Unraveling the microbial community interactions in anaerobic digesters with Common Components Analysis

Francesc Puig-Castellví, PhD
francesc.puigcastellvi@agroparistech.fr
Microbial communities & anaerobic digesters

Anaerobic digester

Microbial community

\[ \text{CH}_4 \text{ (biogas)} \]

\[ \text{biomass (sludge)} \]
16S rRNA sequencing

- **Data generation**

  Microbial community → 16S rRNA sequences

- **Every unique 16S rRNA is represented by an OTU (Operational Taxonomic Unit)**

  4 OTUs

- **OTU count is proportional to species’ population:**

  16S rRNA sequences → OTU count → POPULATION

  - Orange: x 13
  - Green: x 16
  - Blue: x 6
  - Yellow: x 8
16S rRNA sequencing data

- **Relative to species’ population:**
  - If some species dominate the microbiome over the others, the data will be strongly skewed.
16S rRNA sequencing data

- **Relative to species’ population:**
  - If some species dominate the microbiome over the others, the data will be strongly **skewed**.
  - If some species are detected in a few samples, the data will be very **sparse**.
Objectives

- To establish a chemometric pipeline for the analysis of 16S rRNA sequencing data.

- To evaluate the changes in the microbial populations in anaerobic digesters derived from using different substrate composition.
1) SUBSTRATE COMPOSITION in ANAEROBIC DIGESTERS:

**BINARY MIXTURES AT DIFFERENT PROPORTIONS**

Fish waste  Sludge  Vegetal waste

2) TIME-COURSE EXPERIMENT

The same inoculum was introduced at day 0 in all digesters.

- Samples collected at the 21st and 28th days
- Samples collected at the 14th and 21st days
### DATASET

#### 1) rRNA extraction
2) 16S rRNA sequencing
3) OTU count

<table>
<thead>
<tr>
<th>Day</th>
<th>Samples</th>
<th>OTUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>28</td>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>28</td>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>28</td>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>28</td>
<td>28</td>
<td>21</td>
</tr>
</tbody>
</table>

19 samples x 1145 OTUs
Each subset was analyzed separately

Fish waste subset

10 samples x 1145 OTUs

Vegetal waste subset

10 samples x 1145 OTUs
1) OTUs found in less than two samples (< 20%) were excluded
   
   To reduce the sparsity of the dataset.
   To avoid resolving components descriptive of few (one or two) samples.

2) OTU counts for each sample were normalized to 100%
   
   To remove yield differences between samples in the extraction of 16S rRNA.

3) Pareto-scaling
   
   To reduce the skewness of the dataset.
   Auto-scaling was not used to avoid increasing excessively the variable importance of the lowest populated species.

4) Common Component Analysis (CCA) [1]

5) Interpretation of Scores and Loadings
   
   Variable selection using S-plot.
   Interpretation of the selected variables using phylogenetic data.

RESULTS AND DISCUSSION

1) OTUs found in less than two samples (< 20%) were excluded

To reduce the sparsity of the dataset.
To avoid resolving components descriptive of few (one or two) samples.

Less species were detected in digesters containing fish waste:

Fish substrate is more selective
RESULTS AND DISCUSSION

4) Common Component Analysis (CCA)

• It is an unsupervised method

• CC components are orthogonal

• CC weights are the greatest for variables with the same dispersion of the observations

• Each CC will group together the largest number of variables with the greatest effect on a particular dispersion of the individuals
RESULTS AND DISCUSSION

4) Common Component Analysis (CCA) Using 2 components on each subset

**Fish waste subset**

**Vegetal waste subset**

<table>
<thead>
<tr>
<th>%sludge</th>
<th>100</th>
<th>75</th>
<th>50</th>
<th>25</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>%fish</td>
<td>0</td>
<td>25</td>
<td>50</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>%sludge</th>
<th>100</th>
<th>75</th>
<th>50</th>
<th>25</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>%vegetal</td>
<td>0</td>
<td>25</td>
<td>50</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Scores**

**Samples**

- CC1
- CC2

**Legend**
- Early time point
- Late time point
RESULTS AND DISCUSSION

5) Interpretation of Scores

Using 2 components on each subset

Fish waste subset

Vegetal waste subset

CC1 is representative of digesters containing only one source of substrate.

• Positive scores: samples rich in sludge.

• Negative scores in fish waste subset: samples rich in fish waste substrate.

• Negative scores in vegetal waste subset: samples rich in vegetal waste substrate.
RESULTS AND DISCUSSION

5) Interpretation of Scores

Using 2 components on each subset

Fish waste subset  Vegetal waste subset

CC2 is representative of digesters containing substrate binary mixtures.

- Positive scores in fish waste subset: binary mixture of fish and sludge.
- Positive scores in vegetal waste subset: binary mixture of vegetal waste and sludge.

- Negative scores in fish waste subset: samples rich in sludge or in fish waste.
- Negative scores in vegetal waste subset: samples rich in sludge or in vegetal waste.
RESULTS AND DISCUSSION

5) Interpretation of Scores

In the fish subset, the microbial community described by positive CC2 scores is less active at late time points.
RESULTS AND DISCUSSION

5) Interpretation of Loadings

5.1 Variable selection using S plots.

S-plot CC1
CORR(CC1, X) vs. COV(CC1, X)

S-plot CC2
CORR(CC2, X) vs. COV(CC2, X)

484 variables → 59 variables
484 variables → 42 variables

71 unique variables

5.2 Interpretation of the selected variables using phylogenetic data
RESULTS AND DISCUSSION

5) Interpretation of Loadings

Phylogenetic map with the 71 selected OTUs
RESULTS AND DISCUSSION

5) Interpretation of Loadings

5.1 Interpretation of the selected variables using phylogenetic data

Selected variables on CC1 for FISH

Phylogenetic map with the 71 selected OTUs
RESULTS AND DISCUSSION

5) Interpretation of Loadings

5.1 Interpretation of the selected variables using phylogenetic data

**CC1**

Positive loadings: more in sludge
Negative loadings: more in fish waste
WHITE: not selected for CC1
### Biological Interpretation

<table>
<thead>
<tr>
<th>Sludge</th>
<th>Microorganism</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td><em>Anaerolineae spp.</em></td>
<td>Habitat → sludge</td>
</tr>
<tr>
<td>X</td>
<td><em>Phycisphaerae spp.</em></td>
<td>Fermentation of carbohydrates</td>
</tr>
<tr>
<td>X</td>
<td><em>Cloacimonadida spp.</em></td>
<td>Degradation of organic acids</td>
</tr>
<tr>
<td>X</td>
<td><em>Methanomicrobia spp.</em></td>
<td>Methanogenesis</td>
</tr>
<tr>
<td>X</td>
<td><em>Clostridia spp.</em></td>
<td>Highly polyphyletic (<em>diverse</em>) class</td>
</tr>
<tr>
<td>X</td>
<td><em>Synergistia spp.</em></td>
<td>Syntrophic metabolism (“symbiosis”)</td>
</tr>
<tr>
<td>X</td>
<td><em>Cloacimonadida spp.</em></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td><em>Clostridia spp.</em></td>
<td>Saprophytic metabolism</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(decompose organic matter)</td>
</tr>
</tbody>
</table>

**Fish waste**
CONCLUSIONS

- 16S RNA datasets are **sparse** and **skewed**.
  - Sparsity → Filtering of variables mainly containing zero-values.
  - Skewness → Pareto-scaling

- CCA allows for the resolution of **underlying microbial communities** linked to the substrate preference

- **Phylogenetic representation** of the loadings simplifies their interpretation

- Microorganisms were grouped within communities on basis to:
  - The natural **habitat**
  - The availability of the **preferred substrate**
  - The presence of related microorganisms (**syntrophy** or **symbiosis**)
ACKNOWLEDGMENTS

Laurent Mazeas
Olivier Chapleur
Laëtitia Cardona

Douglas. N. Rutledge
Delphine Jouan-Rimbaud Bouveresse
Christophe B. Y. Cordella

F. Puig-Castellví
francesc.puigcastellvi@agroparistech.fr
5) Interpretation of Loadings

Variable selection using S plots.

S-plot CC1

\[ \text{CORR(CC1, X) vs. COV(CC1, X)} \]

828 variables \( \rightarrow \) 109 variables

S-plot CC2

\[ \text{CORR(CC2, X) vs. COV(CC2, X)} \]

828 variables \( \rightarrow \) 104 variables

161 unique variables

Vegetal waste subset
RESULTS AND DISCUSSION

5) Interpretation of Loadings

CC1

Positive loadings: more in sludge
Negative loadings: more in vegetal waste
WHITE: not selected for CC1

CC2

Vegetal waste subset

Positive loadings: more in sludge+vegetal
Negative loadings: more in sludge or vegetal
WHITE: not selected for CC2
## BIOLOGICAL INTERPRETATION

<table>
<thead>
<tr>
<th>Sludge</th>
<th>Microorganism</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td><em>Anaerolineae spp.</em></td>
<td>Habitat → sludge</td>
</tr>
<tr>
<td>X</td>
<td><em>Phycisphaerae spp.</em></td>
<td>Habitat → soil rich in S(^{-2}) and SO(_4)(^{-2})</td>
</tr>
<tr>
<td>X</td>
<td><em>Cl. syntrophomonadaceae</em></td>
<td>Degradation of organic acids</td>
</tr>
<tr>
<td>X</td>
<td><em>Cl. clostridiales 1</em></td>
<td>Highly polyphyletic (<em>diverse</em>) class</td>
</tr>
<tr>
<td>X</td>
<td><em>Alphaproteobacteria spp.</em></td>
<td><em>Acidolactic fermentation</em></td>
</tr>
<tr>
<td>X</td>
<td><em>Cl. christensenellaceae</em></td>
<td>Grows on a high variety of sugars</td>
</tr>
<tr>
<td>X</td>
<td><em>Bacilli spp.</em></td>
<td><em>Gut microbiota</em></td>
</tr>
<tr>
<td>X</td>
<td><em>Cl. clostridiales (Ruminococccaceae)</em></td>
<td><em>Cellulose decomposition</em></td>
</tr>
<tr>
<td>X</td>
<td><em>Cl. Christensenellaceae</em></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td><em>Methanomicrobia spp.</em></td>
<td>CH(_4) from CO(_2) (from sugar?)</td>
</tr>
<tr>
<td>X</td>
<td><em>Cl. clostridiales 1</em></td>
<td>Highly polyphyletic (<em>diverse</em>) class</td>
</tr>
</tbody>
</table>
RESULTS AND DISCUSSION

1) OTUs found in < 20% of the samples were excluded

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Number</th>
<th>Number</th>
<th>Number</th>
<th>Number</th>
<th>Number</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>domain</td>
<td>phylum</td>
<td>class</td>
<td>order</td>
<td>family</td>
<td>genus</td>
<td>species</td>
</tr>
<tr>
<td>Fish waste subset</td>
<td>10 x 484</td>
<td>2</td>
<td>25</td>
<td>40</td>
<td>65</td>
<td>112</td>
<td>215</td>
</tr>
<tr>
<td>Vegetal waste subset</td>
<td>10 x 828</td>
<td>2</td>
<td>29</td>
<td>54</td>
<td>101</td>
<td>177</td>
<td>342</td>
</tr>
</tbody>
</table>

- Phylogenetic differences exist at the order, family, genus, and species levels.
- Important similarity at the phylum and class level.
- *Classes* from species in the fish waste subset exist in the vegetal waste subset.
RESULTS AND DISCUSSION

5) Interpretation of Loadings

5.1 Variable selection using S plots.

484 variables $\rightarrow$ 59 variables

828 variables $\rightarrow$ 109 variables

132 unique variables

5.2 Interpretation of the selected variables using phylogenetic data
RESULTS AND DISCUSSION

5) Interpretation of Loadings

**CC1 FISH SUBSET**

More in sludge
More in fish

**CC1 VEGETAL SUBSET**

More in sludge
More in vegetal
RESULTS AND DISCUSSION

5) Interpretation of Loadings

CC1 FISH SUBSET

More in sludge
More in fish

CC1 VEGETAL SUBSET

More in sludge
More in vegetal
5) Interpretation of Loadings

**CC1 FISH SUBSET**

More in sludge
More in fish

**CC1 VEGETAL SUBSET**

More in sludge
More in vegetal
RESULTS AND DISCUSSION

5) Interpretation of Loadings

5.1 Variable selection using S plots.

484 variables $\rightarrow$ 42 variables

828 variables $\rightarrow$ 104 variables

129 unique variables

5.2 Interpretation of the selected variables using phylogenetic data
RESULTS AND DISCUSSION

5) Interpretation of Loadings

CC2 FISH SUBSET

More in sludge+fish
More in sludge or fish

CC2 VEGETAL SUBSET

More in sludge+grass
More in sludge or grass
RESULTS AND DISCUSSION

5) Interpretation of Loadings

**CC2 FISH SUBSET**

More in sludge + fish
More in sludge or fish

**CC2 VEGETAL SUBSET**

More in sludge + grass
More in sludge or grass
RESULTS AND DISCUSSION

4) Principal Component Analysis (PCA) Using 2 components on each subset

Fish waste subset

**PC1**

Scores

Samples

PC 1 Exp. Var.: 52.9149

**PC2**

Scores

Samples

PC 2 Exp. Var.: 17.405

Vegetal waste subset

**PC1**

Scores

Samples

PC 1 Exp. Var.: 34.5472

**PC2**

Scores

Samples

PC 2 Exp. Var.: 22.6285

Legend:
- Early time point
- Late time point

<table>
<thead>
<tr>
<th>%sludge</th>
<th>100</th>
<th>100</th>
<th>75</th>
<th>75</th>
<th>50</th>
<th>50</th>
<th>25</th>
<th>25</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>%fish</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>25</td>
<td>50</td>
<td>50</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>%vegetal</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>25</td>
<td>50</td>
<td>50</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
RESULTS AND DISCUSSION

4) Independent Component Analysis (ICA) Using 2 components on each subset

Fish waste subset

Vegetal waste subset

IC1

Scores

Scores

IC2

Samples

Samples

Early time point

Late time point

%sludge 100 100 75 75 50 50 25 25 0 0
%fish 0 0 25 25 50 50 25 25 100 100
%vegetal 0 0 25 25 50 50 25 25 100 100
RESULTS AND DISCUSSION

4) Common Component Analysis (CCA)

Using 4 components on global dataset
ComDim

\[ X_i = \frac{X_i - \bar{X_i}}{\|X_i\|} \]

\[ W = X.X^T \]

\[ \lambda_1 \cdot W_1 \]
\[ \lambda_2 \cdot W_2 \]

\[ W_G = U_w \cdot S_w \cdot V_w \]

\[ \text{PCA} \]

\[ \text{Aux} = I - q \cdot q^T \]

\[ X_1 = \text{Aux} \cdot X_1 \]

\[ X_2 = \text{Aux} \cdot X_2 \]

\[ \text{Dif} = (W_1 - \lambda_1 \cdot q \cdot q^T) + (W_2 - \lambda_2 \cdot q \cdot q^T) \]

\[ \text{Dif}_n^2 - \text{Dif}_{n-1}^2 < \text{limit} \]

\[ \lambda_1 = \frac{q^T \cdot W_1 \cdot q}{\|q\|^2} \]

\[ \lambda_2 = \frac{q^T \cdot W_2 \cdot q}{\|q\|^2} \]
Outline

Introduction
- Microbial communities & anaerobic digesters
- 16S rRNA sequencing

Objectives

Results and Discussions

Conclusions