Machine learning assisted identification of bioindicators in metagenomics

Afonso de Oliveira Santos Goulart
afonso.goulart@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

July 2021

Abstract

Transcription factors are proteins essential in the control of gene expression and a proper mapping and profiling of these proteins could prove invaluable for the understanding and control of gene regulation in a microbiome. Directly analyzing the abundance and presence of different transcription factors may be a tool that can help provide insight on the external factors that drive transcription factor abundance.

The metagenomic assemblies originating from a benzene-degrading nitrate-reducing succession experiment were analyzed with the newly developed PredicTF tool in order to identify which transcription factors were present in the metagenome. This data was used to conduct a diversity analysis which detected shifts in the abundance of different transcription factor families over time. Machine learning algorithms, such as Random Forest, and statistical tests were then employed to identify potential bioindicators among the different transcription factor families, with 2 different families being identified as potential bioindicators. This work demonstrates that this technique has potential for determining the impact of external factors on biological samples, motivating further exploration of the proposed approach in broader datasets.

Keywords: Random Forest; Transcription Factor; Microbial Diversity; Machine Learning.

1. Introduction

The inspiration for this work is the combination of 3 main factors: The progress and availability of new tools to determine the abundance of transcription factors (TFs); the interest in better modeling of microbial communities' transcription networks; and the question of whether the regulatory and expression patterns of a microbial community are related to the abundance of transcription factors within that community.

Given that one of the major objectives in microbiome manipulation for the ecology and biotechnology fields is to control microbiome function and expression[1][2], and that TFs are essential in the control of gene expression[3], a proper mapping and profiling of TFs could allow for better understanding and control of gene regulation in a microbiome. Improving our understanding of these regulatory networks and their many components permits a superior level of manipulation, control, and monitoring of these microbiomes. This could prove valuable in many facets, providing benefits in the biotechnology fields, but also allowing better prevention of potential environmental hazards in communities, given that understanding the specific methods of regulation of a microbiome would allow us to better predict the environmental impact of external factors introduced to this environment[1].

When studying TFs, there is a lot of work published on their effect on gene expression [3]. However, research on which external factors directly impact TF abundance and how these factors integrate transcription regulatory networks is often overlooked. With the advent on new tools geared specifically towards the detection of TFs such as PredicTF[2], and with the expansion of TF databases like UniProt[4] and CollectTF[5], analysis of TF abundance and the environmental stresses impacting it can be more properly undertaken.

The main objective of this work is to determine the impact of environmental stresses on the abundance of bacterial TFs and whether these differences in abundance levels could be used to distinguish between samples at different time points of a succession experiment, thus providing direct insight into the regulatory networks of the community. In order to achieve this objective, 4 main steps are taken, according to a proposed methodology graphically summarized in fig 1: The elaboration of a succession experiment on a bacterial, benzene degrading, nitrate reducing microcosm performed by a collaborator[6], the sequencing of metagenomic data of these samples and it's assembly, the use of a tool to predict TFs in the assemblies (PredicTF[2]) and an extensive diversity analysis and search for bioindicators in the TF abundance results, using multiple diversity metrics, statistical tests and machine learning approaches.

2. Materials and Methods

The proposed methodology and associated main steps, summarized in figure 1, are detailed next.

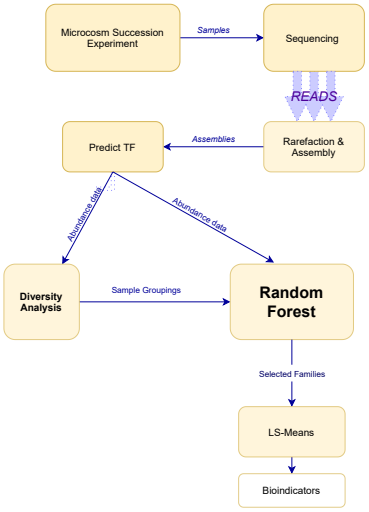

Figure 1: Schematic description of the proposed methodology.

2.1. Microcosm setup and sampling

A controlled mineralization experiment was set up with a benzene-mineralizing nitrate reducing culture with concentration of 1% (v/v) benzene, using 2,2,4,4,6,8,8-Heptamethylnonane as a carrier phase. Samples for DNA extraction were taken as 5 mL of liquid. Technical triplicates were obtained per each sampling time (0, 70, 76, 96 and 124 days) and were immediately stored at -80℃ until DNA extraction. Microbial community diversity was analyzed by paired-end sequencing of 16S RNA amplicons on the Illumina MiSeq platform using the MiSeq Reagent Kit v3 (2 x 300 bp). The V3-V4 regions of the 16S rRNA genes were amplified using the primers according to Klindworth and colleagues[7].

2.2. Rarefaction and assembly

Following sequencing and quality control,the initial reads were normalized by rarefaction to the lowest sample size of $6{\cdot}10^6$. After the rarefaction step, assembly was performed using

the Metaspades[8] algorithm, with 30GB of allocated memory over 10 hours. This was performed in order to remove library size bias from our analysis[9].

2.3. PredicTF

PredicTF was used to determine the abundance of TF families and subfamilies in the metagenomic assemblies. Following the assembly process, the assemblies are used as input for PredicTF. PredicTF outputs the predicted TF subfamily and family, their query position, closest hit in the database, probability of a real match, alignment length, e-value and in which bin/assembly it was predicted. Any output with a probability lower than 97% is disregarded as well as any e-value (probability due to chance, that there is another alignment with a similarity greater than the given score) higher than 10^{-10}.

2.4. Diversity analysis

For the diversity analysis, 4 different indices are used: absolute abundance of TF subfamilies, relative abundance of TF subfamilies, absolute abundance of TF families and relative abundance of TF families. Given the fact that, to the best of our knowledge, no research has yet been published using the PredicTF tool or measuring TF abundance for a diversity analysis, these multiple metrics were used in an attempt to ascertain which could provide the most meaningful results.

Data analyses were completed using the phyloseq[10], vegan[11] and RandomForest[12] packages in R software [13]. In an attempt to assess richness and evenness of the samples, the alpha diversity indices (observed number of TFs and Shannon Index) were estimated using the rarefied data. The statistical significance of the differences in alpha diversity measures between samples was then estimated using a pairwise t-test ("rstatix" R package)[14]. The beta diversity of the samples was studied using both NMDS and PCoA ordination, using the Bray-Curtis dissimilarity method to calculate distance [14].

PCoA is an ordination method which preserves dissimilarity measures between objects[15]. It is used to represent all points in an Euclidean space and can produce 2-dimensional (2D) reduced ordinations of multi-dimensional objects[15]. NMDS is also an ordination technique, however, unlike PCoA where many axes are calculated but only a few are viewed, in NMDS, a small number of axes are explicitly chosen prior to the analysis and the data is fitted to those dimensions; there are no hidden axes of variation[16].

Given that PredicTF identified 27 families and well over 100 subfamilies, in order to represent them graphically in a manner that could be understandable there was need for the dimensionality reduction of the PCoA and/or NMDS methods, allowing the visualization of the most relevant features of our data in a 2-dimensional plot. This visualization is used in order to determine whether or not clusters or decision surfaces can be observed, separating the different time points at which samples were collected. They may also be used to cluster multiple time points. For example, the visualizations may show a clear division in samples taken before a certain date vs those taken at later time points.

Following the visualization of the results, the samples were divided between Early (taken at 0,70 and 76 days) and Late (taken at 96 and 124 days).

The statistical significance of the principal coordinates in the different time points and groupings created was then assessed using a PERMANOVA[17], as this statistical test has been shown to be very powerful as a tool to detect changes in community structures[14].

In order to analyze variance, the Bray-Curtis distance of each sample within the same time point was graphed. This is done as it reveals which time points had the greatest variance of abundance level between samples indicating potential sampling or analysis errors.

Overall this diversity analysis will allow the determination of whether there really is a statistically significant difference in the abundance levels of certain TFs between the different time points and groupings, allowing for a further study forward on which specific TF (sub)families have differences in their abundance level.

2.5. Identifying bioindicators using machine learning

A Random Forest[18] is constructed to distinguish the samples between Early and Late, with the families that contribute the most GINI information gain to the creation of this forest being selected. Those with statistically significant differences in their abundance levels are considered potential bioindicators and the difference in abundance between their Early and Late samples is tested using LS Means tests.

The Gini impurity metric is the probability of an incorrect classification of a new input of a random variable, if that classification were done according to the current distribution of the data set's class labels[19]. It is calculated using equation 1, with G representing the Gini impurity, c being the total number of classes and p(i) the probability of picking a data point with class i.

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i)) \tag{1}$$

A process of feature selection can be performed by observing the impact each feature has on the Gini value of a node and establishing a threshold value of Gini impact above which the features are selected[20].

This analysis was performed using the "RandomForest" package[21]. The number of variables tried at each split (also known as tree depth) is generally recommended to be close to the square root of total, so in this case $\sqrt{27} = 5$. In order to analyze the impact of tree depth in the final performance of the model and associated study, two models were implemented, corresponding to a depth of 5 and 20. Forests were created with the number of trees ranging from 2000 to 800000.

Following this, a plot of the mean decrease GINI value for each family was obtained from the Random Forests with the lowest out-of-bag error rate and all families that caused an increase of +5% in the mean Gini value were selected. This was done since distinctive features identified by random forests are visualized by their Gini score[12]. Prediction success was estimated with the out-of-bag error (how often a subsample was misclassified)[22].

Following the Random Forest generation and selection of the relevant TF families by using their Mean GINI value, a Least Squares Means test is performed on all relevant families, comparing their early and late abundance values. Least-squares means are predictions on a linear model and can be used in unbalanced datasets[23]. Following a pairwise t-test on the LS means results, with false discovery rate adjustment, the families with statistically significant ($p<0.01$) differences between these two time groups were considered to be bioindicators.

After the selection of these bioindicator families, one family (NtrC.DctD) was chosen as it had the most significant results and few subfamilies. These subfamilies were also tested with a pairwise LS-means statistical test, in order to determine whether their abundance varied significantly between the Early and Late time groups.

3. Results and Discussion

3.1. PredicTF

Upon using the PredicTF tool to predict transcription factors on the 15 metagenomic assemblies, a total of 1569 TFs were found, spread out over 159 subfamilies and 27 families.

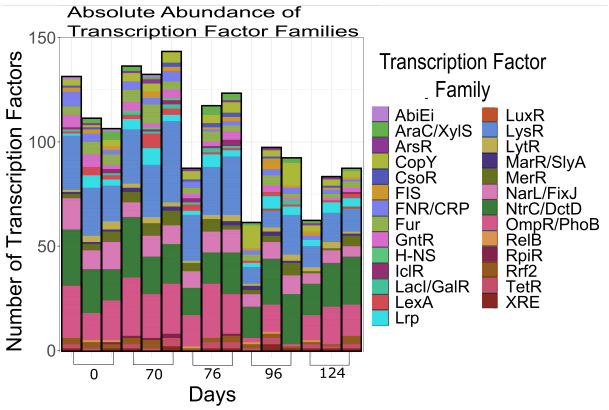

Figure 2: Absolute abundance of Transcription Factor families for samples collected at the different days of the experiment. Each column represents one sample. The divisions (line) inside each family represent subfamilies.

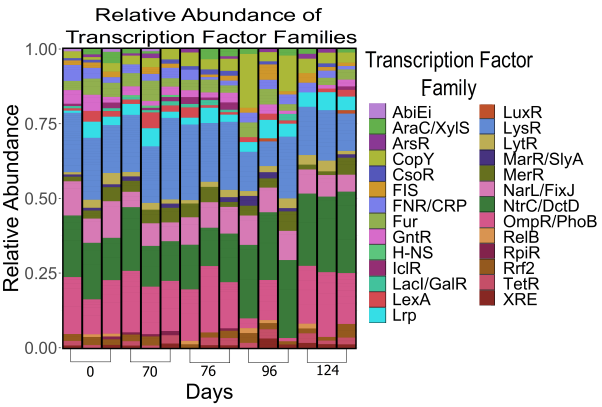

Figure 3: Relative abundance (%) of Transcription Factor families for samples collected at the different days of the experiment. Each column represents one sample. The divisions (line) inside each family represent subfamilies.

3.2. Alpha Diversity

The alpha diversity analysis reveals that there are statistically significant differences in the alpha diversity of sub-families among certain samples, for both the observed and Shannon metrics, but no statistically significant differences between the alpha diversity of the families.

Days	Days	Adjusted p value (Observed)	Adjusted p value (Shannon)
0	70	1	1
0	76	1	1
0	96	1	1
0	124	1	0.969
70	76	0.897	**0.019**
70	96	1	1
70	124	0.184	**0.026**
76	96	1	1
76	124	**0.033**	0.057
96	124	1	1

Table 1: Pairwise t-test on the Alpha diversity results for TF sub-families (Observed Number of TF families and Shannon Index), using Bonferroni correction for adjusted p-values. Statistically significant (adjusted p-value$<$0.05) values in bold.

Days	Days	Adjusted p value (Observed)	Adjusted p value (Shannon)
0	70	1	1
0	76	1	1
0	96	1	1
0	124	0.848	1
70	76	1	1
70	96	0.848	1
70	124	0.848	1
76	96	1	1
76	124	0.198	1
96	124	1	1

Table 2: Pairwise t-test on the Alpha diversity results for TF families (Observed Number of TF families and Shannon Index), using Bonferroni correction for adjusted p-values.

This indicates that the overall species diversity seems to be the same among all samples, for the families of TFs. Since these are alpha diversity measures, the results indicate that over time in this succession experiment, the number of different transcription factor families does not seem to change significantly. This does not however mean that the abundance of certain families has not changed. In order to observe whether different families are being expressed more or less, we'll have to resort to beta diversity metrics, comparing the diversity between samples and time points.

3.3. Beta Diversity

The PCoA results using relative abundance of TF families and the absolute abundance of subfamilies can be seen in figure 4 and 5, respectively. Following analysis of the graphs, there was a decision to split the samples into two distinct groupings. These groupings are the Early (0, 70 and 76 days) samples and the Late (96 and 124 days) samples. The reason for these groupings came from realizing that this separation of samples could always be divided by a decision surface in all PCoA and NMDS graphs, but most importantly, this split is very apparent in the family plots and given that in these graphs the 2 first principal coordinates account for the greatest amount of variance(60% for the family analysis vs 30% for the subfamily analysis) and they also have the most statistically significant results in the PERMANOVA analysis, they were considered the most relevant. The lower variance of the 2 main principal coordinates and the less statistically significant values of the subfamily analysis are believed to be due to the somewhat limited nature of this dataset and the large number of different subfamilies with very low abundance present, which makes it difficult to meaningfully discern between different samples.

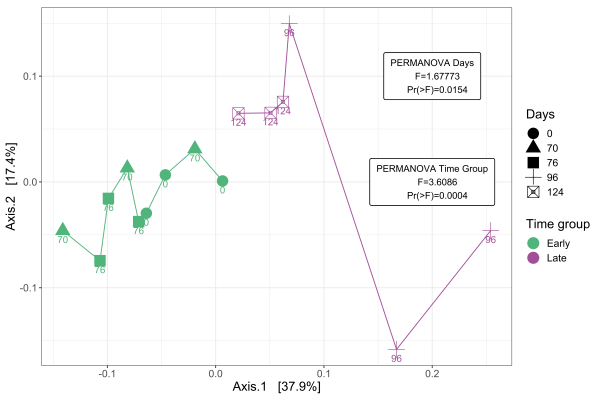

Figure 4: Principal Coordinate analysis of Bray distances comparing the relative abundance of TF families in different samples. We defined two time groups as Early and Late. Permutational multivariate analysis of variance (PERMANOVA) of both the individual Days and the time groups are shown in the figure.

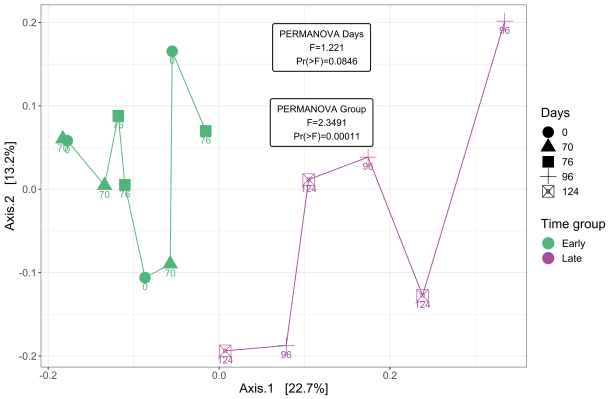

Figure 5: Principal Coordinate analysis of Bray distances comparing the absolute abundance of TF subfamilies in different samples. We defined two time groups as Early and Late . Permutational multivariate analysis of variance (PERMANOVA) of both the individual Days and the time groups are shown in the figure.

Principal coordinates analysis using Bray-Curtis distances revealed contrasts associated with the sampling time, both for the individual time samples and time groups. Samples collected before and at 76 days differed significantly from those collected afterwards. These differences among the different time points and time groups were confirmed in a PERMANOVA analysis on Bray distances, for the different time samples (Pseudo-F=1.88;P=0.01149) and for the time groups (Pseudo-F=3.61;P=0.0005). This indicates there is a clear difference in the abundance levels of certain families in the samples among the different time points/group, potentially as a result of the stress of the environment altering abundance levels over time. Following these results, a decision was made to, for the most part, disregard the abundance of subfamily values in the proceeding analyses, as these are not considered as relevant, for the reasons previously stated.

By classifying the samples as late or early, as previously established, a random forest model was obtained with the confusion matrix shown in table 3. This confusion matrix corresponds to an out-of-bag error of 6,7%, due to a missclassification of 1 late sample as early. Interestingly, this exact confusion matrix (1 missclassification of a late sample) was always obtained with either 5 or 20 variables per tree, with the number of trees ranging from 2000 to 800000. This shows that, at least for this dataset, corresponding the number of trees in a forest to the number of possible combinations of features was excessive and a lower value should be picked as it provides the same results while being less computationally intensive. It should however be noted that this effect may be due to the very high GINI index value of the NtrC.DctD family, as it is possible to determine the class of a sample by simply taking into account the relative abundance of this family, something which will be further discussed below. In addition, several different divisions of the samples were created in an attempt to create a RF classifier with a confusion matrix with a smaller OOB error rate than the Early vs Late division RF in an attempt to determine whether a more optimal division of the samples could be found. One such example was the grouping of all samples except the ones taken at 96 days and comparing them against the samples taken at 96 days, given that, as demonstrated in **??** and the pairwise t-tests in table **??**, the samples taken at this time point seem to have higher variance than at any other period. All different subdivisions, however, proved unsuccessful as all forests generated OOB errors above 10%.

The mean decrease GINI values for the RF with Early and Late time groups and tree depth of 5 is plotted in figure 6, and the relevant families are determined by evaluating which families contribute more then 5% to the sum of Mean GINI values so far. This results in 7 families being selected as relevant(NtrC.DtcD, LysR, XRE, GntR, LacI.GalR, CopY and Fur).

Table 3: Confusion Matrix following the Random forest analysis of the time groups. Out-of-bag error= 6,7%. [a] Early time group, consisting of the samples taken at 0, 70 and 76 days, [b] Late time group, consisting of the samples taken at 96 and 124 days, [c] The proportion of instances misclassified over the whole set of instances.

Confusion Matrix	Early[a]	Late[b]	Classification error[c]
Early	**9**	**0**	**0**
Late	**1**	**5**	**0,17**

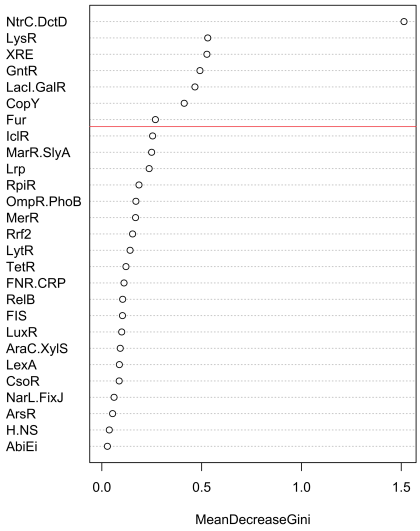

Figure 6: Mean Gini Decrease values for all 27 Transcription Factor families, representing their impact on the Random Forest algorithm, using 5 variables per tree. The Transcription Factor families below the red line add less than 5% for the sum of the Mean Gini Decrease values.

3.4. Statistical Test

Following the selection of the families in the previous section, the LS means test was applied, with results being shown in table 4.

Table 4: Least square mean analysis of selected Transcription Factor families between the Early and Late time groups. Values in bold indicate TF families with relative abundance values statistically different between the Early and Late time groups.

TF family	p-value	SE	t ratio
NtrC.DctD	**0.00004**	0.01417	-6.08016
LysR	**0.00644**	0.02002	3.24074
XRE	0.07771	0.00388	-1.91529
GntR	0.02975	0.00678	2.44025
LacI.GalR	0.13747	0.00352	1.58285
CopY	0.05081	0.02058	-2.15160
Fur	0.20420	0.00853	1.33683

The 2 families selected as bioindicators are the NtrC.DctD and LysR families.

In order to visualize the difference in abundance levels of these 2 families between early and late, a box plot was constructed, as can be seen in figure 7. Interestingly, the change in the number of variables tried at each tree led to no difference in the final results as to which families are considered bioindicators, since the families added by lowering this value did not have statistically significant ($p < 0.01$) differences in their early vs late abundance levels.

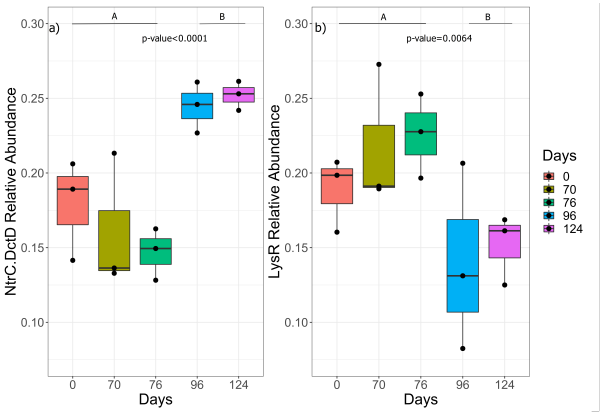

Figure 7: Relative abundance of the (a) NtrC.DctD and (b) LysR transcription factor families over 5 different days. Different capital letters (A or B) represent a statistically significant difference between the two time groups (Early and Late), with the respective p-value represented in the graph.

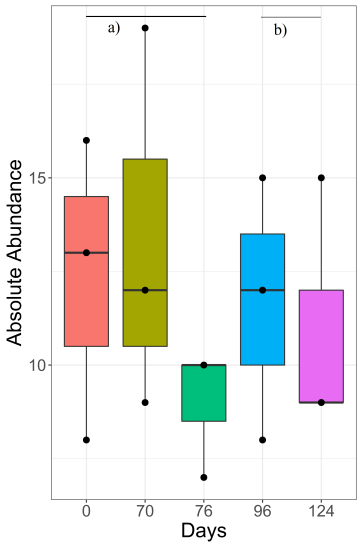

Figure 8: Absolute abundance of the PILR transcription factor subfamily over 5 different days. Different letters (a or b) represent a statistically significant difference between the two time groups (Early and Late).

As can be seen in figure 7, the relative abundance of the NtrC.DctD family rises in the late time group while decreasing for the LysR family. Of note, there is a clear cutoff for the abundance of the NtrC.DctD family since all early samples have a relative abundance below 22%, while all late samples have a relative abundance above 22%. This clear cutoff may account for why this family had such a high GINI value following the construction of the Random Forest, as a clear cutoff would have made any node which took into account the relative abundance of this family be able to immediately classify a sample as early or late. This effect is only exacerbated in the forest which had 20 variables per tree, as most of the trees would have this family as a node, thus increasing its impact in the decision making even further, showing why a smaller number of variables should be used in each tree, as it allows them to be further decorrelated[18].

This effect is further confirmed by the building of a decision tree using all families selected by the GINI plot (before the LS-means test), in order to be able to visualize the decision making process and potentially visualize some patterns regarding whether a family was up or downregulated over time. However, due to the clear cutoff previously mentioned, a decision tree which includes the NtrC.DctD family is just a single node, using that family's abundance value to distinguish between samples, with 100% accuracy.

The NtrC.DctD family is not only relevant in the RFs, but also had relatively few subfamilies detected (7) when compared to families like LysR which had around 40 subfamilies. Therefore, an LS-means analysis on the absolute abundance values of NtrC.DctD subfamilies with more than 5 instances comparing the Late and Early time groups was conducted in an attempt to study which specific subfamilies were up- or downregulated over time. The results can be found in table 5.

NtrC.DctD Subfamily	p-value	SE	t-ratio
PILR	**0.01201**	0.68563	-2.91703
GNFM	0.16114	0.48603	1.48595
XYLR	0.71193	0.58875	0.37745
CBRB	0.05081	0.02058	-2.15160

Table 5: Least square mean analysis of selected Transcription Factor families between the Early and Late time groups. Values in bold indicate TF families with absolute abundance values statistically different between the Early and Late time groups.

Only one subfamily had a statistically significant($p<0,05$) difference in it's abundance between Early and Late time groups, the PILR subfamily. This subfamily is responsible for activation of the pilin gene. This gene has been implicated in playing a key role during the initial stages of colonization of a host by the pathogen *Pseudomonas aeruginosa*. In order to visualize it's abundance, the following graph was produced 8.

Although not as statistically significant as the results from the 2 bioindicator families, the abundance of the PILR subfamily seems to decrease over time. This could be due to a multitude of reasons: the increased stress of this environment led to the microorganisms having to "focus" their expression to other TFs which would help them deal with the pollutant in the ecosystem; it's also possible the motility conferred by the *piln* gene activated by the PILR subfamily was not as advantageous in the new benzene-degrading conditions of the community, leading to either it not being expressed as much or even to a decrease in the population of microorganisms which express this TF, given that it has only been identified in certain species[24][25].

The NtrC.DctD family represents transcription factors which are involved in nitrogen regulation[2]. As previously mentioned, the family NtrC.DctD has a higher relative abundance in the late time group. Considering these samples originate from a microcosm with nitrate-reducing conditions, this result may be partly attributed to environmental conditions. Due to the use of nitrate as a substrate, the pathways that regulate nitrogen consumption/nitrate reduction would be expected to have higher importance among the different organisms present in this ecosystem. It is important to note that this family contains TFs that both upregulate and downregulate multiple different steps in the nitrogen regulation pathways and therefore an increase in the abundance levels of NtrC.DctD TFs can have multiple causes[26].Although, it should be noted that most TFs have been shown to upregulate genes[27].

One potential cause for this increase could be that, due to the increased importance of nitrogen regulation in a nitrate-reducing environment, the organisms present within the microcosm had their NtrC.DctD transcription factors' abundance increased in order to more accurately and better regulate the nitrate-reducing pathways. This increase in abundance can be observed gradually over time, indicating why the abundance levels of the TFs are higher in samples taken at later times.

Another potential bioindicator is the LysR family, which has a statistically significant difference in its relative abundance when comparing the early and late time groups. Unlike the NtrC.DctD family, the relative abundance of the LysR family decreases for the later time group, possibly indicating a decline in the transcription or regulation of genes regulated by this family in a nitrate reducing/benzene present environment.

The LysR family of transcriptional regulators is the most abundant family in the prokaryotic kingdom and regulates a very diverse set of genes, involved in functions such as virulence, metabolism, quorum sensing and motility[28]. Due to the varied amount of gene functions that the LysR family regulates,

it is hard to determine whether the regulation of certain specific functions within this family is being suppressed or activated due to the environmental stresses and it is therefore harder to draw conclusions as to what specific functions could be affecting the abundance levels of TFs of this family. One way to more precisely map out which functions' TFs are being affected is to analyze the abundance levels of the LysR subfamilies, as each will have more precise functions who's importance in this system may be more easily studied. However, due to the size of the dataset and, in the LysR family, the very large number of subfamilies, there isn't enough data to draw any significant conclusions. This could potentially be remedied with a larger dataset to work with, less specific subfamily groupings, or better-quality reads.

It should always be noted that an increase or decrease in the abundance of a certain TF in the genome does not necessarily equate to that TF being more or less translated (though the 2 are often correlated)[29]. Therefore an analysis of the transcriptome and proteome (as was originally planned to be performed in this work) can accompany this sort of work in order to more accurately determine whether or not these TFs are indeed being transcribed into RNA and proteins.

The techniques performed in this work for TF analysis have proved useful, as they allow a better understanding of regulatory networks and the mechanisms which govern TF abundance, helping in the modeling and description of these complex networks, by providing insight that techniques such as phylogenetic or gene expression analyses do not, namely TF abundance values and how they are affected.

4. Conclusions and Future Work

The main objectives of this work were to determine the impact of environmental stresses on the abundance of bacterial TFs and whether these differences in abundance levels could be used to distinguish between samples at different time points of a succession experiment, thus providing direct insight into the regulatory networks of the community. With this objective in mind, data from a microcosm experiment was used as input for the newly developed PredicTF tool in order to detect TF abundance. These abundance values are then used to perform a diversity analysis and to determine bioindicators using Random Forests and statistical tests.

The diversity analysis results, mainly the beta diversity analysis using PCoA, show that the TF family abundance levels of the community are indeed changing over time, indicating that external factors do indeed affect bacterial TF abundance, as had been shown in literature before[30]. In addition, this analysis also showed that there appears to be a clear contrast between samples taken earlier than 96 days and those taken later.

When searching for bioindicators, 2 TF families were found who's abundance altered significantly over time: NtrC.DctD and LysR. While the NtrC.DctD change in abundance may be caused by the nitrate reducing environment of the microcosm experiment the differences in abundance of LysR TFs may be harder to explain, due to the very wide variety of functions different TFs in this family regulate. Regardless, these results show that not only does TF abundance change as a result of external factors, but this fact may potentially be used in the realms of ecology, given that if the abundance of specific families changes with specific external factors (e.g. overabundance of NtrC.DctD TFs in the presence of a nitrate reducing environment), then the abundance of these families may be used as a tool for better monitoring of the regulatory networks of microbial communities.

As previously stated, just because a TF is present in the metagenome, this does not mean it is necessarily transcribed into a protein. In future work one could complement the metagenomic samples with transcriptomic (determine whether the TF genes are being transcribed) and proteomic (determine whether the TFs are being translated into proteins) samples in order to have a more accurate understanding of which TFs are in fact being expressed. In addition, in this work due to the limited nature of the dataset, the abundance values of the TF subfamilies were not extensively used, namely in the search for bioindicators. Despite this, the abundance values of the subfamilies have the potential to be more relevant than the more generalized family groupings and should not be disregarded immediately in any future work, especially if a larger dataset with greater abundance of subfamilies can be produced. This is due to the subfamilies affecting the expression of a smaller amount of genes, often with similar functions, so a variation on the abundance of a specific subfamily can be more easily linked to a specific function or characteristic of the community.

Another analysis that may be performed alongside this work is a phylogenetic analysis of the species present in the microbiome. Such an analysis provides insight into the microbial composition of the community over time. When coupled with the PredicTF tool, it may allow for determination of, for instance, whether a decrease of the abundance of a specific TF is due to that TF not being upregulated in the conditions of microcosm, or, due to the population of microorganisms in which that TF is present decreasing or disappearing over time[31]. A phylogenetic analysis could have provided such insight into the causes of the decreased abundance of the PILR subfamily over time.
The work presented in this thesis was performed at Helmholtz Centre for Environmental Research (Leipzig, Germany), during the period of September 2020 - February 2021, under the supervision of Dr. Ulisses Rocha. The thesis was co-supervised at Instituto Superior Técnico by Prof. Ana Luísa Nobre Fred.

5. Acknowledgments

The work presented in this thesis was performed at Helmholtz Centre for Environmental Research (Leipzig, Germany), during the period of September 2020 - February 2021, under the supervision of Dr. Ulisses Rocha. The thesis was co-supervised at Instituto Superior Técnico by Prof. Ana Luísa Nobre Fred.

References

[1] S. Widder, R. J. Allen, T. Pfeiffer, T. P. Curtis, C. Wiuf, W. T. Sloan, O. X. Cordero, S. P. Brown, B. Momeni, W. Shou, *et al.*, "Challenges in microbial ecology: building predictive understanding of community function and dynamics," *The ISME journal*, vol. 10, no. 11, pp. 2557–2568, 2016.

[2] L. M. O. Monteiro, J. Saraiva, R. B. Toscan, P. F. Stadler, R. Silva-Rocha, and U. N. da Rocha, "Predictf: a tool to predict bacterial transcription factors in complex microbial communities," *bioRxiv*, 2021.

[3] L. J. Hawkins, R. Al-Attar, and K. B. Storey, "Transcriptional regulation of metabolism in disease: From transcription factors to epigenetics," *PeerJ*, vol. 6, p. e5062, 2018.

[4] U. Consortium *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 46, no. 5, p. 2699, 2018.

[5] S. Kılıç, E. R. White, D. M. Sagitova, J. P. Cornish, and I. Erill, "Collectf: a database of experimentally validated transcription factor-binding sites in bacteria," *Nucleic acids research*, vol. 42, no. D1, pp. D156–D160, 2014.

[6] D. Metze, D. Popp, L. Schwab, N.-S. Keller, U. N. da Rocha, H.-H. Richnow, and C. Vogt, "Temperature management potentially affects carbon mineralization capacity and microbial community composition of a shallow aquifer," *FEMS Microbiology Ecology*, 2020.

[7] A. Klindworth, E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, and F. O. Glöckner, "Evaluation of general 16s ribosomal rna gene pcr primers for classical and next-generation sequencing-based diversity studies," *Nucleic acids research*, vol. 41, no. 1, pp. e1–e1, 2013.

[8] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaspades: a new versatile metagenomic assembler," *Genome research*, vol. 27, no. 5, pp. 824–834, 2017.

[9] R. Lande, "Statistics and partitioning of species diversity, and similarity among multiple communities," *Oikos*, pp. 5–13, 1996.

[10] P. J. McMurdie and S. Holmes, "phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data," *PloS one*, vol. 8, no. 4, p. e61217, 2013.

[11] J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, B. O'Hara, G. Simpson, P. Solymos, H. Stevens, and H. Wagner, *Vegan: Community Ecology Package*, vol. 1. 01 2010.

[12] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[13] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[14] M. J. Anderson and D. C. Walsh, "Permanova, anosim, and the mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing?," *Ecological monographs*, vol. 83, no. 4, pp. 557–574, 2013.

[15] S. Dray, P. Legendre, and P. R. Peres-Neto, "Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (pcnm)," *Ecological modelling*, vol. 196, no. 3-4, pp. 483–493, 2006.

[16] S. M. Holland, "Non-metric multidimensional scaling (mds)," *Department of Geology, University of Georgia, Athens, Tech. Rep. GA*, pp. 30602–2501, 2008.

[17] M. J. Anderson, "Permutational multivariate analysis of variance (permanova)," *Wiley statsref: statistics reference online*, pp. 1–15, 2014.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] R. I. Lerman and S. Yitzhaki, "A note on the calculation and interpretation of the gini index," *Economics Letters*, vol. 15, no. 3-4, pp. 363–368, 1984.

[20] P. M. Rosado, D. C. Leite, G. A. Duarte, R. M. Chaloub, G. Jospin, U. N. da Rocha, J. P. Saraiva, F. Dini-Andreote, J. A. Eisen, D. G. Bourne, *et al.*, "Marine probiotics: increasing coral resistance to bleaching through microbiome manipulation," *The ISME journal*, vol. 13, no. 4, pp. 921–936, 2019.

[21] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[22] S. Janitza and R. Hornung, "On the overestimation of random forest's out-of-bag error," *PloS one*, vol. 13, no. 8, p. e0201904, 2018.

[23] R. Lenth, "Least-squares means. r package 'lsmeans'. 2016," 2016.

[24] K. S. Ishimoto and S. Lory, "Identification of pilr, which encodes a transcriptional activator of the pseudomonas aeruginosa pilin gene.," *Journal of Bacteriology*, vol. 174, no. 11, pp. 3514–3521, 1992.

[25] D. Sakai and T. Komano, "The pill and piln genes of inci1 plasmids r64 and colib-p9 encode outer membrane lipoproteins responsible for thin pilus biogenesis," *Plasmid*, vol. 43, no. 2, pp. 149–152, 2000.

[26] M. R. Atkinson, E. S. Kamberov, R. L. Weiss, and A. J. Ninfa, "Reversible uridylylation of the escherichia coli pii signal transduction protein regulates its ability to stimulate the dephosphorylation of the transcription factor nitrogen regulator i (nri or ntrc).," *Journal of Biological Chemistry*, vol. 269, no. 45, pp. 28288–28293, 1994.

[27] M. J. Mann, "Transcription factor decoys: a new model for disease intervention," *Annals of the New York Academy of Sciences*, vol. 1058, no. 1, pp. 128–139, 2005.

[28] S. E. Maddocks and P. C. Oyston, "Structure and function of the lysr-type transcriptional regulator (lttr) family proteins," *Microbiology*, vol. 154, no. 12, pp. 3609–3623, 2008.

[29] P. Hugenholtz and G. W. Tyson, "Metagenomics," *Nature*, vol. 455, no. 7212, pp. 481–483, 2008.

[30] D. F. Browning, M. Butala, and S. J. Busby, "Bacterial transcription factors: regulation by pick "n" mix," *Journal of molecular biology*, vol. 431, no. 20, pp. 4067–4077, 2019.

[31] J. P. Huelsenbeck, J. Bull, and C. W. Cunningham, "Combining data in phylogenetic analysis," *Trends in Ecology & Evolution*, vol. 11, no. 4, pp. 152–158, 1996.