Motivation

A new approach to microbial identification

Goals

Genotypic feature based taxonomy
Visualization

Methodologies

Self organizing Topographic map
Deterministic annealing

Results

A methodology to create a visualization and classification tool

Conclusions
Motivation

A new approach to microbial identification

Goals

Genotypic feature based taxonomy
Visualization

Methodologies

Self organizing
Topographic map
Deterministic annealing

Results

A methodology to create a visualization and classification tool

Conclusions
Motivation

- Microbial identification is crucial for the study of infectious diseases.
- Bacterial taxonomy is usually based on phenotypic characters.
- A new approach based on bacteria genotype is under development.
- 16S rRNA “housekeeping” gene for taxonomic purposes.
Motivation

A new approach to microbial identification

Goals

Genotypic feature based taxonomy
Visualization

Methodologies

Self organizing
Topographic map
Deterministic annealing

Results

A methodology to create a visualization and classification tool

Conclusions
Goal

- Genotypic features based taxonomy
- Topographic representation of the bacteria clusters
  - Finding misclassification = discovery of new pathogens
  - Classifying organisms with an unusual phenotype
Motivation
A new approach to microbial identification

Goals
Genotypic feature based taxonomy
Visualization

Methodologies
Self organizing Topographic map
Deterministic annealing

Results
A methodology to create a visualization and classification tool

Conclusions
Methodologies

- General framework
- Building Dataset
- Sequence Alignment
- Evolutionary Distance
- Soft Topographic Map Algorithm
General framework

1. downloading and filtering gene sequences from NCBI databases
2. sequence alignment (Needleman-Wunsch)
3. computing dissimilarity matrix (evolutionary distance)
4. clustering (SOM on pairwise distances) and visualization (UMatrix style map)

Diagram:
- Sequence DB
- Results Filtering
- Taxonomy Retrieval
- Labeling
- Sequence Alignment
- Computing Distance
- Clustering and mapping
Phylum BXII (Proteobacteria)  
Class III (Gammaproteobacteria) 

<table>
<thead>
<tr>
<th>Order Name</th>
<th>Number of Families</th>
<th>Number of Type Strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromatiales</td>
<td>3 Families</td>
<td>25 Type Strains</td>
</tr>
<tr>
<td>Xanthomonodales</td>
<td>1 Family</td>
<td>11 Type Strains</td>
</tr>
<tr>
<td>Thiotrichales</td>
<td>3 Families</td>
<td>11 Type Strains</td>
</tr>
<tr>
<td>Methylcoccales</td>
<td>1 Families</td>
<td>7 Type Strains</td>
</tr>
<tr>
<td>Pseudomonadales</td>
<td>2 Families</td>
<td>7 Type Strains</td>
</tr>
<tr>
<td>Vibrionales</td>
<td>1 Family</td>
<td>3 Type Strains</td>
</tr>
<tr>
<td>Enterobacteriales</td>
<td>1 Family</td>
<td>39 Type Strains</td>
</tr>
<tr>
<td>Acidithobacillales</td>
<td>2 Families</td>
<td>2 Type Strains</td>
</tr>
<tr>
<td>Cardiobacterales</td>
<td>1 Family</td>
<td>3 Type Strains</td>
</tr>
<tr>
<td>Legionellaes</td>
<td>2 Families</td>
<td>2 Type Strains</td>
</tr>
<tr>
<td>Oceanospirillales</td>
<td>4 Families</td>
<td>11 Type Strains</td>
</tr>
<tr>
<td>Alteromonadales</td>
<td>1 Family</td>
<td>13 Type Strains</td>
</tr>
<tr>
<td>Aeromonadales</td>
<td>2 Families</td>
<td>7 Type Strains</td>
</tr>
<tr>
<td>Pasteurellales</td>
<td>1 Families</td>
<td>6 Type Strains</td>
</tr>
</tbody>
</table>

147 16S rRNA gene sequences downloaded from GenBank database.
Sequence Alignment

- Sequence alignment allows to compare homologous sites of the same gene between two different species
- Two well known alignment algorithms used:
  - ClustalW: multiple-alignment
  - Needleman-Wunsch: pairwise alignment
Evolutionary Distance

- The simplest type of distance is the number of nucleotide substitutions per site.
  - Warning: it underrates real distances
- Jukes and Cantor method was used: it provides a better estimate of evolutionary distances
- Evolutionary distances are elements of the symmetric dissimilarity matrix:

<table>
<thead>
<tr>
<th>Type strain</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06286</td>
<td>0.11215</td>
<td>0.06482</td>
<td>0.05128</td>
<td>0.09451</td>
<td>0.06785</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.10608</td>
<td>0.0579</td>
<td>0.065</td>
<td>0.07196</td>
<td>0.04682</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.1224</td>
<td>0.11418</td>
<td>0.10279</td>
<td>0.11538</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.06082</td>
<td>0.10224</td>
<td>0.06764</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.10595</td>
<td>0.07362</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.08232</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Soft Topographic Map Algorithm (1)

- Extension of Kohonen's SOM for pairwise data
- The position of bacteria clusters in the topographic maps is based on the optimization, through deterministic annealing technique, of a cost function that takes its minimum when each data point is mapped to the best matching neuron

Dissimilarity matrix

<table>
<thead>
<tr>
<th>Type strain</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.06286</td>
<td>0.11215</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.10608</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Topographic map showing relationships among Bacteria clusters
Soft Topographic Map Algorithm (2)

1) Initialization Step:
   a) put $e_{tr} \leftarrow \eta_{tr}, \forall \ t, r, \in [0, 1]$
   b) compute lookup table for $h_{rs}$
   c) choose initial value of $\beta$, $\beta_{final}$, increasing temperature factor $\eta$, threshold $\epsilon$

2) Training Step:
   a) while $\beta < \beta_{final}$ (Annealing cycle)
      i. repeat (EM cycle)
         A) E step: compute $P(x_t \in C_r) \ \forall \ t, r$
         B) M step: compute $a_{tr}^{new}, \forall \ t, r$
         C) M step: compute $e_{tr}^{new}, \forall \ t, r$
      ii. until $\|e_{tr}^{new} - e_{tr}^{old}\| < \epsilon$
      iii. put $\beta \leftarrow \eta \beta$
   b) end while

Neighborhood function:
$$h_{rs} = \exp \left( -\frac{|r-s|^2}{2 \sigma^2} \right), \forall \ r, s$$

Assignment probability:
$$P(x_t \in C_r) = \frac{\exp(-\beta e_{tr})}{\sum_u \exp(-\beta e_{tu})}, \forall \ t, r$$

Partial assignment cost:
$$e_{tr} = \sum_s h_{rs} \sum_t a_{t's} \left( d_{tt'} - \frac{1}{2} \sum_{t''} a_{t''s} d_{t't''} \right), \forall \ t, r$$

Weighting factor:
$$a_{tr} = \frac{\sum_s h_{rs} P(x_t \in C_s)}{\sum_t \sum_s h_{rs} P(x_t \in C_s)}, \forall \ t, r$$
Soft Topographic map

The topographic map is a lattice (two dimensional in our case) that self organize in the pattern space.
Soft Topographic Map Algorithm (3)

- The algorithm that “moves” this lattice is called deterministic annealing
  - The advantage of deterministic annealing is to find a global minimum of the approximation error

\[ e_{tr} = \sum_{s} h_{rs} \sum_{t} a_{tr}s \left( d_{tt} - \frac{1}{2} \sum_{i} a_{tr}d_{i}\right), \forall t, r \]

Neighborhood function:
\[ h_{rs} = \exp\left(-\frac{|r-s|^2}{2\sigma^2}\right), \forall r, s \]

Assignment probability:
\[ P(x_{t} \in C_{r}) = \frac{\exp(-\beta e_{tr})}{\sum_{u} \exp(-\beta e_{tu})}, \forall t, r \]

Partial assignment cost:
\[ e_{tr} = \sum_{s} h_{rs} \sum_{i} a_{r}d_{i}\left( d_{tt} - \frac{1}{2} \sum_{i} a_{tr}s \right), \forall t, r \]

Weighting factor:
\[ a_{tr} = \frac{\sum_{s} h_{rs} P(x_{t} \in C_{r})}{\sum_{r} \sum_{s} h_{rs} P(x_{t} \in C_{r})}, \forall t, r \]
Motivation
A new approach to microbial identification

Goals
Genotypic feature based taxonomy
Visualization

Methodologies
Self organizing
Topographic map
Deterministic annealing

Results
A methodology to create a visualization and classification tool

Conclusions
Experimental Results

- From 8x8 up to 45x45 map dimensions
- We trained 20 maps of each geometry in order to avoid the dependence from the initial conditions
- The results obtained using the two alignments methods do not present any significant difference
Experimental tests

- Hardware resources
  - 16 nodes cluster, dual processor Xeon 3.4 GHz, 4 GB RAM, 6 TB storage, Myrinet-Fiber communication

- Software
  - Languages: Java, Python
  - Libraries: BioJava, Jama ....

<table>
<thead>
<tr>
<th>Map size</th>
<th>Average processing time (min.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8x8</td>
<td>0.6</td>
</tr>
<tr>
<td>9x9</td>
<td>1</td>
</tr>
<tr>
<td>10x10</td>
<td>2</td>
</tr>
<tr>
<td>11x11</td>
<td>2</td>
</tr>
<tr>
<td>12x12</td>
<td>4</td>
</tr>
<tr>
<td>13x13</td>
<td>4</td>
</tr>
<tr>
<td>14x14</td>
<td>6</td>
</tr>
<tr>
<td>15x15</td>
<td>9</td>
</tr>
<tr>
<td>16x16</td>
<td>12</td>
</tr>
<tr>
<td>17x17</td>
<td>16</td>
</tr>
<tr>
<td>18x18</td>
<td>17</td>
</tr>
<tr>
<td>19x19</td>
<td>23</td>
</tr>
<tr>
<td>20x20</td>
<td>30</td>
</tr>
<tr>
<td>25x25</td>
<td>90</td>
</tr>
<tr>
<td>30x30</td>
<td>240</td>
</tr>
<tr>
<td>35x35</td>
<td>360</td>
</tr>
<tr>
<td>40x40</td>
<td>660</td>
</tr>
<tr>
<td>45x45</td>
<td>1380</td>
</tr>
</tbody>
</table>
Map Evaluation

Mixed Clusters
We only have distances between patterns, and no metrics!

Usually topology measures are considered, but in our case there is not a space that contains the patterns (sequences).

Map Evaluation

We only have distances between patterns, and no metrics!
Map evaluation

- We take rows and columns of the maps and compare the order of the elements in map with the order obtained from the dissimilarity matrix.
Map evaluation

This sequence...

...is compared with...

The sequence of the same objects obtained from the dissimilarity matrix

<table>
<thead>
<tr>
<th>Type strain</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.06286</td>
<td>0.11215</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.10608</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

This comparison is made using the Spearman coefficient in order to obtain a similarity value among the two sequences.

Of course the two sequences should be the same in a good map.
Map evaluation

- We have an index for each map and we can see that some geometry are better than others.
The "Best" Map

<table>
<thead>
<tr>
<th>Order Name</th>
<th>Number of Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromatiales</td>
<td>3 Families</td>
</tr>
<tr>
<td>Xanthomonodales</td>
<td>1 Family</td>
</tr>
<tr>
<td>Thiotrichiales</td>
<td>3 Families</td>
</tr>
<tr>
<td>Methylcoccales</td>
<td>1 Families</td>
</tr>
<tr>
<td>Pseudomonadales</td>
<td>2 Families</td>
</tr>
<tr>
<td>Vibrionales</td>
<td>1 Family</td>
</tr>
<tr>
<td>Enterobacteriales</td>
<td>1 Family</td>
</tr>
<tr>
<td>Acidithiobacillales</td>
<td>2 Families</td>
</tr>
<tr>
<td>Cardiobacteriales</td>
<td>1 Family</td>
</tr>
<tr>
<td>Legionellales</td>
<td>2 Families</td>
</tr>
<tr>
<td>Oceanospirillales</td>
<td>4 Families</td>
</tr>
<tr>
<td>Alteromonadiales</td>
<td>1 Family</td>
</tr>
<tr>
<td>Aeromonadiales</td>
<td>2 Families</td>
</tr>
<tr>
<td>Pasteurellales</td>
<td>1 Families</td>
</tr>
</tbody>
</table>
Comparison with the phylogenetic tree
Motivation

A new approach to microbial identification

Goals

Genotypic feature based taxonomy
Visualization

Methodologies

Self organizing Topographic map
Deterministic annealing

Results

A methodology to create a visualization and classification tool

Conclusions
Conclusions

- Soft Topographic Map for clustering and classification of bacteria
- Genotype based taxonomy
- Detecting singular situations
- Further analysis with other housekeeping genes or using other distance algorithms, e.g. Normalized Compressed Distance