Aalto University
School of Science
Department of Mathematics and Systems Analysis
Degree Programme in Engineering Physics and Mathematics

Inferring *Trichoderma reesei* Gene Regulatory Network

Bachelor's thesis

Oskari Vinko

27.10.2013

Aalto University

AALTO-UNIVERSITY
SCHOOL OF SCIENCE
PL 11000, 00076 Aalto
http://www.aalto.fi

ABSTRACT OF BACHELOR'S DEGREE THESIS

Author: Oskari Vinko

Title: Inferring Trichoderma reesei Gene Regulatory Network

Degree programme: Engineering Physics and Mathematics

Major: Systems Sciences **Major code:** F3010

Supervisor: Harri Ehtamo, **Director:** Merja Oja

Significant fraction of cellulosic biofuel production costs are caused by expensive enzyme production. Researchers have been able to develop super-producing strains of Trichoderma reesei, the most widely used industrial cellulase and hemicellulase producer. The purpose of this thesis was to apply a module networks approach to this organism to form new, testable hypotheses for genes regulating cellulases and other carbohydrate active enzymes. A type of module networks algorithm was used to infer regulation programs for co-regulated clusters. This algorithm was run repeatedly and the outputs were scored and summarized to identify regulators that the algorithm consistently inferred to regulate cellulases. This resaulted to 95 cellulase regulator candidates that can be tested in laboratory. The module networks approach is able to predict potential regulators genome-wide although the quality of the hypotheses needs to be evaluated. Discovering new regulators and understanding their mechanisms would enable further improvement of cellulase production in T. reesei and lowering biofuel costs.

Date: 27.10.2013 **Language:** English **Number of pages:** 23

Keywords: gene regulatory network, module networks, Trichoderma reesei, cellulases, gene regulation, biofuels, expectation maximization

Contents

Introduction

One of the main reasons for the slow adaptation to the use of renewable fuels is the production cost compared to the refining costs of the fossil alternatives. Second generation biofuels, such as bioethanol, can be produced from sugars hydrolysed from cellulose and hemicellulose of lignocellulosic materials. The hydrolysis, however, is a very costly process because the required cellulase enzymes are responsible for approximately 0.35 €/L of current bioethanol production costs (Klein-Marcuschamer et al. 2012). *Trichoderma reesei* (teleomorph of Sordariomycete *Hypocrea jecorina*) is the most commonly used industrial producer of these enzymes due to its superior ability to produce cellulose and hemicellulose degrading enzymes. Classical mutagenesis has led to super-producing strains, however, the causes of the enhanced production are just beginning to unveil (Portnoy et al. 2011; Karimi-Aghcheh et al. 2013; Derntl & Gudynaite-Savitch 2013; Seiboth et al. 2012). Understanding the regulation mechanisms of the cellulase production is crucial to be able to make controlled and focused improvements to current super-producing strains.

Carbohydrate Active Enzymes (CAZymes) consist of glycoside hydrolases, glycosyltransferases, carbohydrate esterases, polysaccharide lyases and carbohydrate-binding proteins (Cantarel et al. 2009) including the aforementioned cellulases used in the hydrolysis of biomass. This thesis examines the *Trichoderma reesei* gene regulatory network and attempts to form new testable hypotheses for CAZyme regulating genes based on transcriptome data and a set of genes with predicted regulatory abilities. The effect of the predicted regulatory genes can be verified in the laboratory by knocking out or overexpressing the candidate regulators.

Regulation of cellulase production

The previous research of *T. reesei* can be divided in two categories: the earlier studies focused in improving the cellulase production by classical mutagenesis (Peterson & Nevalainen 2012). More recent studies examine the beneficial mutations (Le Crom et al. 2009) and enhance the effect of key regulators on cellulase production. (Kubicek et al. 2009) summarizes the five key transcription factors of the cellulase genes: the transcriptional activators include xylanase regulator 1 (XYR1), HAP complex of three proteins and ACE2 and the reported repressing factors are ACE1 and CRE1.

T. reesei cellulase genes are subject to carbon catabolite repression (CCR), that is, they are not expressed in the presence of more favourable energy source such as glucose. CRE1 was found to mediate the effect of CCR on cellulases as mutation in CRE1 binding site leads to a constitutive expression of the corresponding gene (Mach et al. 1996). Moreover, cellulase overproducing strain RUT-C30, which is released from CCR, has a truncated *cre1* (Ilmén et al. 1996).

XYR1 regulates the xylose metabolism and it is one of the main factors in cellulase gene activation (Stricker et al. 2006) and a specific point mutation in this gene causes glucose blind expression of cellulases (Derntl & Gudynaite-Savitch 2013). Although *xyr1* transcription is not induced by cellulose, deletion of *xyr1* causes depletion of all cellulases. It is unclear whether enhanced expression of *xyr1* would lead to increase in cellulases (Kubicek et al. 2009). ACE2's binding site is similar to that of XYR1 and it has been found in multiple cellulase genes. Deletion of ACE2 led to reduced cellulase activity (Stricker et al. 2008).

Furthermore, the putative protein methyltransferase LAE1 is recently reported to have a major effect on cellulase production (Seiboth et al. 2012) along with the effects on secondary metabolism and sporulation. However, its regulation mechanisms are not yet understood (Karimi-Aghcheh et al. 2013). In addition, a recent study (Jun et al. 2013) suggests that XYL1P, a D-xylose reductase involved in xylose and lactose catabolism, is a positive factor in cellulase gene transcription on lactose although *xyl1p* overexpression impairs cellulase expression on xylose. The study predicts *xyl1p* to be potential target for metabolic engineering in cellulase production.

To express cellulases, *T. reesei* requires an inducer as it would be ineffective to produce them without presence of cellulose in its natural habitat. However, as cellulose is not soluble in water, cellulase production is induced by sugars cleaved from cellulose and hemicellulose by specific cellulases expressed at low basal level (Kubicek et al. 2009). Interestingly, lactose is most widely used industrial inducer although it is only found in mammal milk. Enzymatic cleavage of lactose in *T. reesei* leads to β-d-galactose epimer that is involved in cellulase induction. Presumably, the same compound is acquired naturally in hydrolysis of hemicellulose, which contains different galactose compounds. As lactose metabolism is slow in *T. reesei*, the yields could be increased by enhancing lactose catabolism pathways (Kubicek et al. 2009).

Hypothesis

If majority of the genes can be organized into co-regulated clusters and we are able to estimate their regulation programs, there should be enough information to predict CAZyme regulating genes. However, as this thesis is based on transcriptome data, only the genes whose effect correlates with their transcription level can be evaluated. Thus, we cannot detect transcription factors with post-translational activation nor non-protein factors.

Basic Concepts

Genes are expressed at a certain level depending on transcription factors and many other factors involved in the regulation of the genes. Altering the expression of a transcription factor can lead to a cascade of responses in other genes. These genes are called regulators as they regulate several other genes, gene clusters and even organism-wide functions (see appendix A for basic biological concepts).

The regulators can together form switches and logical ports that control the expression of the target genes very accurately as a response to change in environment, life cycle and other factors. Related genes are often co-regulated, that is, their expression is increased or decreased at the same rate as they may be needed in the same set of reactions. The co-regulation is often a result of the genes having several mutual transcription factors or belonging to same co-regulated cluster where the genes are physically close to each other.

Purpose of this Thesis

In addition to understanding the mechanism of currently identified regulators, we need to explore other possible regulatory genes that affect to cellulase production as well as the factors that limit the production capabilities. Functionally related genes form groups that are co-regulated, that is, the genes’ expression is depended on the same regulators as they often have similar set of binding motifs for transcription factors. Thus, it is reasonable to study the gene regulatory network as a system of regulators and modules (group of co-regulated genes).

Segal et al. (2003) used a type of module networks approach for *Saccharomyces cerevisiae* to successfully identify regulators for modules with distinct functions. This thesis focuses in applying the module networks approach to find new cellulase regulator candidates in *T. reesei* and test the method’s ability to recognize already identified regulators. It is not reported that the module networks approach would have been applied on this problem previously.

Module networks algorithm by Segal et al. (2003) has been implemented in a computer program Genomica. It is used in this thesis to cluster the genes into modules and infer regulation programs for each module. As Genomica uses a probabilistic algorithm that can only find a local maximum in the state space, the output varies significantly between runs. To tackle this problem, the algorithm was run repeatedly and the results were summarized using a scoring method designed for this purpose. The genes were clustered again based on the scoring information collected from multiple algorithm runs (i.e. Genomica runs). Thus, this score based clustering should provide more general information about the gene regulation. In addition, Genomica is able to construct regulatory programs that may provide valuable qualitative information in modelling the regulatory roles for potential regulators.

As the method is used to examine the whole genome, it is capable of identifying co-regulated clusters and their regulators regardless of their functions. Thus, our methods and results can be used in identifying the regulatory patterns for other functions related to metabolism, cellulose production and excretion.

Impact

Examining the regulatory roles and effects of random genes with known regulatory capabilities is time consuming and difficult. Thus, well-grounded hypotheses for cellulase regulating genes provide excellent starting point for further experiments as the hypotheses can be tested in laboratory and their actual effects on cellulase production can be identified. In the most fortunate case, new cellulase regulators may emerge in the laboratory experiments. Discovering new regulators and understanding their mechanisms would enable further improvement of cellulase production in *T. reesei* and lowering biofuel costs.

Materials and Methods

Transcriptome and Data

The information about the transcription levels of all the genes in the organism's genome is called a transcriptome. To get the transcriptome data, the RNA is isolated and the complementary DNA (cDNA) is synthesized and marked with fluorescent dye. The information about the transcription level of the gene (i.e. gene activity) can be retrieved by using a cDNA microarray. It is a solid surface with thousands of spots that can identify a specific cDNA sequence based on DNA hybridization. The amount of cDNA attached with the spot can be detected by fluorescence of the attached to the cDNA and it is proportional to the activity of the corresponding gene at the moment of measurement. The methods used in retrieving the transcriptome data used in this thesis are described in detail in (Häkkinen et al. 2012).

Module Networks Approach

We are applying a module networks approach by (Segal et al. 2003) that simplifies the gene regulatory network compared to methods using Bayesian network (Pe'er et al. 2001). The fundamental idea is to cluster similar genes into modules according to their expression profiles using a Bayesian scoring function and finding the common regulatory elements, that is, the regulation program, for each module separately.

Each gene of the organism belongs to a single module (group of co-regulated genes). For each module the algorithm infers a regulation program, which may have several regulators (see figure 1). Each node divides the module in two parts according to the corresponding regulator's expression. For example, if right child includes all the experiments where the regulator is upregulated, the left child includes all the rest. Thus, each path from root to leaf contains information of all the regulators' levels. Each leaf of this regulatory program is called a regulatory context that explains the module's expression in certain conditions. According to the probabilistic

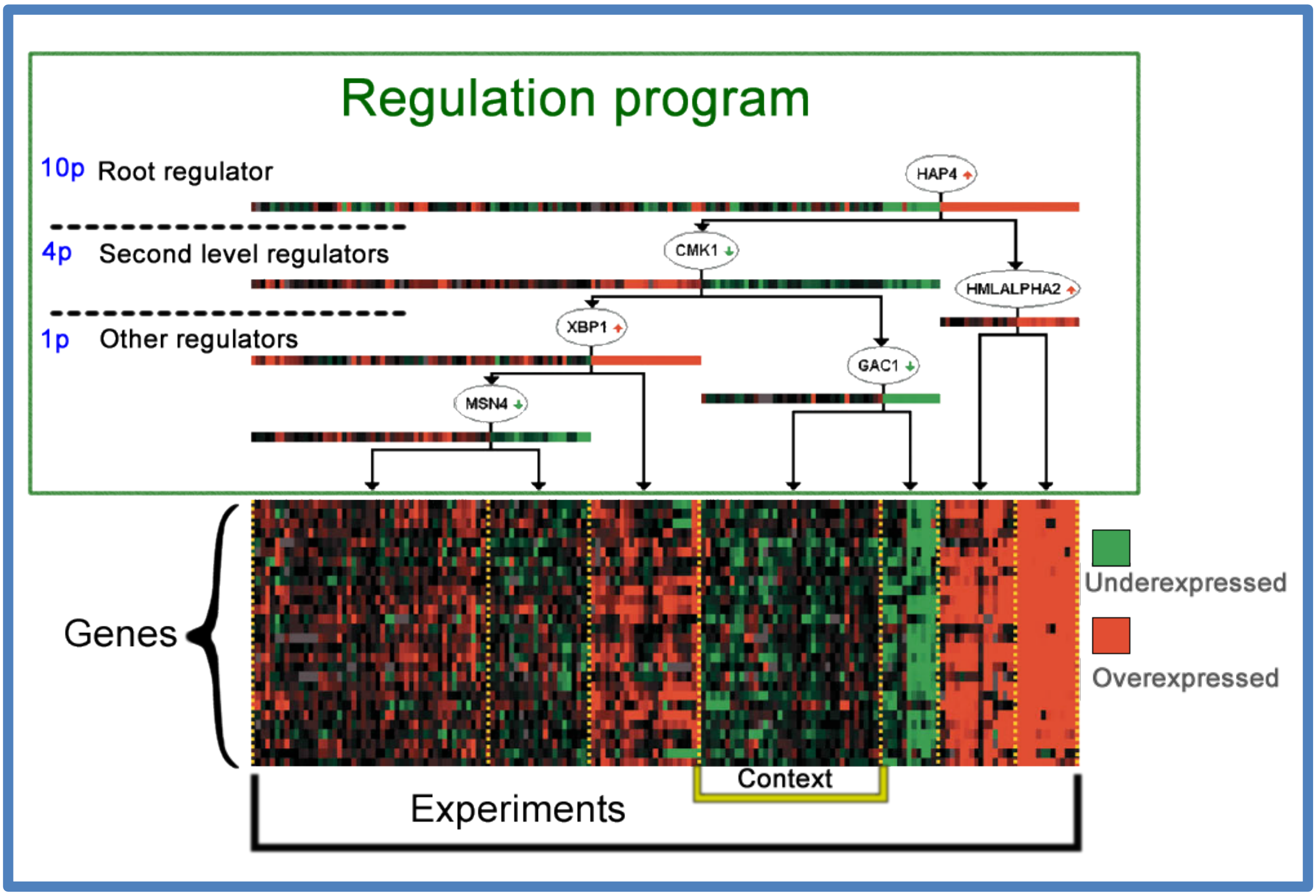

Figure 1: Module structure

Structure of a module can be divided in two parts: regulation program is the section in the figure separated by green rectangle and the below part includes all the genes and their expression. The regulation program includes the inferred regulators and their expression information. Each row in the expression view represents a single gene whereas columns represent different experiments and the cell colours are $\log_2$ ratios of comparisons between experiments. Red colour indicates elevated expression compared to control experiment and green colour stands for decreased expression.

Each regulator's expression is divided in two different parts that attempt to explain the variation in the genes. Each regulation level (separated by dashed lines on the left) brings more resolution but lower level regulators represent only small segment of experiments and thus their significance is lower. Hence, each regulator level is scored differently. Root level regulator is often able to explain the most of the variance and it governs all experiments and it has the highest score, 10 points. Second level regulators and lower level regulators get 4 and 1 points, respectively.

Each lead of the regulation program defines a context indicated by vertical, yellow, dashed lines in the picture. According to the module networks model, each expression value in same context is generated by approximately same normal distribution. The parameters for the normal distribution are context specific.

Figure is modified from original figure in (Segal et al. 2003). The figure does not represent the actual experiments in this thesis.

model of module networks all the genes in a specific context are required to follow the same conditional probability distribution. That is, the gene expression levels are generated by the same probability distribution specified by the context itself. This probability distribution is modelled by normal distribution parameterized by context-specific variance and mean (Segal et al. 2002). A detailed visual presentation of the module structure is presented in Figure 1.

To construct the module network for given expression data, we use a probabilistic algorithm that requires an initial clustering structure (i.e. starting point) and set of candidates for regulatory genes (Segal et al. 2002). This iterative Expectation Maximization (EM) algorithm consists of two steps that are repeated until the structure converges.

The first step is the structure search part that attempts to find best possible regulatory program for every module. It searches the list of candidate regulators and chooses the genes that can explain the most of the variation in a given module (Segal et al. 2002). This also affects to the parameterization of the regulatory contexts within the cluster to better match the expression profile. Whereas a gene can belong to only one module, the regulatory genes can regulate multiple modules as long as they do not contain the regulator itself.

In the module assignment step the algorithm iterates through all the genes in every module and attempts to find another module with regulation program that matches better to the expression profile of the given gene. The matching is evaluated by Bayesian scoring function and any changes leading to higher score will be performed (Segal et al. 2002). In this way the iteration of these two steps continues until the scoring converges, that is, the process terminates when these reassignments do not improve the score.

There are multiple local maxima in the Bayesian scoring function and there isn't currently any reliable way to find the global maximum. However, to avoid converging to local lower maxima the algorithm takes a random step instead of highest scoring step with certain probability that decays exponentially as the algorithm proceeds (Segal et al. 2001).

The module networks algorithm has been implemented in Genomica program developed by Y. Lubling and E. Segal from Weizmann Institute of Science. In this thesis, Genomica version 3.040710 is used to infer the regulatory network based on provided *T. reesei* transcriptome data.

Experiments

The transcriptome data was transformed to log2 ratios:

$$v = \log_2 \frac{a}{b}$$

where *a* is the gene's measured transcription level in given experiment and *b* is the gene's transcription level in corresponding control experiment, which can be a certain time point, strain, pH, temperature, carbon source etc.

For the purpose of running Genomica, the experiments in available *T. reesei* transcriptome data were divided in two datasets; data set A contains 66 experiments comparing different conditions such as pH, carbon source and temperature whereas data set B contains 49 experiments comparing different time points during perturbation. The module network was constructed separately for both datasets.

Genomica offers various options for creating the module network, such as maximum number of modules, maximum number of iterations, scoring prior options, such as maximum tree depth, minimum experiments per context, regulator split constrains etc. Generally, the default values were used with the following exceptions.

The regulation programs of different modules were compared and it was concluded that the first three levels of the regulation program are the most informative as lower level regulators are able to explain the variance only in a subset of experiments. Thus, the maximum regulation depth was set to 3.

To find an appropriate parameter for the maximum number of runs, a sample of 23 scores was taken from separate Genomica runs for data set A after 70 iterations. The average value was 761000 and standard deviation was 14000. The average score improvement per iteration between 50 and 70 iterations was approximately 130 which is very low compared to the standard deviation between the runs with 70 iterations. The score improvement per iteration gradually decreases while the algorithm advances. This kind of characteristic was also observed for data set B. In conclusion, it was reasonable to set the maximum number of iterations to 70 in all experiments.

In this thesis, maximum number of modules was set to 70 instead of default value 50. Minimum experiments per context was set to 3 so that the option wouldn't often limit the formation of regulation program when the depth is set to 3. For smaller data set B this results in 7 conditions on average for the largest possible regulation program. Regulator split constrains were set to -0.4 and 0.4 because the values were deemed to be significant difference in the log2 ratio.

The list of candidate regulators includes known *Trichoderma reesei* regulators and genes homologous to known transcription factors in other species as well as signalling proteins. Known regulator domains (INTEPRO) were used as a criterion.

As the EM algorithm may converge to different local maxima when the procedure of inferring module network is repeated, the outcome of the procedure can be significantly different between the runs. To tackle this problem, we created a scoring method that rates individual algorithm runs and compiles this information into one table of scores.

As mentioned earlier in this section, the higher level regulators in a regulation program can generally explain the gene expression variance in the module more effectively compared to lower level regulators. Based on this fact, a scoring method was designed to compile the multiple module networks from repeated Genomica runs to a single table. For each gene, a score was given to each regulator based on the regulation program. If the regulator was on root level, the score was set to 3 whereas second and third level regulators were given 2 and 1 points. An alternative stronger method scored the root level, second level and third level regulators with 10, 4 and 1 points, respectively (see figure 1). The scores for each Genomica run were summed up to single table of scores including all the genes as rows and the regulators as columns.

Three regulation characteristics were identified based on the score data. Each gene was labelled with a regulation role based on two criteria: combined score of the five strongest regulators (top5 score) and the ratio of the strongest and the third strongest regulators (one-to-three ratio, $r_{1/3}$). The first category is named *distinct* and it includes all the genes with distinct regulator that stands out in comparison with the scores of the other regulators. The requirements for this category are: $top5 > \frac{1}{2} top5_{\max}$ and $r_{1/3} > 2$. If $top5 > \frac{1}{2} top5_{\max}$ and $r_{1/3} < 2$, the gene belongs to *strong* category (see figure 4). This category includes the genes with high score but low one-to-three ratio. This indicates that the gene has several strong regulator candidates. *Weak* category includes the rest of the genes, that is, the genes with many low scoring regulators. These genes do not have distinguishable regulators as the scores for their regulators are too low to make any conclusions. Low score indicates that these genes have different regulators in different Genomica runs.

The genes were clustered based on two different data sources. The first source was the Euclidean distance between the genes in N dimensional space where N was the number of candidate regulators and the corresponding scores represented the coordinates. The second measure was the co-occurrence of the genes, which was based on how often a pair of genes appeared in the same module. It was calculated using the following formula:

$$1 - \frac{k}{n}$$

Where *k* is the number of times two genes appeared in the same cluster and *n* is the total amount of Genomica runs in current data set. K-means clustering and hierarchical agglomerative clustering methods were used to cluster the genes. The aforementioned distance measures were used in the clustering.

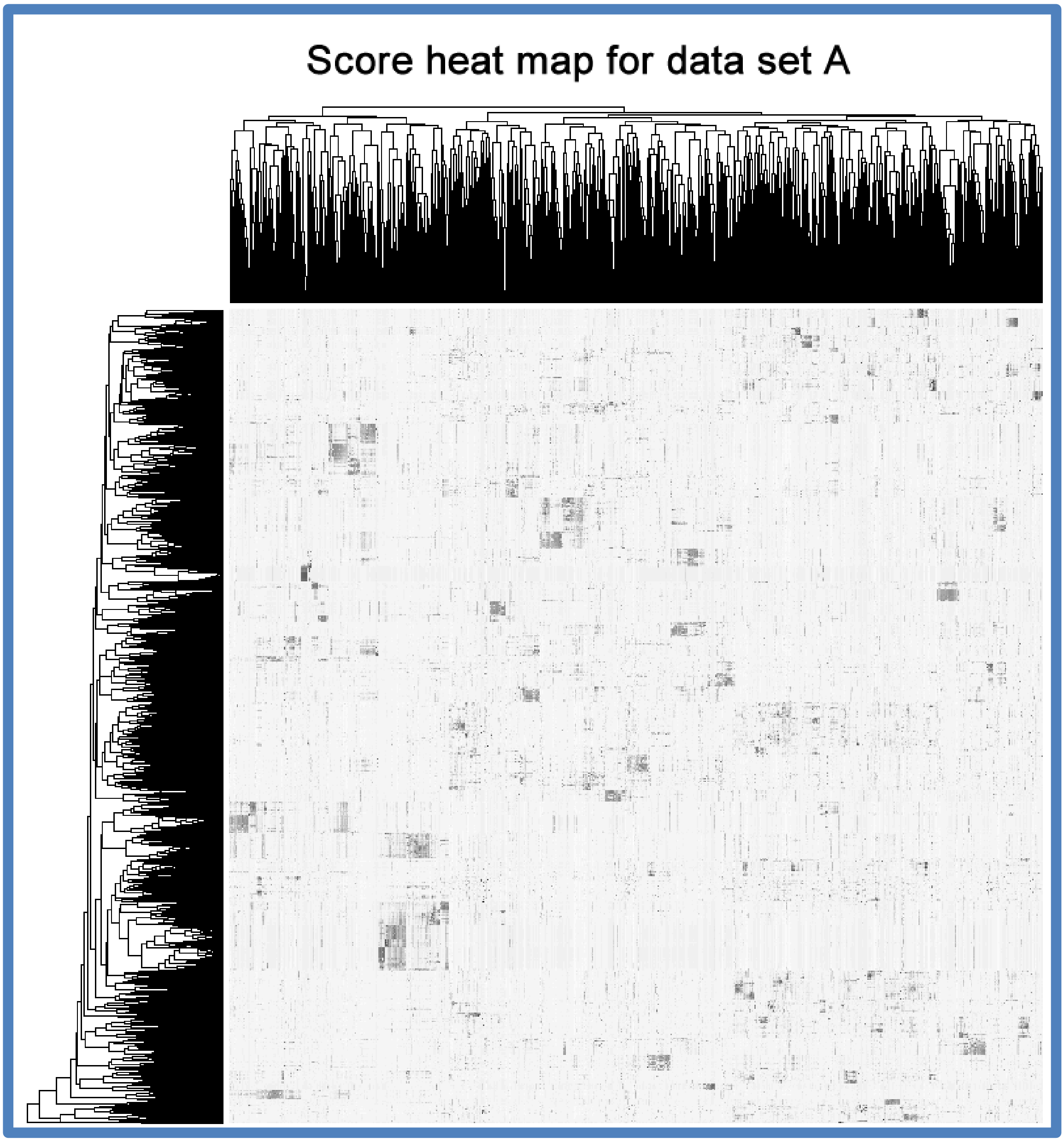

Figure 2: Score heat map

The heat map is generated for scored and summarized Genomica runs for data set A. Rows represent genes and columns represent regulators and grey colour stands for higher score. Distinct clusters of similarly regulated genes are clearly visible as grey boxes and vertical lines in the heat map. The figure is scaled to square although the number of genes is 10376 and number of regulators is 946.

Results

General regulator results

1224 Genomica runs were carried out for data set A containing 10376 genes and 1050 runs were carried out for data set B containing 10429 genes. The runs were summarized into table of scores where rows represent genes and column represent the regulator candidates provided for Genomica. Each cell tells the regulator's score (i.e. importance) for the given gene. Figure 2 presents a heat map for the score table of data set A with hierarchical clustering. The heat map shows multiple coherent clusters with varying sizes and clear vertical lines. Figure 3 is a heat map for co-occurrence table (gene vs. gene) of data set B with hierarchical clustering. The figure shows clearly how the genes concentrate in distinct clusters, that is, gene groups that always appear together in the same module.

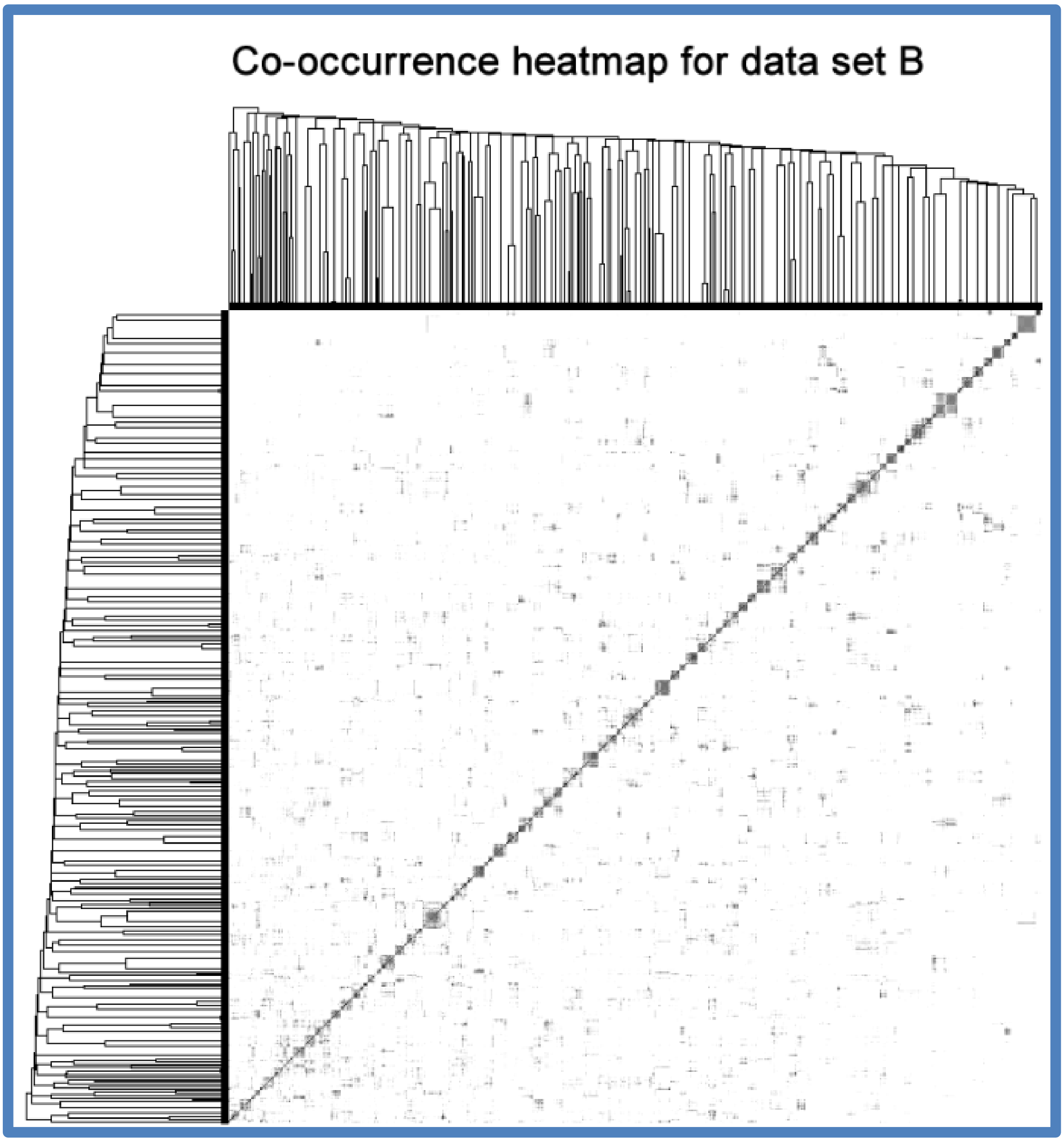

Figure 3: Co-occurrence heat map

The heat map is generated for summarized co-occurrence data over all Genomica runs for data set B. Grey colour stands for more frequent occurrence. The heat map is symmetrical and thus the clusters of frequently occurring genes are concentrated on the diagonal. Noise is present but low. High density of dendrogram near the leaves confirms that the genes are remarkably close to each other within a cluster.

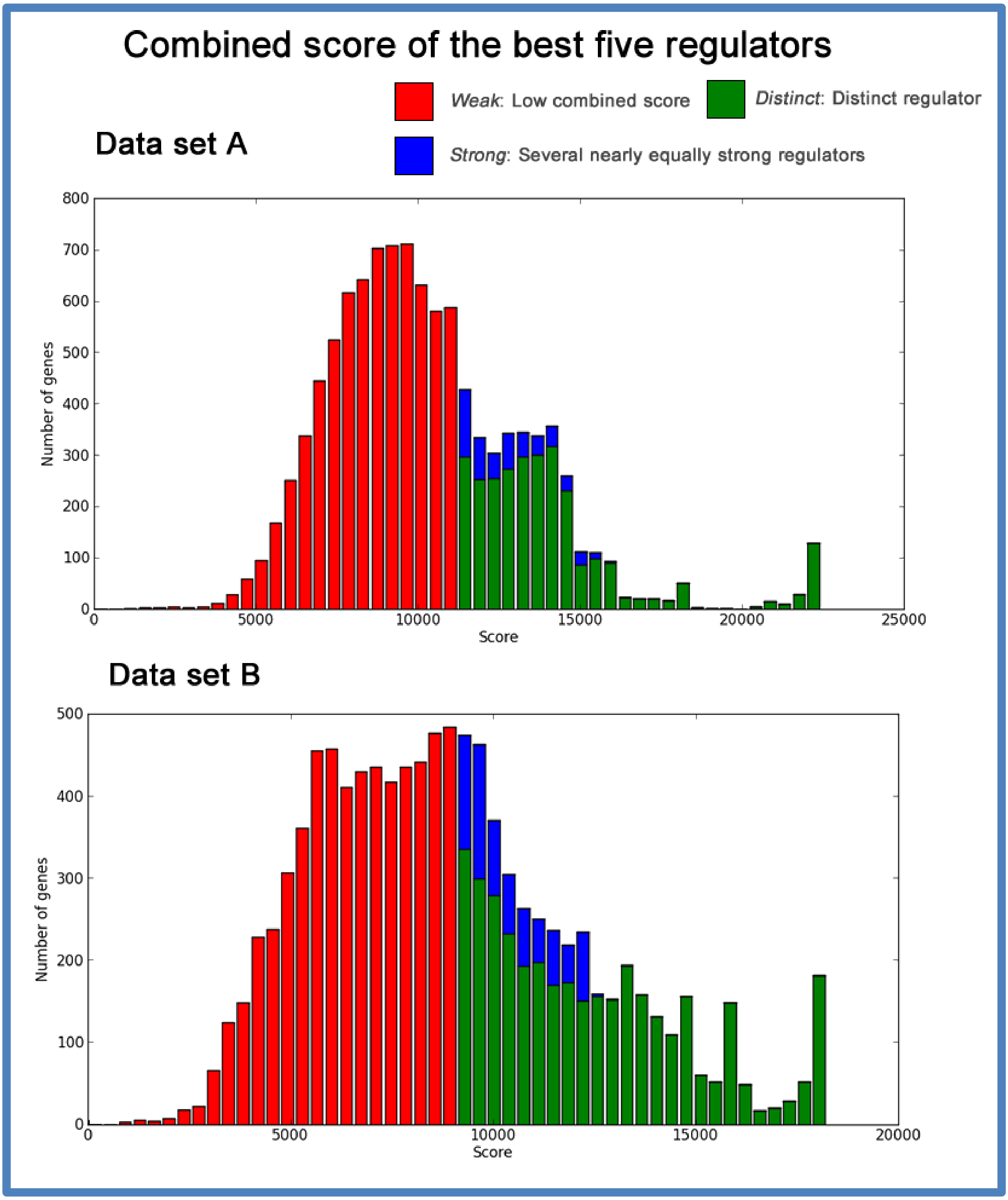

Figure 4: Score distribution

The diagrams here present the distribution of combined score of the five strongest regulators (top5 score) and regulation categories for the genes in data sets A and B. The scores in data set A are slightly higher as more runs were conducted for that data set.

If the combined score of the top 5 regulators is lower than ½ of the highest observed score, it is catecorized as weak (red color in the diagrams): it does not have high score and thus the regulators are somewhat unclear. If a gene’s score is high enough and the ratio of the best scoring regulator and third best scoring regulator is over 2.0, the gene is categorized as distinct (green color). Genes categorized as distinct have one clear high scoring regulator that stands out with its score compared to other top regulators. Blue color stands for strong: several high scoring nearly equal regulators.

Figure 4 shows score distributions with different gene regulation categories for both data sets. The tail is on the right for both distributions and a significant rise in the high-scoring genes is observed in both histograms. The scores are generally higher for data set A as the total score is directly proportional to the amount of runs. Both distributions resemble Gaussian distribution. In data set A, more genes have been categorized as *weak* gene compared to data set B because the tail for data set A is clearly longer and the decision value for weak genes is derived from the highest observed score.

The genes were clustered into 120 gene clusters using k-means and hierarchical clustering algorithms. Both methods were able to construct coherent gene clusters in which the genes have very similar scores for mutual regulators. Clusters were ranked according to average score and number of CAZy genes included in the cluster. The clusters with high average score consisted mostly of genes with *distinct* category, that is, they have single strong regulator that has significantly higher score when compared with other top regulators.

Identified CAZy regulator candidates

The analysis revealed a number of gene clusters with a significant CAZy gene enrichment. 95 CAZy gene regulator candidates were found in the analysis including the following four genes that have been studied earlier: *env1 (trire0081609), cre4 (trire0081690)* and *gene1* and *gene2* (personal communication with T. Pakula, manuscript submitted). Table 1 shows the number of CAZy genes potentially regulated by these four candidate regulators based on the implemented module networks approach.

Candidate regulator	*Env1*	*Cre4*	*Gene1*	*Gene2*
Regulated CAZy genes	3	1	17	21

Table 1: Identified known CAZy regulators

The table shows four candidate regulators and the number of potentially regulated CAZy (Carbohydrate Active Enzyme) coding genes.

Discussion

Interpreting the results

Based on the heat maps in figure 2 and figure 3, hierarchical clustering of the genes was quite successful as distinct clusters are present in both heat maps. The dendrogram was cut at appropriate height so that the larger clusters remain intact but most of the clusters are still separate. The dendrogram for co-occurrence heat map (figure 3) is very dense near its leaves, which means that the genes are remarkably close to each other within the clusters. Because only little noise is present in the figure we can conclude that the genes were often located in the same clusters between the Genomica runs. Distinct clusters are present in the score heat map but significantly more noise is present compared to co-occurrence heat map. The clusters in score heat map cannot be identified as easily as in figure 3 because the clusters are dispersed. This dispersion is caused by common regulators shared by multiple clusters. Thus, the common regulators in the clusters may not be placed next to each other but rather spread out. However, clear vertical lines can be distinguished in figure 2, which stands for co-regulated clusters.

Some vertical lines and clusters may result from duplications (caused presumably by overlapping spot sequences in the used microarrays). Duplication of some genes would also explain the exceptionally high frequency in the high scoring end of the score distributions for both data sets in figure 4. Indeed, this is the most logical explanation as the scores for the genes appeared to be identical or nearly identical. Multiple very similar gene expression profiles were observed in the initial expression data. If this expression profile matches well to any of the candidate regulators, these duplicates can form together very uniform clusters that have high score in the analysis as the same regulators will be inferred almost without exception. Thus, the duplicates would have very high score in the analysis. However, the amount of different duplicated genes was low so that the results are generally reliable.

Different clustering methods and clustering criteria (co-occurrence vs. score) resulted in slightly different cluster structures although the common cluster-specific regulators were often preserved. Thus, inferred co-regulation depended on the clustering method. For example, in score-based clustering, genes X and Y were in the same cluster regulated by Z that had high score for both X and Y. In co-occurrence-based clustering the genes X and Y belonged to different clusters both regulated by Z. Thus gene Z regulates X and Y in both methods although they were co-regulated only in the score-based clustering.

The identified CAZy gene regulator candidates in data set A were nearly complementary to the candidates found in data set B. This underlines the importance of collecting logical set of experiment comparisons for the regulator analysis.

Results compared to earlier research

Surprisingly, none of the reported CAZy gene regulators discussed earlier in this thesis were found in this analysis. The regulation mechanisms of these genes are not yet fully understood and thus the function may not be depended on transcription level. Furthermore, if the changes in the transcription level of these regulators do not vary in the experiments, the regulators' effect cannot be detected. In addition, overexpression of some of the regulators has not led to elevated cellulase production although their deletion would have an effect. This kind of genes may be necessary factors that are required in the cellulase synthesis but they might not actually regulate the rate of the cellulase gene transcription.

However, it is known that *cre1* and *xyr1* regulate cellulase gene expression and the mechanism is strictly depended on their transcription level. The strains used in this analysis were derived from RUT-C30, which has a truncated version of *cre1* which is not functional. Thus the effect of this regulator cannot be detected. *Xyr1* is reportedly a strong regulator that controls many genes (including CAZy genes) and this effect is highly correlated with its expression (Derntl & Gudynaite-Savitch 2013). *Xyr1* might regulate other regulators that strictly correlate with *xyr1*'s transcription. If the changes in such mediating regulator's expression are stronger than in *xyr1*, the mediating gene is detected by the module networks approach instead of the actual regulator. *Lae1* is also reported to regulate CAZy genes with its expression (Seiboth et al. 2012) but in this study, the differences in *lae1* expression were small and thus its regulatory capabilities could not be detected.

It has been reported (Schmoll et al. 2005) that *env1*, which was detected in our analysis, is involved in light depended cellulase regulation in *T. reesei* and the gene is similar to *Neurospora crassa* light response modulator *vvd*. *Env1* responds to light and it is transcribed under cellulase inducing conditions. *Cre4* is the homologue of *N. crassa creD* and (Denton & Kelly 2011) suggests that its overexpression of this gene could have a positive effect on cellulase production.

Gene1 and *gene2* have been overexpressed in *T. reesei* in our laboratory without changes in cellulase gene expression. Although these genes may not be able to regulate cellulase expression alone, the genes may require additional factors to be able to influence cellulase production. It is possible, that these unknown co-regulators can be identified in further analyses of the data generated in this thesis. Examining the generated clusters and initial regulation programs could provide valuable hints.

Evaluation of the approach

Analysing large number of Genomica runs confirmed that the variation between single Genomica runs is remarkable. Figure 5 shows a heat map for a gene and its regulatory program in every run. The heat map confirms that individual runs are not informative as such because they vary greatly. However, repetition of the algorithm enables discovery of significantly enriched regulator candidates that cannot be reliably identified from single Genomica runs.

For some genes, the main regulators were not present simultaneously but they appeared one at a time in the regulation programs. Although these regulators were rarely present at the same time, all they may be inferred as main regulators. Presumably, this kind of competitive pattern is probably caused by very similar regulator expression profiles so that the algorithm randomly chooses one of the matching regulators. Thus, as one expression profile is already used to explain the variance in the module, the other similar regulators would not be able to explain the remaining variance. Thus, they would be excluded from the regulation program. This behaviour is also present in the figure 5 as regulators 2 and 4 from left do not appear when other common regulators are present. In addition, they appear only as root regulators.

The module networks approach was able to predict regulators related to very different cellular functions. Even though the method was able to find very strong regulator candidates, the CAZy gene regulator candidates did not generally have high score in the analysis. In other words, this method was able to generate significantly stronger hypotheses for some non-CAZy gene regulators.

This approach is useful in searching general genome-wide information about gene regulatory network and co-regulated clusters. It was capable of predicting regulators for CAZy genes and thus it was suitable for the purpose of this thesis. The capability of this approach and generated hypotheses can be evaluated when actual regulatory functions of the predicted regulators have been tested in the laboratory.

Restrictions

The analysis is based on transcriptome data and thus it is only capable of recognizing regulators whose function correlates with its transcription rate. Furthermore, it is required that the regulator expression correlates with target genes' transcription. Hence, post-translational activation, non-protein transcription factors, chromatin structure and other levels of regulation are outside of this scope. We are able to examine only a subset of the complete gene regulatory network. As the different levels of regulation are interconnected, our method can make only constricted predictions.

The module networks algorithm implemented in Genomica was not customizable although some parameters could be changed. Thus, the algorithm and its scoring method could not be improved or tailored for the purpose of this analysis. Furthermore, the algorithm is not capable of recognizing conditional expression of the regulators, that is, logical conditionals for the regulators that are required for the genes' expression. For example, let's say that regulator A and regulator B is required for a gene's expression. The algorithm that was used in this expression might recognize only regulator B or even regulator C that is not even related to the gene but its expression profile matches better than regulators A or B individually.

This approach is only able to identify regulators that are provided in the list of candidate regulators. Thus, it is essential that the list includes as many regulatory genes as possible. Having a false regulator in the list should not be a problem as long as it does not appear in the crucial results. However, if false regulators appear in the results they will can override the actual regulator. For example, if a cellulase gene ended up in the regulator list, it would most probably be the best scoring regulator for most of the other cellulase genes as its expression explains the cellulase genes' variance very well.

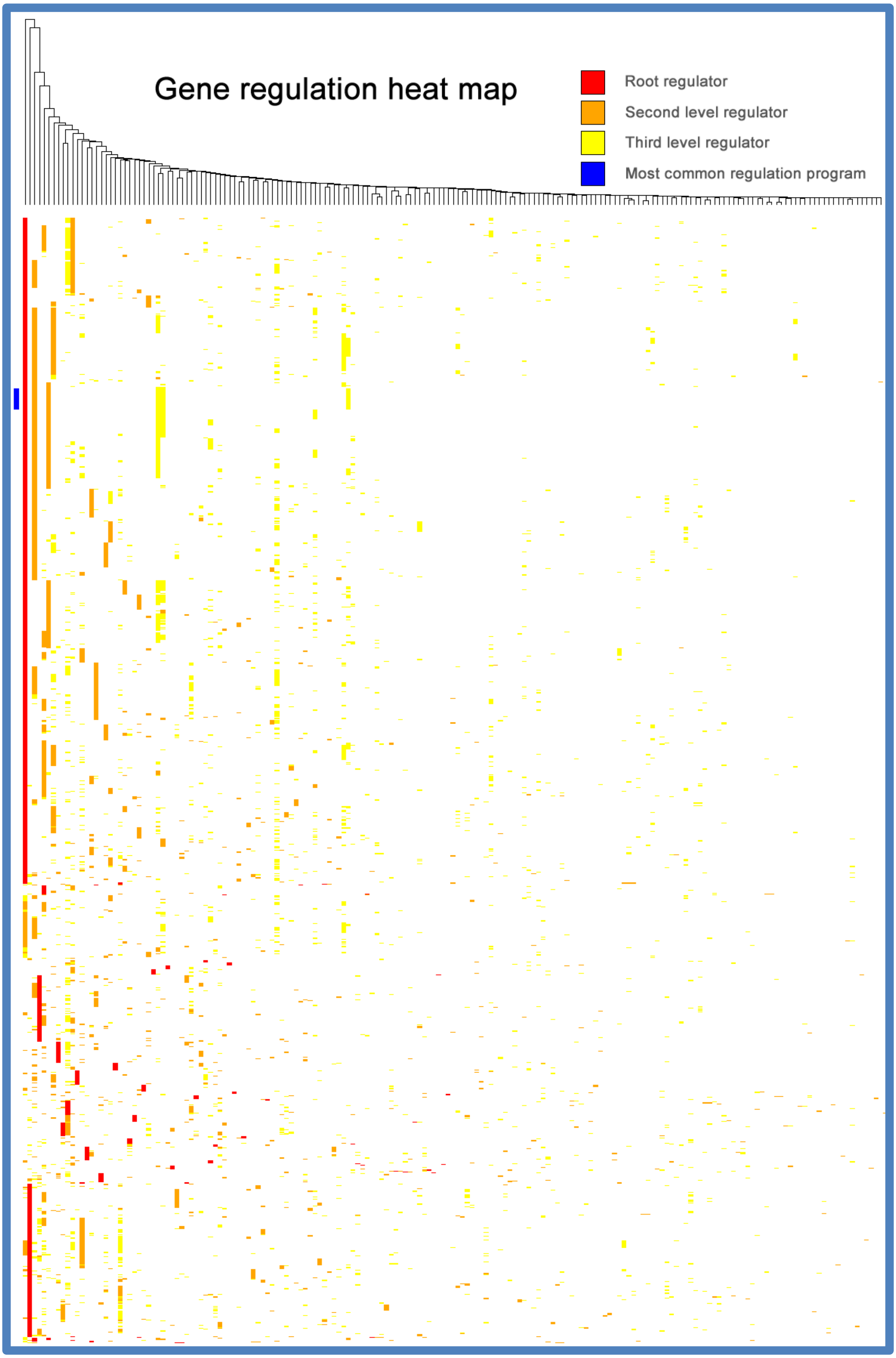

Figure 5: Gene regulation heat map

This heat map is generated for a single gene and it presents the regulation programs in each Genomica runs. Each column represents a regulator (all regulators are not included) and each row represents the inferred regulation program for the gene in question. Red colour stands for root level regulator, orange colour stands for second level regulator and yellow colour stands for third level regulator. Regulation programs marked with blue colour are estimated to be the best representatives.

Conclusions

The module networks approach is potential method for generating genome-wide hypotheses for gene regulators in *T. reesei*, but repetition is essential for the module networks algorithm to reach more reliable results. The used scoring method was sufficient and a number of distinct regulator candidates were identified.

95 CAZy gene regulator candidates were found in our analysis and some of them are quite promising. Available annotation information about these genes was scarce and little is known about these genes. Majority of the regulator candidates were not previously reported to be connected with CAZy gene regulation which makes the generated hypotheses tempting. However, the novelty of these hypotheses may also indicate the method's inability to find the actual regulators. The method was not able to identify the reported regulators and it supports this possibility. In order to evaluate the capability of this approach, the candidate regulators' functions would have to be tested in laboratory.

Further research

As a number of the generated hypotheses were new, it is possible to discover novel cellulase regulating genes among the strongest and most interesting candidates. However, testing all 95 candidates would be unnecessary as many of them had weak score or only few CAZy genes were predicted to be regulated by them. Some of the predicted CAZy regulators may also be co-regulating factors that may function with other already hypothesized regulators.

Identifying the most common Genomica run for a single gene could provide hypotheses for the regulatory network directly related to the gene without expensive laboratory experiments. This information would help to explore the gene regulatory network and plan the laboratory experiments, such as testing conditional co-regulation of two regulator candidates identified in the regulation programs instead of overexpressing or knocking out a single gene at a time.

Related genes are often enriched in co-regulated clusters. Thus, we might be able to identify genes that are co-regulated with cellulase genes and discover non-regulatory genes with important functions in cellulase production. Identifying these genes is essential for understanding how the cellulose degradation system works in *T. reesei*. Identifying central regulators in metabolism and other relevant functions might shed light to mechanisms supporting cellulase production. Enhancing these supporting mechanisms could increase the cellulase production capabilities and it will have great importance when cellulase regulation is understood well enough.

Cluster analyses such as annotation enrichments and counting the related genes within the inferred clusters could shed light into genome-wide gene regulatory network. In the earlier study applying module networks to *Saccharomyces cerevisiae* (Segal et al. 2003) the clusters represented different mechanisms or responses in the organism. Similar distinct functions could be identified in the clusters generated in this analysis.

As the expectation maximization algorithm implemented in Genomica has its own restrictions, implementing a new, more customizable algorithm could be useful. The necessary repetition step could be integrated in the algorithm and an ability to recognize simple conditional expression might lead to better hypotheses. However, these modifications would increase the computation time of single algorithm run significantly.

Acknowledgements

I would like to thank Tiina Pakula for reviewing my paper and explaining the biological background in *T. reesei*. I also want to thank Lauri Kuutti for sharing his computer for running the time-consuming Genomica analyses and Katri Olkkonen for helping to find the opportunity to write this thesis at VTT. Above all, I want to thank Merja Oja, my instructor and R expert, for everything she helped me with –from finding the references to scripting and writing this thesis.

Appendix

A: Basic Biological Concepts

Gene regulation and the gene regulatory network

Genes are stretches of deoxyribonucleic acid (DNA) located in cell nucleus in Eukaryotes. They contain the formulas for building all the organism's proteins, the amino acid polymers, which are the basic building blocks of the cell and have many vital functions along with the enzymatic capabilities (Alberts et al. 2008 chapter 3).

Eukaryotic DNA is winded around histone proteins which are further organized into secondary chromatin structure. Tight structure (referred as heterochromatin) suppresses gene expression whereas loose structure (euchromatin) enables more efficient gene expression. This chromatin structure is regulated by methylation and acetylation of the tails of the proteins that form the histones. These changes are performed by methyl- and acetyltransferases, enzymes that attach the methyl and acetyl groups to proteins (Alberts et al. 2008 chapter 4).

Each Eukaryote gene has a promoter region that regulates the gene expression rate. A gene is expressed when the coding sequence following the promoter region is transcribed into RNA. The promoter region itself does not pass information. Controlling the transcription initiation and transcription rate happens in the promoter region, which contains sequences recognized by transcription factors, that is, activators, repressors and facilitating complexes (Alberts et al. 2008 chapter 6).

The synthesized RNA (transcript) is further translated into proteins in endoplasmic reticulum and cytosol. The amount of the protein depends often on the transcription rate as well as transcript degradation rate, which is controlled by various mechanisms. The protein might require activation by methylation, phosphorylation, glycosylation or other mechanisms. This post-translational modification is important factor in regulating the protein production along with the control of the chromatin structure, transcription, translation and protein recycling (Alberts et al. 2008 chapter 7).

References

Alberts, B. et al., 2008. *Molecular Biology of the Cell* 5th Edit.,

Cantarel, B.L. et al., 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic acids research*, 37(Database issue), pp.D233–8. Available at: http://nar.oxfordjournals.org/content/37/suppl_1/D233.short [Accessed August 7, 2013].

Le Crom, S. et al., 2009. Tracking the roots of cellulase hyperproduction by the fungus Trichoderma reesei using massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38), pp.16151–6. Available at: http://www.pnas.org/content/106/38/16151.short [Accessed August 2, 2013].

Denton, J.A. & Kelly, J.M., 2011. Disruption of Trichoderma reesei cre2, encoding an ubiquitin C-terminal hydrolase, results in increased cellulase activity. *BMC biotechnology*, 11, p.103. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3226525&tool=pmcentrez&rendertype=abstract [Accessed August 30, 2013].

Derntl, C. & Gudynaite-Savitch, L., 2013. Mutation of the Xylanase regulator 1 causes a glucose blind hydrolase expressing phenotype in industrially used Trichoderma strains. *Biotechnology for …*, 6(1), p.62. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3654998&tool=pmcentrez&rendertype=abstract [Accessed August 2, 2013].

Gibson, D.G. et al., 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329(5987), pp.52–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20488990.

Häkkinen, M. et al., 2012. Re-annotation of the CAZy genes of Trichoderma reesei and transcription in the presence of lignocellulosic substrates. *Microbial cell factories*, 11, p.134. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3526510&tool=pmcentrez&rendertype=abstract [Accessed August 6, 2013].

Ilmén, M. et al., 1996. Functional analysis of the cellobiohydrolase I promoter of the filamentous fungus Trichoderma reesei. *Molecular & general genetics : MGG*, 253(3), pp.303–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9003317 [Accessed August 19, 2013].

Jun, H., Guangye, H. & Daiwen, C., 2013. Insights into enzyme secretion by filamentous fungi: Comparative proteome analysis of Trichoderma reesei grown on different carbon sources. *Journal of proteomics*, 89C(null), pp.191–201. Available at: http://dx.doi.org/10.1016/j.jprot.2013.06.014 [Accessed August 7, 2013].

Karimi-Aghcheh, R. et al., 2013. Functional analyses of Trichoderma reesei LAE1 reveal conserved and contrasting roles of this regulator. *G3 (Bethesda, Md.)*, 3(2), pp.369–78. Available at: http://www.g3journal.org/content/3/2/369.short [Accessed August 2, 2013].

Klein-Marcuschamer, D. et al., 2012. The challenge of enzyme cost in the production of lignocellulosic biofuels. *Biotechnology and bioengineering*, 109(4), pp.1083–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22095526 [Accessed August 2, 2013].

Kubicek, C.P. et al., 2009. Metabolic engineering strategies for the improvement of cellulase production by Hypocrea jecorina. *Biotechnology for biofuels*, 2, p.19. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2749017&tool=pmcentrez&rendertype=abstract [Accessed August 2, 2013].

Mach, R.L. et al., 1996. Carbon catabolite repression of xylanase I (xyn1) gene expression in Trichoderma reesei. *Molecular microbiology*, 21(6), pp.1273–81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8898395 [Accessed August 19, 2013].

Pe'er, D. et al., 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(Suppl 1), pp.S215–S224. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/17.suppl_1.S215.

Peterson, R. & Nevalainen, H., 2012. Trichoderma reesei RUT-C30--thirty years of strain improvement. *Microbiology (Reading, England)*, 158(Pt 1), pp.58–68. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21998163 [Accessed August 2, 2013].

Portnoy, T. et al., 2011. Differential regulation of the cellulase transcription factors XYR1, ACE2, and ACE1 in Trichoderma reesei strains producing high and low levels of cellulase. *Eukaryotic cell*, 10(2), pp.262–71. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3067402&tool=pmcentrez&rendertype=abstract [Accessed August 2, 2013].

Schmoll, M., Franchi, L. & Kubicek, C.P., 2005. Envoy, a PAS/LOV domain protein of Hypocrea jecorina (Anamorph Trichoderma reesei), modulates cellulase gene transcription in response to light. *Eukaryotic cell*, 4(12), pp.1998–2007. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1317494&tool=pmcentrez&rendertype=abstract [Accessed August 30, 2013].

Segal, E. et al., 2002. Learning module networks. *Proceedings of the Nineteenth …*. Available at: http://dl.acm.org/citation.cfm?id=2100648 [Accessed August 2, 2013].

Segal, E. et al., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2), pp.166–76. Available at: http://dx.doi.org/10.1038/ng1165 [Accessed August 2, 2013].

Segal, E. et al., 2001. Rich probabilistic models for gene expression. *Bioinformatics (Oxford, England)*, 17 Suppl 1, pp.S243–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11473015.

Seiboth, B. et al., 2012. The putative protein methyltransferase LAE1 controls cellulase gene expression in Trichoderma reesei. *Molecular microbiology*, 84(6), pp.1150–64. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3370264&tool=pmcentrez&rendertype=abstract [Accessed August 2, 2013].

Stricker, A.R. et al., 2008. Role of Ace2 (Activator of Cellulases 2) within the xyn2 transcriptosome of Hypocrea jecorina. *Fungal genetics and biology : FG & B*, 45(4), pp.436–45. Available at: http://dx.doi.org/10.1016/j.fgb.2007.08.005 [Accessed August 9, 2013].

Stricker, A.R. et al., 2006. Xyr1 (xylanase regulator 1) regulates both the hydrolytic enzyme system and D-xylose metabolism in Hypocrea jecorina. *Eukaryotic cell*, 5(12), pp.2128–37. Available at: http://ec.asm.org/content/5/12/2128.short [Accessed August 9, 2013].

Yhteenveto

Korkeat tuotantokustannukset hidastavat uusiutuvien polttoaineiden yleistymistä, sillä kuluttajat sekä eri teollisuudenalat suosivat usein halvempia fossiilisia polttoainevaihtoehtoja. Fossiiliset polttoaineet vapauttavat kasvihuonekaasuihin kuuluvaa hiilidioksidia maaperän hiilinieluista ilmakehään, ja siksi niiden käyttö edistää ilmastonmuutosta. Biopolttoaineet sen sijaan valmistetaan biomassasta, jonka sisältämä hiili on peräisin ilmakehästä. Siksi biopolttoaineiden käyttö ei lisää ilmakehän kasvihuonekaasujen määrää.

Biopolttoaineita, kuten bioetanolia ja biodieseliä, valmistetaan usein mikrobien avulla sokereista, joita tuotetaan pilkkomalla biomassan polysakkarideja, pääosin selluloosaa ja hemiselluloosaa. Polysakkaridit voidaan pilkkoa oikeanlaisilla entsyymeillä, mutta näiden valmistus on melko kallista. Arvioiden mukaan bioetanolin valmistuksessa entsyymikustannukset ovat noin 0,35€/l. Entsyymien tuotantokustannusten pienentämisellä olisi siten suuri vaikutus biopolttoaineiden hintaan ja käyttöasteeseen.

Trichoderma reesei on home, jota käytetään yleisimmin selluloosaa hajottavien sellulaasientsyymien tuotannossa, sillä se kykenee tuottamaan huomattavan suuria määriä näitä hyödyllisiä entsyymejä muihin tunnettuihin organismeihin verrattuna. Siitä on lisäksi tehty geneettisesti muunneltuja kantoja, jotka kykenevät tuottamaan sellulaaseja huomattavasti alkuperäistä kantaa tehokkaammin. Nämä kannat on valmistettu klassisella mutageneesillä eli aiheuttamalla satunnaisia mutaatioita ja valitsemalla ne yksilöt, joiden sellulaasituotanto on tehostunut. Nykyinen tutkimus pyrkii ymmärtämään *T. reesein* säätelymekanismeja sekä ylituottajakannoissa tapahtuneita muutoksia. Tämän ymmärryksen myötä voidaan tehdä harkittuja sellulaasituotantoa tehostavia muutoksia.

Geenien aktiivisuutta voidaan mitata mikrosiruilla, joissa on oma tunnistuskohtansa kullekin tutkittavalle geenille. Kustakin tunnistuskohdasta voidaan koneellisesti lukea geenin aktiivisuus mittaushetkellä. Transkriptomiksi kutsutaan koko genomin kattavaa dataa geenien aktiivisuudesta ja sitä tutkimalla voidaan saada arvokasta tietoa geenien toiminnasta eri olosuhteissa ja niiden vaikutuksista toisiinsa. Geenit vaikuttavat aktiivisuudellaan lukuisiin muihin geeneihin ja sitä kautta koko organismin toimintaan. Tätä monimutkaista vuorovaikutusten verkkoa kutsutaan geenisäätelyverkostoksi.

Tämän tutkielman tarkoituksena on tutkia *T. reesein* transkriptomia sekä ehdottaa sen perusteella kandidaatteja sellulaasituotantoa sääteleville geeneille. Näiden ehdokasgeenien vaikutusta voidaan tutkia laboratoriossa ylituottamalla tai alituottamalla kutakin geeniä. On siis mahdollista, että ehdokassäätelijöiden joukosta löytyy entuudestaan tuntemattomia säätelygeenejä, jotka ohjaavat sellulaasituotantoa.

T. reesein geenisäätelyverkostoa pyritään tässä tutkielmassa hahmottamaan Module Networks –menetelmällä. Tämän lähestymistavan pääperiaatteena on, että geenit voidaan jakaa ryhmiin, joiden aktiivisuutta ohjaavat yhteiset säätelygeenit. Tähän tarkoitukseen käytettiin suurimman uskottavuuden algoritmia (eng. expectation maximization algorithm), joka koostuu kahdesta toistettavasta vaiheesta. Ensimmäinen vaihe tutkii muodostettuja geeniryhmiä ja pyrkii etsimään niille sopivimpia säätelygeenejä annetusta listasta geenejä, ja toinen vaihe tarkastelee jokaista geeniä erikseen ja sijoittaa kunkin tämän aktiivisuutta parhaiten vastaavaan geeniryhmään.

Algoritmissa käytetty suurimman uskottavuuden algoritmi pyrkii maksimoimaan tietyn bayesiläisen kohdefunktion arvoa ohjaamalla sellaisiin muutoksiin, jotka kasvattavat kohdefunktion arvoa. Tällä kohdefunktiolla on kuitenkin lukuisia paikallisia ääriarvoja eikä tällä hetkellä tunneta menetelmää, jonka avulla voitaisiin löytää globaali ääriarvo. Siksi tässä työssä käytetty algoritmi ottaa joitakin satunnaisaskeleita välttääkseen suppenemisen pieniin paikallisiin ääriarvoihin. Näiden satunnaisaskelten vuoksi algoritmin tarjoama säätelyverkosto vaihtelee huomattavasti eri ajokertojen välillä.

Algoritmia ajettiin toistuvasti noin 2000 kertaa, tulokset pisteytettiin ja koottiin yhdeksi taulukoksi. Taulukkoon on merkitty kunkin säätelygeenin pistemäärä jokaiselle geenille. Tämä pistemäärä kertoo, kuinka tärkeä säätelijä on kyseiselle geenille keskimäärin kaikissa toistokerroissa. Tällä tavoin ajokerroista saatiin esille selkeät pääpiirteet satunnaisten yksityiskohtien sijaan. Tämän lisäksi geenit ryhmiteltiin uudelleen pistemäärien perusteella. Siten saadaan myös informaatiota siitä, mitkä geenit saattavat ohjautuvat samojen säätelijöiden vaikutuksesta.

Kiinnostaviksi geeneiksi luokiteltiin kaikki sellaiset geenit, jotka liittyvät jollain tapaa polysakkaridien pilkkomiseen (CAZy-geenit). Tähän geeniluokkaan kuuluu lukuisia hydrolaaseja (mm. sellulaaseja, hemisellulaaseja) sekä muita hydrolyysiä katalysoivia entsyymejä. Saatujen tulosten perusteella löytyi 95 säätelygeeniä, jotka saattavat säädellä näitä CAZy-geenejä. Joukossa on joitakin hyvin mielenkiintoisia ehdokkaita, jotka säätelevät suurta joukkoa kiinnostavia entsyymejä, mutta osa säätelyehdokkaista liittyy tulosten perusteella vain yksittäisiin kiinnostaviin geeneihin.

Näiden ehdokasgeenien joukosta etsittiin myös seitsemää säätelygeeniä, joiden on havaittu vaikuttavan *T. reesein* sellulaasituotantoon aikaisemmissa tutkimuksissa. Yhtäkään näistä geeneistä ei kuitenkaan löytynyt 95 ehdokasgeenin joukosta. Vaikka monen tunnetun säätelijän puuttuminen on perusteltavissa, on tulos melko yllättävä. Tämä kertoo osaltaan käytetyn menetelmän soveltumattomuudesta tietyntyyppisten säätelijägeenien etsimiseen. Käytetty data perustuu yksinomaan geenin transkriptiotasojen mittaukseen, vaikka geenien säätelyyn liittyy myös monia muita mekanismeja. Sen vuoksi organismin transkriptomiin keskittyvä tutkimus kykenee tavoittamaan vain pienen osan geenien säätelyverkostosta ja siksi monet säätelijägeenit ovat menetelmän ulottumattomissa.

Suurin osa 95 ehdokasgeenistä on kuitenkin aiemmin huonosti tunnettuja, eikä niiden tiedetä liittyvän sellulaasien tuotannon säätelyyn. Tämän vuoksi potentiaalisten säätelijöiden joukosta voi hyvinkin löytyä entuudestaan tuntemattomia säätelygeenejä, jotka puolestaan tarjoavat uusia mahdollisuuksia *T. reesein* entsyymituotannon tehostamiseen.

Ehdokasgeenien joukosta löytyi kuitenkin muutama geeni, joiden on ajateltu liittyvän sellulaaseihin aikaisemmissa tutkimuksissa. *Env1*:n on todettu reagoivan valoon ja aktivoituvan sellulaaseja indusoivissa olosuhteissa. *Cre4* on puolestaan *Neurospora crassan* geenin *creD* homologi ja sen on arveltu tehostavan sellulaasituotantoa.

Saadut tulokset ovat koko genomin kattavia ja näin ollen niiden joukosta löytyy myös lukuisia muihin *T. reesein* toimintoihin liittyviä säätelyehdokkaita. Itse asiassa monien geenien osalta säätelijöitä vastaavat pistemäärät ovat huomattavasti korkeampia löydettyjen sellulaaseja säätelevien ehdokasgeenien pistemääriin verrattuna. Tämä tarkoittaa sitä, että joidenkin säätelygeenien havaittiin säätelevän tiettyjä geenejä lähes jokaisessa algoritmiajossa. Nämä hyvin vahvat ehdokkaat ovat erittäin kiinnostavia *T. reesein* muihin toimintoihin liittyvässä tutkimuksessa, mutta valitettavasti ne eivät ole yhtä kiinnostavia sellulaasituotannon kannalta.

Sellulaasituotannon kasvattaminen edellyttää myös organismin valmiuksien kehittämistä niin aineenvaihdunnan, proteiinisynteesin kuin proteiinierityksenkin osalta. Kerätystä datasta voi löytyä hyödyllisiä ehdokassäätelijöitä, joista saattaa olla merkittävää hyötyä näiden toimintojen kehittämisessä.

Koska tämän tutkielman analyysi perustui pitkälti samalla tavalla säädeltyjen geenien ryhmittelyyn, voi tämän rakenteen tarkempi tutkiminen paljastaa uusia johtolankoja. Jatkossa olisi siis syytä kiinnittää huomiota myös niihin geeneihin, joita sellulaaseja säätelevät geenit ohjaavat. Osa näistä geeneistä voi liittyä oleellisestikin polysakkaridien hajottamiseen tai sitä tukeviin toimintoihin. Tällaisten toimintojen tunteminen on hyvin tärkeää *T. reesein* arvokkaiden ominaisuuksien tehostamisessa.

Tulosten perusteella Module Networks –menetelmä vaikuttaa soveltuvan hyvin *T. reesein* sellulaasituotantoa säätelevien geenien etsimiseen. Löydettyjen säätelyehdokkaiden todellisen roolin selvittämiseksi tarvitaan kuitenkin lukuisia laboratoriokokeita. Ehdokasgeenejä ylituottamalla tai poistamalla voidaan tutkia geenien vaikutusta sellulaasituotantoon sekä muihin toimintoihin. Tätä ennen ei hypoteeseja voida osoittaa oikeiksi. Toistaiseksi voidaan siis vain todeta, että menetelmä kykenee löytämään hyviä ehdokkaita, mutta niiden todellinen vaikutus jää myöhempien tutkimuksien arvioitavaksi. Toisaalta, mikäli lupaavimpien säätelyehdokkaiden joukossa on yksikin ennalta tuntematon sellulaasituotantoa tehostava geeni, osoittautuu menetelmä toimivaksi ja sovelluskelpoiseksi myös muihin samankaltaisin ongelmiin.