED4 and *Synechocystis* PCC6803, are added into the initial model. The predicted regulation process is about iron homeostasis network.57 genes in *synchocystis* PCC6803 involved this regulation process are predicted by CINPER. Two Fur-like transcription regulators predicted has been confirmed by a published paper.[13] The prediction result also have been validated by public transcript omic data under iron limitation conditions.[14.15] One gene from the initial model and two genes from protein-protein interactions are experimentally verified.

2.3 Recent research work

The main process of sporulation can be divided into five stages. Figure 2.2 shows a simplified sporulation cascade in *Bacillus subtilis* . The main events will be described in the next several sections.

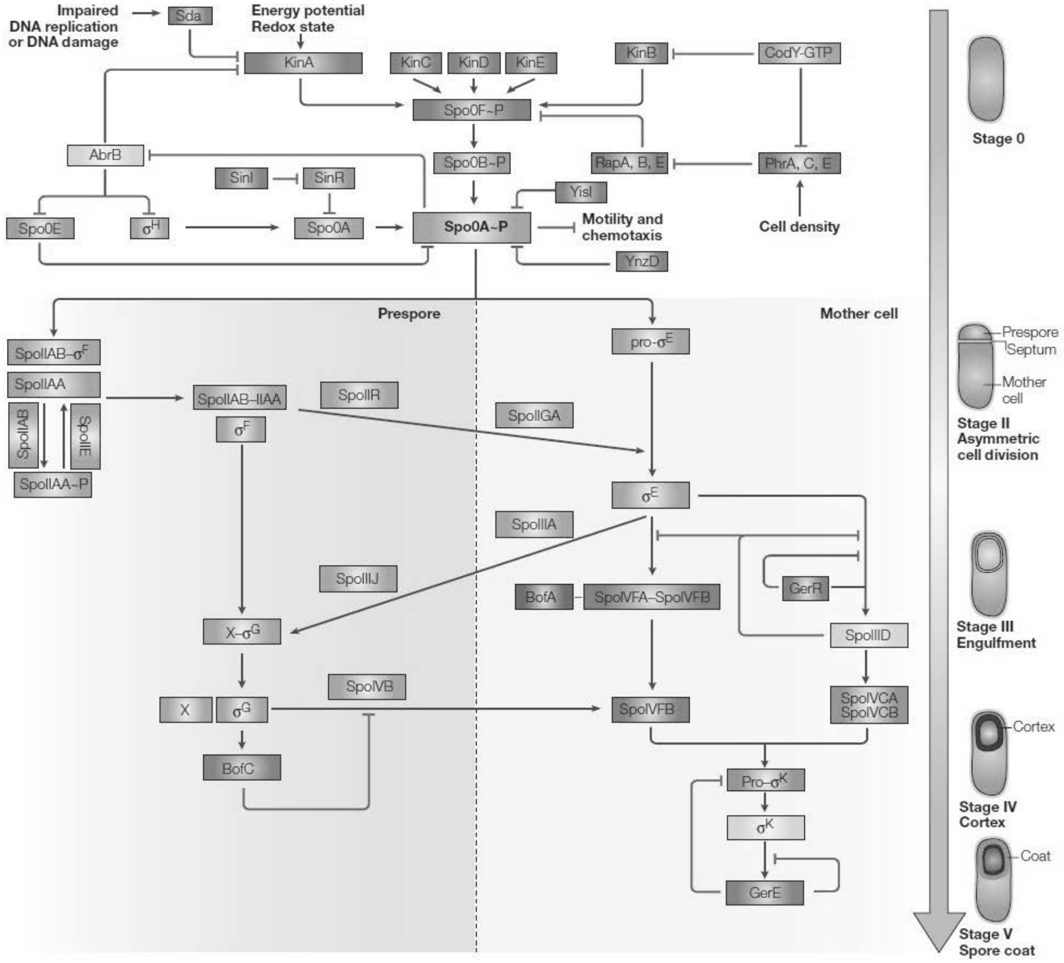

Fig 2.2 The sporulation cascade in *Bacillus subtilis* and selected clostridia[21]

2.3.1 Main events during the sporulation

Bacillus subtilis will initialize the sporulation process by phosphorylating the master regulator Spo0A　. Five histidine kinases(KinA, KinB, KinC, KinD and KinE) work as sensors, which respond to hazards from the intracellular or extracellular environment.[16]These kinases will phosphorylate the phosphotransferase Spo0F and the single-domain response regulator Spo0B. Then the phosphoryl group will be transferred to the master gene *spo0A*,which plays an important role over the whole sporulation process. This phosphorylation process is related to these histidine kinases, which means, KinA and KinC can directly phosphorylate Spo0A or its mutants, even though the efficiency is quite low. The other kinase KinB, however ,cannot phosphorylate Spo0A. All five kinds of phosphorly kinases are orphan kinases. Here, the orphan kinases mean they do not have an adjacent response regulator. Three of them are related to the function of environmental sensing. The other two are not. For *Clostridium acetobutylicum*, there are 35 histidine kinases. Six of them are orphan. Among the orphan kinases, three of them are involved with the environment sensing. Two of them are not. The last one is CheA (chemotaxis histidine kinase A), which is related to the chemotaxis system. Microarray expression data of *Clostridium acetobutylicum* shows 4 kinases (CAC 0437,CAC 0323,CAC2730 and CAC 0903) has a clear correlation, with the kinases in *Bacillus subtilis*, which indicates they might be phosphorylate Spo0A. And then the phosphorylated Spo0A protein(Spo0A~P) seems to regulate the sporulation process in all kinds of clostridia. However, there exists differences in the sporulation between different bacillus and clostridium. The components of the *Bacillus subtilis* sporulation phosphorelay are not identified in the *Clostridium acetobutylicum*. After the master gene Spo0A is phosphorylated, the Spo0A~P can down regulate the expression of Abrb, which will lift the Abrb restriction on Spo0H, and increase the expression of sigma factor H (σ^H) and meanwhile, the expression of Spo0A will also increase. The next event is asymmetric cell division, and it is also the first time shape-structure change occurs over the whole spore-forming process. The genes needed in symmetric cell division are also needed (PbpB may be ruled out). But there are several differences. Firstly, a chromosome is needed to form axial filament when asymmetrical cell division occurs. Secondly, the division device is not located in the middle of the cell. Thirdly, the septum is thicker and with fewer peptidoglycan.

After the asymmetric cell division and before the prespore gets a whole chromosome, sigma factor F will be activated. SpoIIE is a phosphatase. It will dephosphoralate and active SpoIIAA (anti-anti-sigma factor). The activated spoIIAA will apply on spoIIAA-σ^F complex and release σ^F. The activation of sigma factor E in the mother cell will follow the activation of sigma factor F in the prespore side.[17]It is believed that protein SpoIIR should work as a signal transferring media. The product of SpoIIR will interactive with SpoIIGA, which will make pro-σ^E into activated σ^E form. Biological chips analysis shows that σ^E directs the expression of 253 genes in 157operons.[18] It indicates that, the way that genes being expressed has been greatly changed when the mother cell is under the control of sigma factor E.

The activation of σ^F and σ^E initializes the modified septum, and the splitting of peptidoglycan on it. Two cells do not separate like normal cell division. They act like that the large mother cell engulfs the little prespore. Three proteins, SpoIID, SpoIIM and SpoIIP are involved in this process. Their function may be related to prevent the second time cell division. All of them are synthesized in the mother cell. For SpoIIQ, another necessary engulfment related protein, is synthesized in the prespore. It will be inserted into the septum and kept there over the whole engulfment process. However, the function of it is still unclear.

The encoding genes of sigma factor G and K are transcribed by the RNA polymerase formed by sigma factor F and sigma factor E respectively. Sigma factor G exists before the engulfment, but it is inactive. The activation of sigma factor G need SpoIIIJ, SpoIIIA and SpoIIIA which can only be transcribed in the mother cell.[12]Sigma factor K is the last one to be activated . Like sigma factor E, sigma factor K exists as an inactive precursor. After receiving a signal protein SpoIVB transcript by the sigma factor G in the prespore, a SpoIVFB,SpovIVFA and BofA combined complex will be broken and release the SpoIVFB. Then, sigma factor will be activated once the restriction of SpoIVFB is removed.

2.3.2 Chemotaxis and motility

Chemotaxis and motility are necessary abilities for bacteria to survive. This phenomenon that bacteria direct their movements due to the oncentration of a certain chemical is very important when they are finding food or avoiding the hostile environments. Research shows that the motility and chemotaxis machinery of the Bacillus and Clostridium are similar[19], and they are both related to the sporulation process in *Bacillus subtilis*. Also, recent gene expression information indicates that chemotaxis genes are directly negatively regulated by Spo0A[20] and it seems that this regulation also works on *Clostridium acetobutylicum*.

Chapter 3 Methods

3.1 Data

The program used to build the model is CINPER. (http://csbl.bmb.uga.edu/CINPER/)

All the sequences, microarray data and pathway information were retrieved from NCBI(http://www.ncbi.nlm.nih.gov/) ,BioCyc(http://biocyc.org/) ,KEGG(http://www.kanehisa.jp/), DOOR(http://csbll.bmb.uga.edu/OperonDB_10142009), SEED (ftp://ftp.theseed.org/genomes/SEED/) and STRING database (http://string-db.org/).[2]

3.2 Modeling workflow

3.2.1 Template building and mapping

Through a literature search [21],55 genes in two organisms were collected to build the initial template, including 42 genes in *Bacillus subtilis* and 13 genes in *Clostridium acetobutylicum*. Compared to the sporulation research on *Bacillus subtilis*, there were less study focused on *Clostridium acetobutylicum*. Therefore, only a few of the genes and pathways were added into the initial template. These 13 genes were all related to the sigma factor pathways. 69 interactions were summarized and added into the model. 55 *Bacillus subtilis* pathways were derived from literature and BioCyc database.14 *Clostridium acetobutylicum* pathways were mainly about the interactions between sigma factors.

The actual operation process on CINPER is consists of three steps. Firstly, by clicking the New template button, a new template was created. Then, genes and pathways are added through the "add genes" and "add interactions" option. The last step is examining the added information and revising the result.

After the initial template was built, use the "derive a new model" under the "Go! Model" menu to map the initial template into the target organism, which is *Clostridium thermocellum* ATCC 27405. The E-value used for the mapping method,which is P-MAP ,is less than 10^{-5}.

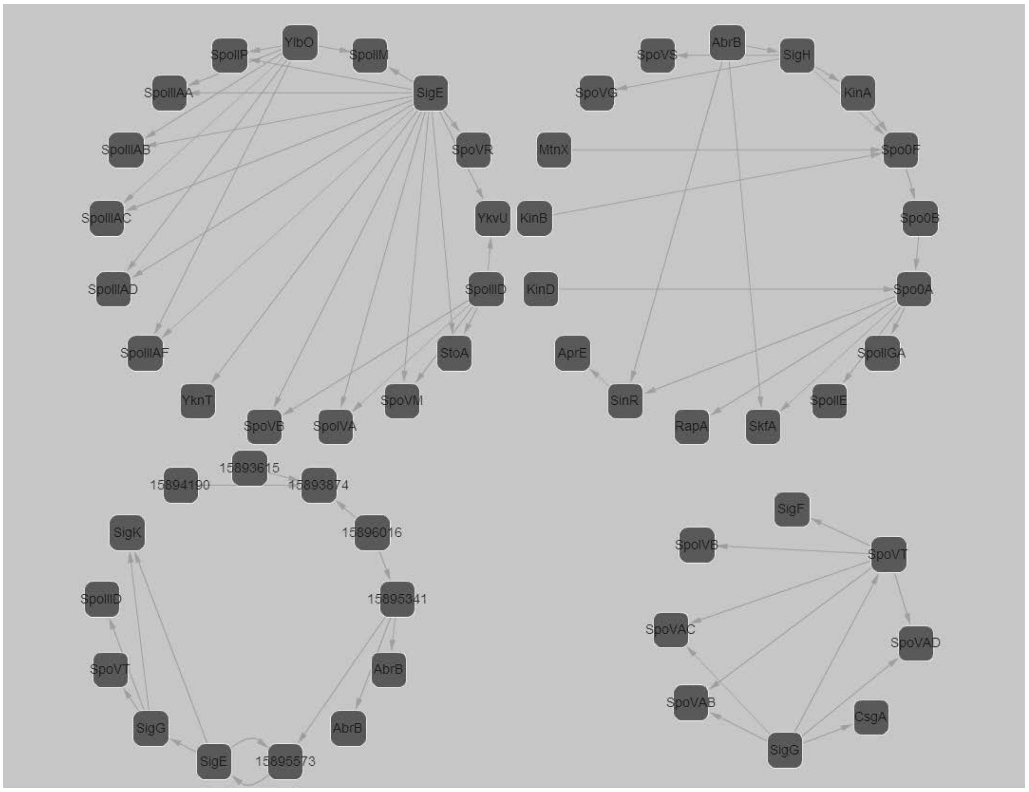

Figure 3.1 Initial template without mapping

Figure 3.1 is a display of the initial template which has not been mapped into the target organism. There are 4 separate sections, and Spo0A is a very important node which has many connections with other genes.

After mapped into the *Clostridium thermocellum*, as we can see in Figure 3.2 that all the genes are connected. It should be noted that there are two arrows pointed the same direction   between two genes such as Cthe_1287 and Cthe_0812, which may be due to the reason that two kinds of evidence can both prove this pathway.

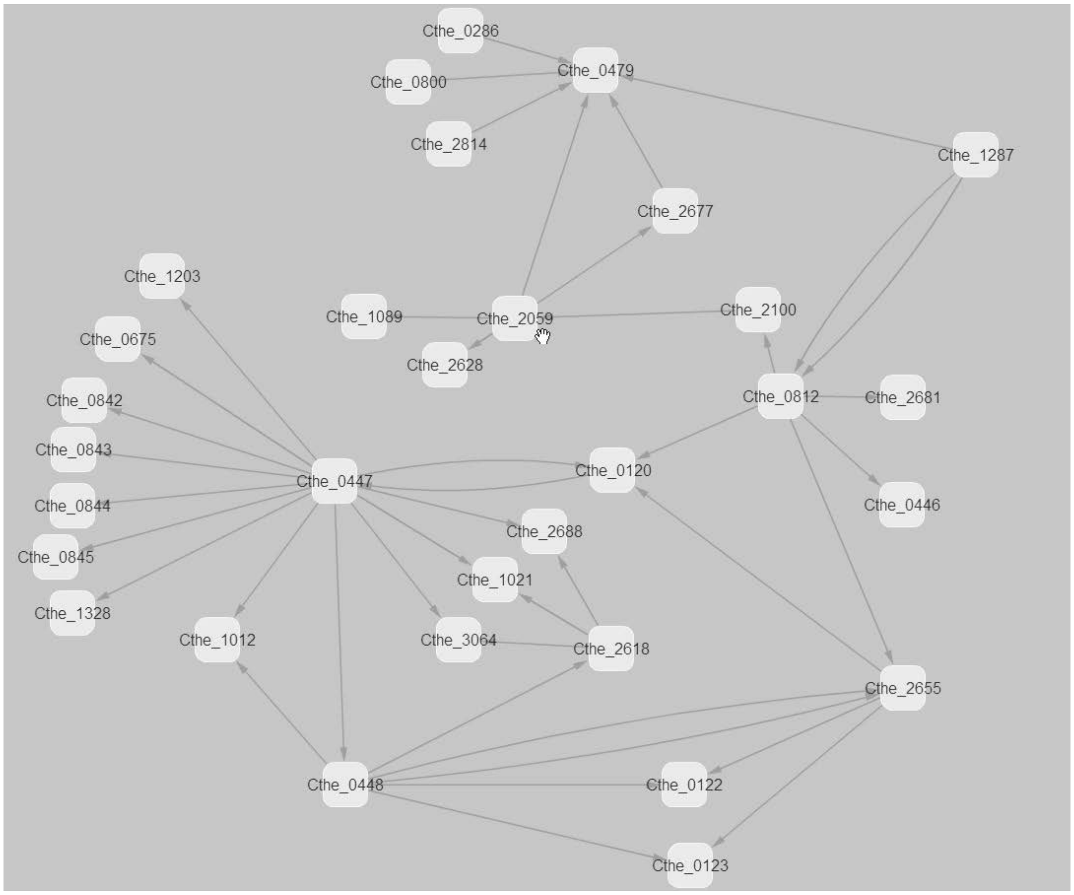

Figure 3.2 A screen shot of CINPER User-Interface (Final network model)

3.2.2 Initial model expanding

The initial model can be expanded by the following:

1. Expand by co-location information

2. Expand by co-regulation information

3. Expand by co-evolution information

The core principle of expanding is adding the missing protein based on the "co-" information analyses. If two proteins could be found of "co-relationship", one is in the initial model, and the other one is not, then, the missing protein would be added to supplement and perfect the model.

It is believed that genes in the same operon are always linked or correlated. For instance, genes sharing the same operon may all related to form a protein or transfer a signal. Based on the point of view above, 22 genes were added into the model by the operon expanding.

The operon information (co-location) used was retrieved from DOOR database. And the regulation and evolution information was provided by STRING database for users to search and add related genes. STRING database contains plenty of information about protein-protein interaction , co-expression and co-regulation in 534 bacterial genomes.[22] For actual operation, there is pull down menu under the "Expand" option for users to choose which database to use. The selectable options are DOOR operon database, STRING functional relation database or uploading users' own regulon information.

Table 3.1 Expanded result

Gene	Symbol	Function	Stage
Cthe_0118	Cthe_0118	anti-sigma-factor antagonist	operon
Cthe_0119	Cthe_0119	anti-sigma F factor ; K06379 stage II sporulation protein AB (anti-sigma F factor)	
Cthe_0121	Cthe_0121	hypothetical protein	
Cthe_0126	Cthe_0126	CheC-like protein	
Cthe_0287	Cthe_0287	GAF sensor hybrid histidine kinase ; K00936	
Cthe_0690	Cthe_0690	hypothetical protein	
Cthe_0799	Cthe_0799	two component transcriptional regulator	
Cthe_0974	*murG*	undecaprenyldiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase	
Cthe_0976	*mraY*	phospho-N-acetylmuramoyl-pentapeptide-transferase	
Cthe_0977	Cthe_0977	UDP-N-acetylmuramoyl-tripeptide--	

		D-alanyl-D-alanine ligase	
Cthe_0978	Cthe_0978	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase	
Cthe_1011	Cthe_1011	peptidoglycan glycosyltransferase	
Cthe_1070	Cthe_1070	metal dependent phosphohydrolase; K07037	
Cthe_1071	Cthe_1071	PhoH-like protein	
Cthe_1073	Cthe_1073	hypothetical protein	
Cthe_1204	Cthe_1204	hypothetical protein	
Cthe_1205	Cthe_1205	putative serine protein kinase, PrkA; K07180 serine protein kinase	
Cthe_1288	Cthe_1288	two component transcriptional regulator	
Cthe_1916	Cthe_1916	two component transcriptional regulator	
Cthe_2105	Cthe_2105	DNA polymerase III subunit delta' ; K02341 DNA polymerase III subunit delta'	
Cthe_2378	Cthe_2378	chromosome segregation DNA-binding protein; K03497 chromosome partitioning protein, ParB family	
Cthe_2813	Cthe_2813	two component transcriptional regulator	
Cthe_0091	Cthe_0091	peptidoglycan glycosyltransferase ; K05515 penicillin-binding protein 2	String
Cthe_0444	Cthe_0444	cell division protein FtsA; K03590	

		cell division protein FtsA	
Cthe_0445	Cthe_0445	cell division protein FtsZ; K03531 cell division protein FtsZ	
Cthe_0466	*fliG*	fliG; flagellar motor switch protein G; K02410 flagellar motor switch protein FliG	
Cthe_0472	Cthe_0472	flagellar hook capping protein; K02389 flagellar basal-body rod modification protein FlgD	
Cthe_0490	Cthe_0490	CheA signal transduction histidine kinase; K03407 two-component system, chemotaxis family, sensor kinase CheA	
Cthe_0492	Cthe_0492	CheC, inhibitor of MCP methylation; K03410 chemotaxis protein CheC	
Cthe_0895	Cthe_0895	RNA polymerase sigma factor RpoD; K03086 RNA polymerase primary sigma factor	
Cthe_1095	Cthe_1095	cell divisionFtsK/SpoIIIE	
Cthe_1321	Cthe_1321	chaperone protein DnaJ; K03686 molecular chaperone DnaJ	
Cthe_1322	*dnaK*	dnaK; molecular chaperone DnaK; K04043 molecular chaperone DnaK	
Cthe_2163	Cthe_2163	anti-sigma-factor antagonist	
Cthe_2284	Cthe_2284	CheA signal transduction histidine kinase; K03407 two-component system, chemotaxis family, sensor kinase CheA	

Cthe_2367	Cthe_2367	rotein translocase subunit yidC; K03217 YidC/Oxa1 family membrane protein insertase	
Cthe_2371	*dnaA*	dnaA; chromosomal replication initiation protein; K02313 chromosomal replication initiator protein	
Cthe_3039	Cthe_3039	cell divisionFtsK/SpoIIIE	
Cthe_3047	Cthe_3047	peptidoglycan glycosyltransferase	
Cthe_3087	Cthe_3087	response regulator receiver protein	

Table2 shows 40 new genes were added into the initial sporulation model. 22 of them were expanded by operon.18 were expanded by STRING evidence.` After finishing the expanding process, we got the final model of sporulation in *Clostridium thermocellum.*

3.2.3 Validation

Two methods were used to validate the final modeling result. One is comparing it with recent literature related to the sporulation in *Clostridium thermocellum*. Another way was using whole-genome microarray gene expression data of *Clostridium themorcellum*. Through comparing the expression patterns of the *Clostridium themorcellum* under different conditions, consistency between the differential expression of microarray data and modeling result was found. Yet it may not be compelling evidence, since the expression data can validate the predicted genes somehow. Five sets of microarray data were collected by searching the NCBI database. (www.ncbi.nlm.nih.gov/geo/)

Then, they were combined, and an expression matrix for *Clostridium thermocellum* was generated. 3455 genes were included in this matrix no matter if they were related with sporulation. 74 expanded genes predicted by CINPER were summarized and listed as query genes. All of the following analysis was based on these two data sets.

Chapter 4 Results and Discussion

4.1Template network

As mentioned in Chapter 3, 55 genes and 69 interactions in two species were summarized and added into the template for mapping. The components are listed in Table4.1.

Table 4.1 Components in the initial template

Organism	Gene	Symbol	Definition
B.subtilis	BSU00370	*abrB*	transition state regulatory protein AbrB

	BSU00560	*spoVT*	stage V sporulation protein T
	BSU00980	*sigH*	RNA polymerase sigma-H factor
	BSU13600	*mtnX*	2-hydroxy-3-keto-5-methylthiopentenyl-1-phosphate phosphatase
	BSU13660	*kinD*	sporulation kinase D; K13532 two-component system, sporulation sensor kinase D
	BSU13990	*kinA*	sporulation kinase A
	BSU15090	*ylbO*	hypothetical protein
	BSU15320	*sigE*	RNA polymerase sigma-E factor; K03091 RNA polymerase sporulation-specific sigma factor
	BSU15330	*sigG*	RNA polymerase sigma-G factor; K03091 RNA polymerase sporulation-specific sigma factor
	BSU24220	*spo0A*	Stage 0 sporulation protein A; K07699 two-component system, response regulator, stage 0 sporulation protein A
	BSU24610	*sinR*	HTH-type transcriptional regulator SinR
	BSU27930	*spo0B*	sporulation initiation phosphotransferase B; K06375 stage 0 sporulation protein B (sporulation initiation phosphotransferase)
	BSU31450	*kinB*	sporulation kinase B ; K07697 two-component system, sporulation sensor kinase B

	BSU36420	*spoIIID*	stage III sporulation protein D; K06283 putative DeoR family transcriptional regulator, stage III sporulation protein D
	BSU37130	*spo0F*	sporulation initiation phosphotransferase F; K02490 two-component system, response regulator, stage 0 sporulation protein F
	BSU01910	*skfA*	sporulation-killing factor SkfA
	BSU23420	*spoVAC*	stage V sporulation protein AC; K06405 stage V sporulation protein AC
	BSU23410	*spoVAD*	stage V sporulation protein AD; K06406 stage V sporulation protein AD
	BSU24230	*spoIVB*	spoivb peptidase; K06399 stage IV sporulation protein B
	BSU23430	*spoVAB*	stage V sporulation protein AB; K06404 stage V sporulation protein AB
	BSU23450	*sigF*	RNA polymerase sigma-F factor; K03091 RNA polymerase sporulation-specific sigma factor
	BSU16980	*spoVS*	stage V sporulation protein S; K06416 stage V sporulation protein S
	BSU00490	*spoVG*	septation protein SpoVG; K06412 stage V sporulation protein G
	BSU25530	*spoIIP*	stage II sporulation protein P; K06385 stage II sporulation

			protein P
	BSU24430	*spoIIIA A*	stage III sporulation protein AA; K06390 stage III sporulation protein AA
	BSU24420	*spoIIIA B*	stage III sporulation protein AB; K06391 stage III sporulation protein AB
	BSU24410	*spoIIIA C*	stage III sporulation protein AC; K06392 stage III sporulation protein AC
	BSU24400	*spoIIIA D*	stage III sporulation protein AD; K06393 stage III sporulation protein AD
	BSU24380	*spoIIIA F*	stage III sporulation protein AF; K06395 stage III sporulation protein AF
	BSU23530	*spoIIM*	stage II sporulation protein M; K06384 stage II sporulation protein M
	BSU27670	*spoVB*	stage V sporulation protein B; K06409 stage V sporulation protein B
	BSU22800	*spoIVA*	stage IV sporulation protein A; K06398 stage IV sporulation protein A
	BSU15810	*spoVM*	stage V sporulation protein M; K06414 stage V sporulation protein M
	BSU14250	*yknT*	sporulation protein cse15; K06437 sigma-E controlled sporulation

			protein
	BSU13840	*stoA*	sporulation thiol-disulfide oxidoreductase A
	BSU13830	*ykvU*	sporulation protein YkvU
	BSU09400	*spoVR*	stage V sporulation protein R; K06415 stage V sporulation protein R
	BSU02070	*csgA*	sigma-G-dependent sporulation-specific SASP protein
	BSU12430	*rapA*	response regulator aspartate phosphatase A; K06359 response regulator aspartate phosphatase A (stage 0 sporulation protein L)
	BSU00640	*spoIIE*	stage II sporulation protein E ；K06382 stage II sporulation protein E
	BSU15310	*spoIIGA*	sporulation sigma-E factor-processing peptidase
	BSU10300	*aprE*	subtilisin E
C.acetobutylicum	CA_C0585	*cheY*	chemotaxis protein CheY
	CA_C1689	*sigK*	sporulation sigma factor SigK
	CA_C2859	*spoIIID*	stage III sporulation protein D; K06283 putative DeoR family transcriptional regulator, stage III sporulation protein D
	CA_C3649	*spoVT*	stage V sporulation protein T

	CA_C3647	*abrB*	transition state regulatory protein AbrB; K06284 transcriptional pleiotropic regulator of transition state genes
	CA_C0310	*abrB*	stationary/sporulation gene regulator; K06284 transcriptional pleiotropic regulator of transition state genes
	CA_C0323		sensory transduction histidine kinase
	CA_C0903		sensory transduction histidine kinase
	CA_C1695	*sigE*	sporulation sigma factor SigE; K03091 RNA polymerase sporulation-specific sigma factor
	CA_C1696	*sigG*	sporulation sigma factor SigG; K03091 RNA polymerase sporulation-specific sigma factor
	CA_C2071	*spo0A*	K07699 two-component system, response regulator, stage 0 sporulation protein A
	CA_C2306	*sigF*	sporulation sigma factor SigF; K03091 RNA polymerase sporulation-specific sigma factor
	CA_C2760		membrane-associated methyl-accepting chemotaxis protein

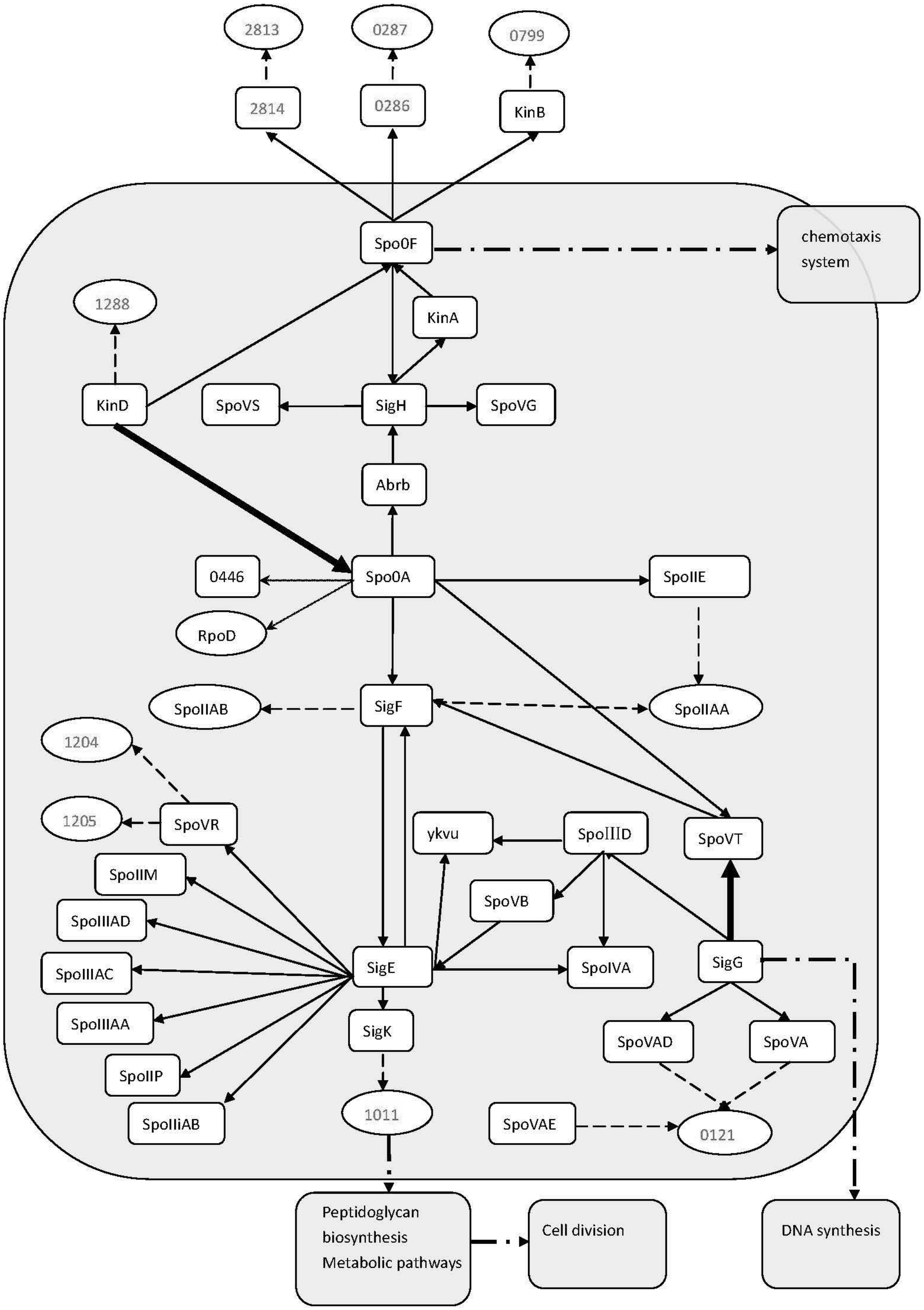

Figure 4.1 Pathway model after expanding

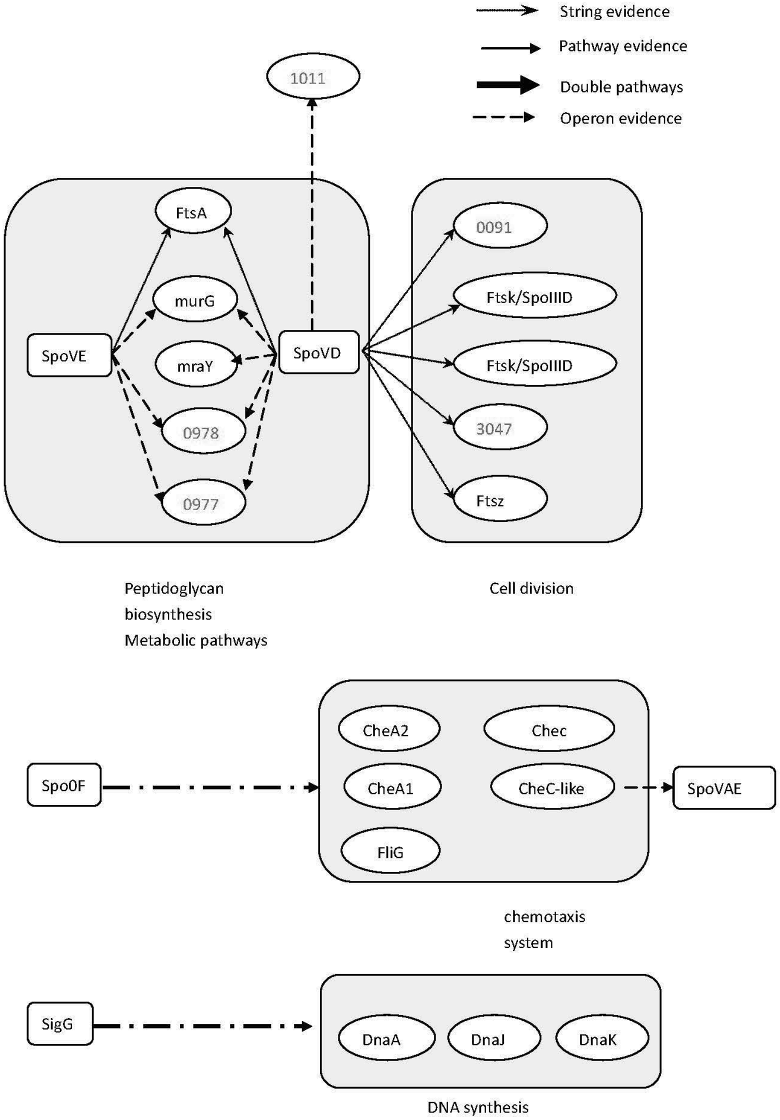

Figure 4.2 Functional groups

Figure 4.1&4.2 shows a working model of predicted sporulation network in *Clostridium thermocellum*. The round corner rectangle represents an initial gene and elliptical one means expanded genes. 4 kinds of arrows indicate that which method was used during the expanding step. The short dashed arrow means the pathway is expanded by functional relationship evidence from STRING database. The normal arrow means this pathway is mapped based on initial interaction provided. The bold arrow marked with double pathways indicates this interaction (pathway) is expanded based on two kinds of evidence at the same time, which normally is more convincing. And for the long dashed arrow, it means the pathway is expanded by the co-location analysis(operon evidence).

There are 5 sigma factors in this network. They are all linked under a certain sequence. One sigma factor is under the control of a previously activated one. It also regulates another sigma factor which will take charge during the following stage of sporulation. This working mechanism will make sure the whole sporulation occurs in the right fixed order. There will be only one sigma factor functional in a certain stage. For instance, sigma factor G will only be activated in prespores during the engulfment stage.

Four functional group are predicted in this network, which are chemotaxis system group, peptidoglycan biosynthesis metabolic pathways group, cell division group and DNA synthesis group.

4.2 Functional groups

4.2.1 Chemotaxis system

It was previously demonstrated that the chemotaxis and motility phenomena are related to sporulation. It seems plausible that some inhibitor will inhibit both sporulation and chemotactic behavior as well as another known inhibitor of chemotaxis in *Enerobacteria* [23] .

In this predicted network, chemotaxis system contains 5 genes, Cthe_0466, Cthe_0126, Cthe_2284, Cthe_0490 and Cthe_0492, which were all confirmed to exist in the *Clostridium thermocellum* pathway by searching the KEGG database. Sporulation actually is a bacterium's behavior respondingto the change of the environment. It makes sense that the chemotaxis behavior which is also an environmental response would use a similar signal transfer mechanism.

4.2.2 Peptidoglycan biosynthesis metabolic pathways & cell division

Peptidoglycan, which is also known as murein, is an important polymer that is the main component of the cell wall. Hence, the peptidoglycan biosynthesis occurs as an accompaniment of cell division, no matter if is a symmetric division in vegetative growth or an asymmetric division during the sporulation.

The first step of cell division is FtsZ protein forming a ring-like structure. At the same time, ATPase FtsA is inserted into this structure. The trigger factor of this process is still unclear. In Figure 4.2, the modeling result indicates that Peptidoglycan biosynthesis pathways and cell division pathway are both related to sporulation-specific stage V protein D (SpoVD). Recent research work [24] indicate that *Bacillus subtilis* SpoVD gene is located in the upstream of the Mur operon. PbpB, which is confirmed only one protein needed in asymmetric division, may be related with it for this reason. Since PbpB is involved with the synthesis of septal peptidoglycan, it seems that SpoVD should also be related to the process of sporulation-specific peptidoglycan synthesis. It has been proved that SpoVD has no effect on vegetative growth or symmetric division by an insertion disruption experiment.

In *Bacillus subtilis*, SpoIIID is a sporulation-specific, DNA binding protein. Its main function is activating or repressing the transcription of genes. Research shows that SpoIIID can increase the transcription of sigma factor K. And it also has an effect on the transcription of GerE by regulating the sigma factor K polymerase. In general terms, SpoIIID shows its connections with sigma factor E and K. Then prespore-specific sigma factor G will be activated following sigma factor E and K. But according to the modeling result, this relationship seems to be a little different in *Clostridium thermocellum*. SpoIIID is still related with Sigma factor E, however, the sigma factor G takes the place of sigma factor K, and it was directly connected with SpoIIID.

At the beginning of the sporulation, the histidine kinases are responsible for environment sensing and phosphorylate the master gene Spo0A to initialize the sporulation. In the modeling result, only 3 histidine kinases are predicted. It indicates that CINPER cannot find the corresponding orthology genes during the mapping step. It is probably due to the reason that *Clostridium thermocellum* can also transform into another state to avoid the hostile environments. This so called "L-form" will recover faster when the environment is proper for growing. However, it is also of a lower resistance to the bad survival conditions compared with the spore form. Under a certain condition, t *Clostridium thermocellum* prefers to turn into the L-form rather than a spore form. As a result, some histidine kinases might lose their original functions during the evolution and work as L-form condition sensors.

4.2.3 DNA synthesis

DnaA is a chromosomal replication initiator protein. DnaK is a molecular chaperone. Cthe_1321 is chaperone protein DnaJ. It makes sense for taht reason that it must be accompanied by chromosomal replication and non-covalent folding or unfolding and the assembly or disassembly of micromolecular structures during the sporulation process.

In Figure 4.2 Sigma factor G is predicted to be related with the DNA synthesis. This result is quite confusing. The sigma factor G should be inactive until the engulfment is completed, even though it has been tanscribed at the beginning of engulfment. This pathway cannot be validated until further experiments are done.

4.3 Microarray data analysis

The co-expression modules are identified only in 74 query genes. Then modules are expanded in the matrix with all the genes. The P-value can evaluate the significance of identifying such co-expressed genes in current conditions. If the P-value is less than 0.01, you can say current bicluster is statistically significant under the corresponding conditions.

74 genes are distributed into 7 modules . If the two genes are in the same module, it means that they share the same expression pattern. Since the P-values of every module are less than 0.01, we can infer that the predicted result is reasonable.

Table 4.2 Microarray data analysis result

Modules	Number of genes included	P-value
1	44	0.00127132
2	16	0.00395124
3	17	3.29069E-06
4	11	3.29069E-06
5	10	3.29069E-06
6	10	0.00692595
7	8	3.29069E-06

Chapter 5 Conclusions and Future Work

5.1 Conclusions

A predicted network of sporulation in *Clostridium thermocellum* is developed by using the web-based platform CINPER. The modeling process is based on the known information on two well studied related organisms. 40 genes are indentified and 4 functional groups are confirmed by checking against the literature. Sigma factors are found to be a sequence control factor during the whole sporulation. And SpoVD is of importance during the division step.

5.2 Future work

The future work will focus on increasing the precision of the modeling result. To achieve this goal, the initial model template should be more accurate. For this reason, two methods can be applied to enhance the accuracy of the initial template. One is to increase the number of resource genes and interactions of one certain organism added into the template. However, the amount of initial resources added does not mean that a final result would be satisfied. The quality is also of great importance. As a result, the genes and interactions in the template need to be precise so that the predicted result will be relatively convincing. Another way is by collecting genes and interactions from more kinds of bacteria that are well studied. Two organisms are collected in this model. Future work could be done such as adding genes from bacteria which also has sporulation behavior and supplement the resource information used now.

Also, validation of the model network needs to be extended further. Experiment data related to the *Clostridium thermocellum* is needed to verify the result.

Besides, further experiments should be done to find the sporulation inhibitor based on the predicted network in *clostridium thermocellum*. If biologists can find ways to suppress some key genes during the initialization of sporulation, the whole process will be stopped. Hence, the sporulation problem during the industrial production can be solved

Chapter 6 References

[1] Nicholson WL, Munakata N, Horneck G, Melosh HJ, Setlow P (2000) Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments. Microbiol M1ol Biol Rev. 2000 Sep;64(3):548-72

[2] Mao X, Olman V, Stuart R, Paulsen, Palenik B, Xu Y(2010) Computational prediction of the osmoregulation network in synechococcus sp.WH8102,11:291

[3] Waditee R, Hibino T, Tanaka Y, Nakamura T, Incharoensakdi A, Takabe T: Halotolerant cyanobacterium Aphanothece halophytica contains an Na(+)/H(+) antiporter, homologous to eukaryotic ones, with novel ion specificity affected by C-terminal tail. J Biol Chem 2001,276(40):36931-36938.

[4] Waditee R, Tanaka Y, Aoki K, Hibino T, Jikuya H, Takano J, Takabe T, Takabe T: Isolation and functional characterization of N-methyltransferases

that catalyze betaine synthesis from glycine in a halotolerant photosynthetic organism Aphanothece halophytica. J Biol Chem 2003,278(7):4932-4942.

[5] Moker N, Reihlen P, Kramer R, Morbach S: Osmosensing properties of the histidine protein kinase MtrB from Corynebacterium glutamicum. J Biol Chem 2007, 282(38):27666-27677.

[6] Lu W-D, Chi Z-M, Su C-D: Identification of glycine betaine as compatible solute in Synechococcus sp. WH8102 and characterization of its Nmethyltransferase genes involved in betaine synthesis. Arch Microbiol 2006, 186(6):495-506.

[7] Mao F, Su Z, Olman V, Dam P, Liu Z, Xu Y: Mapping of orthologous genes in the context of biological pathways: An application of integer programming. Proc Natl Acad Sci USA 2006, 103(1):129-134.

[8] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389-3402.

[9] Mao X, Chen X, Zhang Y, Pangle S, Xu Y (2012) CINPER: An Interactive Web System for Pathway Prediction for Prokaryotes. PLoS ONE 7(12): e51252.doi:10.1371/journal.pone.0051252

[10] Su Z, Mao F, Dam P, Wu H, Olman V, et al. (2006) Computational inference and experimental validation of the nitrogen assimilation regulatory network in cyanobacterium Synechococcus sp. WH 8102. Nucleic Acids Res 34: 1050–1065.

[11] Su Z, Olman V, Mao F, Xu Y (2005) Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. Nucleic Acids Res 33: 5156–5171.

[12] Su Z, Olman V, Xu Y (2007) Computational prediction of Pho regulons in cyanobacteria. BMC Genomics 8: 156.

[13] Kobayashi M, Ishizuka T, Katayama M, Kanehisa M, Bhattacharyya-Pakrasi M, et al. (2004) Response to oxidative stress involves a novel peroxiredoxin gene in the unicellular cyanobacterium Synechocystis sp. PCC 6803. Plant Cell Physiol 45: 290–299.

[14] Shcolnick S, Summerfield TC, Reytman L, Sherman LA, Keren N (2009) The mechanism of iron homeostasis in the unicellular cyanobacterium synechocystis sp. PCC 6803 and its relationship to oxidative stress. Plant Physiology 150:2045–2056.

[15] Morel FM, Price NM (2003) The biogeochemical cycles of trace metals in the oceans. Science 300: 944–947.

[16] Piggot, P. J. & Hilbert, D. W. Sporulation of *Bacillus subtilis*. Curr. Opin. Microbiol. 7, 579–586 (2004).

[17] Piggot, P. J. & Losick, R. in *Bacillus subtilis* and its Closest Relatives (eds Sonenshein, A. L., Hoch, J. A. & Losick, R.) 483–517 (ASM Press, Washington DC, 2002).

[18] Eichenberger P, Jensen S T, Conlon E M, et al. The σ E regulon and the identification of additional sporulation genes in *Bacillus subtilis*. J Mol Biol, 2003, 327: 945~972

[19] Aizawa, S.-I., Zhulin, I. B., Márquez-Magaña, L. & Ordal, G. W. in *Bacillus subtilis* and its Closest Relatives (eds Sonenshein, A. L., Hoch, J. A. & Losick, R.) 437–452 (ASM Press, Washington DC, 2002).

[20] Molle, V. et al. The Spo0A regulon of *Bacillus subtilis*. Mol. Microbiol. 50, 1683–1701 (2003).The mapping of the Spo0A regulon using ChIP-onchip and transcriptional profiling together with mobility-shift assays and bioinformatics. It shows how new technologies help unravel the mysteries of cell regulation.

[21] Carlos J. Paredes*, Keith V. Alsaker* and Eleftherios T. Papoutsakis A comparative genomic view of clostridia sporulation and physiology Nature Reviews Microbiology | AOP, published online 24 October 2005

[22] Santangelo, J. D., Kuhn, A., Treuner-Lange, A. & Dürre, P. Sporulation and time course expression of σ-factor homologous genes in *Clostridium acetobutylicum*. *FEMS Microbiol. Lett.* 161, 157–164 (1998).

[23] Weir, J., Predich, M., Dubnau, E., Nair, G. & Smith, I. Regulation of *spo0H*, a gene coding for the *Bacillus subtilis* σ-H factor. *J. Bacteriol.* 173, 521–529 (1991).

[24] Wolfe, A. J. *et al.* Evidence that acetyl phosphate functions as a global signal during biofilm development. *Mol. Microbiol.* 48, 977–988 (2003).