出國報告 (出國類別 ：研習)

# 赴荷蘭研習生物資訊課程

服務機關：行政院農業委員會農業試驗所
姓名職稱：曾馨儀助理研究員
派赴國家：荷蘭
出國期間：108 年 10 月 25 日至 109 年 2 月 07 日
報告日期：109 年 3 月 13 日

**摘要**

本計畫為科發基金補助計畫-農業科研國際化與產業化人才培育，工作項目係赴荷蘭瓦荷寧恩大學 (Wageningen University & Research，WUR) 研習生物資訊相關課程。已完成「資料分析與視覺化（Data Analysis and Visualization）」，研習利用 R 語言進行基因表現量分析與視覺化，包含群集分析、回歸、檢測、分類等，期末報告完成番茄外表型與成份分析報告。並完成「生物資料探勘 (Biological Discovery through Computation)」課程，研習細菌基因體資料分析，分析菌株與植物賀爾蒙之關聯性。研習本課程有助於數據資料分析技術提昇及資料視覺化呈現，未來可應用於我國種原之基因體資料建立，協助有用基因之開發及快速精準育種平台之建置。

# 目錄

## 一、研習目的

台灣擁有豐富的種原資源，但對種原基因體的了解有限，仍需以傳統育種方式進行篩選，且種原數量龐大需耗費大量時間與人力進行育種。目前可透過次世代定序結合生物資訊分析，快速取得大量基因體資料，可發展分子標誌或進行基因型選拔，生物資訊分析成為未來精準育種不可或缺的工具。本研習為前往荷蘭瓦荷寧恩大學(Wageningen University & Research，WUR) 研習生物資訊相關技術及課程，學習基因體資料分析，了解序列與基因間的相關性及外表性狀之關聯性，並透過電腦運算比對資料庫，進行基因體資料探勘及視覺化呈現。

## 二、出國行程

| 日期 | 地點 | 內容 |
|---|---|---|
| 108 年 10 月 25 日 ~108 年 10 月 26 日 | 臺灣→荷蘭 | 啟程搭機至荷蘭 |
| 108 年 10 月 26 日~ 108 年 12 月 20 日 | 荷蘭瓦赫寧恩大學 | 研習「資料分析與視覺化（Data Analysis and Visualization）」課程 |
| 108 年 12 月 21 日 ~109 年 01 月 05 日 | 荷蘭↔臺灣 | 自費返國休假 |
| 109 年 01 月 06 日 ~109 年 02 月 06 日 | 荷蘭瓦赫寧恩大學 | 研習「生物資料探勘 (Biological Discovery through Computation)」課程 |
| 109 年 02 月 07 日 | 荷蘭→臺灣 | 搭機返國 |

## 三、研習內容

### (一)「資料分析與視覺化」課程

課程由 Dr. Aalt-Jan van Dijk 助理教授、Dr. Jos Hageman 助理教授及 Prof. dr. ir. Dick de Ridder 教授共同指導，研習利用 R 語言進行基因表現量分析與視覺化，課程內容包含 R 程式撰寫與圖表視覺化、群集分析、回歸、檢測、分類等，正確使用統計方法尋找關鍵指標，研習過程注重統計方法使用的時機與條件，並非只是盲目的使用，課程中所分析之資料為癌症病人與正常人之基因體微陣列資料(microarray)，分析癌症病人與正常病人間基因體表現之差異。課程需每週完成課程指定考試及分析報告，所學之統計方法不僅可用於生物資料之分析，亦可應用於社會科學問題等數據資料分析。期末報告內容為番茄外表型與成份分析，資料來源為 Center for BioSystems Genomics (CBSG)，提供的資料為 93 個番茄品系，分為 3 個不同種類(cherry tomato, round tomato 與 breef tomato)，其 129 種代謝物的數值及甜味、番茄味、果重的數值，報告內容需以視覺化呈現資料，並利用決策樹找出能夠區分此番茄的指標代謝物，另一方面，利用代謝物資料建立甜味、番茄味、果重之預測模型，並找出影響性狀之重要之代謝物。

**(二)「生物資料探勘」課程**

課程由 Dr. Marnix Medema 助理教授授課，課程內容包含研究計畫撰寫，規劃實驗、執行及群組研究報告簡報及個人研究報告撰寫，過程中需與同組人員討論研究方法、制訂研究題目，並有助教從旁協助。研究計畫的內容為分析沙漠植物 *Indigofera argentae* (豆科)根圈附近所分離出 15 個菌株之基因體資料，本組討論之主題為探討植物益生菌與植物賀爾蒙生合成、固氮相關基因的關聯，試圖找出植物益生菌幫助植物生長的因子。過程中使用學習使用免費軟體 Draw.io 畫製流程圖，將整個計畫架構以視覺化方式呈現。計畫首先先從文獻中搜尋有關植物賀爾蒙生合成、固氮相關基因，並使用 UniProt 資料庫查詢植物賀爾蒙生合成、固氮相關基因之蛋白質資訊及 Pfam IDs 及 TIGRFAM IDs，並使用 HmmSearch 及 BLAST 搜尋在菌株中具有相同編號之候選基因，並使用 EMBL-EBI MAFFT 進行候選基因與目標基因的序列比對與分群，找出候選基因中與目標基因最接近之分群，篩選出在菌株中可能參與植物賀爾蒙生合成、固氮相關基因，並討論植物益生菌如何透過不同植物賀爾蒙或固氮途徑來幫助植物提升耐旱及耐鹽能力。

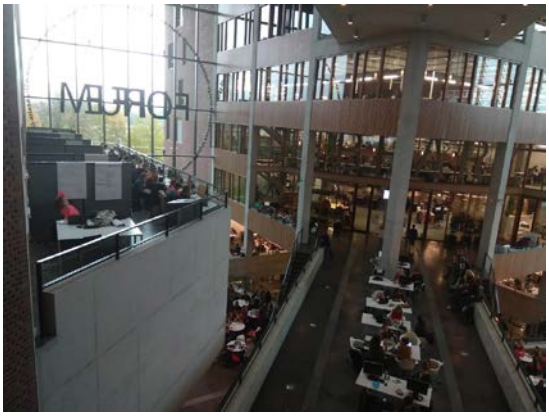
**四、心得及建議事項**

荷蘭瓦赫寧恩大學，是荷蘭研究型大學之一，根據 2019 年《美國新聞與全球報導》(U.S. News & World Report) 發表世界大學排名，瓦赫寧恩大學在農業科學領域排名第一，學校主要分成大學部門與研究中心，大學部門從事教學與基礎研究，研究中心主要從事應用研究、技術擴散，並與許多公部門、學術單位、私人公司進行合作，致力於將研究成果應用到各領域，如協助政府制定政策、提升企業技術等，校園有多家私人公司設立實驗室或辦公室，形成產業聚落及技術中心，校園內也設有新創中心，協助畢業學生或企業人員進行創新技術開發，產官學密切合作的關係，有助於提升整體農業發展。

學校教學重視學生的主動學習，上課方式不只講課，更著重在實作與分組合作，課堂上學生與老師互動多，學生亦會主動發問，課堂上很多分組報告，有助於培養與他人溝通、協調及整合的能力，並可將所學的理論應用到實際的例子進行分析，透過課程學習也同時訓練學生思考、表達、寫作、合作能力，使學生具備有就業所需之能力，此外，學生也可透過國內或實習的方式提早了解產業的動向，提升就業機會。校園中有許多討論室及開放空間可供學生討論，可促進彼此交流及腦力激盪。

位於世界貿易中心之一的荷蘭普遍英文程度高，英文已成為生活中的一部分，我國推動雙語化政策可參考荷蘭相關作法。本次研習除有助於專業能力提升外，可增進國際視野及強化英文程度，加強我國研究人員與國際機構的聯繫，因此應值得鼓勵研究人員持續出國進行交流。

研習照片

<table>
<tr><td>  </td><td>  </td></tr>
<tr><td>WUR 校園 Forum 大樓外觀。</td><td>WUR 校園建築內空間。</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>課堂上練習腦力激盪，進行研究計畫規劃。</td><td>學生進行分組報告研究計畫與成果。</td></tr>
</table>

## 五、附錄

## 附錄一、番茄成分分析報告

# Visualization and Analysis of three traits in 93 tomato cultivars.

By Hsin Yi Tseng (Tseng004) and Romee Verhagen (Verha058)

## Introduction

For this project a dataset was provided by the Center for BioSystems Genomics (CBSG) and it contains 93 genotypes of tomato cultivars. The Netherlands and the CBSG together provide roughly 70% of varieties and seeds of the global fresh tomato production. The cultivars in this dataset can be divided into three categories; cherry, round and beef. Of all these cultivars the metabolic profiles were measured which resulted in 129 potentially interesting metabolites. These can also be divided into separate categories; Volatiles (V_), untargeted volatiles (UV_), non-volatiles (NV_) and derivative volatiles (DV_). For all the genotypes multiple sensory and physiological traits were also determined, however, in this case only three traits are of interest; sweetness, tomato taste and fruit weight. For the metabolites models were used to estimate the mean trait values per genotype and corrected for the influence of other factors on individual measurements.

The goal in this project is threefold; first the dataset will be visualized, secondly several classification methods are used to see which best fits the data and to help determine metabolites of interest. And lastly a regression analysis is performed to see how well the traits can be predicted and again which metabolites are of interest or necessary for these predictions.

## Description of Methods

### Visualisation (Description & motivation of method and choices)

To visualise the dataset three boxplots were made, each boxplot was made to compare one of the traits between the different tomato types. This was done to give a clear visual of the distribution of data. Furthermore, a PCA was performed for each category of metabolites to reduce the amount of variables and make the large data set is easy to visualize. To identify which metabolites have a correlation with sweetness, tomato taste and fruit weight the correlation matrices were plotted. For sweetness and fruit weight the correlations were only plotted if they were higher than 0.7, for the tomato taste they were only plotted if they were higher than 0.3.

### Classification (Description & motivation of method and choices)

For the classification multiple methods were used and compared to see which one was

best for this data set. To compare the methods the models were trained with 75% of the data set and subsequently tested with the remaining 25% of the dataset. The methods were compared for their predictive capabilities and their accuracy. The methods that were compared were the k-nearest neighbours, decision tree and random forest. After that one-way ANOVA's were performed on the most important metabolites found to confirm that they were significantly different between the tomato categories and then boxplots were made to see what those differences were.

**Regression (Description & motivation of method and choices)**

For this dataset, due to the large number of variables, it was only possible to use Lasso regression or Ridge regression. For this analysis Lasso regression was used to eliminate any irrelevant or less important data and to only focus on the most important variables for this study.

## Results & Discussion

### Visualisation

From the boxplots (Figure 1), it can clearly be seen that for sweetness and fruit weight cherry tomatoes have very different values than round or beef tomatoes, with cherry tomatoes being much sweeter and lighter. For the round and beef tomatoes no big difference can be seen for sweetness or tomato taste, however the beef tomatoes appear to be heavier than the round tomatoes. For tomato taste, it seems that there is no difference amongst those three tomatoes.

From the PCA most variances can be explained by the first two PCs. (PC1 explains 71.8% of the overall variation and PC2 can explain 17.2%). From PC1 for all traits a clear difference can also be seen with on the one hand the cherry tomatoes and on the other hand the beef and round tomatoes. There is a lot of overlap between the beef and the round tomatoes, with the main difference being that the round tomatoes cover a bigger area (Figure 2).

To determine which metabolites have a correlation with sweetness, tomato taste and fruit weight correlation matrices were plotted for each trait (Figure 3). For sweetness the original matrix was very large so the metabolites which had a correlation lower than 0.7 were excluded from this analysis. From the correlation matrix it is clear that sweetness has a negative correlation with fruit weight and a positive correlation with sugars such as sucrose, fructose and glucose which is to be expected.

Fruit weight also had a large original matrix so for this trait the metabolites with a

correlation lower than 0.7 were again excluded. Since fruit weight has a strong negative correlation with sweetness it is logical to see that is also has a negative correlation with the same metabolites that sweetness has a positive correlation with, such as sucrose, fructose and glucose. Furthermore, the negative correlation between sweetness and fruit weight makes sense since the bigger the fruit is the more the sugars are spread out over the volume of the tomato, making the concentration of sugars lower overall and thus the fruit less sweet.

For tomato taste there were only very small correlations with the metabolites, the highest correlation still being lower than 0.4, therefore only those with a correlation lower than 0.3 were excluded. The biggest correlations were with UV_u_2182_14 and UV_nitrocyclopent, but as said before these were still quite low. Since tomato taste is a trait that is quite hard to define it isn't surprising to find no strong correlations with anything, since it is not clearly defined what tomato taste itself really means.

**Classification**

For the k-nearest neighbour method an optimum value for k was found at k=4. This has a 85.7% accuracy for the test data and an accuracy of 78.3% for the training data (Table 1). In the training data, however, most of the beef tomatoes were categorized as round tomatoes, this was the same for the testing data. Since the goal of classification is to define all the samples in the correct group this method doesn't appear to be the best choice since it has a lot of difficulty differentiating beef tomatoes from round tomatoes.

The decision tree had an accuracy of 91.3% for both the training and the testing data (Table 2). After cross-validation the tree was pruned with an optimum level of three, given that this is also the amount of categories of the data set this was to be expected. This method is better at categorizing beef tomatoes than the k-nearest neighbour method, however there are still a few mistakes in both the training and testing dataset. From the pruned tree two metabolites can be identified that were used to differentiate between the different types of tomato. These metabolites are NV_gtocopherol (<5.564) and NV_rutin (<3.042) (Figure 4).

The third method tested was the random forest method. Here an accuracy of 91.3% was found (Table 3), which is the same accuracy as found in the decision tree method, and random forest also has some mistakes with the beef tomato. From the results the same important metabolites are found as from the decision tree with one addition in between. NV_rutin was found as the most important metabolite, UV_terpineol_91 as

second best and NV_gtocopherol as third (Figure 5).

Overall the decision tree and random forest methods give the highest accuracy, both making 2 mistakes (see Tables 2 & 3), the difference only being in that the random forest made 2 mistakes in classifying the beef tomatoes and the decision tree made 1 mistake for beef tomatoes and 1 for round tomatoes. Based on these results both methods appear to be equally well suited to classify this dataset. What can be said for sure is that they are both better than the k-nearest neighbour method for this dataset. However, if more of the available traits that pertain to the distinction between the round and beef tomatoes were included in the dataset this could change, given that this seems to be the main problem for all the methods.

To further investigate the metabolites of interest two one-way ANOVAs were performed, one on NV_gtocopherol and the other on NV_rutin.(Table 4 & 5). For both metabolites the Anova showed that at least one of the tomato categories is significantly different ($p<2e-16$ for both metabolites, $\alpha =0.01$, see Table 4 & 5) from the other categories. Next a boxplot was made for each metabolite to compare them between the tomato categories and this showed that cherry tomatoes had the highest count for both NV_gtocopherol and NV_rutin, and that round tomatoes showed a higher count than beef tomatoes for Nv_rutin. This leads to the conclusion that the cherry tomatoes are significantly different from round and beef tomatoes for these two metabolites, which is expected since the classification methods had no problem with differentiating cherry tomatoes from round and beef tomatoes.

**Regression**

The lasso regression for sweetness had a $R^2$ value of 63% and a $Q^2$ value of 57.99% (Table 6), which are both reasonable. The highest coefficient was for DV_glucose which, again, is to be expected since this is a sugar and the measured trait is sweetness. Therefore, this is an important metabolite which can be used to predict the sweetness of tomatoes.

The values found for tomato taste from the regression were extremely low, with the $R^2$ being only 4.51% and the $Q^2$ only 1.23% (Table 6). The most important metabolite for this trait appears to be NV_atocopherol, however, with such low values for the $R^2$ and the $Q^2$ the regression is extremely unreliable and can’t really be used to predict anything accurately.

For fruit weight the highest numbers were found, with a $R^2$ value of 84,88% and a $Q^2$

value of 77,47%. Therefore, the regression model for this trait is the most accurate in explaining the found values and is the best prediction model. The most important metabolite from this regression is again NV_rutin with a very high negative coefficient (Figure 7), which means this is an important metabolite to help predict fruit weight in tomatoes. Other metabolites with high negative coefficient values are again NV_gtocopherol and DV_fructose, which again isn't surprising since this is a sugar and bigger fruit are generally less sweet.

## Conclusions

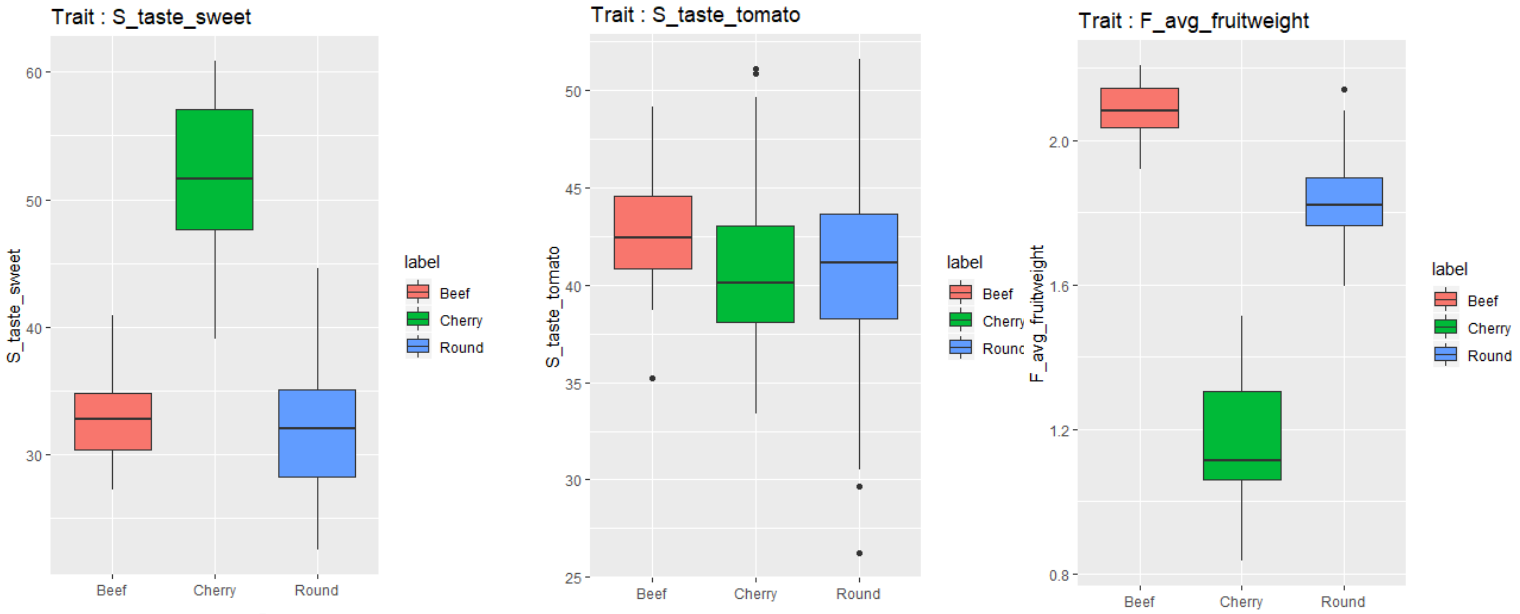
From the boxplots and PCA used for visualization it was clear to see the cherry tomatoes were the easiest to distinguish from the other categories of tomato. It was also evident that it would be harder to distinguish the beef tomatoes from the round tomatoes. This proved true in the classification, where the decision tree and random forest method both performed equally well, better than the k-nearest neighbour method. From the classification it was clear that NV_rutin and NV_gtocopherol were important metabolites to classify the samples into the appropriate categories. From the Anova's and boxplots of these metabolites confirmed that the cherry tomatoes are significantly different from the round and beef tomatoes, however, this can't be said for certain about the difference between round tomatoes and beef tomatoes. From the regression analysis it is clear that no accurate predictions can be done based on tomato taste, however sweetness and mainly fruit weight are quite accurate and reliable predictors. Furthermore, the regression analysis showed that DV_sucrose and DV_fructose are important metabolites which can be used to make reliable prediction models for sweetness and fruit weight respectively.

## Reflection

Overall the project went quite well. In the future it might be a good idea to also do t-tests to see if there was a difference between round and beef tomatoes for NV_rutin, or to also do Anova's for DV_sucrose and DV_fructose to see if they might be able to provide enough of a distinction between round and beef tomatoes.

The cooperation went well, there weren't any disagreements and the workload seems to have been divided fairly equally. The main change to this project that would be beneficial would be if it didn't start in the study week since most students plan on studying for their tests at that time and would prefer to have done this earlier in the period.

# Figures




**Visualization**

Figure 1. Boxplots for comparing three different tomato types on sweetness (left), taste (middle) and fruit weight (right).

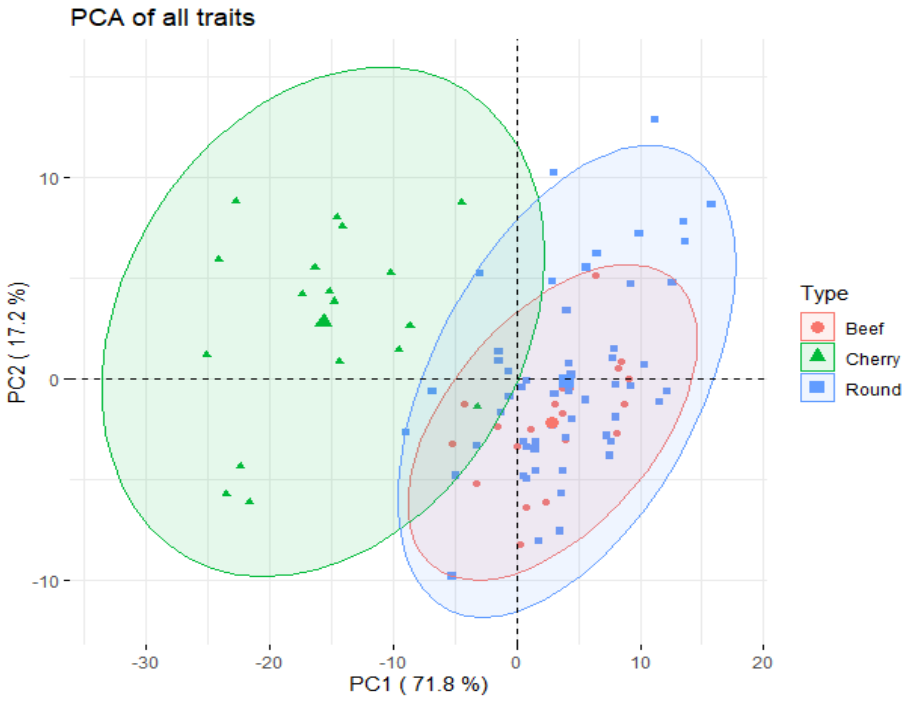
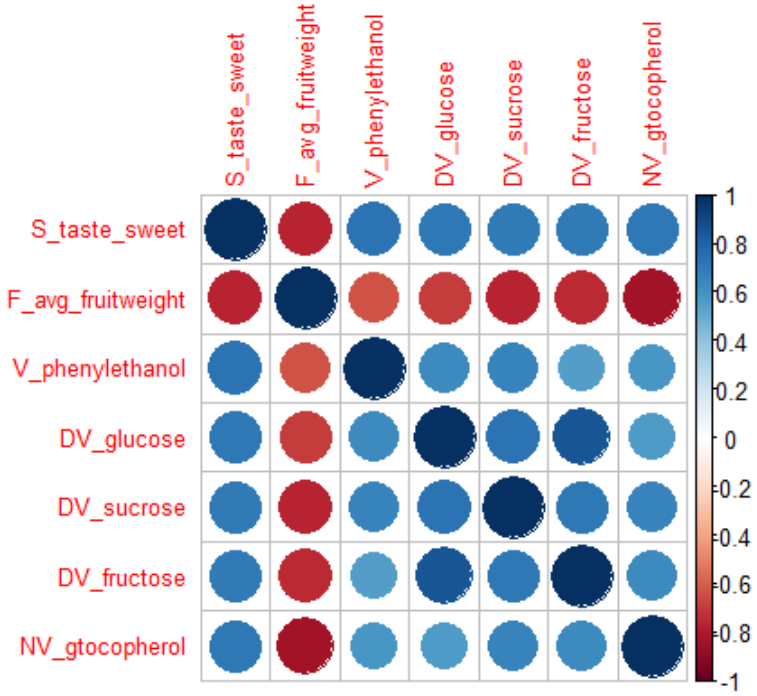
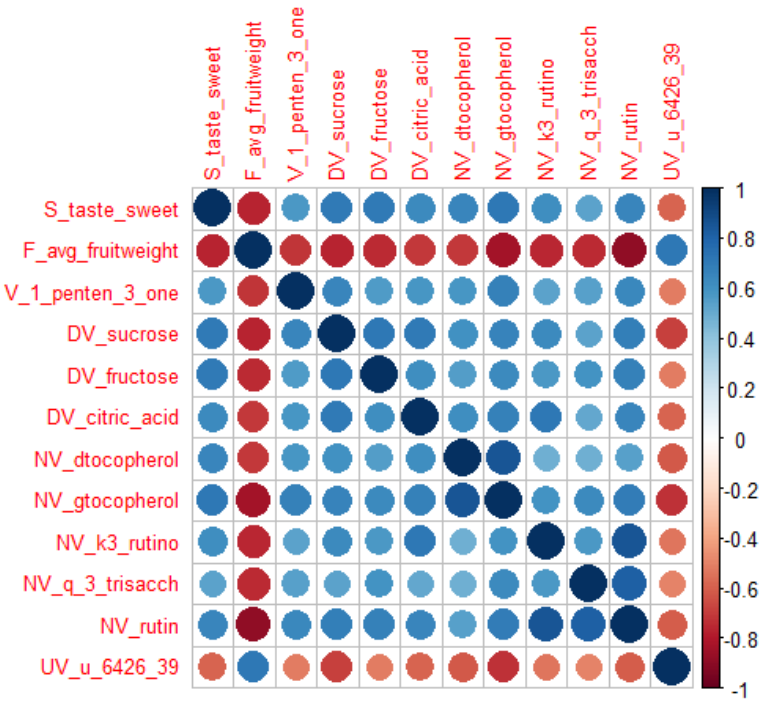



Figure 2. PCA plot for 3 sensory/ physiological traits and 129 metabolites. Different

colors represent different types of tomato.



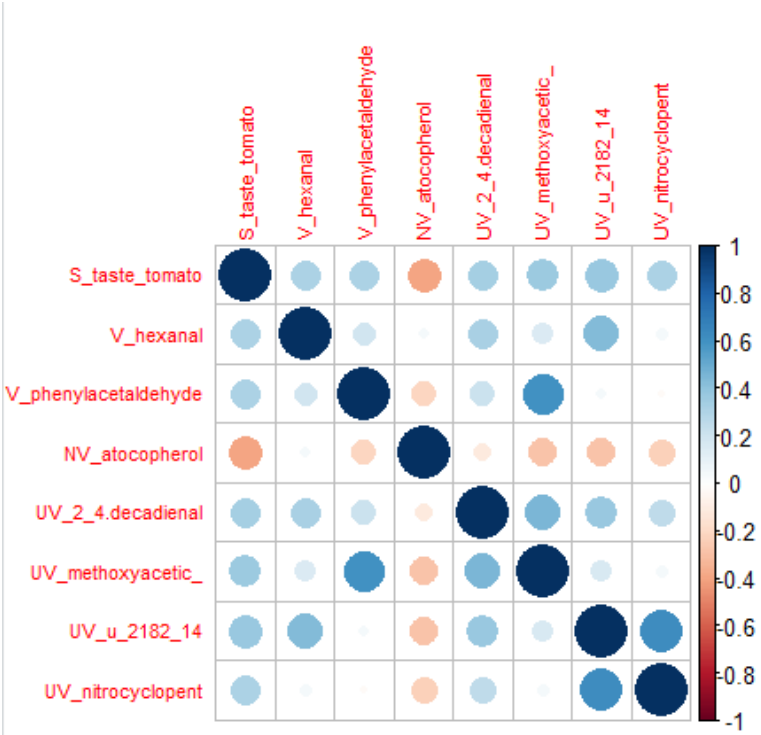



Figure 3. The correlation matrix plots for the metabolites or sensory/physiological traits whose correlation is higher than 0.7 with sweetness (upper left) and fruit weight (upper right), and higher than 0.3 with taste.

**Classification**

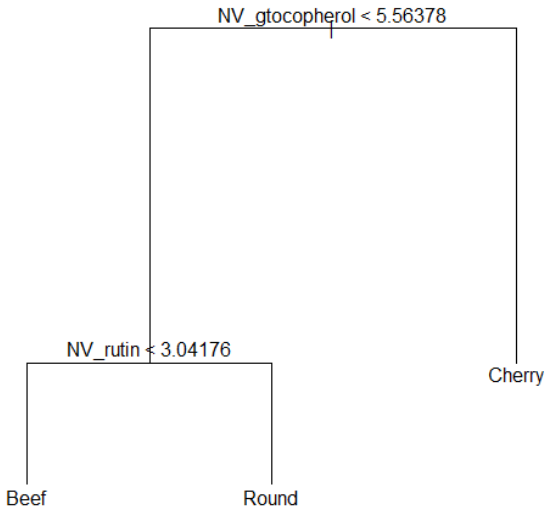



Figure 4. Decision tree plot for the classification of the different types of tomatoes into three groups.

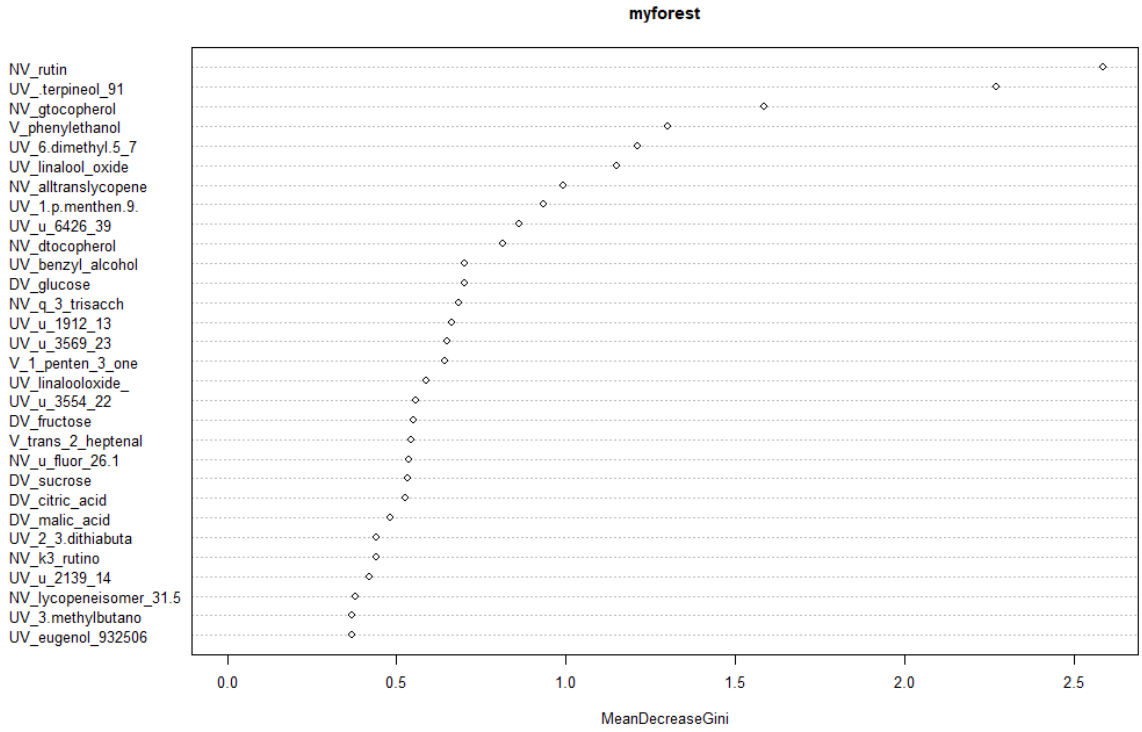



Figure 5. Variable importance plot of the random forest analysis. The variables are ordered top-to-bottom as most-to-least important in classifying.

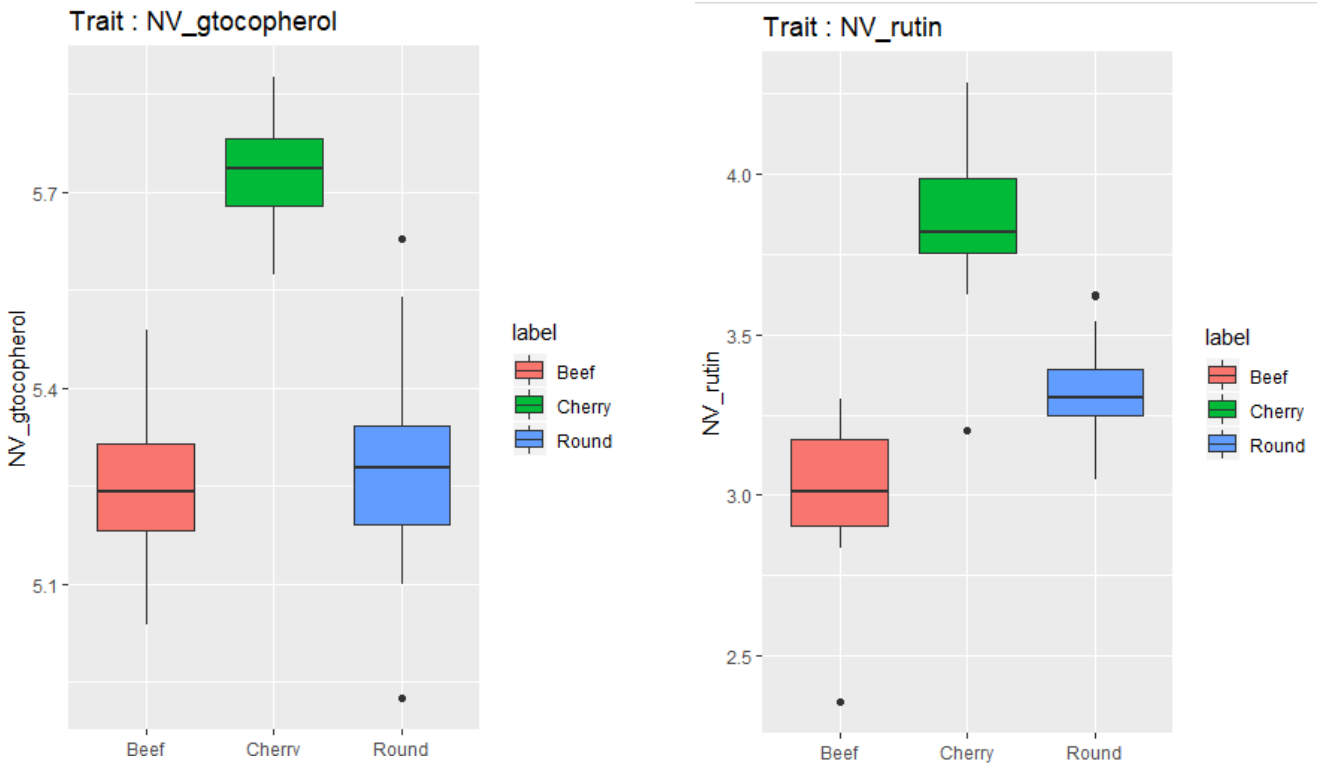



Figure 6. Boxplots for comparing three different tomato types on NV_gtocopherol (left) and NV_rutin (right).

## Regression

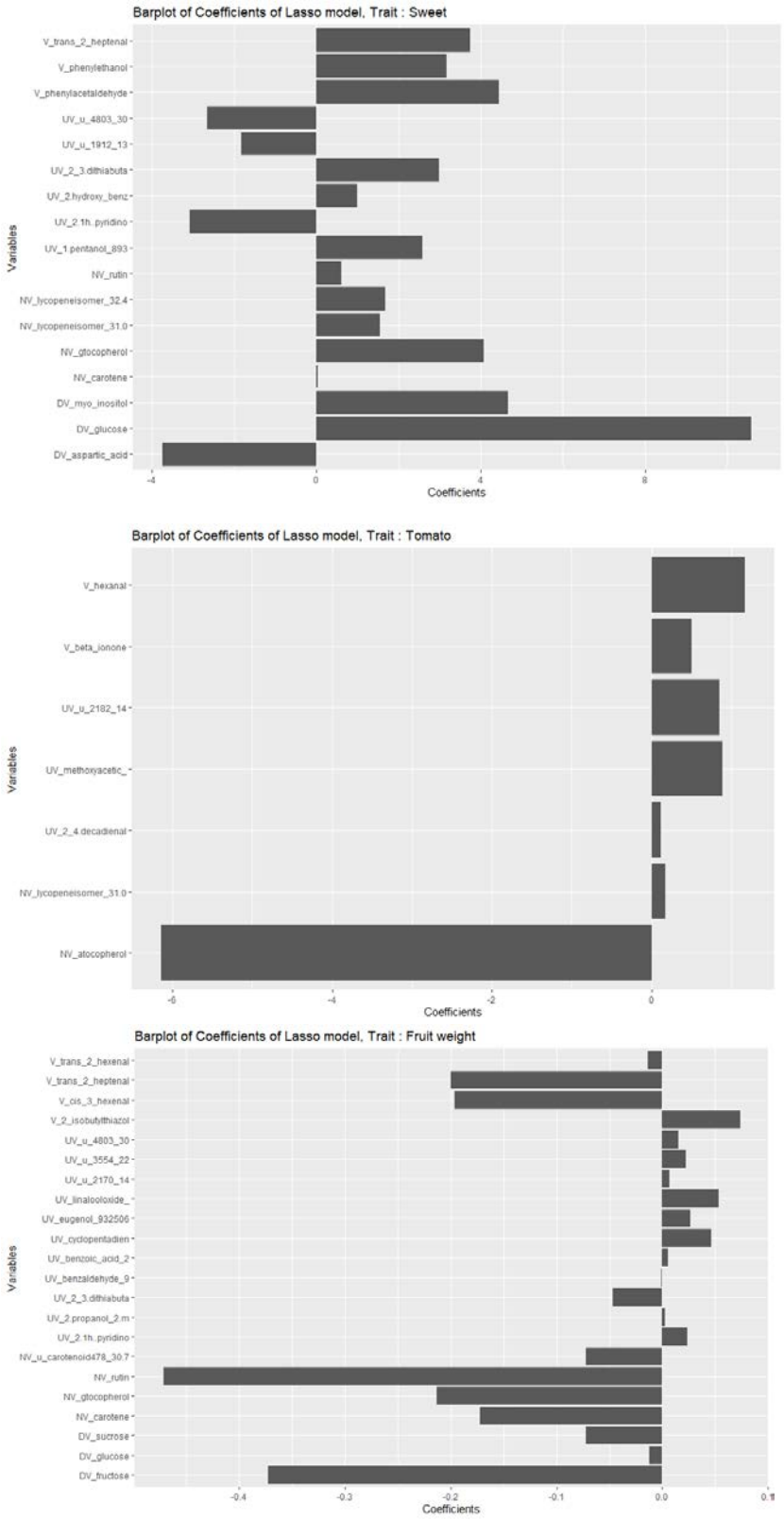



Figure 7. The coefficient of most significant metabolites after Lasso regularization on sweetness (top), taste (middle) and fruit weight (bottom).

## Tables

Table 1. The classification table shows the prediction results of using the k-nearest neighbour classifier

| | Observed | Beef | Cherry | Round | Overall Percentage |
|---|---|---|---|---|---|
| Predicted | Beef | 2 | 0 | 0 | |
| | Cherry | 0 | 3 | 0 | |
| | Round | 5 | 0 | 13 | |
| | Percentage correct | 28.6% | 100% | 100% | 78.3% |

Table 2. The classification table shows the prediction results of using the decision trees classifier

| | Observed | Beef | Cherry | Round | Overall Percentage |
|---|---|---|---|---|---|
| Predicted | Beef | 6 | 0 | 1 | |
| | Cherry | 0 | 3 | 0 | |
| | Round | 1 | 0 | 12 | |
| | Percentage correct | 85.7% | 100% | 92.3% | 91.3% |

Table 3. The classification table shows the prediction results of using the random forests classifier

| | Observed | Beef | Cherry | Round | Overall Percentage |
|---|---|---|---|---|---|
| Predicted | Beef | 5 | 0 | 0 | |
| | Cherry | 0 | 3 | 0 | |
| | Round | 2 | 0 | 13 | |
| | Percentage correct | 71.4% | 100% | 100% | 91.3% |

Table 4. One-way ANOVA table to assess effects on NV_gtocopherol

| Source | Df | Sum of squares | Mean square | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| label | 2 | 3.091 | 1.5454 | 113.4 | <2e-16 | *** |
| Residuals | 90 | 1.226 | 0.0136 | | | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’1

Table 5. One-way ANOVA table to assess effects on NV_rutin

| Source | Df | Sum of squares | Mean square | F value | Pr(>F) |
|---|---|---|---|---|---|
| label | 2 | 6.668 | 3.334 | 110 | <2e-16 *** |
| Residuals | 90 | 2.727 | 0.03 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6. The $R^2$ and $Q^2$ of Lasso regression on sweetness, taste, and fruit weight.

| | Sweet | Tomato | Fruit.Weight |
|---|---|---|---|
| R^2 | 63% | 4.51% | 84.88% |
| Q^2 | 57.99% | 1.23% | 77.47% |

## 附錄二、植物益生菌與植物賀爾蒙生合成、固氮相關基因的關聯報告

# Project Report-Group 2

Weeren van Emma, Hsinyi Tseng, Irene Huizing, Rijst van der Jasper, Lars Essink

## Introduction

Plant growth promoting rhizobacteria (PGPR) are a group of bacteria living in rhizosphere and benefiting plant growth by different mechanisms. To understand the interactions between plants and PGPR can help us utilize PGPR effectively in agriculture, such as used as bio-fertilizers. Some possible biosynthetic pathways of phytohormones and nitrogen fixation have been proposed. However, many mechanisms of PGPR characteristics which enhance the plant growth or stress tolerance are still not clear. Therefore, it deserves to discover more PGPR characteristics related to plant growth.

Bioinformation analysis is a useful tool to provide insight into the molecular level. Here, we use proteomics data of 15 PGPR strains collected form desert plants, and try to figure out how these bacteria help plants growth under drought or salinity stress. Phytohormone regulations and nitrogen fixation are common ways in beneficial bacteria to promote the plant growth or deal with abiotic stress. Therefore, we want to identify the strains which are capable of producing phytohormones or involving in nitrogen fixation and link these characteristics to growth promotion of plants. In addition, we also use phenotype to associate the possible orthologous and to find if there is other possible mechanism. Under this project, I will focus on identifying the pathways which are involved in the biosynthetic pathways of phytohormones in bacteria genomes and comparing the performance of different strains.

## Materials and Methods

### Bacteria materials

Dr. Rene Geurts and his teams collected about 1600 isolates form *Indigofera argentae*, and selected 15 strains that represent the most abundant OTUs in rhizosphere and endophytic compartment of *Indigofera argentae*. They also provided the phenotype data of tomato growth under non-sterile saline conditions. Including individual data and treated with different synthetic communities (SynCom).

### Pfam IDs identification

In order to identify the characteristics in 15 strains, we can analysis their protein function to know the possible role they play in metabolic pathways. We can compare

the sequence similarity with known proteins or domains in database to recognize their protein function in 15 strains. Therefore, we searched possible biosynthetic pathways of phytohormones and nitrogen fixation form previous studies, and listed key enzymes/genes involved in the pathways. We can use these known information as reference to find the protein domains with the same functions in bacteria genomes. The aim of Pfam IDs identification is to search if there are proteins involved in biosynthetic pathways of phytohormones and nitrogen fixation in 15 strains.

The Pfam database is a database of protein families. Each Pfam entry is defined by multiple sequence alignments and hidden Markov models (HMMs) [1]. According to the known proteins, we can search their specific Pfam accession numbers from UniProt database (https://www.uniprot.org/) and find their major domains (Figure 1). After listing all known IDs, the Pfam-A HMM library (version 32.0) was download from EMBL-EBI website (ftp://ftp.ebi.ac.uk/pub/databases/Pfam) to search the proteins which have the same Pfam IDs with referential proteins in 15 strains by HmmSearch (version 3.1b2)[2]. The default parameters were used. In this way, we can get candidate proteins in 15 strains.

In addition, we also use BLAST and TIGRFAMs ID to identify if referential proteins exist in 15 strains. We use the proteins of 15 strains to do protein-protein BLAST against the protein sequences of referential proteins for searching protein sequence similarities. TIGRFAMs is a database of protein families described by hidden Markov models (HMMs) [3]. TIGRFAMs focus on not only sequence similarity but specific function of proteins.

We use some strategies to filter proteins. Trusted cutoff scores is the HMM score. That can help us decide if the proteins belong to the same members. Proteins is above trusted cutoff scores which means no false positive hits. Both trusted cutoff scores of Pfam and TIGRFAMs provided from TIGRFAMs database (http://tigrfams.jcvi.org/cgi-bin/index.cgi) were used to filter out the proteins which are below the score. After searching genomes by HmmSerach in 15 strains, we select the first 5 proteins which are significant proteins and ascending ranked by E-value of full sequence. E-value is the statistical significance. Therefore, the lower E-value, the more confidence we can say that this is a homologous domain. According to the results of BLAST, we also include the proteins which are above 60 % identity. We use these proteins to do multiple alignment and plot the phylogenetic tree. The results of BLAST and the identifications of TIGRFAMs ID were provided by other members.

**Phylogenetic tree construction**

Although sequences with the same Pfam IDs, they may not have the same function. For example, the Amino_oxidase entry (Pfam: PF01593) consists of various

amine oxidases, including maize polyamine oxidase (PAO), L-amino acid oxidases (LAO) and various flavin containing monoamine oxidases (MAO). Therefore, we can use phylogenetic tree to subdivide them into subfamilies and identify the most related group.

Except referential proteins, some similar protein with the same Pfam IDs and annotation obtained from UniProt were also used as referential proteins to do multiple alignment with candidate proteins, and then plotted phylogenetic tree. According to the results of cluster, we can identify the most related proteins. Generally, these referential proteins would cluster closely to each other. The multiple alignment and Phylogenetic tree construction were performed on EMBL-EBI MAFFT website (https://www.ebi.ac.uk/Tools/msa/mafft/) [4].

After plotting the phylogenetic tree, we selected the proteins which clustered with referential proteins as candidate proteins. If the most related cluster contains more than one protein of each strain, the lowest E-value protein of each strain was retained.

## Results

### Pfam IDs identification and candidate proteins selection in 15 strains

We searched genes involved in phytohormone production and nitrogen fixation from literature, used known proteins as referential proteins to do BLAST and searched their Pfam IDs and TIGRFAM IDs (Table 1). Afterwards, we used HmmSearch to find the proteins which had the same Pfam ID with referential proteins in 15 strains (Supplementary data S1). However, the same Pfam IDs might contain different functional proteins. Therefore, we can use multiple alignment and phylogenetic tree to help us find the most related proteins (Supplementary data S2-S6). The referential proteins usually group nearby, so we can select the proteins which are in the same cluster with referential proteins to decide the most related proteins in 15 strains (Figure 1). The candidate proteins are shown in Table 2.

For auxin, we selected 3 possible pathways : iaaM/iaaH pathway, ipdc/ IAAld dehydrogenase pathway, and nitrilase pathway. None of iaaH was identified in 15 strains, so iaaM/iaaH might be not possible. The annotations in most of proteins are involved in tryptophan metabolism. That is because tryptophan is a main precursor for the biosynthesis of indole-3-acetic acid (IAA). According to the results of Pfam IDs identification, most of strains can produce IAA via ipdc/IAAld dehydrogenase or nitrilase pathway. On the other hand, ipdc contains multiple domains (PF02775, PF00205 and PF02776), so we can use TIGRFAMs ID identification to find the most related proteins. Unfortunately, no proteins hit the TIGRFAMs TIGR03393

(Supplementary data S7). IAA can regulate various plant growth processes. According to the results, 12 strains are identified to be capable of producing IAA. This indicates that IAA is an important phytohormone in bacteria and most of stains can produce IAA. Auxin can help plant growth by root-growth regulation, and root elongation is also an important characteristic under drought or salinity stress.

For abscisic acid (ABA), there is not a clear ABA biosynthetic pathway described in bacteria. In *Arabidopsis*, abscisic-aldehyde oxidase, AAO3, catalyzes the final step in abscisic acid biosynthesis [5], so we use AAO3 as a key protein for ABA biosynthesis. AAO3 include 6 domains. Among them, PF01315 relates to aldehyde oxidase and xanthine dehydrogenase, so we use PF01315 as the main domain to represent AAO3. Here, we found 8 strains might contain AAO3. Nevertheless, if AAO3 also can work in bacteria is unknown. It needs further experiments.

Iso-pentyl transferase (ipt) is a key enzyme which is responsible for the synthesis of cytokinins (CK) [6]. The main domain, PF01715, belongs to tRNA dimethylallyltransferase. The results show that there are 3 strains, SA403, SA436 and SA444, annotated as tRNA dimethylallyltransferase. We assume these strains might contain ipt and could produce CK.

Ethylene can stimulates leaf senescence under stress and affect the plant growth. 1-aminocyclopropane-1-carboxylate (ACC) is a precursor of ethylene. Bacteria can produce ACC deaminase to help degrade ethylene. The Pfam ID of ACC deaminase is PF00291. Members of this family are all pyridoxal-phosphate dependent enzymes. The protein identified as PF00291 in SA188, SA424, and SA613 was annotated as tryptophan synthase beta chain, which was also one of the members of PF00291 family. However, we can just assume that tryptophan synthase beta chain are related to ACC deaminase. Whether it could act as ACC deaminase is not clear.

Bacteria can use 2-oxyglutarate as substrate and the ethylene-forming enzyme (EFE) to synthesis ethylene [7]. However, the results show that the most related proteins in SA148, SA188, SA244, SA403, and SA613 are not annotated as EFE. Again, we can just assume they are related to EFE. Whether they can act as EFE needs further confirmation.

The gibberellin (GA) biosynthetic pathway in plants need cytochrome P450 monooxygenases (CYPs) to synthesis GA. In bacteria, CYP112, CYP114, and CYP117 are found in *Bradyrhizobium japonicum* and are involved in GA production [8]. Here, we found some candidate proteins annotated as Cytochrome P450 related proteins and grouped with referential proteins, which means they share high identity. Therefore, we can infer that these candidate proteins might have similar function with referential proteins involved in GA production.

Biological nitrogen fixation also can help plants for enhancing plant growth.

Nodulation factors (Nod factors) act as signal chemicals between symbiotic bacteria and plants to form root nodules in leguminous plants. In addition, rhizobia have the ability to fix nitrogen via nitrogenase and nitrogen fixation (nif) gene. The results show that the candidate proteins of nod factors and nitrogen fixation found in SA403 are related to nodulation or nitrogen fixation. Therefore, SA403 might exist nod factors and nitrogen fixation proteins to help plant growth.

**Comparing the possible pathways in SynComs with the phenotype of tomato**

Table 4 shows the number of strains involved in the biosynthetic pathway of phytohormones in SynComs. We compared the performance of tomato growth with different SynComs under saline conditions, and combined the possible pathways which SynComs are involved. The result shows that SynCom E, F, and G contain all possible pathway. SynCom A, B, and C lack of ACC degradation pathway. Nod factor and nitrogen fixation pathway is absent in SynCom D and H. It can suppose that SynCom E, F, and G can help the plant growth mostly. However, from Rene's *et. al*. data, the highest dry weight of tomato is with SynComs C.

## Discussion

According to previous studies, many rhizobacteria can produce IAA [9] and different biosynthetic pathways have been proposed [5]. SA087, SA187, and SA188 might be involved in both ipdc/ IAAld and nitrilase pathway, so multiple pathways to produce IAA might be possible. Only SA113, SA436 and SA670 seems like not involved in ipdc or nitrilase pathway. However, there are other possible biosynthetic pathway of IAA which are not included in this study, such as amine oxidase pathway [5].

Some proteins contain multiple domains, so using one Pfam domain to find the related proteins seems like unreliable. However, although NifA contains 3 different domains (PF01590, PF02954 and PF00158), the most related proteins are the same (Figure 2). That is, when we used these 3 different domains to perform clusters separately, ICELOAJG_05683, Nif-specific regulatory protein, always grouped with referential proteins. Consequently, for some proteins containing multiple domains, they still would cluster with similar proteins for each domain, and using phylogenetic tree can help identify the most related groups.

Even if we have no ideas about the species of 15 strains, according to the results of candidate protein identification, SA403 contains the most nod factors and nitrogen fixation proteins, so we can suppose that SA403 is a rhizobium, which can nodulate legumes or fix nitrogen. Indeed, SA403 is *Ensifer* sp, a rhizobial species [10]. In addition, SA403 contains the most candidate proteins involved in the biosynthetic

pathway of phytohormones and nitrogen fixation than other strains, which supposes that it can promote the plant growth better than others. SA403 only lacks of ACC deaminase to degrade ACC (Table 3). Besides, from the phenotype of individual data of tomato, the dry weight of tomato living with SA403 performs the best as well. Therefore, SA403 might help plant growth via phytohormone production and nitrogen fixation.

Although the candidate proteins cannot explain the performance of SynCorm completely, it is possible that those bacteria can produce intermediates or precursors of phytohormones to enhance the plant growth. Therefore, some strains may be involved in the intermediate steps but we didn't search every intermediates. In addition, the interaction of phytohormones between bacteria and plants is still unclear. It needs more evidence to discover. Moreover, bacteria can help the plant growth under salinity stress via other non-phytohormones mechanisms, such as osmotic adjust or nitrogen fixation, but they are not been studied in this project.

## Conclusions and Recommendations

We identified some candidate proteins involved in biosynthetic pathway of phytohormones and nitrogen fixation. According to the results, auxin production might be the common way that bacteria enhance plant growth and stress tolerance. On the other hand, CK and GA production might be strain specific, which means only a few strains are involved. SA403 contains proteins involved in nod factors and nitrogen fixation, and it might help plants absorb nutrition under stress. It can suppose that these strains can enhance the plant growth via these pathways, but the interaction between these strains is still unclear.

In the future, the candidate proteins need more experimental designs to validate their functions. Some possible experiments could be adopted. It can detect if these strains can produce phytohormones or induce root nodule to confirm the pathways which the strains are involved. In addition, gene expression data of candidate proteins for the stress response can also provide evidence to validate the results. Moreover, genome-wide association study is an approach to identify genes involved in phytohormone production and nitrogen fixation. For SynComs experiments, it can reduce the number of strains of each SynComs because the interaction of bacteria is very complicate.

## Author Contribution

Hsinyi searched the Pfam ID of known genes and identified the Pfam ID in the 15 strains. Emma compared the candidate proteins with the phenotype of the plants. Irene annotated all the genes of the important orthlogous groups for plant growth using

eggNOG. Jasper ran the orthofinder to make orthologous groups and convert the files for PhenoLink. Lars performed BLAST against the related genes in the 15 strains, and helped with finding the Pfam ID and TIGRFAMs ID of known genes. All members searched the genes and pathways involved in the biosynthetic pathway of phytohormones and nodulation.

## Supplementary material

Supplementary material was provided on google drive. (https://drive.google.com/drive/folders/1FlgVcTmw3An7elAFifRlRe6x6a3xe-r_?usp=sharing)

Supplementary data S1. The results of HmmSearch in the 15 strains.

Supplementary data S2. The protein sequence used in multiple alignment.

Supplementary data S3. The results of multiple alignment.

Supplementary data S4. The results of phylogenetic tree.

Supplementary data S5. Summary BLAST results-provided by Lars.xlsx

Supplementary data S6. Trusted cutoff scores.xlsx

Supplementary data S7. Identification of TIGRFAMs in 15 strains

## References


[1] S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, and A. Smart, The Pfam protein families database in 2019. Nucleic acids research 47 (2019) D427-D432.

[2] L.S. Johnson, S.R. Eddy, and E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. BMC bioinformatics 11 (2010) 431.

[3] D.H. Haft, J.D. Selengut, and O. White, The TIGRFAMs database of protein families. Nucleic acids research 31 (2003) 371-373.

[4] F. Madeira, Y.M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A.R. Tivey, S.C. Potter, and R.D. Finn, The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic acids research 47 (2019) W636-W641.

[5] S. Spaepen, and J. Vanderleyden, Auxin and plant-microbe interactions. Cold Spring Harbor perspectives in biology 3 (2011) a001438.

[6] M. Kamínek, V. Motyka, and R. Vaňková, Regulation of cytokinin content in plant cells. Physiologia Plantarum 101 (1997) 689-700.

[7] K. Nagahama, T. Ogawa, T. Fujii, M. Tazaki, S. Tanase, Y. Morino, and H. Fukuda, Purification and properties of an ethylene-forming enzyme from Pseudomonas syringae pv. phaseolicola PK2. Microbiology 137 (1991) 2281-2286.

[8] R. Tully, P. van Berkum, K. Lovins, and D. Keister, Identification and sequencing of a cytochrome P450 gene cluster from Bradyrhizobium japonicum. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1398 (1998) 243-255.

[9] R. Hayat, S. Ali, U. Amara, R. Khalid, and I. Ahmed, Soil beneficial bacteria and their role in plant growth promotion: a review. Annals of Microbiology 60 (2010) 579-598.

[10] P. DE LAJUDIE, A. Willems, B. Pot, D. DEWETTINCK, G. MAESTROJUAN, M. NEYRA, M.D. COLLINS, B. DREYFUS, K. Kersters, and M. Gillis, Polyphasic Taxonomy of Rhizobia: Emendation of the Genus Sinorhizobium and Description of Sinorhizobium meliloti comb. nov., Sinorhizobium saheli sp. nov., and Sinorhizobium teranga sp. nov. International Journal of Systematic and Evolutionary Microbiology 44 (1994) 715-733.

**Table 1. The genes involved in the biosynthetic pathway of phytohormones and nitrogen fixation from literature.**

| Phytohormones | Genes | Protein | Pfam ID | TIGRFAMs | Referenced UniProt Protein Entry |
|---|---|---|---|---|---|
| **Auxin** | iaaM | tryptophan 2-monooxygenase | PF01593 | | P06617 |
| | iaaH | Indoleacetamide hydrolase | PF01425 | | Q04557 |
| | dhaS, aldA | indole-3-acetaldehyde dehydrogenase, IAAld dehydrogenase | PF00171 | | O34660 |
| | ipdC | Indole-3-pyruvate decarboxylase | PF02775, PF00205, PF02776 | TIGR03393 | P23234 |
| | Nitrilase | Nitrilase | PF00795 | TIGR04048 | A0A380UK73 |
| **Abscisic acid** | AAO3 | Abscisic-aldehyde oxidase | PF01315, PF02738, PF03450, PF00941, PF00111, PF01799 | | Q7G9P4 |
| **Cytokinin** | ipt | iso-pentyl transferase | PF01715 | TIGR00174 | Q94ID3 |
| **ACC degradation** | acdS | ACC deaminase | PF00291 | | Q5PWZ8 |
| **Ethylene** | EFE | 2-oxoglutarate-dependent dioxygenase | PF03171, PF14226 | | P32021 |
| **GA** | CYP112 | Cytochrome P-450 BJ-1 | PF00067 | | P55544 |
| | CYP114 | Cytochrome P-450 BJ-3 | PF00067 | | P55543 |
| | CYP117 | cytochrome P-450 BJ-4 | PF00067 | | P55540 |
| **Nod factors** | NodD2 | Nodulation protein D2 | PF00126 | | P23719 |
| | NodD1 | Nodulation protein D1 | PF00126 | | P23718 |
| | NodA2 | Nodulation protein A2 | PF02474 | | L0LQ69 |
| | nodB | Chitooligosaccharide deacetylase | PF01522 | | P24150 |
| **Nitrogen fixation** | NifH | Nitrogenase iron protein | PF00142 | TIGR01287 | |
| | NifA | Nif-specific regulatory protein | PF01590, PF02954, PF00158 | TIGR01817 | P54930 |

**Tabble 2. Candidate proteins in the 15 strains involved in the biosynthetic pathway of phytohormones.**

| Gene | ID | Strain | Protein annotation | Gene | ID | Strain | Protein annotation |
|---|---|---|---|---|---|---|---|
| iaaM | LJPOONFC_05823 | SA087 | hypothetical protein | AAO3 | CJEMLAAP_02801 | SA148 | Putative xanthine dehydrogenase molybdenum-bi |
| iaaM | CGAAFPPJ_01728 | SA113 | Tryptophan 2-monooxygenase | AAO3 | NKLKAAAI_03419 | SA244 | putative xanthine dehydrogenase subunit D |
| iaaM | OCHNELDH_01781 | SA188 | hypothetical protein | AAO3 | ICELOAJG_00629 4- | SA403 | hydroxybenzoyl-CoA reductase subunit alpha |
| iaaM | HOJGJPCK_02680 | SA424 | L-amino acid dehydrogenase | AAO3 | HOJGJPCK_02340 | SA424 | Putative xanthine dehydrogenase molybdenum-bi |
| iaaM | MJNAGANF_00540 | SA444 | Tryptophan 2-monooxygenase | AAO3 | MJNAGANF_01140 | SA444 | Aldehyde oxidoreductase |
| iaaM | OKDGACPM_02389 | SA613 | L-amino acid dehydrogenase | AAO3 | OKDGACPM_03645 | SA613 | putative xanthine dehydrogenase subunit D |
| iaaM | CFELDAFL_01821 | SA619 | Tryptophan 2-monooxygenase | AAO3 | JILMHJOK_07934 | SA670 | hypothetical protein |
| iaaM | JILMHJOK_04884 | SA670 | Tryptophan 2-monooxygenase | ipt | ICELOAJG_07515 | SA403 | tRNA dimethylallyltransferase |
| iaaM | HGEOGIFL_04713 | SA681 | Tryptophan 2-monooxygenase | ipt | CADKOPAA_00940 | SA436 | tRNA dimethylallyltransferase |
| dhaS, aldA | CGAAFPPJ_03198 | SA113 | Aldehyde dehydrogenase | ipt | MJNAGANF_05482 | SA444 | tRNA dimethylallyltransferase |
| dhaS, aldA | CJEMLAAP_02976 | SA148 | Aldehyde dehydrogenase | acdS | OCHNELDH_01957 | SA188 | Tryptophan synthase beta chain |
| dhaS, aldA | LDGNIBBK_03314 | SA187 | Aldehyde dehydrogenase | acdS | HOJGJPCK_04122 | SA424 | Tryptophan synthase beta chain |
| dhaS, aldA | OCHNELDH_02086 | SA188 | Long-chain-aldehyde dehydrogenase | acdS | OKDGACPM_02632 | SA613 | Tryptophan synthase beta chain |
| dhaS, aldA | ICELOAJG_01723 | SA403 | Aldehyde dehydrogenase | EFE | CJEMLAAP_03688 | SA148 | hypothetical protein |
| dhaS, aldA | CADKOPAA_00670 | SA436 | Phenylacetaldehyde dehydrogenase | EFE | OCHNELDH_00679 | SA188 | hypothetical protein |
| dhaS, aldA | CFELDAFL_04401 | SA619 | Aldehyde dehydrogenase | EFE | NKLKAAAI_02823 | SA244 | Validamycin A dioxygenase |
| dhaS, aldA | JILMHJOK_04088 | SA670 | Aldehyde dehydrogenase | EFE | ICELOAJG_05209 | SA403 | hypothetical protein |
| dhaS, aldA | HGEOGIFL_02133 | SA681 | Aldehyde dehydrogenase | EFE | OKDGACPM_03647 | SA613 | Validamycin A dioxygenase |
| ipdc | LJPOONFC_03988 | SA087 | hypothetical protein | CYP112, … | CGAAFPPJ_03313 | SA113 | Cytochrome P450 107B1 |
| ipdc | LDGNIBBK_03181 | SA187 | Pyruvate-flavodoxin oxidoreductase | CYP112, … | OCHNELDH_01282 | SA188 | Carnitine monooxygenase reductase subunit |
| ipdc | OCHNELDH_00508 | SA188 | Alpha-keto-acid decarboxylase | CYP112, … | NKLKAAAI_06535 | SA244 | Mycinamicin IV hydroxylase/epoxidase |
| ipdc | CKMIMADB_03670 | SA190 | hypothetical protein | CYP112, … | ICELOAJG_05607 | SA403 | Pentalenolactone synthase |
| ipdc | CKMIMADB_00408 | SA190 | 1-deoxy-D-xylulose-5-phosphate synthase | CYP112, … | HOJGJPCK_05782 | SA424 | Pentalenic acid synthase |
| ipdc | NKLKAAAI_04503 | SA244 | 2-oxoglutarate oxidoreductase | CYP112, | CADKOPAA_00829 | SA436 | Cytochrome P450(BM-1) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | subunit KorB | … | | | |
| Nitrilase | LJPOONFC_00801 | SA087 | Apolipoprotein N-acyltransferase | CYP112, … | CFELDAFL_00881 | SA619 | Cytochrome P450-SU2 |
| Nitrilase | CJEMLAAP_03368 | SA148 | Apolipoprotein N-acyltransferase | CYP112, … | HGEOGIFL_00316 | SA681 | Cytochrome P450-SU2 |
| Nitrilase | LDGNIBBK_02294 | SA187 | Apolipoprotein N-acyltransferase | NodD2, NodD1 | ICELOAJG_05277 | SA403 | Nodulation protein D 2 |
| Nitrilase | OCHNELDH_03044 | SA188 | Apolipoprotein N-acyltransferase | NodA2 | ICELOAJG_04065 | SA403 | Nodulation protein A |
| Nitrilase | CKMIMADB_00063 | SA190 | Apolipoprotein N-acyltransferase | nodB | CGAAFPPJ_02920 | SA113 | hypothetical protein |
| Nitrilase | NKLKAAAI_00609 | SA244 | Apolipoprotein N-acyltransferase | nodB | ICELOAJG_04064 | SA403 | Peptidoglycan-N-acetylglucosamine deacetylase |
| Nitrilase | ICELOAJG_04010 | SA403 | hypothetical protein | nodB | MJNAGANF_01241 | SA444 | hypothetical protein |
| Nitrilase | HOJGJPCK_04466 | SA424 | Apolipoprotein N-acyltransferase | nodB | CFELDAFL_03399 | SA619 | hypothetical protein |
| Nitrilase | OKDGACPM_00261 | SA613 | Apolipoprotein N-acyltransferase | nodB | JILMHJOK_02077 | SA670 | hypothetical protein |
| Nitrilase | CFELDAFL_03203 | SA619 | Apolipoprotein N-acyltransferase | nodB | HGEOGIFL_02821 | SA681 | hypothetical protein |
| Nitrilase | HGEOGIFL_01397 | SA681 | Apolipoprotein N-acyltransferase | NifH | ICELOAJG_05821 | SA403 | Nitrogenase iron protein |
| AAO3 | LJPOONFC_01147 | SA087 | Aldehyde oxidoreductase | NifA | ICELOAJG_05683 | SA403 | Nif-specific regulatory protein |

**Table 3. The result of 15 strains involved in the biosynthetic pathway of phytohormones**

**Table 4. Number of strains involved in the biosynthetic pathway of phytohormones in SynComs.**

| | IAA | ABA | CK | Ethelene | ACC degration | GA | Nod | N fixation | Total |
|---|---|---|---|---|---|---|---|---|---|
| SA087 | v | v | | | | | | | 2 |
| SA113 | | | | | | v | | | 1 |
| SA148 | v | v | | v | | | | | 3 |
| SA187 | v | | | | | | | | 1 |
| SA188 | v | | | | v | v | | | 3 |
| SA190 | v | | | | | | | | 1 |
| SA244 | v | v | | v | | v | | | 4 |
| SA403 | v | v | v | v | | v | v | v | 7 |
| SA424 | v | v | | | | v | | | 3 |
| SA436 | | | v | | | v | | | 2 |
| SA444 | v | v | v | | | | | | 3 |
| SA619 | v | | | | | v | | | 2 |
| SA613 | v | v | | v | v | | | | 4 |
| SA670 | | v | | | | | | | 1 |
| SA681 | v | | | | | v | | | 2 |
| Total | 12 | 8 | 3 | 4 | 2 | 8 | 1 | 1 | |

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| IAA | 4 | 3 | 4 | 4 | 5 | 5 | 5 | 4 |
| ABA | 3 | 3 | 3 | 3 | 4 | 4 | 3 | 2 |
| CK | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 |
| Ethelene | 1 | 1 | 1 | 2 | 4 | 4 | 3 | 1 |
| ACC degration | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 2 |
| GA | 3 | 2 | 2 | 1 | 3 | 3 | 3 | 2 |
| Nod factor | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| N fixation | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |

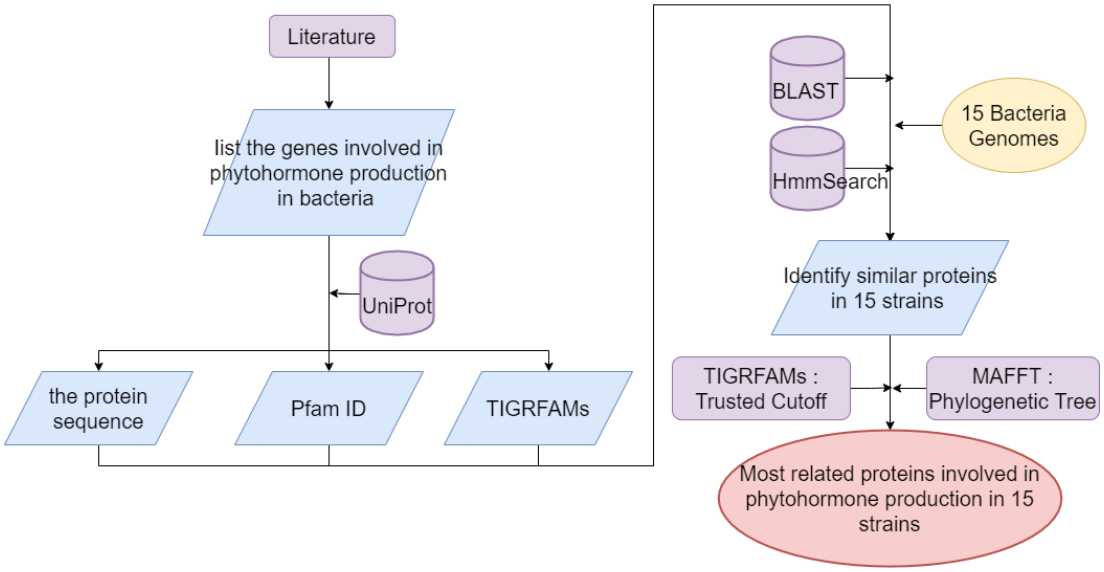
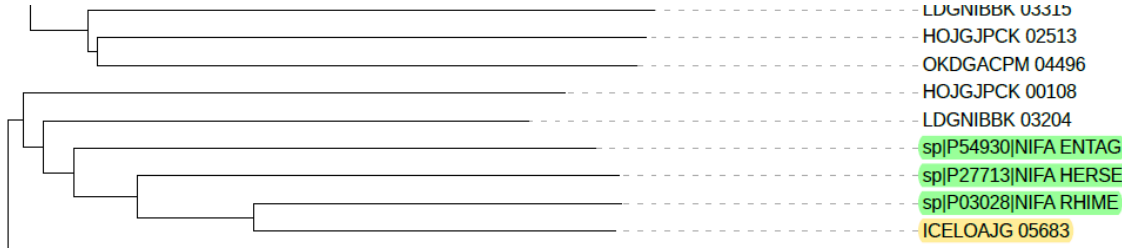
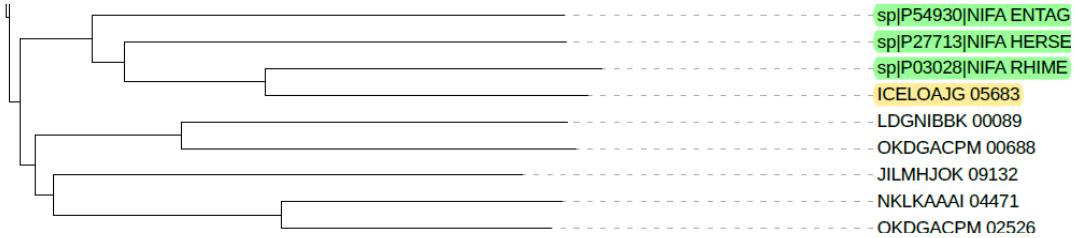



**Figure 1. Strategies to identify the most related proteins in 15 strains.**


**(a) Part of phylogenetic tree of PF01590**




**(b) Part of phylogenetic tree of PF02954**




**(c) Part of phylogenetic tree of PF00158**

**Figure 2. Part of phylogenetic tree of PF01590 (a), PF02954 (b), and PF00158 (c) in NifA. The protein with green means referential proteins, and the protein with yellow means the most related protein.**

## 附錄三、WUR 研習學分成績單

WAGENINGEN
UNIVERSITY & RESEARCH

This is to certify that the student has completed the following subject(s) at Wageningen University.

| | Course/Description | Mark date | Mark | Credits |
|---|---|---|---|---|
| | | | | 860509-843-050 |
| | Tseng, H | | | |
| BIF51306 | Data Analysis and Visualization | 2020-01-09 | 8.0 | 6.00 |
| BIF51806 | Biological Discovery through Computation | 2020-02-13 | 7.0 | 6.00 |

With kind regards,

Ingrid Hijman
Head Student Service Centre

WAGENINGEN UNIVERSITEIT
WAGENINGEN UNIVERSITY

Grading table Transcript of records:

The grading table requires universities to keep track of their grading practice and culture, which is good practice in many institutions across Europe. The ECTS grading table allows for simple, transparent interpretation and conversion of grades from one system or context to another, and therefore does justice to the level of academic performance of all learners. Used correctly, it bridges different grading systems as well as different cultures in the European Higher Education Area and beyond. The grading table gives the distribution of grades for this specific study period. It presents how many students (in percentages) receive a specific grade. This provides all necessary information to convert the grade in any local grading system.

| National / Wageningen University Grade | Total number awarded in reference group | Grading percentages * |
|---|---|---|
| 10 | 461 | 0.60 |
| 9,5 | 1242 | 1.62 |
| 9 | 3797 | 4.95 |
| 8,5 | 7177 | 9.36 |
| 8 | 13147 | 17.14 |
| 7,5 | 14497 | 18.90 |
| 7 | 14278 | 18.62 |
| 6,5 | 11008 | 14.35 |
| 6 | 11082 | 14.45 |
| Total | 76689 | *100* |

* Based on the total number of grades awarded in the last academic year.