University of Padua

Department of mathematics Tullio-Levi Civita

Master Degree in Data Science

Metagenomics data analysis:
Development and application of AI methods for gut microbiome community characterization in IBD

Supervisor
prof. Francesco Rinaldi
University of Padua

Co-supervisor
Loris Bertoldi, PhD
BMR Genomics

Master Candidate
Veronica Pederiva

Academic Year
2021-2022
To all my beloved ones.
Abstract

The gut microbiome plays a crucial role in maintaining the host homeostasis. On the contrary, a dysregulation in the gut microbial composition can seriously affect the host health leading to a condition known as Inflammatory bowel disease (IBD), one of the most prevalent forms of dysbiosis. The serious impact the microbiome regulation has on the human health has led researchers to investigate which microbial and metabolite elements constitute a healthy core microbiome. In particular, understanding which species must be protected from pathogens proliferation in order to ensure a healthy functional environment may help with the definition of targeted therapies that can be either prebiotic or probiotic-based. Differentially abundant analysis (DAA) is usually applied to investigate species and metabolic pathways that are enriched or depleted in the dysbiotic condition compared to the healthy one. However, since interactions play a major role in the microbiome regulation, an innovative approach based on community detection was proposed in this thesis to identify communities characterizing a healthy or IBD-affected microbiota. Results of this latter approach were compared to the DAA outcomes and interestingly the IBD enriched Phasolactobacterium succinatutens species emerged as an IBD community leading bacteria, too. Consequently, this succinate-consumer bacteria might be studied as a potential target of new therapies. Notwithstanding, marker-based approaches as DAA are still valid to identify features that can be used for the definition of machine learning models. Indeed, the integration of data-driven models in the medical practice might provide a reliable evaluation of the IBD risk avoiding invasive procedures. In this work, a Random Forest classifier was successfully designed and trained to discriminate between healthy and IBD samples.
Contents

Abstract v
List of figures ix
List of tables xi
Listing of acronyms xiii

1 Introduction 1
  1.1 Microbiome .................................................. 1
  1.2 Dysbiosis ..................................................... 3
    1.2.1 IBD ..................................................... 4
    1.2.2 IBD and Machine Learning ............................. 5
  1.3 Networks in biology ......................................... 7

2 Objectives 9

3 Data 11
  3.1 Italian cohort ................................................ 11
  3.2 Previous studies cohorts .................................... 13

4 Methods 15
  4.1 Reads preprocessing ......................................... 15
  4.2 Taxonomic and functional profiles ......................... 15
  4.3 Rarefaction and diversity analysis ......................... 16
  4.4 Differential abundance analysis ........................... 19
  4.5 Community detection ....................................... 20
    4.5.1 Girvan-Newman algorithm .............................. 25
    4.5.2 Louvain algorithm ................................... 26
    4.5.3 FAST-ATVO ............................................ 28
    4.5.4 Louvain Multiobjective Method ....................... 29
  4.6 Random forest classifier .................................. 31

5 Results and discussion 37
  5.1 Italian cohort characterization ............................. 37
  5.2 Characterization of the mixed dataset .................... 47
5.3 Differential abundance analysis ............................................. 50
  5.3.1 Italian cohort ............................................................ 50
  5.3.2 Mixed dataset ........................................................... 54
5.4 Community detection .......................................................... 58
  5.4.1 Single-layer analysis ..................................................... 58
  5.4.2 Multi-layer analysis ..................................................... 64
  5.4.3 Single-layer - Multi-layer comparison ............................. 67
5.5 Random forest classifier ....................................................... 75
  5.5.1 Feature selection ....................................................... 75
  5.5.2 Classification results .................................................. 77

6 Conclusion ................................................................. 85

References ................................................................. 87

Acknowledgments .......................................................... 97
### Listing of figures

3.1 EDA visualization comparing the healthy and non-healthy (IBD) group .......................... 13  
3.2 IBD group relevant EDA information ............................................................................... 13  
3.3 Number of subjects per age class ................................................................................... 14  
4.1 Rarefaction curve example ............................................................................................. 17  
4.2 A simple decision tree classifier with 4 features ............................................................ 32  
5.1 Most prevalent phyla in the Italian cohort ........................................................................ 38  
5.2 Most prevalent phyla in the BMR-ITA healthy samples ................................................. 38  
5.3 Most prevalent phyla in the BMR-ITA IBD samples ....................................................... 39  
5.4 Most abundant pathways in the Italian cohort ................................................................. 39  
5.5 Most abundant pathways in the BMR-ITA healthy samples .......................................... 40  
5.6 Most abundant pathways in the BMR-ITA IBD samples ............................................. 40  
5.7 Rarefaction analysis ......................................................................................................... 41  
5.8 Observed number of species per BMR-ITA sample ....................................................... 42  
5.9 Alpha diversity indices of the Italian cohort ..................................................................... 43  
5.10 Most significant alpha diversity indices ......................................................................... 44  
5.11 Alpha diversity indices to compare T0 and T1 IBD BMR-ITA samples .......................... 45  
5.12 Beta diversity PCoA on various distance metrics on the Italian cohort ......................... 46  
5.13 Unweighted UniFrac PCoA on the Italian cohort ............................................................ 46  
5.14 Number of genera, species and pathways per dataset .................................................. 48  
5.15 Beta diversity PCoA on various distance metrics on the Mixed dataset ......................... 48  
5.16 Unweighted UniFrac PCoA on the Mixed dataset ........................................................... 49  
5.17 Beta diversity PCoA on the healthy samples of the Mixed dataset ............................... 49  
5.18 Beta diversity PCoA on the IBD samples of the Mixed dataset ..................................... 50  
5.19 Heatmap of the Spearman correlation between DA species and pathways .................. 52  
5.20 Unweighted UniFrac beta diversity PCoA highlighting samples dysbiosis index computed on BMR-ITA DA features ................................................................. 56  
5.21 Unweighted UniFrac beta diversity PCoA highlighting samples dysbiosis index computed on the Mixed dataset DA features .............................................................................. 57  
5.22 CoOccur Control single-layer network ......................................................................... 58  
5.23 CoOccur IBD single-layer network ................................................................................ 59  
5.24 Pearson Control single-layer network ............................................................................ 61  
5.25 Pearson IBD single-layer network .................................................................................. 61  
5.26 Multi-layer Louvain Control Community 2 ..................................................................... 66
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.27</td>
<td>Pathways supergroups comparison between Control and IBD networks both Single and Multi-Layer</td>
<td>69</td>
</tr>
<tr>
<td>5.28</td>
<td>Multi-layer Louvain Control community 2 related pathway heatmap (100% enzyme coverage)</td>
<td>71</td>
</tr>
<tr>
<td>5.29</td>
<td>Multi-layer Louvain Control community 0 related pathway heatmap (Faecalibacterium prausnitzii)</td>
<td>72</td>
</tr>
<tr>
<td>5.30</td>
<td>Multi-layer Louvain IBD community 3 related pathway heatmap (Faecalibacterium prausnitzii)</td>
<td>72</td>
</tr>
<tr>
<td>5.31</td>
<td>Multi-layer Louvain IBD community 1 related pathway heatmap (Escherichia coli)</td>
<td>73</td>
</tr>
<tr>
<td>5.32</td>
<td>Single-layer GN IBD community 9 related pathway heatmap (Phascolarctobacterium succinatutens)</td>
<td>74</td>
</tr>
<tr>
<td>5.33</td>
<td>Multi-layer Louvain IBD community 4 related pathway heatmap (Phascolarctobacterium succinatutens)</td>
<td>75</td>
</tr>
</tbody>
</table>
Listing of tables

1.1 Summary of Random forest IBD-Control classification results published in recent works ........................................ 6
3.1 Datasets composition and origin ........................................ 14
4.1 Random forest hyperparameters tested values ......................... 34
5.1 Kruskal-Wallis rank sum test on alpha diversity indices p-values .......... 45
5.2 Number of differentially abundant features in the Italian cohort .......... 51
5.3 Differentially abundant features in the Italian cohort ................. 53
5.4 Number of DA features in the mixed dataset and DA features shared with BMR-ITA .................................................. 55
5.5 Single-layer CoOccur networks quality analysis .......................... 60
5.6 Similarity indices capturing differences between healthy and non-healthy Single-layer CoOccur communities ....................... 60
5.7 Single-layer Pearson networks quality analysis .......................... 62
5.8 Similarity indices capturing differences between healthy and non-healthy Single-layer Pearson communities ....................... 62
5.9 Similarity indices comparing Single-layer communities to groups of species participating in the same pathway ....................... 63
5.10 Number of edges per layer in the Multi-Layer network ................... 64
5.11 Number of communities found in the Multi-Layer networks with the different algorithms ........................................ 64
5.12 Multi-layer networks per layer modularity .................................... 65
5.13 Similarity indices capturing differences between healthy and non-healthy Multi-layer communities ...................................... 67
5.14 Similarity indices comparing Multi-layer communities to groups of species participating in the same pathway ....................... 67
5.15 Comparison of pathways enzyme coverage between the Single-Layer CoOccur and Multi-Layer communities in the Control and IBD case ........ 69
5.16 Single to Multi-layer communities mapping .................................... 70
5.17 Top 11 most important features selected by the Random forest on the BMR-ITA cohort .................................................. 76
5.18 Samples stratification per case and cohort in the first Random forest classification .................................................. 77
5.19 Random forest classifier performance metrics ............................. 78
§.20 Random forest classifier best hyperparameters configurations 
§.21 Random forest misclassified samples 
§.22 Samples stratification per case and cohort in the second Random forest classification 
§.23 Random forest classifier performance metrics on tests without ES samples 
§.24 Random forest classifier best hyperparameters configurations on tests without ES samples 
§.25 Random forest misclassified samples on tests without ES samples
## Listing of acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BH</td>
<td>Benjamini-Hochberg correction procedure</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>BMR-ITA</td>
<td>Italian Cohort Dataset identifier</td>
</tr>
<tr>
<td>CD</td>
<td>Chron’s Disease</td>
</tr>
<tr>
<td>CRC</td>
<td>Colorectal Cancer</td>
</tr>
<tr>
<td>DA</td>
<td>Differentially Abundant</td>
</tr>
<tr>
<td>DAA</td>
<td>Differential Abundance Analysis</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
</tr>
<tr>
<td>EDA</td>
<td>Exploratory Data Analysis</td>
</tr>
<tr>
<td>FMT</td>
<td>Fecal Microbiota Transplantation</td>
</tr>
<tr>
<td>GN</td>
<td>Girvan-Newman algorithm</td>
</tr>
<tr>
<td>HMP2</td>
<td>Integrative Human Microbiome Project</td>
</tr>
<tr>
<td>IBD</td>
<td>Inflammatory Bowel Disease</td>
</tr>
<tr>
<td>IBS</td>
<td>Irritable Bowel Syndrome</td>
</tr>
<tr>
<td>MetaHIT</td>
<td>METAgenomics of the Human Intestinal Tract</td>
</tr>
<tr>
<td>PCoA</td>
<td>Principal Coordinates Analysis</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>SCFA</td>
<td>Short-Chain Fatty Acid</td>
</tr>
<tr>
<td>SRA</td>
<td>Sequence Read Archive</td>
</tr>
<tr>
<td>Sar</td>
<td>Adjusted Rand similarity coefficient</td>
</tr>
<tr>
<td>Sfm</td>
<td>Fowlkes-Mallows similarity coefficient</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Sg</td>
<td>Gamma similarity coefficient</td>
</tr>
<tr>
<td>Sj</td>
<td>Jaccard similarity coefficient</td>
</tr>
<tr>
<td>Sm</td>
<td>Minkowski similarity coefficient</td>
</tr>
<tr>
<td>Sr</td>
<td>Rand similarity coefficient</td>
</tr>
<tr>
<td>T2D</td>
<td>Type 2 Diabetes</td>
</tr>
<tr>
<td>UC</td>
<td>Ulcerative Colitis</td>
</tr>
</tbody>
</table>