Machine Learning Techniques for Bacteria Classification

Massimo La Rosa
Riccardo Rizzo
Alfonso M. Urso
S. Gaglio

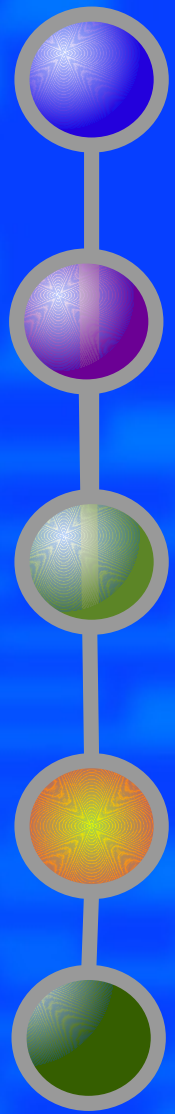
ICAR-CNR

University of Palermo

Workshop on
Hardware Architectures Beyond 2020:
Challenges and Opportunities for Computational Biology and Bioinformatics

Napoli – December 19, 2007

Outline



Motivation

A new approach to microbial identification

Goals

Genotypic feature based taxonomy
Visualization

Methodologies

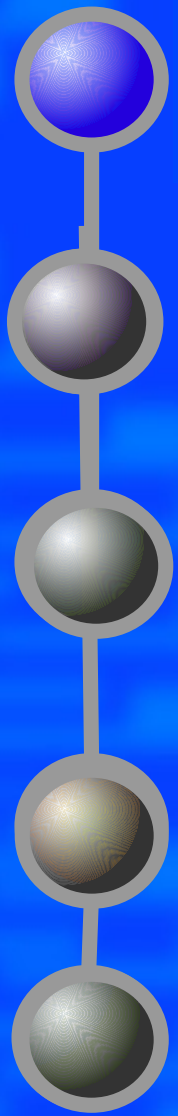
- Self organizing Topographic map
- Deterministic annealing

Results

A methodology to create a visualization and classification tool

Conclusions

Outline



Motivation

A new approach to microbial identification

Goals

Genotypic feature based taxonomy
Visualization

Methodologies

- Self organizing Topographic map
- Deterministic annealing

Results

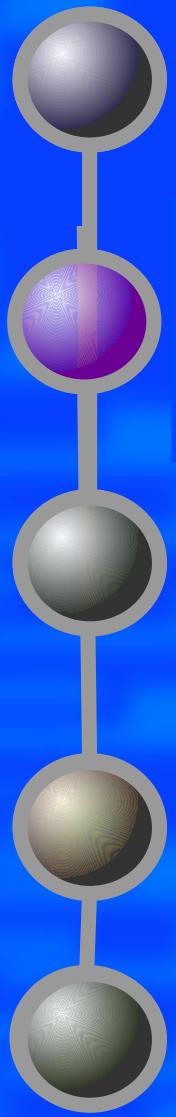
A methodology to create a visualization and classification tool

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

Outline



Motivation

A new approach to microbial identification

Goals

Genotypic feature based taxonomy
Visualization

Methodologies

- Self organizing Topographic map
- Deterministic annealing

Results

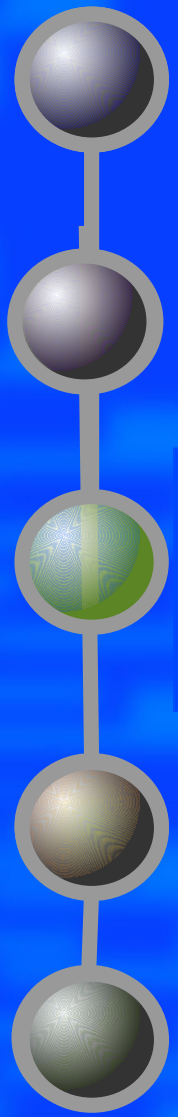
A methodology to create a visualization and classification tool

Conclusions

Goal

- Genotypic features based taxonomy
- Topographic representation of the bacteria clusters
  - Finding misclassification = discovery of new pathogens
  - Classifying organisms with an unusual phenotype

Outline



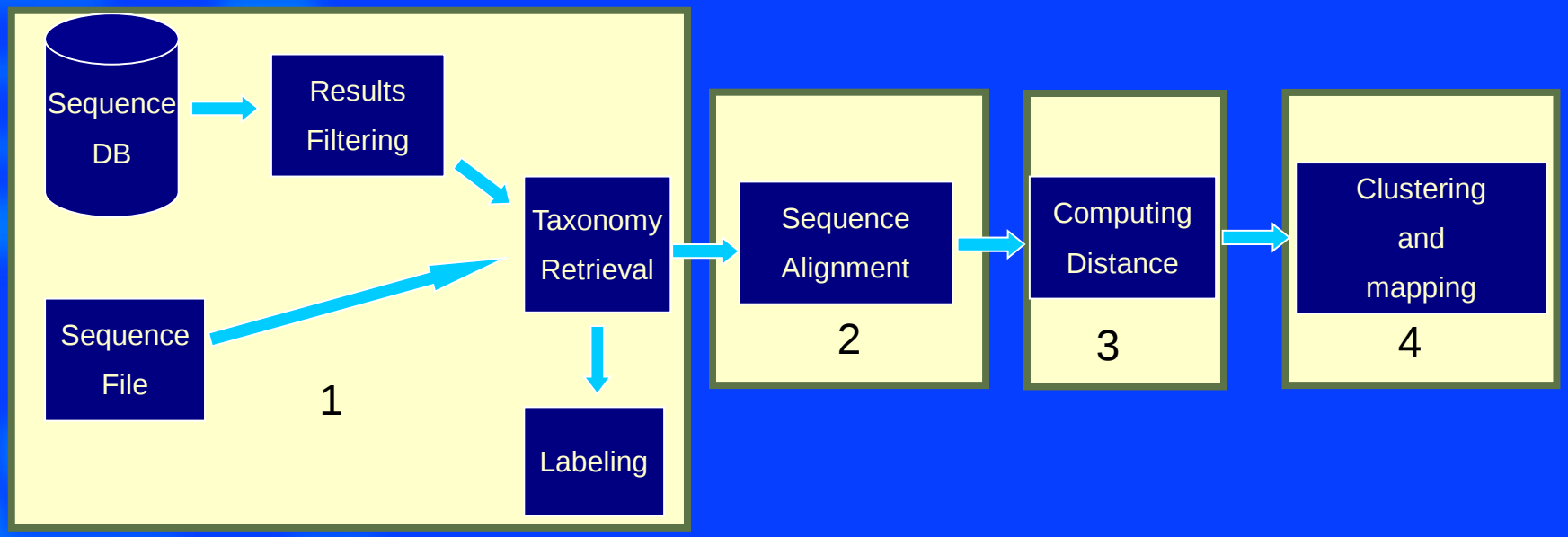
- Motivation
  - A new approach to microbial identification
- Goals
  - Genotypic feature based taxonomy
  - Visualization
- Methodologies
  - Self organizing Topographic map
  - Deterministic annealing
- Results
  - A methodology to create a visualization and classification tool
- Conclusions

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)



Building Dataset

14 Orders

Phylum BXII (Proteobacteria)
Class III (Gammaproteobacteria)

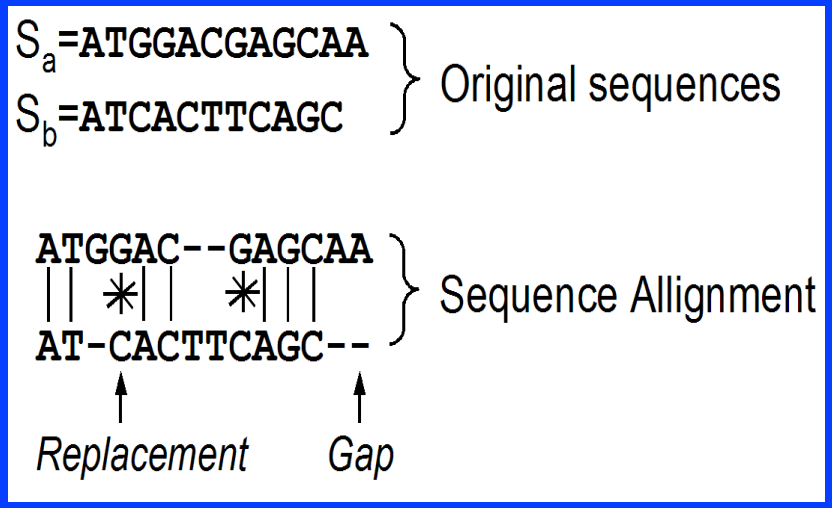
147 16S rRNA gene sequences downloaded from GenBank database

Gammaproteobacteria

Order Name	Number of Families	Number of Type Strains
Chromatiales	3 Families	25 Type Strains
Xanthomonodales	1 Family	11 Type Strains
Thiotrichales	3 Families	11 Type Strains
Methylococcales	1 Families	7 Type Strains
Pseudomonadales	2 Families	7 Type Strains
Vibrionales	1 Family	3 Type Strains
Enterobacteriales	1 Family	39 Type Strains
Acidithobacillales	2 Families	2 Type Strains
Cardiobacteriales	1 Family	3 Type Strains
Legionellales	2 Families	2 Type Strains
Oceanospirillales	4 Families	11 Type Strains
Alteromonadales	1 Family	13 Type Strains
Aeromonadales	2 Families	7 Type Strains
Pasteurellales	1 Families	6 Type Strains
		147 Type Strains

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:

Type strain	1	2	3	4	5	6	7
1	0	0.06286	0.11215	0.06482	0.05128	0.09451	0.06785
2		0	0.10608	0.0579	0.065	0.07196	0.04682
3			0	0.1224	0.11418	0.10279	0.11538
4				0	0.06082	0.10224	0.06764
5					0	0.10595	0.07362
6						0	0.08232
7							0
...							

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

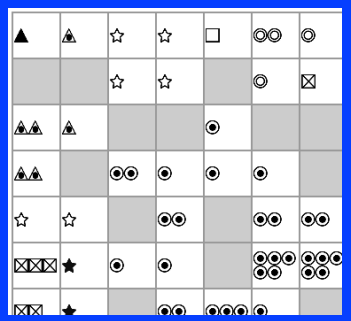
Type strain	1	2	3
1	0	0,06286	0,11215
2		0	0,10608
3			0

input

Soft Topographic Map Algorithm

output

Topographic map showing relationships among Bacteria clusters



Soft Topographic Map Algorithm (2)

1) Initialization Step:
   a) put $e_{tr} \leftarrow n_{tr}, \forall t, \boldsymbol{r}, \in [0,1]$
   b) compute lookup table for h_{rs}
   c) choose initial value of β, β_{final}, increasing temperature factor η, threshold ϵ
2) Training Step:
   a) while $\beta < \beta_{final}$ (Annealing cycle)
      i. repeat (EM cycle)
         A) E step: compute $P(\boldsymbol{x}_t \in \boldsymbol{C}_r) \forall t, \boldsymbol{r}$
         B) M step: compute $a_{tr}^{new}, \forall t, \boldsymbol{r}$
         C) M step: compute $e_{tr}^{new}, \forall t, \boldsymbol{r}$
      ii. until $\|e_{tr}^{new} - e_{tr}^{old}\| < \epsilon$
      iii. put $\beta \leftarrow \eta \beta$
   b) end while

Neighborhood function:

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

Assignment probability:

$$P(\boldsymbol{x}_t \in \boldsymbol{C}_r) = \frac{\exp(-\beta e_{tr})}{\sum_{\boldsymbol{u}} \exp(-\beta e_{t\boldsymbol{u}})}, \forall t, \boldsymbol{r}$$

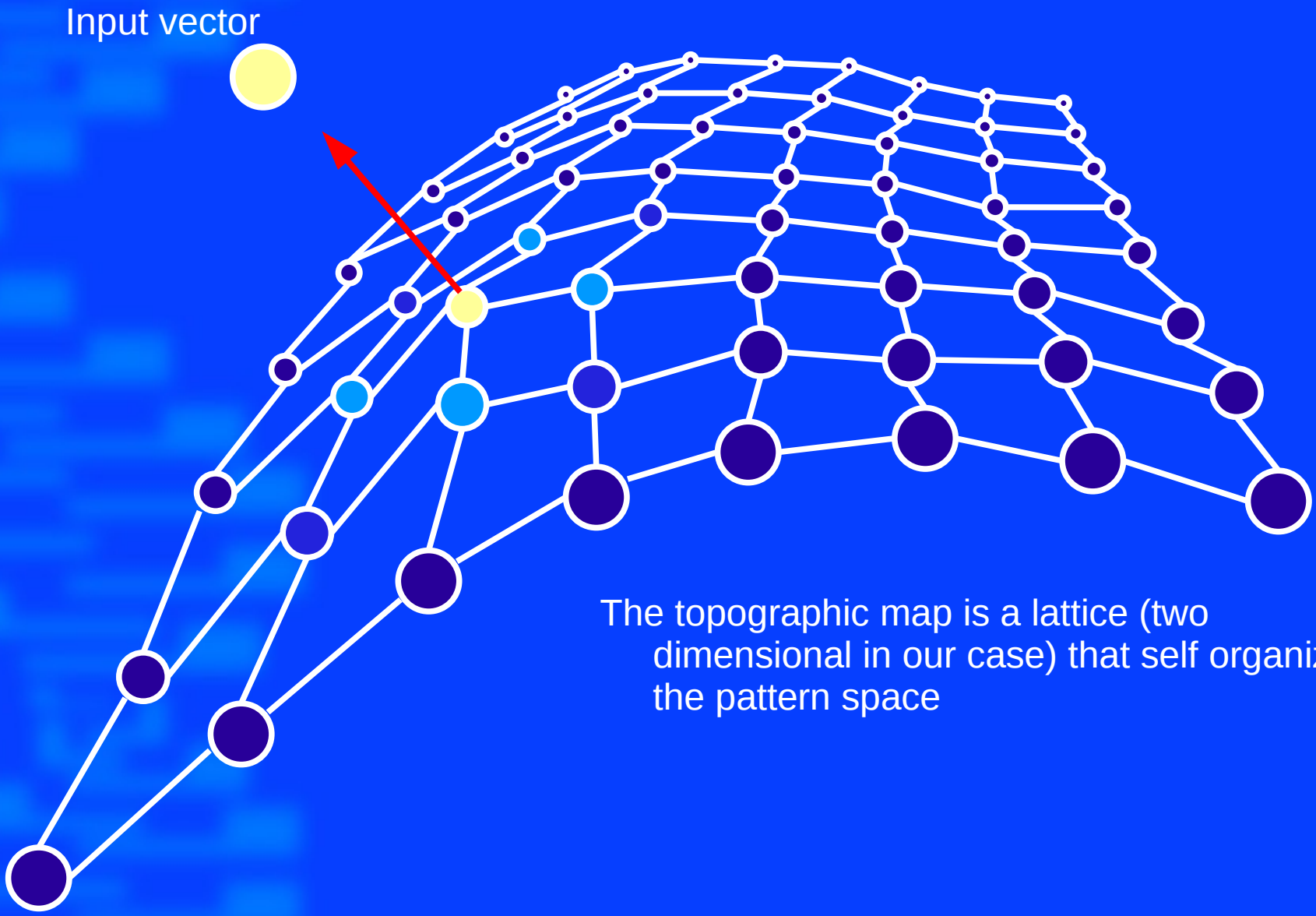
Partial assignment cost:

$$e_{tr} = \sum_{\boldsymbol{s}} h_{rs} \sum_{t'} a_{t's} \left(d_{tt'} - \frac{1}{2} \sum_{t''} a_{t''s} d_{t't''} \right), \forall t, \boldsymbol{r}$$

Weighting factor:

$$a_{tr} = \frac{\sum_{\boldsymbol{s}} h_{rs} P(\boldsymbol{x}_t \in \boldsymbol{C}_s)}{\sum_{t'} \sum_{\boldsymbol{s}} h_{rs} P(\boldsymbol{x}_{t'} \in \boldsymbol{C}_s)}, \forall t, \boldsymbol{r}$$

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

Deterministic annealing

1) Initialization Step:
   a) put $e_{tr} \leftarrow n_{tr}, \forall t, \boldsymbol{r}, \in [0,1]$
   b) compute lookup table for h_{rs}
   c) choose initial value of β, β_{final}, increasing temperature factor η, threshold ϵ
2) Training Step:
   a) while $\beta < \beta_{final}$ (Annealing cycle)
      i. repeat (EM cycle)
         A) E step: compute $P(\boldsymbol{x}_t \in \boldsymbol{C}_r) \forall t, \boldsymbol{r}$
         B) M step: compute $a_{tr}^{new}, \forall t, \boldsymbol{r}$
         C) M step: compute $e_{tr}^{new}, \forall t, \boldsymbol{r}$
      ii. until $\|e_{tr}^{new} - e_{tr}^{old}\| < \epsilon$
      iii. put $\beta \leftarrow \eta \beta$
   b) end while

Neighborhood function:

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

Assignment probability:

$$P(\boldsymbol{x}_t \in \boldsymbol{C}_r) = \frac{\exp(-\beta e_{tr})}{\sum_{\boldsymbol{u}} \exp(-\beta e_{t\boldsymbol{u}})}, \forall t, \boldsymbol{r}$$

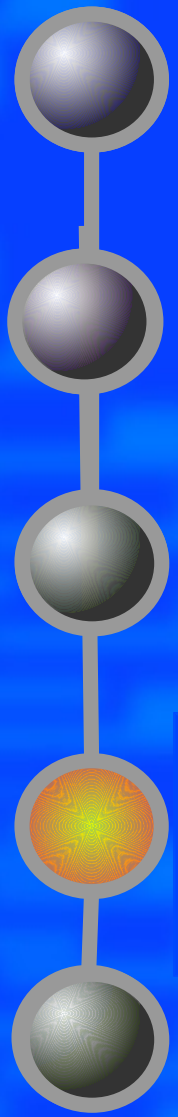
Partial assignment cost:

$$e_{tr} = \sum_{s} h_{rs} \sum_{t'} a_{t's} \left(d_{tt'} - \frac{1}{2} \sum_{t''} a_{t''s} d_{t't''} \right), \forall t, \boldsymbol{r}$$

Weighting factor:

$$a_{tr} = \frac{\sum_{s} h_{rs} P(\boldsymbol{x}_t \in \boldsymbol{C}_s)}{\sum_{t'} \sum_{s} h_{rs} P(\boldsymbol{x}_{t'} \in \boldsymbol{C}_s)}, \forall t, \boldsymbol{r}$$

Outline



Motivation

A new approach to microbial identification

Goals

Genotypic feature based taxonomy
Visualization

Methodologies

- Self organizing Topographic map
- Deterministic annealing

Results

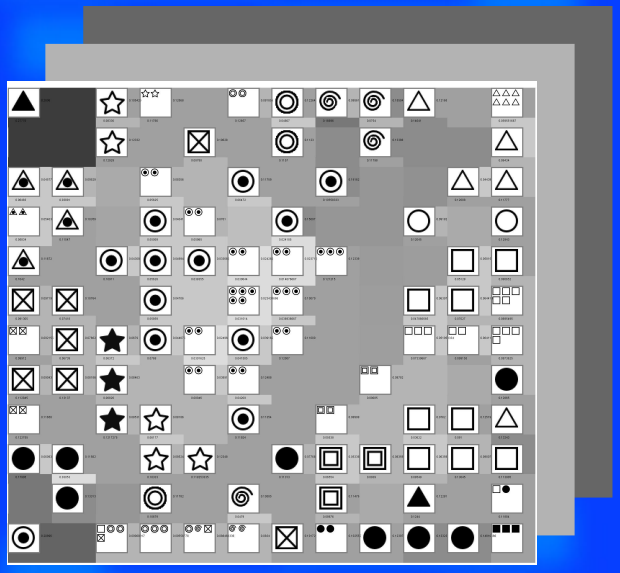
A methodology to create a visualization and classification tool

Conclusions

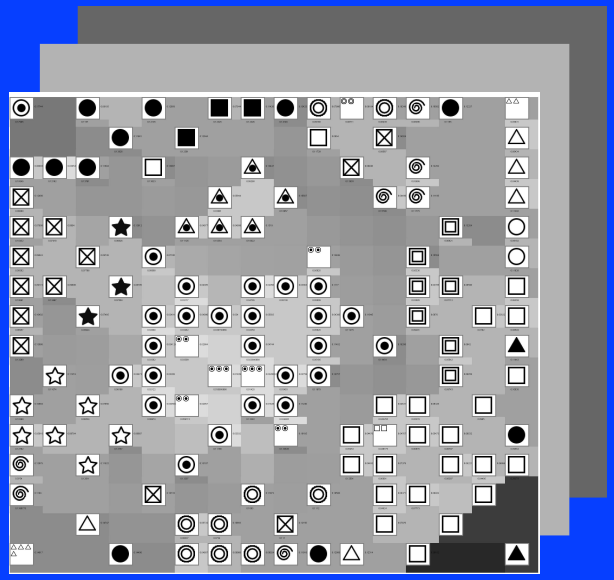
Experimental Results

- 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

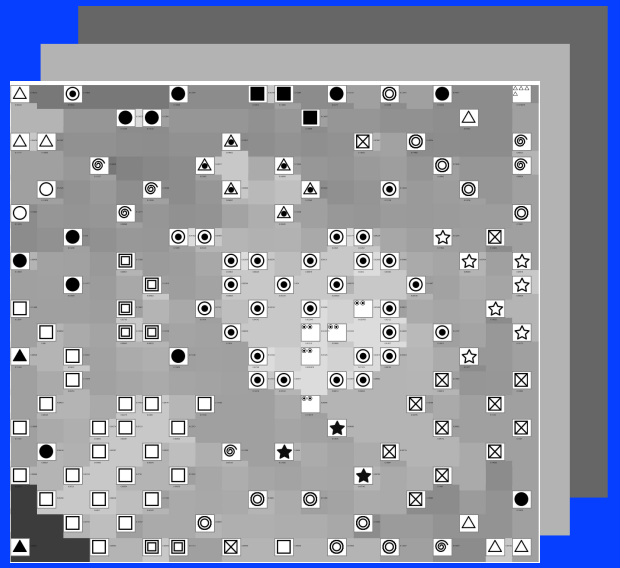
12x12 map



16x16 map



20x20 map

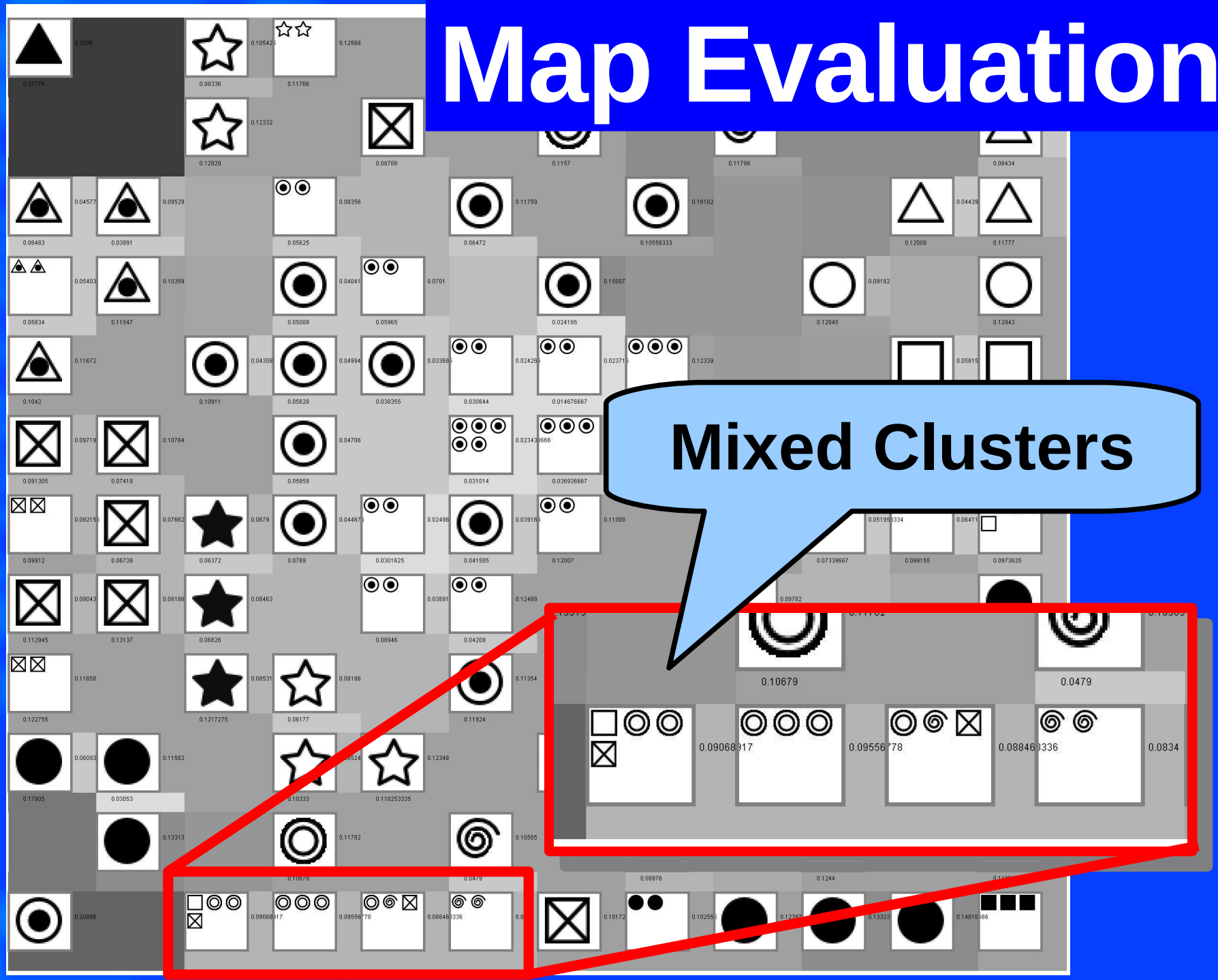


Experimental tests

- 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 ....

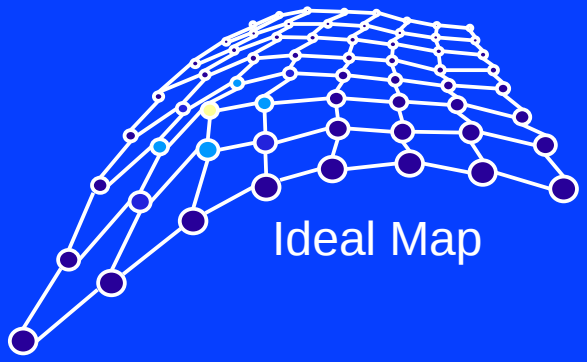
Map size	Average processing time (min.)
8x8	0,6
9x9	1
10x10	2
11x11	2
12x12	4
13x13	4
14x14	6
15x15	9
16x16	12
17x17	16
18x18	17
19x19	23
20x20	30
25x25	90
30x30	240
35x35	360
40x40	660
45x45	1380

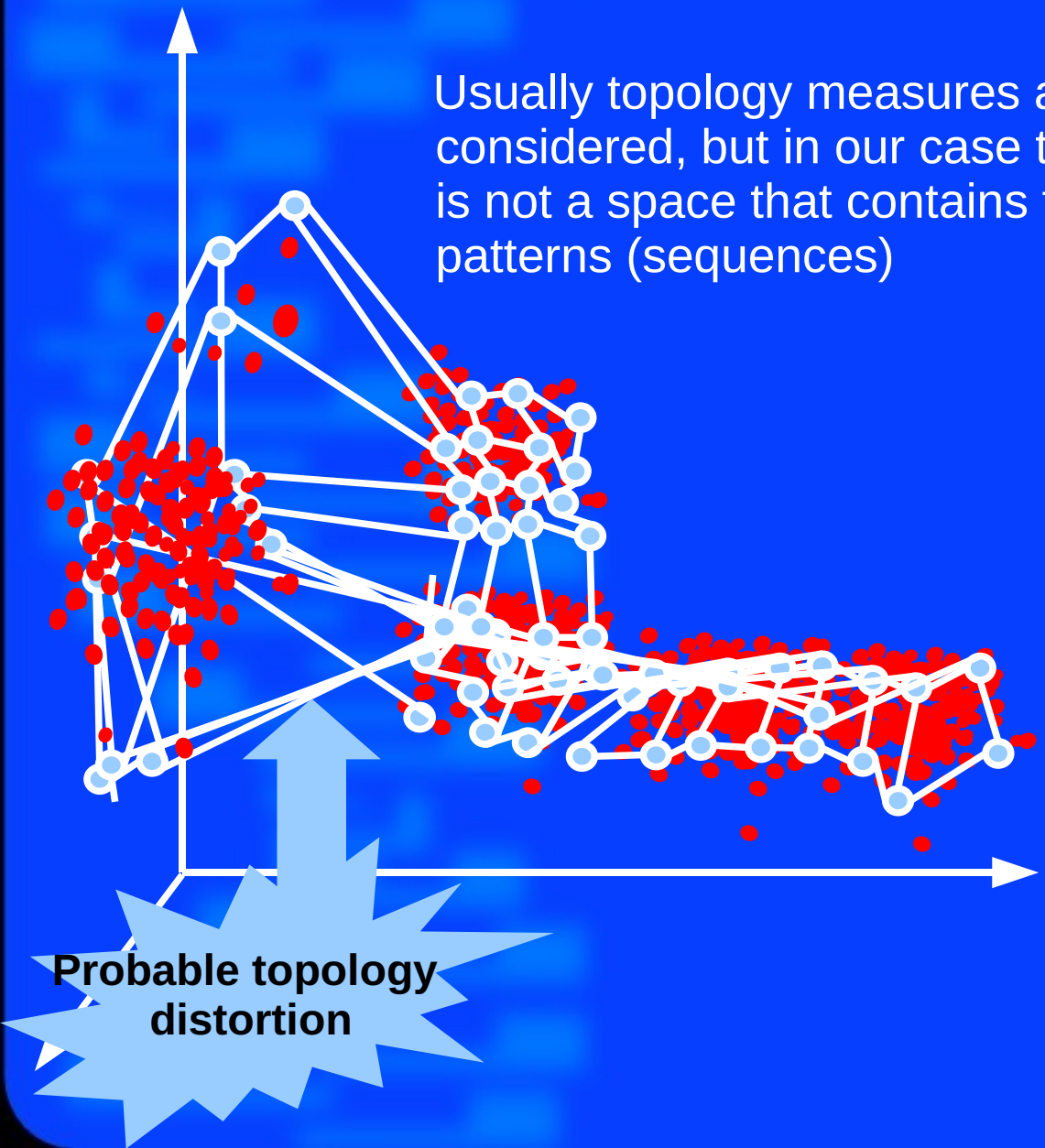
Map Evaluation



Map Evaluation

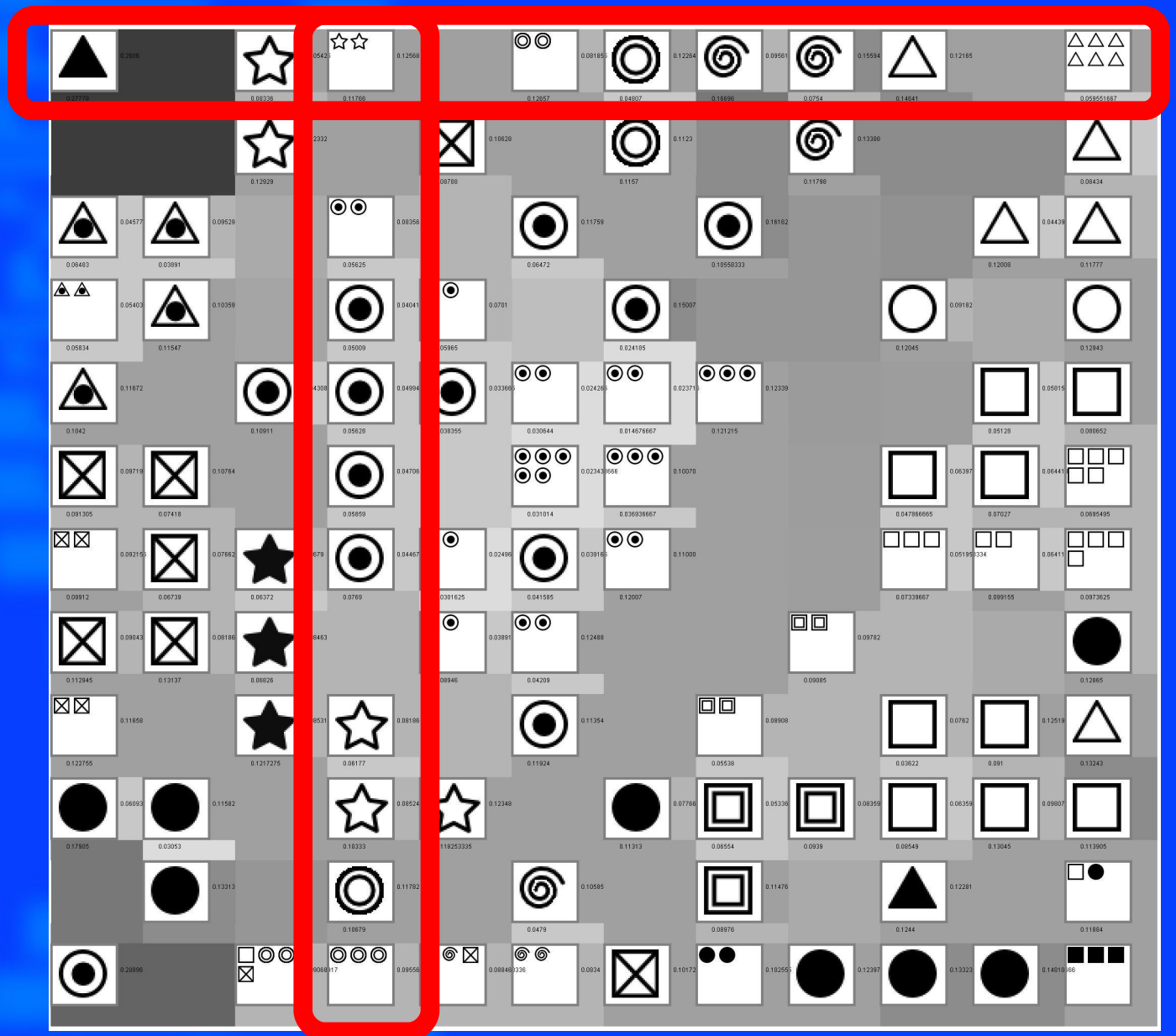
Usually topology measures are considered, but in our case there is not a space that contains the patterns (sequences)





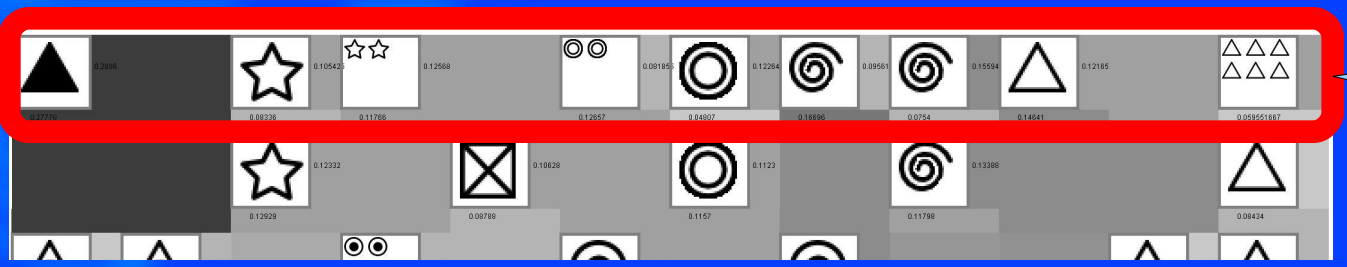
We only have distances between patterns, and no metrics!

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...

Dissimilarity matrix

Type strain	1	2	3
1	0	0.06286	0.11215
2		0	0.10608
3			0

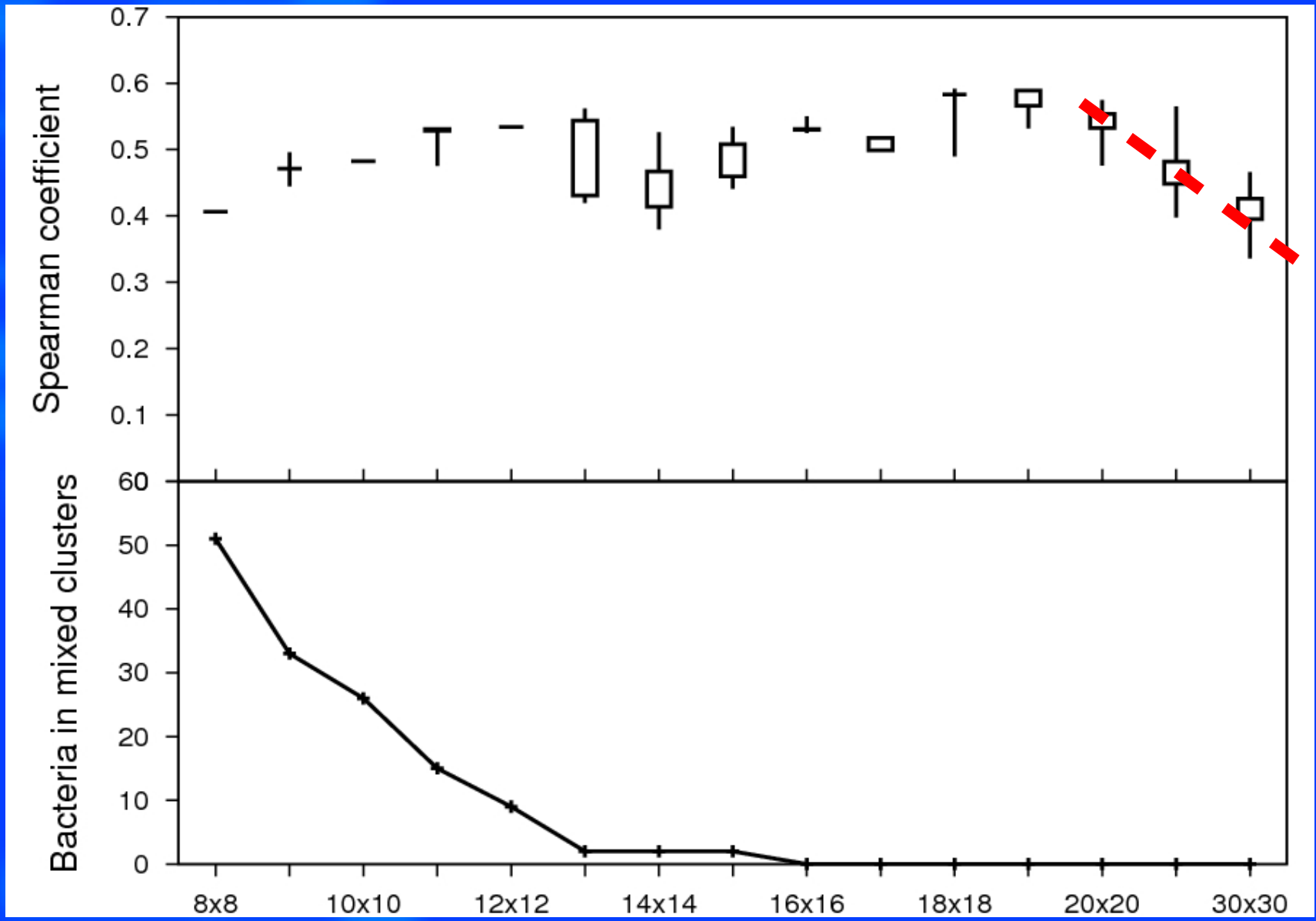
...the sequence of the same objects obtained from the dissimilarity matrix

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

.

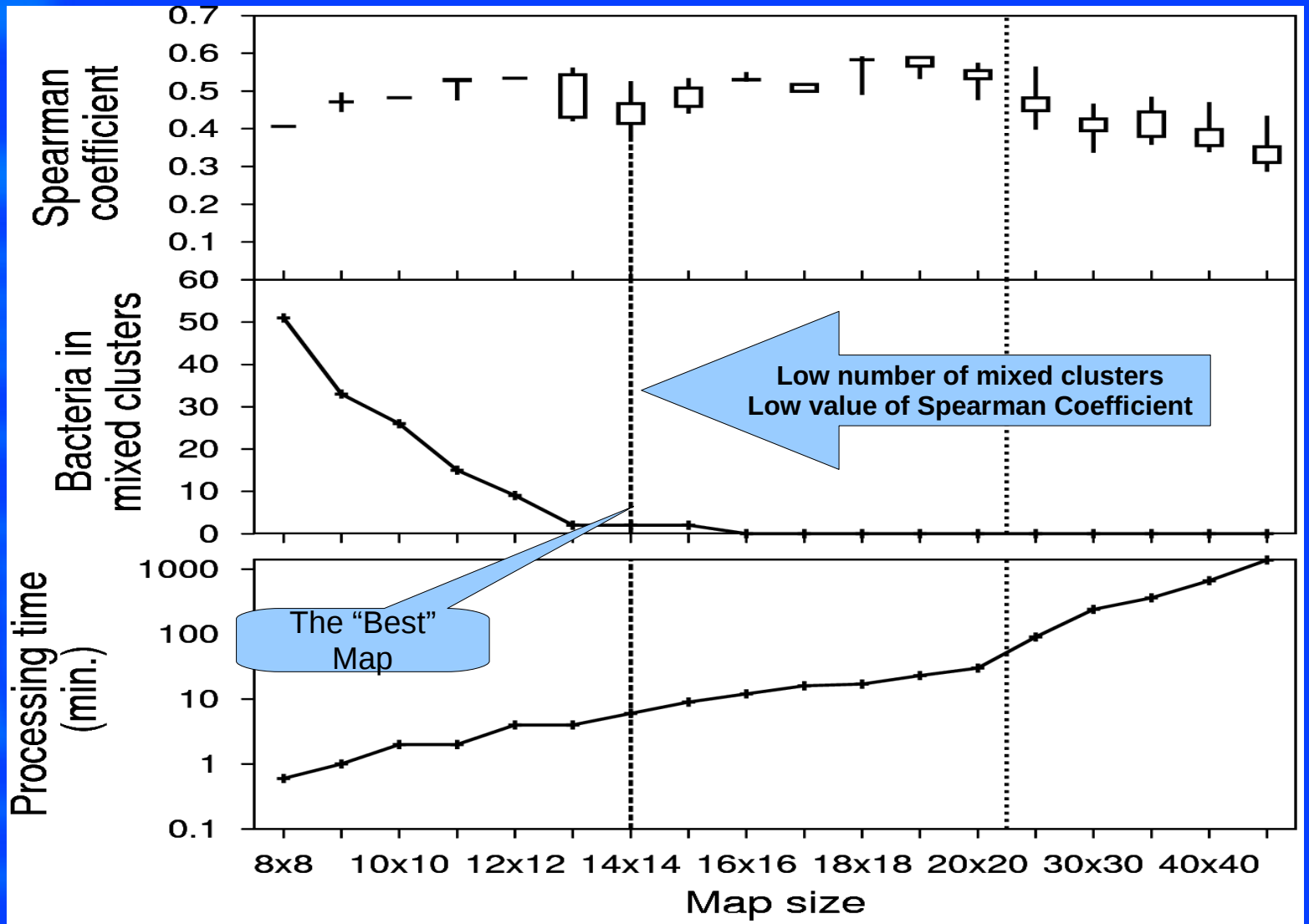
Of course the two sequences should be the same in a good map

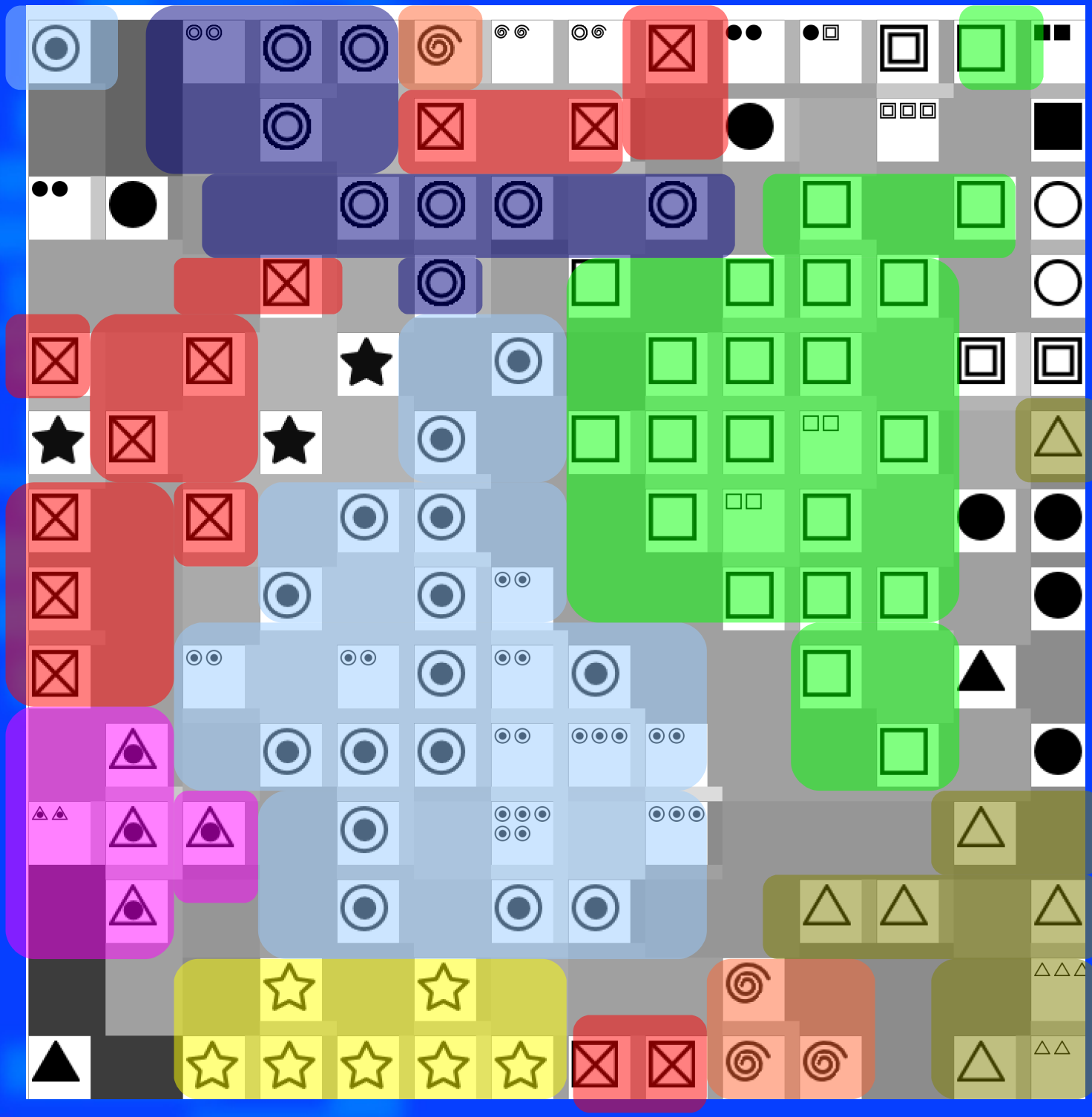
Map Evaluation



Map evaluation

- We have an index for each map and we can see that some geometry are better than other

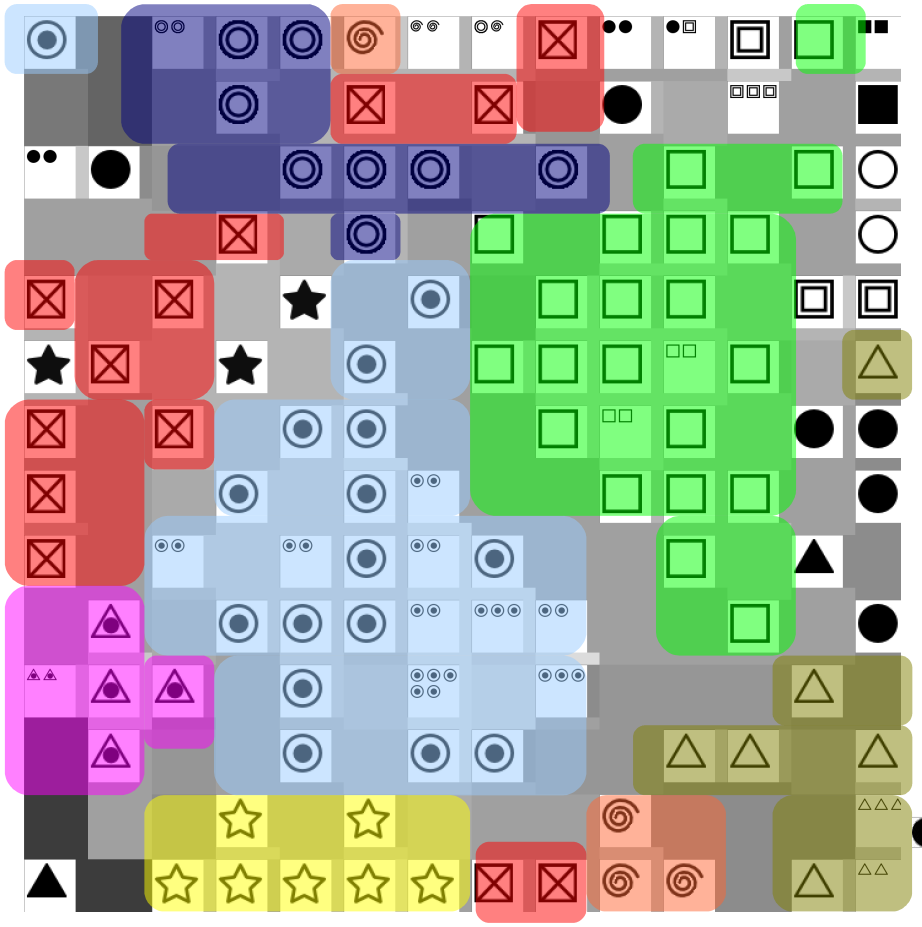


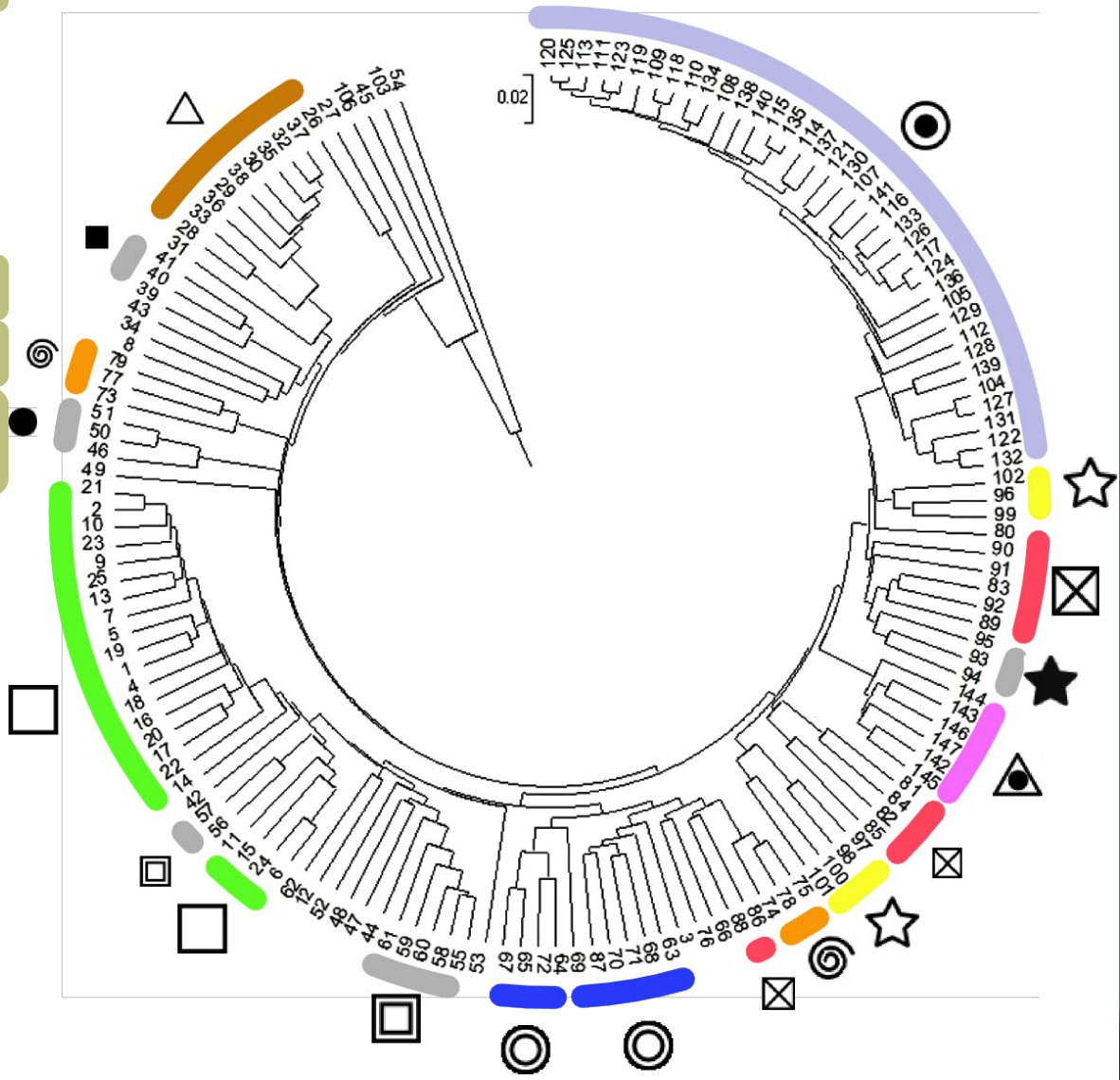


The "Best" Map

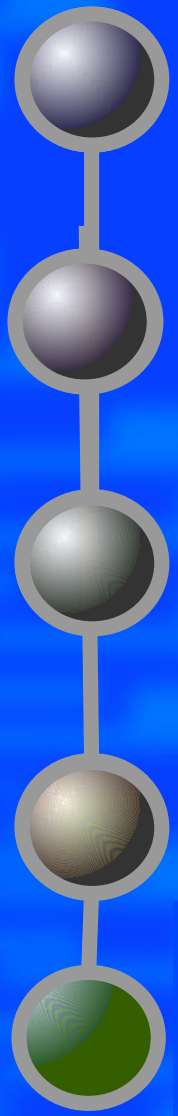
	Order Name	Number of Families
□	Chromatiales	3 Families
△	Xanthomonodales	1 Family
●	Thiotrichales	3 Families
▣	Methylococcales	1 Families
◎	Pseudomonadales	2 Families
★	Vibrionales	1 Family
◉	Enterobacteriales	1 Family
○	Acidithobacillales	2 Families
■	Cardiobacteriales	1 Family
▲	Legionellales	2 Families
◎	Oceanospirillales	4 Families
⊠	Alteromonadales	1 Family
☆	Aeromonadales	2 Families
◬	Pasteurellales	1 Families

Comparison with the phylogenetic tree





Outline



- 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