Predicting PCOS with Gut Microbiome data

Machine Learning Project

Summary

Elizaveta Nikolaeva, Ritudisha Biswas, Eduardo Brito and Jennifer Neumaier

Dr. Elin Org, Head of Microbiome Research group at the Institute of Genomics, Tartu

https://github.com/edurbrito/pcos-ml
Polycystic Ovary Syndrome (PCOS)

Can gut microbiome data individually predict PCOS?

Does gut microbiome data offer higher accuracy when combined with metabolic traits?
The Gut Microbiome in PCOS patients
The Gut Microbiome in PCOS patients

[1]
Research Question and Challenges

Can gut microbiome data individually classify PCOS from the data?

Does gut microbiome data offer higher accuracy when combined with metabolic traits?

202 non-PCOS women

102 PCOS women

+ BMI matching

testosterone

insuline

age

glucose
Ideas

- Validating Model
- Balancing Dataset
- SVM, and Gaussian Naive Bayesian
- Ensemble Learning
- Integrating metabolic data

12/14/2021
. Getting a feel for the data: PCA
Getting a feel for the data: TSNE
Model performances original data
Keeping features with at least 0.1% abundance and 10% appearance in samples
Model performances LASSO features
Model performances Kbest features

![Calibration plots (Reliability Curves)](chart)

- Perfectly calibrated
- Logistic Regression
- Random Forest
- Gradient Boosting
- Extra Trees
Feature selection direct comparison

Filtered Data

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteroidetes</td>
<td>Bacteroidia</td>
<td>Bacteroidales</td>
<td>Bacteroidaceae</td>
<td>Bacteroides</td>
<td>Bacteroides uniformis</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>Bacteroidia</td>
<td>Bacteroidales</td>
<td>Bacteroidaceae</td>
<td>Bacteroides</td>
<td>Bacteroides dorei</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>Melainabacteria</td>
<td>Gastranaerophilales</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

Abundances of selected features

- Bacteria
- Bacteroides_dorei
- Bacteroides_uniformis
- Gastranaerophilales

12/14/2021
Model performances three features
h = 200
model = Sequential()
model.add(Dense(h, input_dim=D, activation='relu'))
model.add(Dense(K, activation='sigmoid'))

# Compile model
model.compile(loss=BinaryCrossentropy(from_logits=True),
              optimizer=SGD(learning_rate = 0.0001),
              metrics=['accuracy'])

# Run the model
history = model.fit(X, truth, epochs=25, batch_size=25,
                     validation_split=0.2)
Lessons Learned

- Feature selection
- Data processing
- Validating Model
- Balancing Dataset
- Ensemble Learning
- SVM, and Gaussian Naive Bayesian
- Integrating metabolic data

12/14/2021
Excerpt: Sequencing data

Various ML steps to sort identified gene sequences to known sequences from a database.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>G___Alistipes</td>
<td>P___Bacteroidetes</td>
<td>O___Bacteroidales</td>
<td>F___Prevotellaceae</td>
<td>G___Prevotella</td>
<td>G___Odoribacter</td>
</tr>
<tr>
<td>2</td>
<td>21Mar11-DL048</td>
<td>4473</td>
<td>3570</td>
<td>5091</td>
<td>879</td>
<td>332</td>
</tr>
<tr>
<td>3</td>
<td>21Mar11-DL056</td>
<td>4278</td>
<td>2020</td>
<td>5986</td>
<td>1119</td>
<td>431</td>
</tr>
<tr>
<td>4</td>
<td>21Mar11-DL064</td>
<td>2539</td>
<td>4531</td>
<td>8719</td>
<td>299</td>
<td>110</td>
</tr>
<tr>
<td>5</td>
<td>21Mar11-DL072</td>
<td>4993</td>
<td>2634</td>
<td>4515</td>
<td>1382</td>
<td>398</td>
</tr>
<tr>
<td>6</td>
<td>21Mar11-DL080</td>
<td>2309</td>
<td>6419</td>
<td>8999</td>
<td>669</td>
<td>425</td>
</tr>
<tr>
<td>7</td>
<td>21Mar11-DL088</td>
<td>4564</td>
<td>1991</td>
<td>3042</td>
<td>95</td>
<td>79</td>
</tr>
<tr>
<td>8</td>
<td>21Mar11-DL140</td>
<td>3922</td>
<td>5534</td>
<td>7222</td>
<td>823</td>
<td>561</td>
</tr>
<tr>
<td>9</td>
<td>21Mar11-DL144</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>21Mar11-DL148</td>
<td>6950</td>
<td>4438</td>
<td>8030</td>
<td>500</td>
<td>157</td>
</tr>
<tr>
<td>11</td>
<td>21Mar11-DL152</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>21Mar11-DL156</td>
<td>2401</td>
<td>1930</td>
<td>2646</td>
<td>568</td>
<td>176</td>
</tr>
</tbody>
</table>
Bacteroides uniformis and dorei

- *Bacteroidetes* is one of the major lineages of bacteria and arose early during the evolutionary process (233).
- *Bacteroides* species are anaerobic, bile-resistant, non-spore-forming, gram-negative rods.
- B. uniformis found in abdominal (peritoneal) and appendix anaerobic infections.

Proportions of Bacteroides species seen clinically

[6]
Melainabacteria class in phylum cyanobacteria
- obligate fermenters missing the genes necessary for aerobic and anaerobic respiration
- all Gastranaerophilales genomes contain the Embden–Meyerhof–Parnas (EMP) pathway, capable of converting glucose, mannose, starch, or glycogen into lactate, ethanol, and/or formate
- All representative genomes have the potential to produce riboflavin, nicotinamide, biotin, dehydrofolate, and pantoate as found previously
- Di Rienzi et al. highlighted the presence of FeFe hydrogenases in their human gut population genomes, speculating that these organisms are hydrogen-producing anaerobes in syntrophic interactions with hydrogenotrophic methanogens or acetogens.
Sources


