Part A Report – June 2012

*Applied Bioinformatics in Lotus japonicus*

Vikas Gupta

Supervisor: Christian N. S. Pedersen
Co-supervisors: Stig U. Andersen and Søren S. Thirup
Bioinformatics Research Centre
Aarhus University
Denmark
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PREFACE</strong></td>
<td>II</td>
</tr>
<tr>
<td><strong>SUMMARY</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>1. PART-I LYSM MODELING</strong></td>
<td>2</td>
</tr>
<tr>
<td>1.1 INTRODUCTION</td>
<td>2</td>
</tr>
<tr>
<td>1.2 COEVOLUTIONARY MODEL</td>
<td>2</td>
</tr>
<tr>
<td>1.3 MULTIPLE SEQUENCE ALIGNMENTS AND INFERRING PHYLOGENY</td>
<td>4</td>
</tr>
<tr>
<td>1.4 EXAMINED INTERACTING PAIRS</td>
<td>4</td>
</tr>
<tr>
<td>1.5 EXPERIMENTAL PROCEDURES</td>
<td>4</td>
</tr>
<tr>
<td>1.5.1 Dataset</td>
<td>4</td>
</tr>
<tr>
<td>1.5.2 Removing redundancy</td>
<td>5</td>
</tr>
<tr>
<td>1.5.3 Creating phylogenetic trees</td>
<td>5</td>
</tr>
<tr>
<td>1.5.4 Calculating independent and correlated scores</td>
<td>5</td>
</tr>
<tr>
<td>1.5.5 Generating controls</td>
<td>5</td>
</tr>
<tr>
<td>1.5.6 Calculating inter-residual distances</td>
<td>5</td>
</tr>
<tr>
<td>1.5.7 Calculating the frequency of residues</td>
<td>6</td>
</tr>
<tr>
<td><strong>1.6 RESULTS AND DISCUSSION</strong></td>
<td>6</td>
</tr>
<tr>
<td>1.6.1 Coevolution score comparison within protein domains</td>
<td>6</td>
</tr>
<tr>
<td>1.6.2 Coevolution score comparison with inter-residual distances</td>
<td>8</td>
</tr>
<tr>
<td>1.6.3 Coevolution score comparison with co-conservation</td>
<td>9</td>
</tr>
<tr>
<td><strong>1.7 NFR5 STRUCTURE</strong></td>
<td>9</td>
</tr>
<tr>
<td><strong>1.8 CONCLUSIONS</strong></td>
<td>11</td>
</tr>
<tr>
<td><strong>1.9 FUTURE PERSPECTIVES</strong></td>
<td>11</td>
</tr>
<tr>
<td><strong>2. PART-II SHORTRAN</strong></td>
<td>12</td>
</tr>
<tr>
<td>2.1 INTRODUCTION</td>
<td>12</td>
</tr>
<tr>
<td>2.2 SHORTRAN PIPELINE</td>
<td>13</td>
</tr>
<tr>
<td>2.2.1 Module-1: Error correction, filtering and profiling</td>
<td>13</td>
</tr>
<tr>
<td>2.2.2 Module-2: Genomic Mapping</td>
<td>14</td>
</tr>
<tr>
<td>2.2.3 Module-3: Cluster analysis</td>
<td>14</td>
</tr>
<tr>
<td>2.2.4 Module-4: miRNA predictions</td>
<td>14</td>
</tr>
<tr>
<td>2.2.5 Module-5: TA-siRNA predictions</td>
<td>14</td>
</tr>
<tr>
<td>2.2.6 Module-6: Annotation</td>
<td>15</td>
</tr>
<tr>
<td>2.2.7 Module-7: Expression plots</td>
<td>15</td>
</tr>
<tr>
<td>2.2.8 Module-8: Pattern search</td>
<td>15</td>
</tr>
<tr>
<td>2.2.9 Module-9: Queries</td>
<td>15</td>
</tr>
<tr>
<td><strong>2.3 DATA ANALYSIS</strong></td>
<td>16</td>
</tr>
<tr>
<td>2.3.1 Data pre-processing</td>
<td>16</td>
</tr>
<tr>
<td>2.3.2 Genomic mapping</td>
<td>18</td>
</tr>
<tr>
<td>2.3.3 Cluster analysis</td>
<td>18</td>
</tr>
<tr>
<td>2.3.4 Annotation</td>
<td>19</td>
</tr>
<tr>
<td>2.3.5 Comparative quantitative analysis and plotting</td>
<td>20</td>
</tr>
<tr>
<td><strong>2.4 CONCLUSIONS</strong></td>
<td>22</td>
</tr>
<tr>
<td><strong>2.5 FUTURE PERSPECTIVE</strong></td>
<td>22</td>
</tr>
</tbody>
</table>
Preface

I am pursuing PhD studies at the Bioinformatics Research Center at Århus University. I started my PhD education July 1st 2010, under the supervision of Christian N. S. Pedersen at the Bioinformatics Research Center (BiRC) and Stig U. Andersen at the Centre for Carbohydrate Recognition and Signaling (CARB), Århus University. The PhD project concerns developing structural models for Nod-factor receptor kinases\(^1\) to understand the interaction with respective ligands, project is coordinated with the supervision of Søren Thirup, a member of Center of Structural Biology (CSB). During my Ph.D. studies, we have also focused on understanding the involvement of small RNAs in the symbiosis between *Lotus japonicus* and rhizobial bacteria\(^2\). I have been involved in developing a toolkit for analyzing small RNA data called *shortRan* [manuscript under revision] under the guidance of Stig U. Andersen, Katharina Markmann and Jens Stougaard at CARB.

In this report, a brief description of my research contribution is provided. A summary highlighting the approaches and results is provided in the start of the report. Research work is divided into three sub-projects pursued during first two years of Ph.D. For each sub-project, I have written concise introduction, methods and results section. The report contains a final section on the conclusions from the research work performed during the first part of Ph.D. education and future perspective for the last two years of Ph.D. studies.

**

Manuscripts, supplement data, tools and posters regarding the results obtained in the report, will be shared with examiners using Drop Box.
Summary

Legume plants form a symbiotic relation with rhizobium bacteria, which fix nitrogen for the plant, many genes and proteins involved in the symbiotic network are well described, while their precise function in the symbiosis is still uncertain. Here our aim is to understand the symbiotic process in details both at level of RNA and protein. At the protein level, we have worked with LysM domain containing proteins. There are over 18,000 LysM domain sequences listed in the Pfam database but the structure of only four single LysM domain containing proteins has been available. Here, we attempt to model the complete extra-cellular domain structure of the Lotus japonicus Nod factor receptors (NFRs) protein using results based on coevolving residues. We implemented a multiple sequence alignment based algorithm, which uses Continuous Time Markov Process (CTMP) sequence substitution to model the coevolution of amino acid position pairs. In this report, I have briefly described the outcomes of our work on three LysM domains containing extracellular region of NFR5.

At the RNA level, we have analyzed small RNA data produced using Next generation sequencing (NGS). These sequencing technologies are capable of producing large data volumes at an affordable cost, revolutionizing our knowledge at genomic and transcriptomic levels. These advancements have also resulted in discovery, categorization and annotation of many small RNA classes, which was not easily achievable with conventional sequencing methods. Efficient analysis of such extensive data amount to support new biological insights and hypothesis based on practical results has always been one of the major challenges in the bioinformatics research field. In this report, we describe shortRan, a pipeline for analyzing small RNA sequencing data. shortRan is capable of pre-processing, normalizing, profiling, categorizing, and annotating small RNA datasets. It produces both graphical and text formatted results, which are compatible with genome browsing tools such as IGV, UCSC and various data analysis tools including edgeR, MeV. We also describe implementation of the shortRan and its application to small RNA sequencing data to gain useful insight in the early symbiotic response of L. japonicus to rhizobial bacteria.

Analyzing small RNA expression requires mapping of small RNA sequences (reads) back on the genome to identify source loci. Therefore it is critical to have a good reference genomic sequence to obtain valuable and accurate results. As the L. japonicus genome is partially assembled and contained multiple uncompleted regions (gaps), we utilized available Illumina sequencing reads pairs that has not been considered in the previous assembly to optimize the genome assembly. We developed an automated pipeline (unpublished) called GapGlu, which can find all the gaps in the assembled reference sequences and then uses the hanging reads to find the potential unplaced contig. Results from GapGlu analysis has been described in the third project section of the report. Using this tool, we were able to find and fill a substantial fraction of gaps in the reference genome. Results from the GapGlu analysis are described in the third section of this report.
1. Part-I LysM Modeling

1.1 Introduction

The phenomenon of coevolution is widely known and generally refers to the mutual changes in the pairs of genes, proteins or species. Often in proteins, co-evolution results from mutual selective pressure on domains. Understanding the intra-molecular coevolution between amino acid residues can provide insights into the structural configuration of a protein. Revealing such coevolving amino acid pairs has been one of the major challenges in bioinformatics\(^5\). Most of the methods have been relying on the examination of multiple sequence alignments (MSA's) to understand the mutational dynamics of proteins. Even though this approach improves our understanding on protein mutation dynamics, the complexity within is beyond the methods utilizing only the linear sequences. Many computational and mathematical tools have been developed in the last decade to explore the spatial, functional and evolutionary dependencies between the amino acid sites within a protein. But the complexity of evolutionary dependencies has been hampering the development of sensitive algorithms. The motivation for developing these algorithms comes from many examples\(^6\) where dependencies between co-evolving amino acid sites have confirmed the functional importance of residues.

Co-evolution between two amino acid sites can be divided into three different categories, stochastic coevolution, functional coevolution and interaction-based coevolution. Each type of coevolution influences predictions depending upon the model and on the quality of the multiple sequence alignment. Methods for detecting co-evolution can be divided into two categories: tree-aware methods and tree-ignorant ones. The former one takes the phylogeny of the sequences into consideration while the latter doesn’t. Tree-ignorant methods often use less computational time but are widely held to be insufficiently accurate, as they do not consider the shared ancestry within coevolution.

In our studies, we have implemented a method for detecting coevolution based on multiple sequence alignment of interacting protein domains and then we have compared results using tree-aware and tree-ignorant methods. As example datasets, we have considered interacting protein datasets such as 2-oxoisovalerate dehydrogenase subunits, elongation factors Ts and Tu, and copper transporting P-type ATPases with copper chaperones then we have applied the method on the LysM domain containing proteins. Our tree-ignorant model is based on a single parameter, while in the tree-aware method we consider a constant molecular clock and use distances between any two leaves-node as second parameter of the model.

1.2 Coevolutionary model

Coevolutionary model\(^7\) presented here is an extension of Continuous Time Markov Process (CTMP) sequence substitution. We have implemented this algorithm and modified parameters so that it could be used in our studies. An overview of the model is present in Figure 1. Instantaneous transition rate matrix (Q) is a 24x24 substitution
matrix, which could have different dimension depending upon the type of substitution matrix used. Q matrix contains transition rates for single amino acid substitutions. Lets consider two protein domains M1 and M2 (Figure 1), each containing alignment of four amino acid sequences. The probability of observing changes in any two amino acids simultaneously can be calculated by using the equation shown on the right-most tree (Figure 1). In this tree we are calculating on last column of M1 and third column of M2. Q\text{indep} is the substitution rate matrix where we consider that each site evolves independently. Therefore in Q\text{indep} substitution rate for double changes is 0 while in Q\text{cov} matrix we penalize the single substitution and reward double substitution.

\begin{equation}
Q_2^e[(a_1, a_2), (b_1, b_2)] = \\
\begin{cases} 
\varepsilon Q_2^e[(a_1, a_2), (b_1, b_2)] & \text{if } (a_1 = b_1) \lor (a_2 = b_2), \\
\tau_{(a_1,a_2)} & \text{if } (a_1 \neq b_1) \land (a_2 \neq b_2), \\
-\sum Q_2^e[(a_1, a_2), (b'_1, b'_2)] & \text{if } (a_1 = b_1) \land (a_2 = b_2). 
\end{cases}
\end{equation}

\text{(Yeang & Haussler 2007)}

In our project, we used the Dayhoff matrix of amino acid substitution. In double substitution model the dimensions for the probability vector and the rate matrix are respectively 1x576 and 576x576. Dependent double substitution matrix (coevolution matrix) is obtained from an independent double substitution matrix (Q_2) by considering three different cases as shown in equation 1. In the equation, a1 and a2 are two initial states of amino acids, which are changed to b1 and b2. For first case where we observe only single amino acid change, the rate of coevolution based double substitution matrix (Q_2^e) can be calculated by the multiplying penalizing value (\varepsilon < 1) with substitution rate (Q_2^i). Q_2^i is equal to substitution rate of either a1 to b1 or a2 to b2 from single substitution matrix. In second case, both amino acids are changed and these changes are considered as potential coevolving residues and we reward this transition with a positive score. Third case is where both amino acid retain their form

---

Figure 1. Framework of coevolutionary model (Yeang & Haussler 2007)
and rate value is equal to \(-1\times(\text{sum of all columns in a1a2 row})\). It is because sum of all the columns in a row has to be 0.

### 1.3 Multiple sequence alignments and Inferring phylogeny

Sequences have to be aligned before being examined for coevolution. Multiple alignments of protein sequences are important in many applications, including phylogenetic tree estimation, secondary structure prediction and critical residue identification. Many multiple sequence alignment (MSA) algorithms have been proposed\(^\text{17}\). Two attributes of MSA programs are of primary importance to the user: biological accuracy and computational complexity. Complexity is of increasing relevance due to the rapid growth of sequence databases, which now contain enough representatives of larger protein families to exceed the capacity of most current programs. Obtaining biologically accurate alignments is also a challenge, as the best methods sometimes fail to align readily apparent conserved motifs.

For aligning our sequences prior to coevolution analysis, we used the multiple alignment software MUSCLE\(^\text{18}\), a MSA program that is considered to provide improvements in accuracy and speed over the more traditional ClustalW\(^\text{19}\). We created trees using ClustalX\(^\text{19}\) based on the neighbour-joining method\(^\text{20}\). It is a bottom-up clustering method used for the construction of phylogenetic tree.

### 1.4 Examined interacting pairs

In this project, we have examined three datasets as controls and then applied model to Nod factor receptor (NFR) proteins. NFRs are receptor kinases involved in the recognition of bacterial nod factor signal molecules during the symbiosis between legume \textit{Lotus japonicus} and nitrogen-fixing bacteria termed rhizobia\(^1\). There are more than 18,000 LysM proteins with known\(^\text{22}\) sequences but only four protein structures (2lp\(^9\text{21}\), 2djp, 1e0g\(^22\), 1y7m\(^23\)) are available each consisting a single LysM domain structure. In \textit{L. japonicus}, NFR proteins have three LysM domains and nature of interaction among domains is not described in previous studies. LysM domains are approximately 40 residues long, and all three LysM domains of NFR proteins are similar in sequence and structure.

### 1.5 Experimental procedures

#### 1.5.1 Dataset

Four datasets of interacting proteins were used for this study but only Elongation factor proteins and LysM proteins are discussed here in detail:

- Elongation factors Ts and Tu (1487 sequences each)
  - Length of aligned sequences:
    - EF-Ts: 862 residues
    - EF-Tu: 640 residues
- 3 LysM domain containing proteins (110 sequences)
  - Length of align sequences:
- LysMs: 250 residues
- Copper pumps and chaperones (255 sequences each)
  - Length of aligned sequences:
    - Copper pump: 1867 residues
    - Copper chaperone: 362 residues
- 2-oxoisovalerate, subunits A and B (205 sequences each)
  - Length of aligned sequences:
    - Subunit A: 988 residues
    - Subunit B: 1301 residues

1.5.2 Removing redundancy

All the sequences were downloaded from Uniprot\textsuperscript{24}, where we found many sequences to be redundant. Duplicates were removed prior to data analysis using a custom python script.

1.5.3 Creating phylogenetic trees

Another python script was created to combine the corresponding sequences based on their species of origin into a format that could be used as an input for constructing a phylogenetic tree. A Neighbor-joining method was used on our datasets using ClustalX\textsuperscript{2}\textsuperscript{19}.

1.5.4 Calculating independent and correlated scores

The independent score $Q_i$, correlated score $Q_c$ and tree-dependent $Q_{ct}$ score were calculated. To calculate significant coevolving residues independent model score $Q_i$ was subtracted from the corresponding dependent model score. Subtraction of independent model from dependent model removes the false positives in the calculations.

1.5.5 Generating controls

Two types of controls were generated in order to test the statistical significance of our data. Random sequences of lengths comparable to the ones analyzed were generated using a method developed by Stothard\textsuperscript{25}, using the software Random Protein Sequence from the Sequence Manipulation Suite. Graphs of analyzed sequences versus control sequences are shown in result section.

1.5.6 Calculating inter-residual distances

Distances were calculated from the PDB structures, which were used in the alignments. Comparison between C\alpha distances and scores was done in order to determine whether residues are directly interacting or not. The distribution of the inter-residual distances against coevolution score is analyzed in detail in the result section.
1.5.7 Calculating the frequency of residues

The occurrence of residues was calculated in order to distinguish the coevolving residues from co-conserved ones. Co-conserved residues generate false positives and must therefore be eliminated. It was observed that few residues with high coevolution score were co-conserved.

1.6 Results and discussion

1.6.1 Coevolution score comparison within protein domains

In our study, we have obtained coevolution scores from comparing two interacting domains: A (mostly EF-Ts) and B (mostly EF-Tu). The results are represented in graphical form. The penalizing e-value is shown in the topmost left corner, scores are on the Z-axis while A and B are on the X and Y-axes, respectively. The data contained in the graph data is locally smoothed using average values of neighboring points.

In Figure 2, it is possible to see the coevolution scores in a variant of EF-Ts and EF-Tu with no gaps removed. The gaps generate large peaks. Tree-ignorant (top layer, multicolor) and tree-aware (bottom layer, red) methods show the same overall pattern, implying that the relationship between peaks does not change even when taking the time factor into account.

![Figure 2. Tree dependent and tree independent scores from Elongation factor proteins](image)

Removing the gaps, which generate many false positives changes the picture drastically, as can be seen in the figures that follow. In Figure 3a and 3b, the Qi (coevolution-independent) and Qc (coevolution-dependent) scores are shown.
Q_i is the score calculated based on the coevolution independent model, which implies that two sites are evolving independently and this score is then subtracted from the Q_c and Q_c (Coevolution dependent models). We could observe from data presented in Figures 4a and 4b that tree dependent model alters the magnitude of scores but the global pattern of coevolution retains its shape. In recent studies^{26}, it has been seen that coevolution is independent of phylogeny and we observe similar results.

In order to test the model, we generated random protein sequences with a similar data size as of copper pump-chaperone pairs as an additional control. Random sequences were divided into two groups of each 800 amino acids long. Figure 5a shows the respective coevolution score values (Q_c - Q_i). A significant reduction in the coevolution score can be easily observed when superimposing this (random sequences) score onto
coevolution score values from the copper pump-chaperone pairs (Figure 5b), it looks like flat sheet indicating very little coevolution in random sequences.

Figure 5a & 5b. Random sequence coevolution compared with copper pump-chaperone pair

Once insured on the quality of coevolution prediction from the control datasets, we applied the model on the LysM structures in order to find the coevolving residues and then modeled the three LysM domains according to the residues predicted. A few potential residues seemed to be involved in the spatial organization of the LysM structure. Figures 6a and 6b give insight in predicted coevolution among amino acid positions of the three LysM domains of NFR5 protein and a detailed description will follow in next section.

Figure 6a and 6b. Coevolution scores among the three LysM domains of NFR5 protein.

1.6.2 Coevolution score comparison with inter-residual distances

We calculated the distribution of coevolving scores against inter-residual distances in order to analyze whether coevolving residues are not always in the contact. Ca
distances within residues were calculated. We plotted two graphs (Figures 7a & 7b) for elongation factors Ts-Tu and 2-oxoisovalerate dehydrogenase subunits A and B, respectively. We can observe that the positions with high coevolution score tend to have lower inter-atomic distance tough the pattern is not very obvious.

![Figure 7a & b. Inter-atomic distance (between Cα atoms) distribution against coevolution score.](image)

**1.6.3 Coevolution score comparison with co-conservation**

Pairs of positions from two different domains might be fully conserved and could lead to false positive coevolving pairs in some cases. In Figure 8a and 8b, we have demonstrated that highly co-conserved residues in Ts-Tu (>95% identity) are not involved in coevolution (score is Qc-Qi). It can be observed that many highly co-conserved sites have low coevolution scores.

![Figure 8a & 8b. A comparison of conservation among the amino acid site with the coevolution scores (Qc - Qi) and Qc scores for Ts-Tu protein domains.](image)

**1.7 NFR5 structure**

NFR5 (Nod factor receptor 5) contains three LysM domains and shares high degree of similarity with NFR1 protein. Only single LysM domain containing proteins have so far been submitted to PDB protein database and three homologous structures are
available. Here we have modeled the individual LysM domains of NFR5 based on the homology modeling\textsuperscript{27}.

NFR proteins contain two highly conserved cystines\textsuperscript{22} which were suggested to form disulfide bonds in order to stabilize three-domain structure. We used de-novo disulfide bond prediction and manually combined three domains using Pymol\textsuperscript{28}. Table 1 shows the potential coevolving residues with high scores. Most of the high scoring residues are in close proximity suggesting that modeled structure is in close agreement with the coevolution results.

<table>
<thead>
<tr>
<th>Contact between Resides</th>
<th>Odd Score Tree Independent</th>
<th>Tree dependent</th>
<th>Intermolecular distances (C\textalpha{}-C\textalpha{})</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>101</td>
<td>102.277</td>
<td>77.607</td>
</tr>
<tr>
<td>74</td>
<td>126</td>
<td>101.536</td>
<td>77.367</td>
</tr>
<tr>
<td>18</td>
<td>101</td>
<td>99.501</td>
<td>76.084</td>
</tr>
<tr>
<td>101</td>
<td>180</td>
<td>98.320</td>
<td>74.540</td>
</tr>
<tr>
<td>101</td>
<td>148</td>
<td>97.670</td>
<td>73.491</td>
</tr>
<tr>
<td>18</td>
<td>82</td>
<td>95.368</td>
<td>73.067</td>
</tr>
</tbody>
</table>

Table 1. Coevolving residues with high score are shown. Distances between atoms are in Å.

Two of the high scoring pairs reside in the LysM domain 2 (Figure 9). These are intra-domain pair (Glycine-74:Asparagine-126 and Glutamate-88:valine-101 ) and therefore it does not provide any useful information about the global structure of the protein. However, the result suggests that coevolution may have had a role in the LysM domain-2 folding. There are pairs such as proline-148 coevolving with valine-101 and valine-18 coevolving with glutamate-88 suggesting that the interatomic distance between these residues could be shorter than in the modeled protein structure.

![Figure 9](image_url)

Figure 9. Coevolution prediction results within the three domains of the NFR5 protein. Predicted coevolving residues are indicated in the orange and pairs are shown by dotted yellow lines.

Taking into the consideration different possible spatial domain orientation of domains as well as the coevolution scores for the respective pairs, we have suggested a potential NFR5 structure shown in Figure 9.
1.8 Conclusions

From the results we conclude that we can predict coevolving residues from the multiple sequence alignment using both tree-aware and tree-ignorant methods, which give a similar coevolution pattern. From both methods we can agree on the coevolving residues predicted for the NFR5 protein. We have observed from the results that coevolving residues are not always in close proximity. In the control dataset Coevolving residues were found as intra- or inter-domain pairs, many of the latter belonging to the hydrophobic protein cores. We also observed that coevolving residues often have a significant amount of co-conservation. On the basis of coevolving residues, it was possible for us to combine three LysM domains together to form a complete extra cellular part of NFR5 protein. Three LysM domains form a symmetrical structure as shown in figure 10.

Figure 10a and 10b. Side and top-view of NFR5 protein structure.

1.9 Future perspectives

Considering the importance and applicability of this method, we would like to continue our work and expand this model with use of the maximum likelihood procedure and then also incorporate weighted parsimony methods to compare the results obtained from the our work. The main aim of the project was to find the coevolving residues NFR5 and ligand (Nod-factors) but we could not achieve as lack of appropriate data.

Though before we start working more on the predicted structure we would like to perform further tests and validate the coevolution calculations. Modeling of structure also considered the disulphide bridges prediction from multiple servers, which is unreliable. Hopefully experimental evidence for disulphide bridges will soon be provided by CARB. The software developed for detecting amino acid coevolution may also be used to identify coevolution between NFR5 sequence and decorations of the cognate nod-factors (ligand) but presently the available data for this calculation is insufficient.
2. Part-II ShortRan

2.1 Introduction

Several noncoding small RNA categories have been described and annotated in both, plants and metazoans\(^2^9\). Small RNAs are often characterized as ranging in size around 19-24 nucleotides. Many small RNAs have been implemented in the regulation of the gene expression. The commonly accepted categories are short interfering (si)RNAs\(^3^0\), small temporal (st)RNAs\(^3^1\), heterochromatic siRNAs\(^3^2\), tiny noncoding RNAs\(^3^3\) and micro (mi)RNAs\(^3^4\). RNA-Seq technologies have been critical in gaining insight into the RNA world. Recent advancements in the high-throughput deep sequencing methods\(^3^5\) have led to the availability of large amount of data with comparatively reduced cost and time investment. Small RNA profiling project such as for Medicago truncatula\(^3^6\) and Brassica napus\(^3^7\) have improved our understanding of various small RNA categories in plants.

The analysis of small RNA projects faces a major bottleneck due to unavailability of flexible informatics tools, which can be used to test complex biological hypotheses. A number of tools for analyzing gene expressions profiles have been developed such as DEGseq\(^3^8\), edgeR\(^1^1\) and MeV\(^1^2\). There has been also an effort on developing combined processing toolkit for sRNA sequences such as the UEA toolkit\(^3^9\) wapRNA\(^4^0\) and Dario\(^4^1\). During the analysis of small RNA data produced in our laboratory, we faced challenges due to numerous dependencies, which need to be installed when setting up such toolkits locally, while web-based tools are often pose limitations to the amount of data that can be uploaded and to the adaptability of analysis parameters. None of the currently available tools offer an easy integration of expression and annotation data while maintaining easy interface parameters to perform querying of complex hypothesis. Here we aim to resolve these issues by implementing an easy to install, python based pipeline. It can be used for pre-processing, profiling, categorizing, annotating and testing of hypotheses on integrated output.

In our analysis, we have utilized small RNA data produced from Lotus japonicus early infection libraries as described in Table-2. The data has been thoroughly processed and biological findings are presented as manuscript [under revision]. In this report, shortRan pipeline modules and results obtained from the each module are presented.

<table>
<thead>
<tr>
<th>Sample #</th>
<th>Genotype</th>
<th>Treatment</th>
<th>Designation in filenames</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>wild type</td>
<td>mock (3 dpi)</td>
<td>wt_mock</td>
</tr>
<tr>
<td>2</td>
<td>wild type</td>
<td>M. loti (3 dpi)</td>
<td>wt_mloti</td>
</tr>
<tr>
<td>3</td>
<td>nfr5</td>
<td>mock (3 dpi)</td>
<td>nfr5_mock</td>
</tr>
<tr>
<td>4</td>
<td>nfr5</td>
<td>M. loti (3 dpi)</td>
<td>nfr5_mloti</td>
</tr>
<tr>
<td>5</td>
<td>nfr1</td>
<td>mock (3 dpi)</td>
<td>nfr1_mock</td>
</tr>
<tr>
<td>6</td>
<td>nfr1</td>
<td>M. loti (3 dpi)</td>
<td>nfr1_mloti</td>
</tr>
<tr>
<td>7</td>
<td>symrk</td>
<td>mock (3 dpi)</td>
<td>symrk_mock</td>
</tr>
<tr>
<td>8</td>
<td>symrk</td>
<td>M. loti (3 dpi)</td>
<td>symrk_mloti</td>
</tr>
<tr>
<td>9</td>
<td>ccamk</td>
<td>mock (3 dpi)</td>
<td>ccamk_mock</td>
</tr>
<tr>
<td>10</td>
<td>ccamk</td>
<td>M. loti (3 dpi)</td>
<td>ccamk_mloti</td>
</tr>
<tr>
<td>11</td>
<td>cyclops</td>
<td>mock (3 dpi)</td>
<td>cyclops_mock</td>
</tr>
<tr>
<td>12</td>
<td>cyclops</td>
<td>M. loti (3 dpi)</td>
<td>cyclops_mloti</td>
</tr>
</tbody>
</table>

Table 2 Tissue samples used to generate small RNA libraries.
2.2 shortRan pipeline

*shortRan* is designed for processing small RNA data produced by deep sequencing. It has been tested on unix based systems and consist of nine independent modules. It is easy to setup and requires minimal dependencies to be installed. *shortRan* is capable of performing filtering, read error correction, expression profiling, mapping, clustering, miRNA/ta-siRNA prediction, pattern search, target prediction and annotation. The pipeline produces a wide range of outputs such as profile tables compatible with other expression analysis tools and BED, IGV format files that can be visualized by UCSC\(^{42}\) and IGV\(^{10}\) browsers etc. It also produces graphical output for easier interpretation and a MySQL database for custom querying on the data produced from each module. Individual sub-folders are produced for each module output. Figure 11 provides an overview of the *shortRan* pipeline.

Figure-11. Overview of *shortRan* pipeline modules.

### 2.2.1 Module-1: Error correction, filtering and profiling

The first module of the pipeline offers pre-processing of the raw data obtained from sequencing (fastq format). Here one can also perform adaptor filtering and size fractionation. One can also correct for sequencing errors by k-mer correction using ECHO\(^{43}\). Module-1 also offers filtering of small reads based on homology search against a user supplied multi-fasta file. It can be useful if one requires filtering of repetitive sequences. The module outputs filtered fastq files, profile tables (in text format and as MySQL table), statistics summary files and graphs for small abundance across different sizes and libraries. In the profile tables, each sequence is associated with a variation score, which represents the degree of differential regulation for a particular sequence among all tested libraries. This variation score is calculated as below:

\[
\text{Variation score}_{i} = \frac{\text{Standard deviation (data}_{i})}{\sqrt{\text{Average (data}_{i})}}
\]

where, data\(_i\) is the list containing all the expression values of row i


2.2.2 Module-2: Genomic Mapping

This module maps all the filtered reads to a given reference using Bowtie software. Here we provide optional parameters to provide flexibility in the analysis and for comparative analysis. A user is asked to provide a reference sequence. One of the outputs from this module is an IGV format file, which can be accessed using genome browser programs such as IGV and contains sequencing reads, respective chromosomal positions, and expression values for individual libraries. A second output is a SAM formatted alignment file that can also be accessed by variety of other tools.

2.2.3 Module-3: Cluster analysis

Here, the term cluster refers to a locus where individual small RNA loci are in close proximity to one another. We wrote a script, which can identify small RNA locus clusters fulfilling a certain set of parameters. With default options, a cluster is defined as a genomic region, containing at least five unique reads and no gap longer than 50 nucleotides between consecutive reads. Also the combined read abundance should sum up to a total of at least 75. Parameters can be changed to adjust the stringency. For each identified cluster, output file lists chromosomal location, small RNA contained, individual small RNA abundances, cluster length and the overlapping genomic annotation such as exonic, intronic etc. An option for a cluster analysis considering only defined small RNA size fractions is available.

2.2.4 Module-4: miRNA predictions

miRNAs\(^{29}\) are small RNAs with an average length of 22 nt. These are post-transcriptional regulators that act on a nearly perfect complementary sequence on the target messenger RNA transcripts containing nearly perfectly complementary sequence stretches. There are a number of tools for predicting miRNAs such as miRCat\(^{39}\), miRNAkey\(^{44}\), miRanalyzer\(^{45}\), miRtools\(^{46}\) etc. Here we use miRdeep-P\(^{47}\), which has been optimize for the plant databases. As all the modules are independent user can use other available tools as control for consistency and for removing false positives. miRNA precursors and predictions of their secondary structures are provided in the output. Optionally, the closely related known miRNA homologs are listed for each small RNA sequence.

2.2.5 Module-5: ta-siRNA predictions

Trans-acting (ta) siRNAs\(^{48}\) represent one form of small interfering RNAs known only from plants. These small RNAs repress gene expression through post-transcriptional gene silencing. ta-siRNAs are produced from the genome to form a polyadenylated, double-stranded precursor. Only few publications are available that present algorithms for de-novo prediction of ta-siRNAs. In shortRan, we have implemented modified version of an algorithm proposed by Chen et al.\(^{49}\).
2.2.6 Module-6: Annotation

Annotation is critical for understanding the possible biological function of a sequence. Annotation can include information of various kinds, such as genomic location, size category or similarity with other annotated sequences. In this module, reads are annotated based on homology to known miRNAs, predicted tasi-RNAs, cluster sequences, repetitive sequences, genomic reference sequence etc. Small RNA sequences also are annotated based on their genomic location such exonic, intronic or intergenic. All the annotations are autonomously added to a back-end MySQL database when the pipeline is executed.

2.2.7 Module-7: Expression Plots

Small RNA expression values represent the basis for evaluating a sequences regulation in a given condition as compared to another one (i.e. particular genotype/treatment combinations). To compare sequences contained in different datasets or sequencing libraries, the expression values of the different libraries are normalized and a correlation co-efficient is calculated to find the degree of similarity between them. Expression plots are generated based on normalized expression counts for two pairs of libraries, for example two genotypes where two treatment conditions are compared, at a time. In this way, global small RNA regulation patterns can be used to evaluate whether and in what way responses in different genotypes are correlated.

2.2.8 Module-8: Pattern Search

Several studies have demonstrated the occurrence of biases introduced by the adaptor ligation to the 5’ end of RNA molecules during sequencing library preparation. Overabundances of sequences with particular starting nucleotides might thus reflect technical artifacts rather than biologically significant phenomena. We have implemented this module to detect any possible biases in 5’ nucleotide abundances by calculating the multi-order Markov probabilities for each nucleotide at the sequences’ 5’ ends.

2.2.9 Module-9: Queries

Facilitating the analysis and validation of complex biological hypotheses based on small RNA sequencing data is one of the most important contributions of our pipeline. MySQL is a structure based query language and a multitude of assumptions can be tested with just a few lines of queries used to extract database information. For example, if we have a MySQL table formed by shortRan containing expression values and miRBase annotation and we would like to retrieve average response values for sRNAs upregulated in one genotype compared to another only for sRNAs mapping to miRNAs annotated in miRBase, we can use the following query:

```
SELECT response_value, AVG(LOG(genotype_1/ genotype_2)) AS genotype,
FROM table
WHERE LOG(genotype_1/genotype_2) > 0
AND `miRBase.fa` !="0";
```
2.3 Data analysis

*shortRan* was used to process ~70 million (~21 million unique) reads produced from deep sequencing of 12 sRNA-seq libraries representing six genotype that were either inoculated with a rhizobial symbiont or mock-treated. These reads were of size classes 19 to 24, other size classes were filtered out as most of the regulatory small RNAs including miRNAs and siRNAs are predicted to belong to this size range. An overview of the data processing is shown in figure 12.

![Flowchart showing data processing](image)

Figure -12. A flowchart summarizing the output counts from each module.

2.3.1 Data pre-processing

All the reads were corrected using a k-mer correction algorithm. This algorithm is aimed to correct unambiguous characters indicated by 'N' where by sequence comparisons they can be unambiguously assigned to an actual nucleotide. Approximately 5.86% of the 21 million reads were corrected. Each of the libraries (genotypes/treatment combinations) had approximately equal distribution of non-redundant reads while there was a slight variation in the total library sizes (Figure 13). It is important to take library size into consideration when comparing across library as normalized library size for each genotype should be equal. Often sequencing data sets contain a large pool of RNA reads which are very lowly abundant. We removed these
reads by applying an abundance cut-off of 20. We then mapped the reads to the TIGR Fabaceae Repeat Database\textsuperscript{51} to remove the reads mapping to transposable elements, centromere-related, telomere-related and rRNA sequences. Following abundance and repeat filtering, 25,559,282 reads (corresponding to 266,979 unique reads) were left. To identify rhizobia-induced regulation of small RNAs, we applied additional variation score filtering on the normalized abundance of each unique sequence across the libraries. This step filtered out most the unique reads and only 18,547 non-redundant reads for the onward analysis.

![Raw Read Counts](image)

Figure-13: Library size overview. Redundant counts represent the total number of reads in each library. The Non-redundant counts represent the total number of unique sequences with each library. This figure is generated to allow checking for consistent library sizes.

We analyzed the fraction of each size classes in all the reads and compared it with differentially regulated reads (Figure 14). Size class 24 contains 58.3% of the total reads while size class 21 encompasses 18.1%. However, when we observing only differentially regulated reads, a big increase in the fraction represented by the 21 nt size was observed. It could be explain by the fact that often 24 nt-size fraction is primarily composed of repetitive sequences and that because the Fabacean Repeat database is not complete, we retained a significant amount of unresponsive 24-size class reads in our dataset.

![All data length distribution](image)

Figure 14a & b. Library length fraction distribution overview.
Small RNA populations were quite distinct when comparing uninoculated and inoculated Lotus japonicus plants. An increase of more than 10 fold can be observed in 971 sequences in inoculated wild type roots as compared to mock treated roots, and 46 sequences were more than 100-fold abundant in infected roots. From the Table 3, we can also observe an even larger pool of 1,810 sequences, which were downregulated by 10-fold or more, with 33 of these sequences showing a 100-fold decrease in abundance upon inoculation in the wild type. These data show that host responses to bacterial infection involve significant changes in small RNA abundance levels at an early stage of symbiosis initiation.

<table>
<thead>
<tr>
<th>genotype(s)</th>
<th>10-fold down</th>
<th>100-fold down</th>
<th>10-fold up</th>
<th>100-fold up</th>
<th>total (10-fold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wt ; nfr5 ; nfr1</td>
<td>1.011</td>
<td>10</td>
<td>875</td>
<td>36</td>
<td>1.886</td>
</tr>
<tr>
<td>wt ; nfr5 ; nfr1</td>
<td>802</td>
<td>37</td>
<td>309</td>
<td>8</td>
<td>1.111</td>
</tr>
<tr>
<td>all genotypes</td>
<td>171</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>178</td>
</tr>
<tr>
<td>wt total</td>
<td>1.810</td>
<td>33</td>
<td>971</td>
<td>46</td>
<td>2.781</td>
</tr>
<tr>
<td>wt only</td>
<td>49</td>
<td>6</td>
<td>54</td>
<td>10</td>
<td>103</td>
</tr>
<tr>
<td>nfr5 only</td>
<td>57</td>
<td>11</td>
<td>49</td>
<td>3</td>
<td>106</td>
</tr>
<tr>
<td>nfr1 only</td>
<td>62</td>
<td>3</td>
<td>82</td>
<td>34</td>
<td>144</td>
</tr>
<tr>
<td>symbk only</td>
<td>62</td>
<td>13</td>
<td>46</td>
<td>7</td>
<td>108</td>
</tr>
<tr>
<td>ccamk only</td>
<td>452</td>
<td>106</td>
<td>106</td>
<td>1</td>
<td>558</td>
</tr>
<tr>
<td>cyclops only</td>
<td>89</td>
<td>20</td>
<td>298</td>
<td>10</td>
<td>387</td>
</tr>
</tbody>
</table>

Table-3 Unique sequences (19-24 nt) responsive to rhizobial inoculation.

2.3.2 Genomic Mapping

Bowtie was used as a mapping tool while allowing a maximum of three mismatches and applying a seed length of 25. Reads were mapped to the L. japonicus genome in two groups; group-1 contained all the reads while group-2 contained differentially regulated reads (variation score > 3). 195,563 (73.25%) unique sequences were mapped to 1,694,666 distinct locations on the genome from first group of 266,979 unique reads and for second group with 18,547 reads, a total of 14,648 (78.97%) reads were mapped on 123,904 locations. As the current L. japonicus reference genome covers approximately 90% of the complete genome, the amount of reads mapped provides indication good sequence quality and coverage.

2.3.3 Cluster analysis

We analyzed the genomic distribution of small RNAs to find the small RNA generating loci specific to the symbiosis. Genomic positions to which small RNAs mapped without mismatches were considered potential source loci. We can observe a significant amount of the intragenic clusters for each size class (Table 4). Many of the unique regulated sequences in the clusters were regulated in symbiotic-dependent manner in particular size class 21. To investigate if the differentially regulated small RNA sequences or clusters have any preferences for genomic region, we calculated fractions mapping to the exonic, intronic, intergenic and overlapping region (fraction of reads which mapped on the junctions). From figure 15, we could observe an increase in the differentially regulated exonic small RNAs. This could be explained by two possibilities. First, many of the transcripts might go through degradation due to particular genotype change. Second, particular genotypic condition might require controlling expression of many
genes therefore cell producing higher amount of exonic small RNA to control the targeted genes. But from the cluster distributions we can observe a reduction in the exonic clusters, which contradicts the first possibility because the transcript degradation should have resulted clustered-exonic small RNAs.

<table>
<thead>
<tr>
<th>sequence length</th>
<th>clusters</th>
<th>intragenic clusters</th>
<th>exonic clusters</th>
<th>unique sequences in clusters</th>
<th>% *</th>
<th>clusters</th>
<th>unique regulated sequences in clusters</th>
<th>% **</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>74</td>
<td>22</td>
<td>17</td>
<td>435</td>
<td>33.33</td>
<td>47</td>
<td>31.97</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>121</td>
<td>24</td>
<td>17</td>
<td>747</td>
<td>30.88</td>
<td>84</td>
<td>42.00</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1181</td>
<td>423</td>
<td>326</td>
<td>7596</td>
<td>39.84</td>
<td>666</td>
<td>49.74</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>3773</td>
<td>1367</td>
<td>1185</td>
<td>11290</td>
<td>40.85</td>
<td>742</td>
<td>44.11</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>2640</td>
<td>258</td>
<td>122</td>
<td>3546</td>
<td>21.06</td>
<td>133</td>
<td>29.89</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>21411</td>
<td>1880</td>
<td>756</td>
<td>46096</td>
<td>38.12</td>
<td>1325</td>
<td>42.58</td>
<td></td>
</tr>
<tr>
<td>19-24 combined</td>
<td>29963</td>
<td>4015</td>
<td>2462</td>
<td>83296</td>
<td>44.26</td>
<td>3373</td>
<td>48.71</td>
<td></td>
</tr>
</tbody>
</table>

1Reference value: total unique sequences 19-24 nt (cutoff 20)
2Reference value: total unique regulated sequences 19-24 nt (cutoff 20; variation score cutoff 3)

Table-4 Compared to total unique sequences, a high proportion of regulated sequences stem from clustered loci.

2.3.4 Annotation

To identify the conserved miRNA in the dataset we mapped small RNA sequences to the mirBase52 version 17 and identified 603 unique small sequences which were mapped (up to 3 mismatches ) to known plant miRNA or miRNA*. Responses of these miRNAs seems to be related to rhizobium inoculation as 68 sequences were more than 10-fold regulated in at least one of the tested genotype. For filtering the false positive candidates we performed further annotation of the small RNA reads by mapping these on Lotus repeats and on the bacterial symbiont M. loti. Reads were also mapped to the host genome to remove any artificial sequences originated during the amplification of the sequences. Further annotation was added by categorization based on de-novo miRNA and ta-siRNA prediction. Using mirDeep-P, we predicted 143 potential miRNAs, out of which 38 candidates were conserved miRNA in closely related species while 105 miRNA candidates were novel. Intriguingly almost half of the predicted miRNAs were differentially regulated (variation score > 3) and 30 conserved miRNAs out of 38 were found in this category. 1,934 unique small RNA sequences were found be generated from predicted 25 TAS-loci and seven of these had successful Blast hit against known defense response related NBS-LRR family genes53.
2.3.5 Comparative quantitative analysis and plotting

To obtain an overview of small RNA expression among the different genotype, we generated pairwise scatter plots. Here, we have presented two different approaches to plot the small RNA expression distribution between the nod factor receptor mutants nfr1 and nfr5 and the similarities between nfr1 and wild type responses. In Figure 16 the top panels show direct comparisons of log (count) values for inoculated samples. The lower panels show comparisons of inoculated/mock changes according to the formula $e = (i-m)/(i+m)$, where $e$= expression score, $i$=inoculated count and $m$=mock count. We could observe that during the early infection of plant with rhizobia, two phenotypes of NFR5 and NFR1 (left side) contain many small RNAs, which are inversely regulated while in the comparison of nfr1 with wildtype (right side) we could observe a strong correlation. shortRan also calculates spearman correlation coefficients for each pair, table 5 summarizes the coefficients for all pairs and from the values we could conclude that targeted five phenotypes can be divided into two separate groups based on their expression patterns, first group NFR5, SYMRK and CYCLOPS genes involved independently of second group NFR5 and CCaMK at the early symbiosis stage.
Figure 16. Inter-library expression comparison.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mock_wild_type</th>
<th>wild_type</th>
<th>Mock_NFR5</th>
<th>NFR5</th>
<th>Mock_NFR1</th>
<th>NFR1</th>
<th>Mock_SYMRK</th>
<th>SYMRK</th>
<th>Mock_CCaMK</th>
<th>CCaMK</th>
<th>Mock_CYCLOPS</th>
<th>CYCLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mock_wild_type</td>
<td>0.204</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wild_type</td>
<td>0.385</td>
<td>0.521</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFR5</td>
<td>0.529</td>
<td>0.538</td>
<td>0.452</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mock_NFR1</td>
<td>0.939</td>
<td>0.225</td>
<td>0.399</td>
<td>0.567</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFR1</td>
<td>0.189</td>
<td>0.946</td>
<td>0.518</td>
<td>0.535</td>
<td>0.223</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mock_SYMRK</td>
<td>0.371</td>
<td>0.519</td>
<td>0.930</td>
<td>0.434</td>
<td>0.383</td>
<td>0.518</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYMRK</td>
<td>0.557</td>
<td>0.497</td>
<td>0.443</td>
<td>0.891</td>
<td>0.582</td>
<td>0.499</td>
<td>0.462</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mock_CCaMK</td>
<td>0.847</td>
<td>0.110</td>
<td>0.349</td>
<td>0.416</td>
<td>0.850</td>
<td>0.109</td>
<td>0.347</td>
<td>0.439</td>
<td>0.130</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCaMK</td>
<td>0.138</td>
<td>0.890</td>
<td>0.462</td>
<td>0.507</td>
<td>0.172</td>
<td>0.893</td>
<td>0.468</td>
<td>0.457</td>
<td>0.130</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mock_CYCLOPS</td>
<td>0.372</td>
<td>0.492</td>
<td>0.921</td>
<td>0.443</td>
<td>0.385</td>
<td>0.495</td>
<td>0.916</td>
<td>0.446</td>
<td>0.358</td>
<td>0.456</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYCLOPS</td>
<td>0.533</td>
<td>0.415</td>
<td>0.363</td>
<td>0.837</td>
<td>0.547</td>
<td>0.413</td>
<td>0.361</td>
<td>0.877</td>
<td>0.405</td>
<td>0.378</td>
<td>0.391</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Spearman rank correlation coefficients for normalized sRNA counts. Values > 0.8 are shown in bold. Tables were generated by Module 7.
2.4 Conclusions

Here we have presented the implementation and application of a small RNA data analysis pipeline. This pipeline has been quite useful for our research project and we will use it also in other small RNA data set analysis in second part of the PhD studies. In table 6, we have compared shortRan with existing small RNA data analysis pipeline. Many modules such as adapter filtering and miRNA predictions are included in existing pipelines but shortRan has unique features, which are not available in the competitive tools. It is only pipeline in comparison that uses MySQL backend to integrate the data in efficient manner. It can also perform multi-sample normalization and quantification. Target gene prediction is not included in our pipeline yet but we intend to add it when we are implementing graphical user interface for shortRan.

<table>
<thead>
<tr>
<th>Component</th>
<th>shortRan</th>
<th>UEA tools</th>
<th>wapRNA</th>
<th>DARIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adapter trimming</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>-</td>
</tr>
<tr>
<td>Repeat and abundance filtering</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>-</td>
</tr>
<tr>
<td>Multisample normalization and quantification</td>
<td>v</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Comparative expression analysis and plotting</td>
<td>v</td>
<td>v</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>miRNA prediction</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>Mapping against reference genome</td>
<td>v</td>
<td>v</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Analysis of sRNA-generating loci</td>
<td>v</td>
<td>v</td>
<td>-</td>
<td>v</td>
</tr>
<tr>
<td>Annotation by homology</td>
<td>v</td>
<td>-</td>
<td>v</td>
<td>-</td>
</tr>
<tr>
<td>Check for 5' sequence composition bias</td>
<td>v</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Target gene prediction</td>
<td>-</td>
<td>v</td>
<td>v</td>
<td>-</td>
</tr>
<tr>
<td>MySQL-based data integration and querying</td>
<td>v</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Web service</td>
<td>-</td>
<td>v</td>
<td>v</td>
<td>-</td>
</tr>
<tr>
<td>Downloadable package</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>-</td>
</tr>
</tbody>
</table>

a: A maximum of two samples. b: also predicts other categories of non-coding RNAs (tRNA, snoRNA, etc.)

**Table 6:** Comparison of sRNA-seq analysis pipelines.

Early infection dataset has been analyzed using shortRan and it has led to important conclusions from biological perspective. We have shown that NFR5 could act in NFR1-independent manner in response to rhizobial infection. From our results we suggest a novel NFR1-independent signaling loop involving NFR5 and the downstream components SYMRK and CYCLOPS, but not CCaMK.

2.5 Future Perspective

Recent study have shown a potential demand for small RNA data analysis tools, which are easy to use for the biologists. Current version of shortRan can be available as a downloadable package and is not a web-based or GUI (graphical user interface) based tool. In part B, we would focus on developing either developing a GUI based downloadable software or web-based interface version of the shortRan. Two additional experiment setups were sequenced for small RNA data. First was the dataset based on infection of *L. japonicus* by *Arbuscular mycorrhiza* and second dataset was sequenced from nodules formed on the *L. japonicus* plant root after the rhizobial bacterial infection. These two datasets will be analyzed in parallel with the improvement on the shortRan pipeline. We will be also focusing on adding more categories of small RNAs in the de-novo prediction, as more details on the nature of novel RNAs are available.
3. Part-III GapGlu

3.1 Introduction

Most of the species contain a wide range of repetitive DNA sequences\textsuperscript{56}. These repetitive sequences have presented challenges for genomic mapping and de-novo assembly. Increased short read production from next generation sequencing technologies has only added to the problem. Capillary-based (Sanger) sequencing methods produced 800-900 bp length that was considerably longer than the NGS read lengths which are currently 100-400 bp. If the repetitive element is longer than the reads, it often results as gaps in genome assembly.

Repeats could be resolved to some extent using pair-end and mate-pair reads. These pairs can be sequenced with range from 200 bp up to 20,000 bp. Even longer fragments could be produced using fosmid clones and bacterial artificial chromosome (BAC) based techniques that can produce approximately 30-40 kb and 140-150 kb of continuous sequence respectively. If a read pair spans a repeat region then assembler can use this information to move from one unique region to another while covering the repeat.

Here we have present the GapGlu, which utilizes the information from paired-end and mate pair sequencing to close up the gaps in a given genome sequence. We implemented this tool in order to improve the quality of the genome in the post assembly process. Even though highly competitive tools are available for assembly they still posses issues in terms of gaps retained in the genome. There have been few attempts to solve this problem caused by repetitive elements but insufficient progress in the subject keeps it open for future research and development.

Generally assembly programs require a pre-assembled genome from the same tool to remove gaps, i.e., GapCloser from SOAP package\textsuperscript{57}. GapGlu requires a set of contigs, some paired-end sequencing data in fastq format, and a list of unplaced sequences to fill the gaps. These contigs can also be part of targeted region where we are not sure about the order and orientations of the contigs within the region. We have tested this property on the 210 kb region around \textit{sym65} gene with 7 unordered fragments, where GapGlu successfully connected fragments in the correct order using mate-pair read information. The current assembled version of the \textit{L. japonicus} genome v2.4\textsuperscript{58} was assembled using 80% of total read pairs. 20% pairs contained at least one read mapping to the Fabaceae Repeat Database\textsuperscript{51} and were filtered prior to the assembly. Genome contains \textasciitilde13,000 gaps and we tried to fill these gaps using filtered reads and unplaced contigs.

3.2 Methods

\textit{GapGlu} is a python based tool and is executed using a shell script. The tool uses formatdb from blast to make a local database for the reference sequence. \textit{GapGlu} scans through the whole reference sequence for the gaps and outputs a list of gaps with
chromosomal positions. It extracts flanking regions from both up and down-stream of the gap. The size of flanking region is determined by the insertion size of the read-pairs.

We performed single-end mapping on the flanking regions and filtered away all the pairs, where both or none reads in the pair mapped on the flanking region of same gap. After this step we were left with the pairs, where only one read maps to the flanking region, here we refer these pairs as hanging-pairs. Unmapped reads of hanging pairs were then mapped to the list contigs and potential candidates were selected from the list. For a particular gap, each potential element was placed in the gap and a score was calculated (equation 2). This score is calculated from mapping output. Here relative insert size refers to the following:

relative insert size = (observed insert size – expected insert size)/expected insert size

$$\text{Score} = \sum_{\text{all libraries}} (1+\frac{1}{(\text{Avg}+\text{Dev})})\times \text{Rew} \quad \text{Equation 2}$$

Where, 
- $\text{Avg}$ = average of all relative insert sizes from read pairs 
- $\text{Dev}$ = deviance of all relative insert sizes from read pairs 
- $\text{Rew}$ = a positive score for observing a read in more than one libraries

The score parameters $\text{Avg}$, $\text{Dev}$ and $\text{Rew}$ were optimized using the control dataset. Many of known $L. japonicus$ repetitive elements are placed in the genome with help of BAC clones, which could cover the repetitive elements. For creating a control dataset, we identified these candidates using Blast and extracted 1000 candidates containing flanking regions with highest alignment score. We replaced the repetitive element by the N's, creating an artificial gap for testing the GapGlu tool. From figure 17, we can observe prediction accuracy higher than 80% in the candidates where have more than 98 % confidence based on blast results.

![Figure 17. GapGlu predictions on positive control. Numbers on x-axis represents the number of the gaps in the reference sequence.](image)
3.3 Results

3.3.1 Simulated Data

Non-uniform coverage creates problems when testing *GapGlu* as many of the elements could not be predicted because of lack of hanging pairs, which could confirm the potential element as a replacement for the gap. *wgsim*\(^5\) was used to generate reads of 100 bp with a uniform coverage of 40 times. In such a scenario, we recovered more than 96% of the correct elements for the data set which 100% confidence on blast result (Figure 18).

![Figure 18. GapGlu predictions for simulated data. X-axis represents the repetitive elements in reference identified with 100% confidence. In red, we have annotated percentage of gaps with potential replacement candidate and in the green column we have shown percentage of true-positives predicted.](image)

3.3.2 *L. japonicus* genome

Many of the gaps could potentially fill with the more than single potential elements. Therefore we predicted potential elements for both 5' and 3' end of the gap. For almost 86% of the gaps for all the chromosomes were predicted to have non-identical elements from 5' and 3' end. Gaps where we observed same elements from both sides could be successfully replaced but these were only 14% of total gap counts (Figure 19).

![Figure 19. Overview of GapGlu predictions on six chromosomes of *L. japonicus* genome](image)
3.4 Conclusions

Using GapGlu we have removed almost 20% of the gaps in genome and we continue to improve the quality of genome and its annotation as we obtain more data from *L. japonicus*. Recently we have sequenced transcriptomics data to improve *L. japonicus* gene models. The GapGlu pipeline is in an early stage of development and can currently process four different library sizes. It also has limited flexibility and if time permits we would like to work further on the solving repeat structure in the eukaryotic genomes.

3.5 Future perspective

In next few months we would like to finish the genome annotation of *L. japonicus* genome and release new version of genome. We would also like to integrate all the data such as genomic, transcriptomic, proteomics, small RNAs and other data. Integration of data is important because it will facilitate the comparison of results at different levels.

4. Discussion

In this report, I have presented the findings of my last two years of studies. At first, I have modeled a potential NFR5 structure using predictions from a coevolutionary model. This resulting modeled protein structure is symmetrical and shows three possible active sites for nod-factor binding. We also used the cysteine bond predictions to find the appropriate di-sulfide bridges. Even though these cysteine bond predictions represent the consensus of multiple methods, we would still prefer to validate using experimental procedures. The next step would be to use the coevolutionary model to find interacting pairs between the nod factor receptors (NFRs) and the nod factors produced by different rhizobial strains. Calculations for coevolution require a multiple sequence alignment of corresponding pairs from the same species but in the case of lotus – rhizobium pairs, not many respective homologs are known. Hence it is unrealistic to perform coevolution prediction for the NFR-nod factor pairs.

In the second project of the report, we discussed implementation and applicability of the *shortRan* pipeline. We found the pipeline very useful for small RNA sequencing data analysis. Modules such as adapter filtering, normalization, and genomic mapping provide basic analysis and at the same time miRNA prediction, ta-siRNA prediction and annotation modules helps to find appropriate candidates for further experimental analysis. We have tried to publish *shortRan* as a small RNA data analysis tool and it is currently under revision for publication in *Bioinformatics* while the data analysis results will be submitted shortly to *Plant Physiology*.

Our work on the *L. japonicus* genome is only partially complete. We have filled up a fraction of the gaps in the genome and currently trying to build a better assembly. A complete genome will be more useful for both small RNA and transcriptomic data analysis in further research work.
5. Expectations for part B

Next generation sequencing techniques are improving and both the data volumes and the number of different applications are growing rapidly. To meet the challenges of sequencing data analysis, we would like to make shortRan more user-friendly by implementing a graphical user interface and we will also update it with the advances in the sequencing techniques. Additional categorization sub-modules will be implemented, as we know more about novel small RNAs. Recently, we have sequenced *L. japonicus* transcriptomes from different genotype corresponding to small RNA data presented in this report. We will be focusing on comparing the small RNA expression with the transcripts across the genome and on expanding the shortRan pipeline to accommodate and automate this integrative approach.

The *L. japonicus* genome is partially completed and we aim to complete a substantial fraction of the genome within the next few months. My part in the analysis will be to carry out quality checks of the genome assembly by aligning Illumina sequencing data from genomic and transcriptomic libraries to the new assembly. I will also use transcriptomic data to increase the accuracy of current gene models and perform a detailed study on genetic variation between lotus accessions based on re-sequencing data from 30 natural Lotus accessions.

Building on the thoughts and results from developing the shortRan pipeline, a major aim of my Part B will be to continue integrating different data types, making both annotation and quantitative data from different sources and experiments easy to correlate and access.

The following publications are planned:

1) Up, down and Around: Identifying recurrent interactions within and between super secondary structures in β-propellers. Søren Thirup, Vikas Gupta and Esben M. Quistgaard (book chapter accepted)
3) Small RNA profiling reveals NFR1-independent symbiosis signalling. Katharina Markmann, Stig U. Andersen, Vikas Gupta, Enrique Kageyama and Jens Stougaard (to be re-submitted within the next two weeks).
5) Combined QTL and local association mapping paper (submission early 2013, co-author)
6) Analysis of correlations between sRNA and mRNA expression levels during early stages of rhizobial infection of *L. japonicus*. (submission 2013, co-author).
7) An integrative database for Lotus japonicus genetics, genomics, and transcriptomics data (main author, submission early 2014).
References