

# Introduction to cluster analysis purpose and parameters and Global Database Management

Benjamin FELIX

Joint training course on the use of BioNumerics software to analyse PFGE data

Rome 3-4 July 2017

# Clustering

The screenshot shows the ER012 (Comparison) software interface. The main window displays a dendrogram for 'PFGE-Apal' with a similarity scale from 85 to 100. A red circle highlights the 'Clustering' icon in the top toolbar, and a red arrow points to the 'Comparison settings' dialog box.

The 'Comparison settings' dialog box is open, showing the following options:

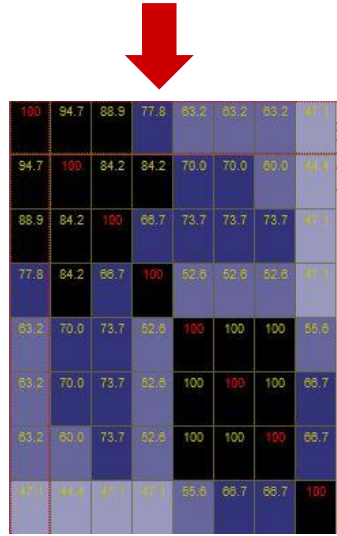
- Page 1: Similarity coefficient
- Keep existing similarity matrix
- Curve based
  - Pearson correlation
  - Cosine coefficient
  - Ranked Pearson correlation
- Including errors
  - Weighted Pearson correlation
- Band based
  - Jaccard
  - Dice
  - Jeffrey's X
  - Ochiai
  - Number of different bands
- Optimization: 1 %
- Band filtering
  - Minimum height: 0 %
  - Minimum surface: 0 %
- Band matching
  - Tolerance: 1 %
  - Tolerance change: 0 %
- Uncertain bands: Ignore
- Relaxed doublet matching
- Area sensitive  Fuzzy logic
- Show all.
- Save as new default to database

Buttons: < Back, Next >, Cancel



# Cluster Analysis

**PFGE Profiles**  
Set of band-based profiles

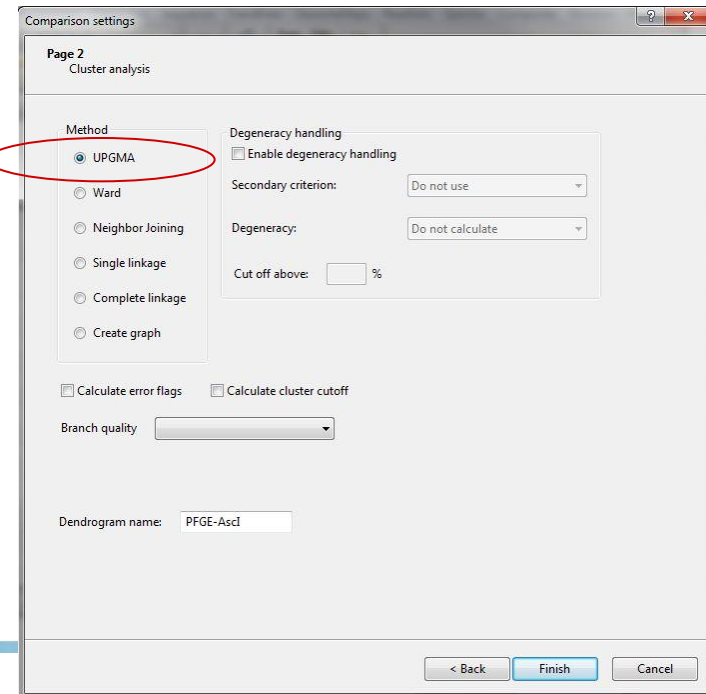
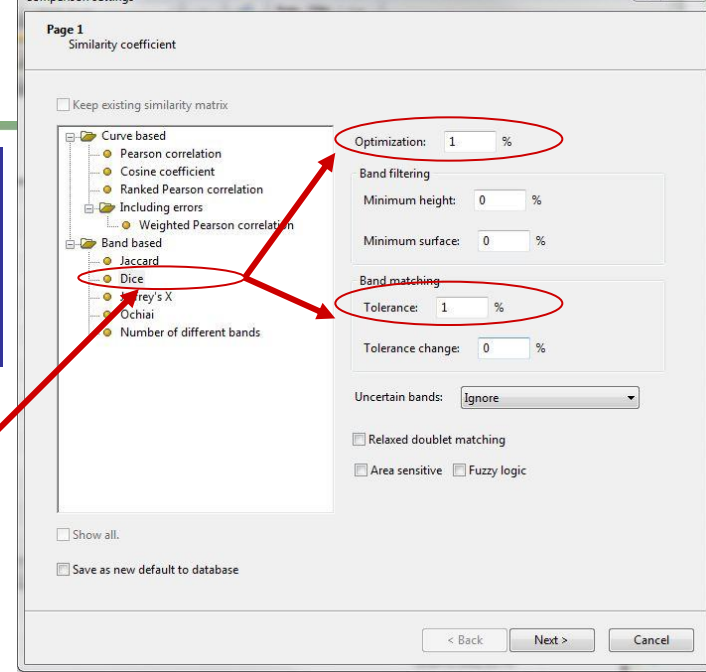


Coefficient calculation:  
Algorithm « Dice »  
Binary: Band absence or presence

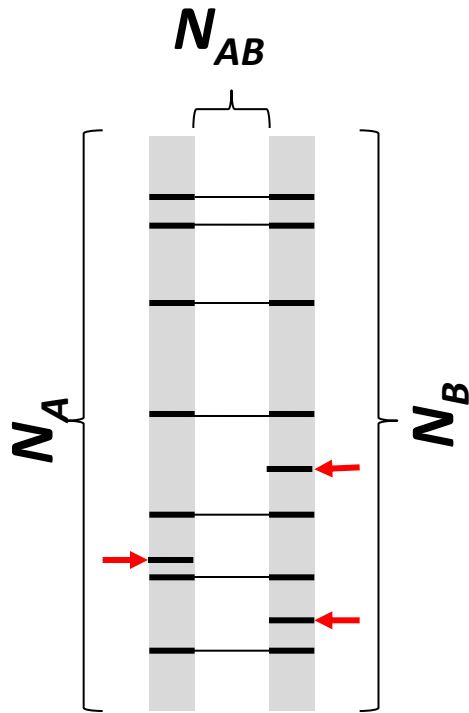
Similarity matrix

Clustering:  
Algorithm « UPGMA »

Dendrogram/clustering



# Dice coefficient: Band-based similarities



3 bands of difference

Dice:

$$S_D = \frac{2N_{AB}}{N_A + N_B}$$

$$S_D = \frac{2N_{common}}{N_{Total}}$$

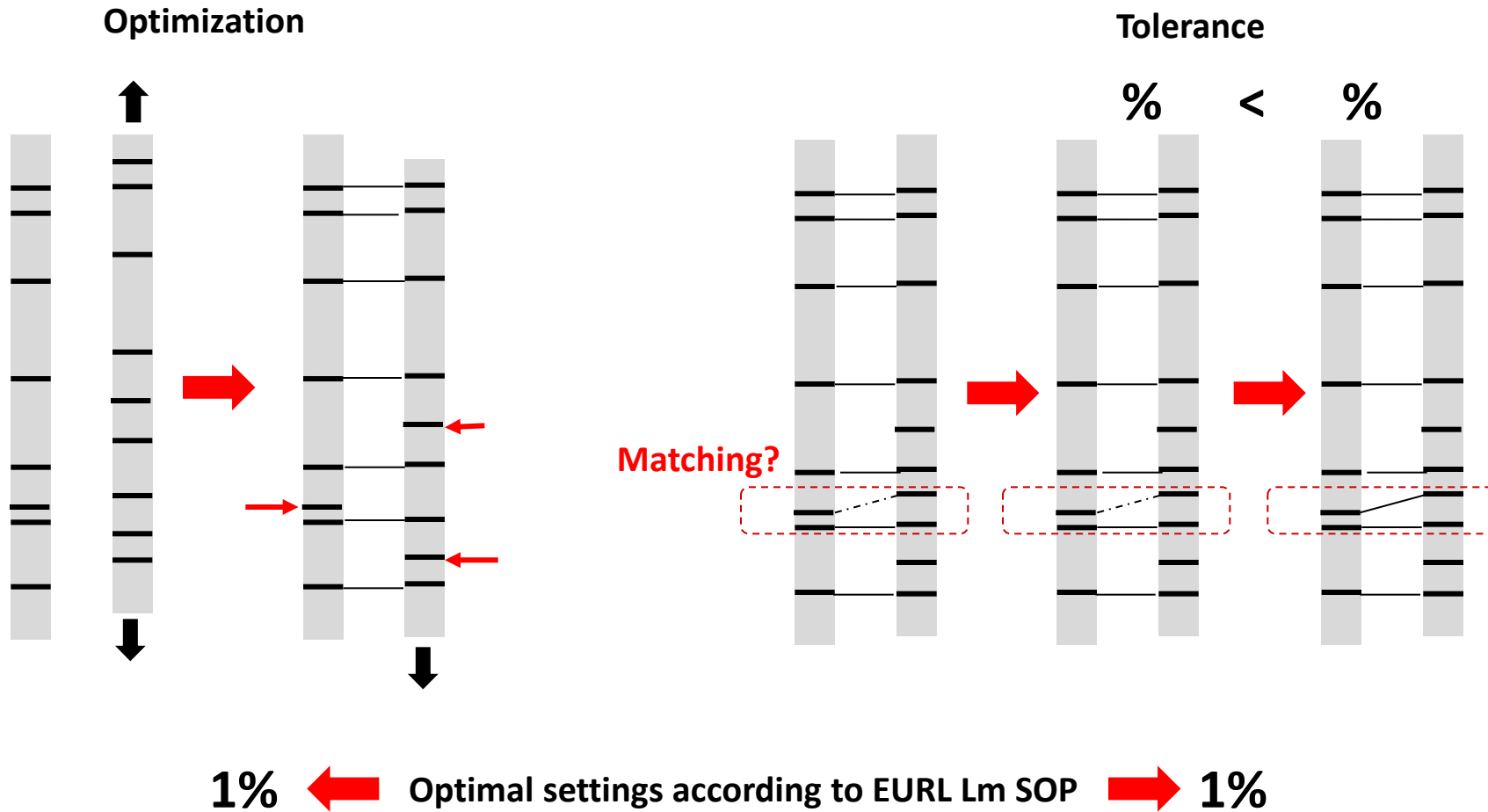
$$S_D = \frac{2 \times 7}{8 + 9} = 82\%$$

Similarity matrix

	A	B	C	...
A	100			
B	82	100		
C	72	90	100	
...				

# Dice coefficient: Optimization and tolerance

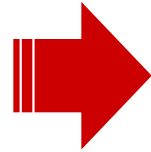
## Band Matching



# UPGMA: Unweighted Pair Group Method using Arithmetic average

Similarity matrix

	A	B	C	...
A	100			
B	82	100		
C	72	98	100	
...				

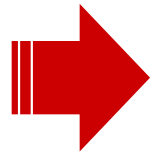


	90	100	B+C	A	D	...
B			98			
C						
A			77	100		
D			70	95	100	
...						

(82+72)/2



	90	100	B+C	A+D
B			98	
C				
A			73.5	95
D				
...				



	70	80	90	100	B+C+A+D
					73.5
B					
C					
A					
D					
...					



...

ER012 (Comparison)

File Edit Layout Groups Clustering Statistics Fingerprints Characters Sequence TrendData GenomeMaps ReadSets Spectra Composite Window Help

PFGE-Apal

Experiments

<All Experiment types>

Name	Aspi
PFGE-Apal	<All bi
PFGE-Ascl	<All bi
Serotypage moleculaire	<All bi
Ascl-Apal	<Defa
Pulsotypes Ascl	<All C
FlaA	<All C
Serogroup	<All C
FlaA PCR result	<All bi
Genome	<Defa
MLST	<All C
MLVA	<All C
MLVA Ascl Apal	<Defa
prfa	<Defa

Analyses

Name
Ascl-Apal
PFGE-Apal

Groups

Size	Name
5	ER012
5	NER012

Dendrogram

Experiment data

Comparison settings

Page 1  
Similarity coefficient

Keep existing similarity matrix

- Curve based
  - Pearson correlation
  - Cosine coefficient
  - Ranked Pearson correlation
- Including errors
  - Weighted Pearson correlation
- Band based**
  - Jaccard
  - Dice
  - Jeffrey's X
  - Ochiai
  - Number of different bands

Optimization: 1 %

Band filtering

Minimum height: 0 %

Minimum surface: 0 %

Band matching

Tolerance: 1 %

Tolerance change: 0 %

Uncertain bands: ignore

Relaxed doublet matching

Area sensitive  Fuzzy logic

Show all.

Save as new default to database

< Back Next > Cancel



# Database management

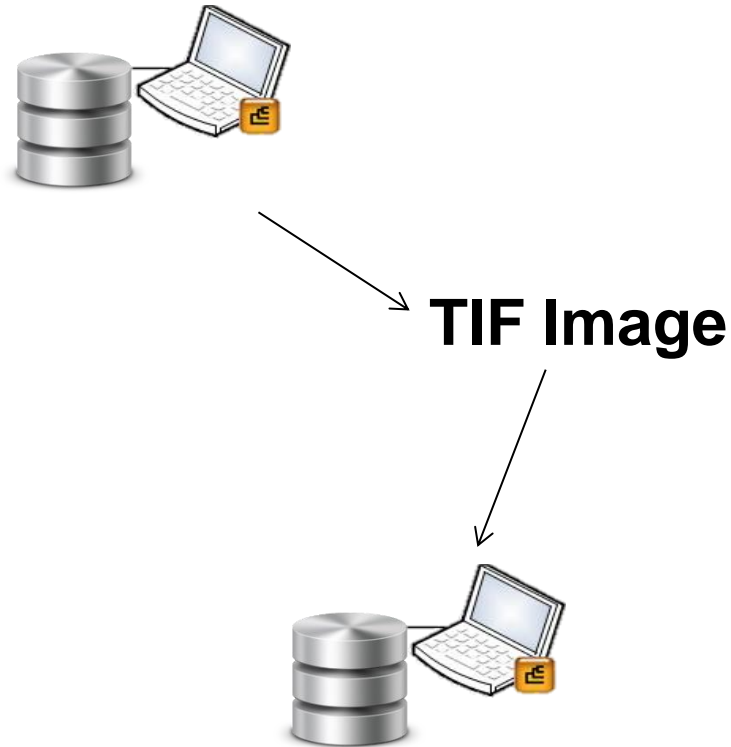




---

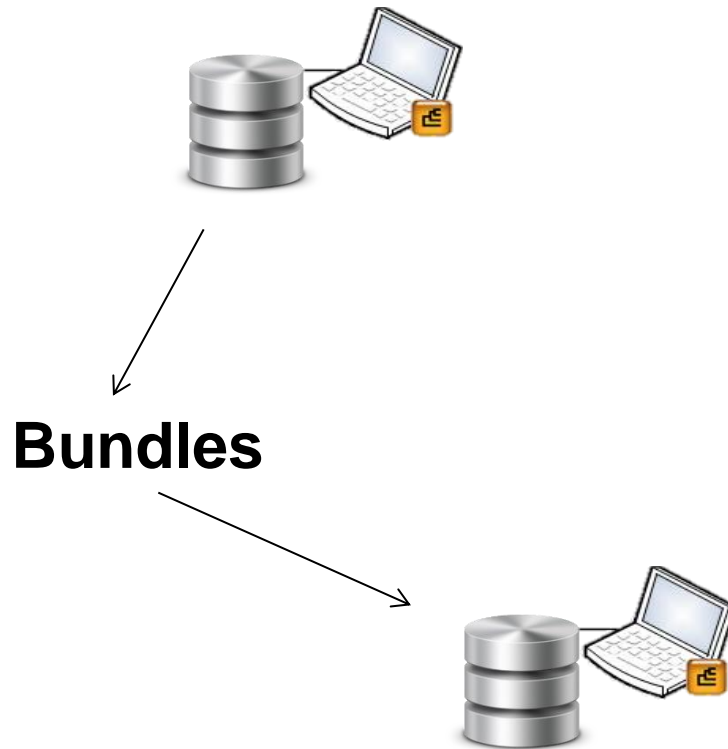
# Communication between databases

# Database exchange



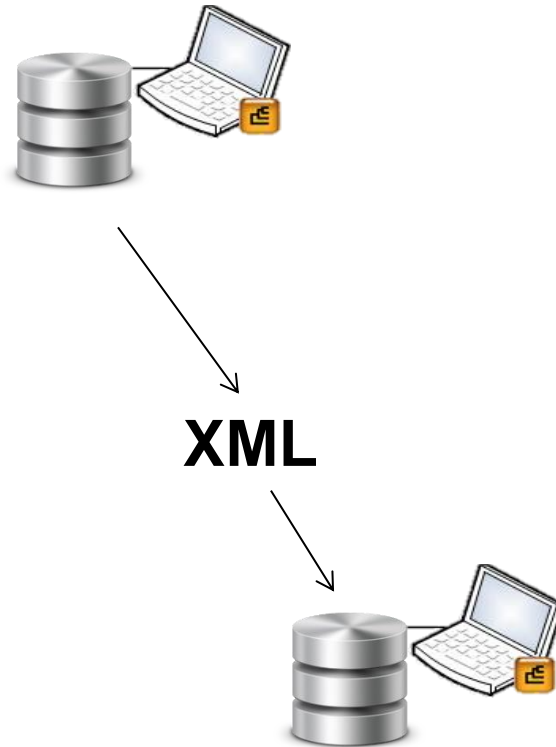
**- Permanent import of fingerprint, need image processing, normalization and analysis**

# Database exchange



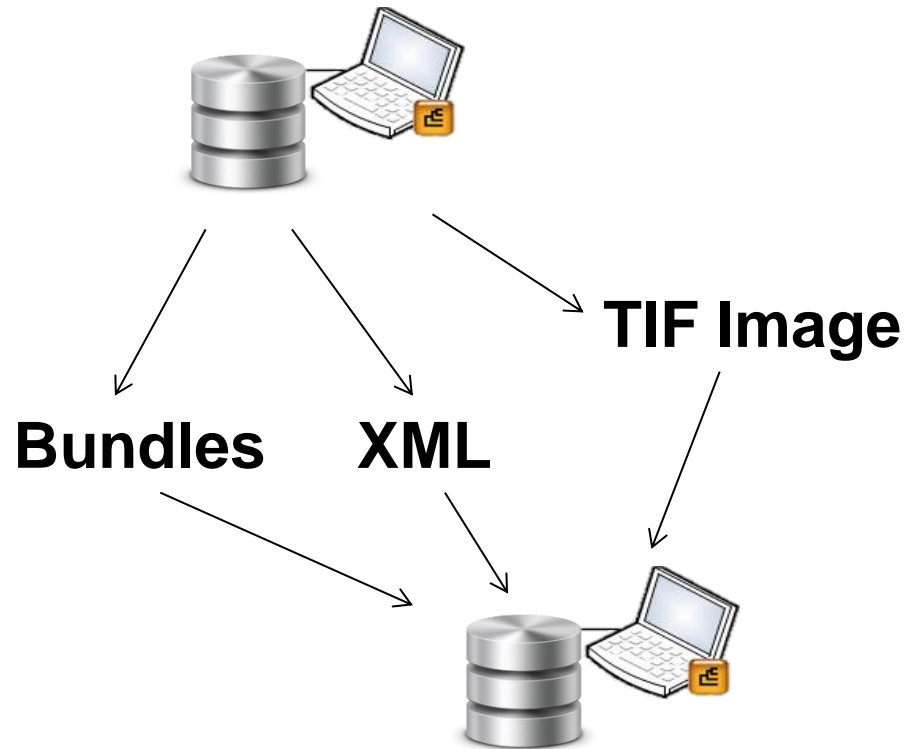
**- Profile are temporally available in the database (profiles, sequences and fields) until the software shut down**

# Database exchange



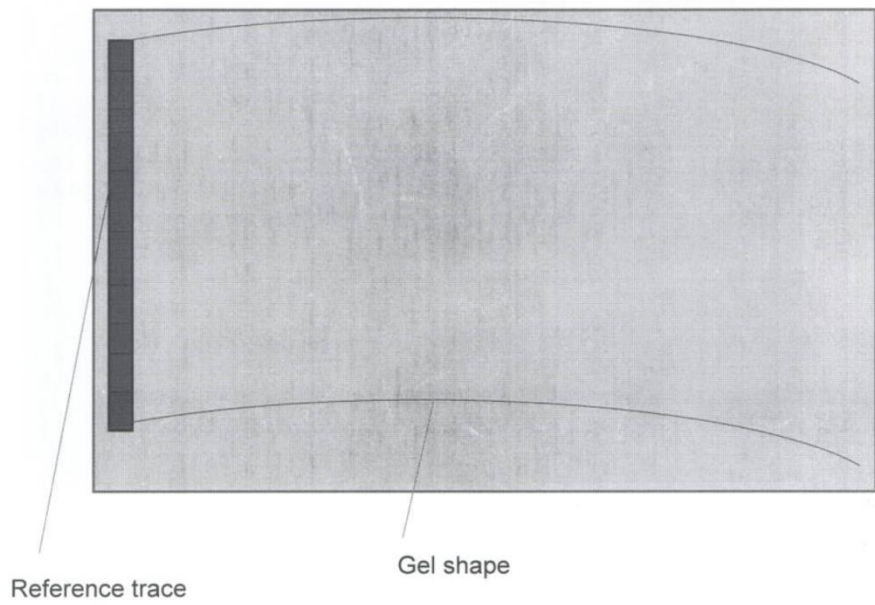
**- Permanent import of data (profiles, sequences and fields) without any analysis.**

# Database exchange

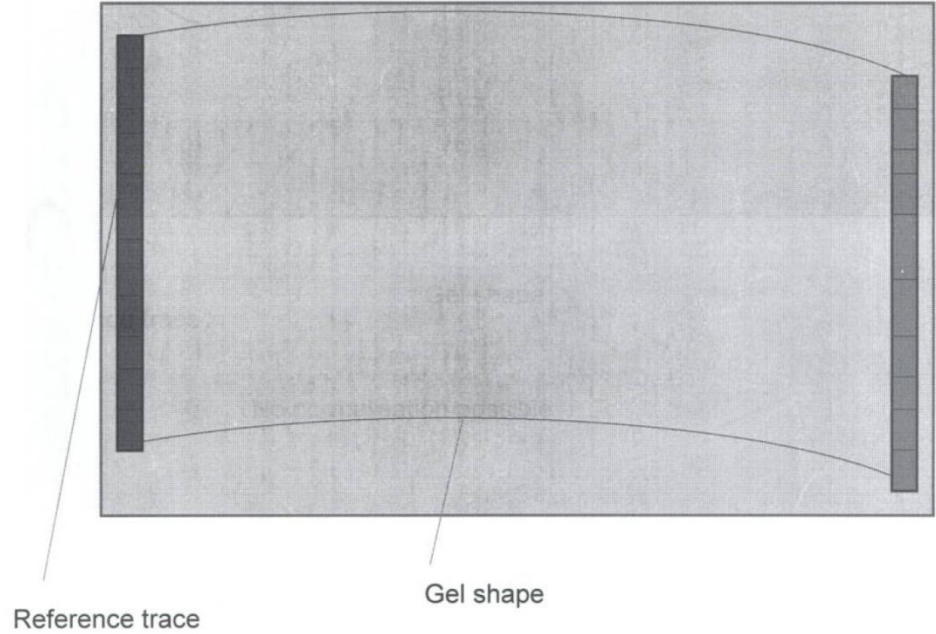


---

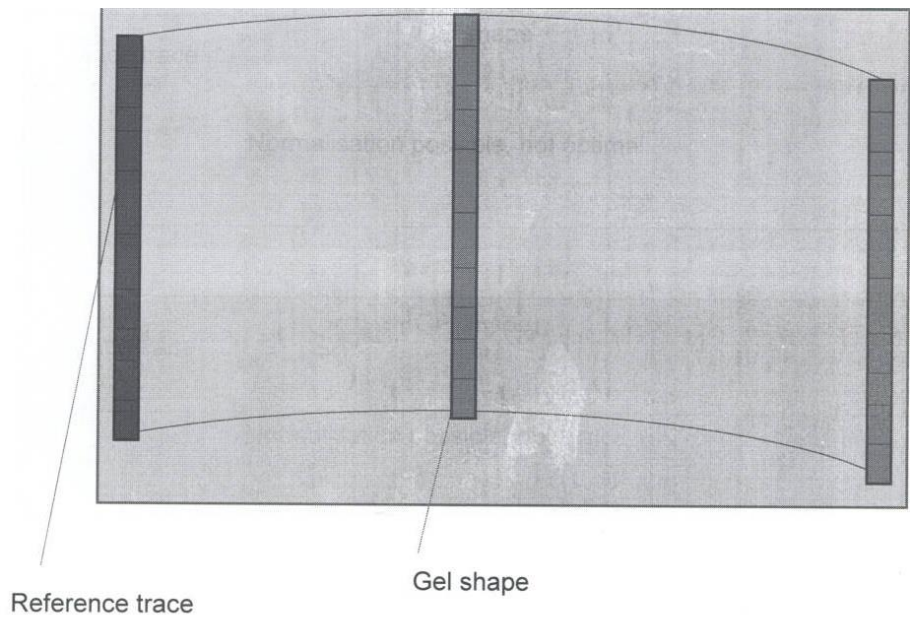
# Database compatibility



No normalisation possible

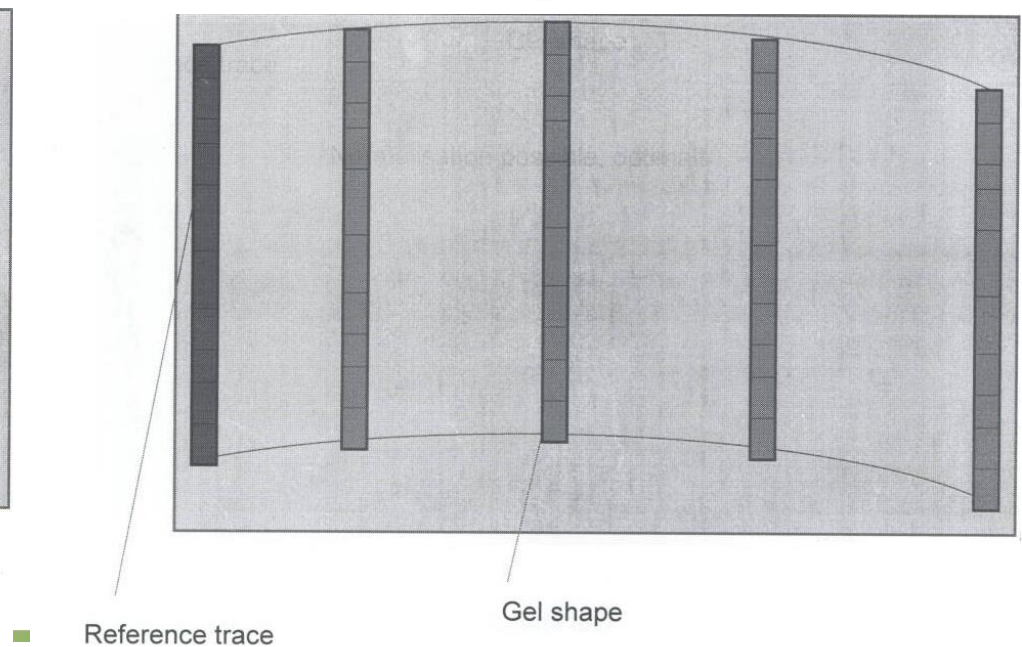


No normalisation possible



15

Normalisation possible, not optimal



Normalisation possible, optimal!

# Reference system

---

## Reference system included in the gel:

- Salmonella Braenderup H9812 *Xba*I restriction profile



# Reference system

## Database main reference system included in the BioNumerics experiment:

- Salmonella Braenderup H9812 *Xba*I restriction profile

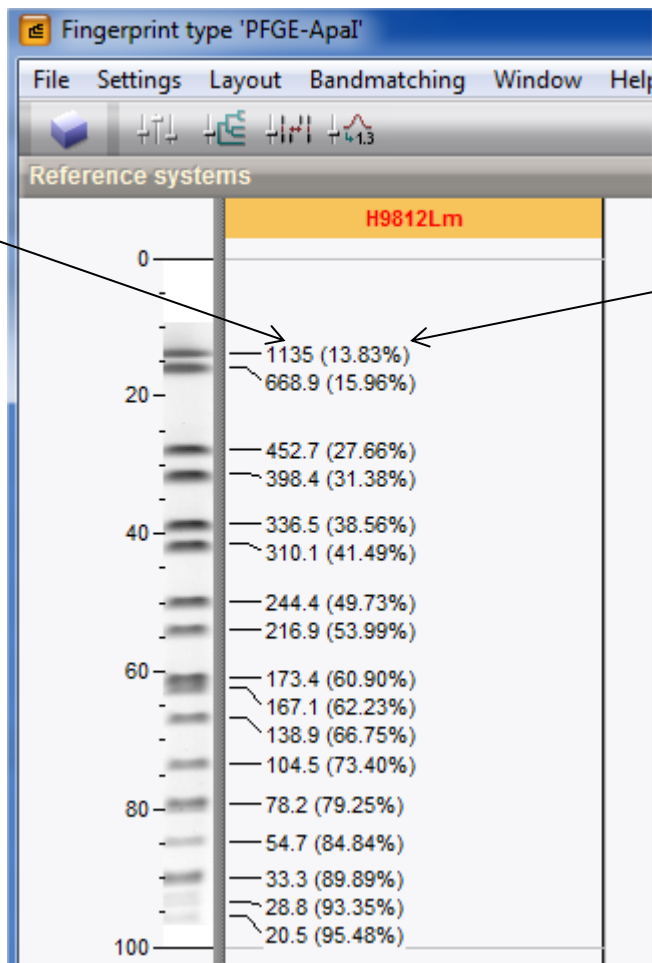
The screenshot displays the BioNumerics software interface. The main window is titled "Listeria\_LSAI\_HQPAP\_IFIP - BioNumerics". The interface is divided into several panes:

- Database entries:** A table listing various Listeria strains with columns for Key, MLST CC, MLST ST, IDGr90\_AscI, and IDGr90\_ApaI. The table contains 2408 entries.
- Experiment types:** A table listing different experiment types with columns for #, Name, and Type. The table contains 3 entries: 1 PFGE-ApaI (Fingerprint types), 2 PFGE-AscI (Fingerprint types), and 3 ApaI-AscI (Composite data sets).
- Identification projects:** A table listing identification projects with columns for Name and Modified date. The table contains 2 entries: Bilan ApaI identification (2016-05-23 17:27:23) and Bilan AscI identification (2016-06-28 15:40:23).
- Comparisons:** A table listing comparisons with columns for Name, Modified date, Level, and Numb. The table contains 10 entries, including Bilan ApaI (2016-06-28 15:34:22, Level 2333), Bilan AscI (2016-06-30 13:22:37, Level 2402), and others.

The status bar at the bottom indicates: "Database: Listeria\_LSAI\_HQPAP\_IFIP (DefaultUser...) Entries: Loaded=2408, View=2408, Selected=0 3 experiments C:\Users\rl.felix\Documents\Publi diverse porc>Listeria\_LSAI\_HQPAP\_IFIP".

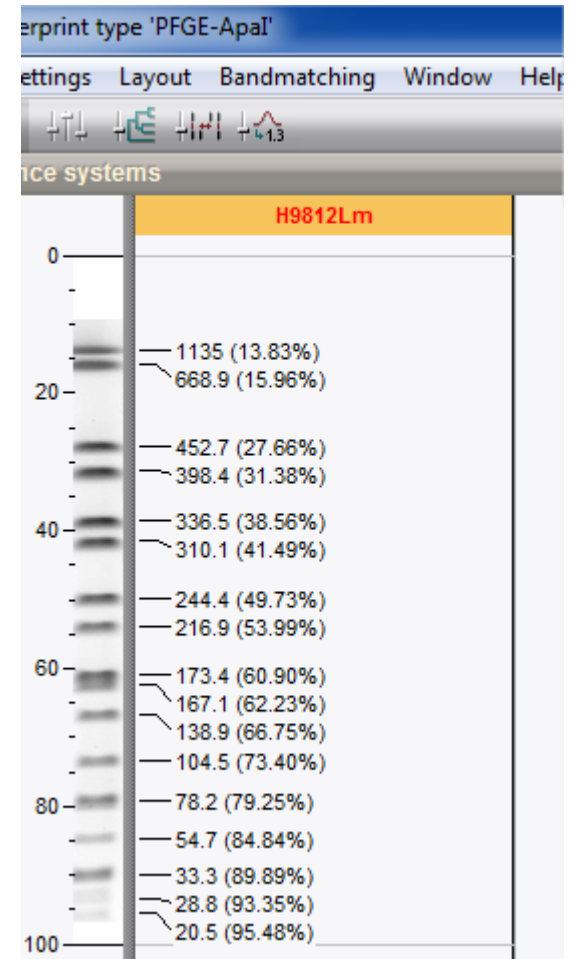
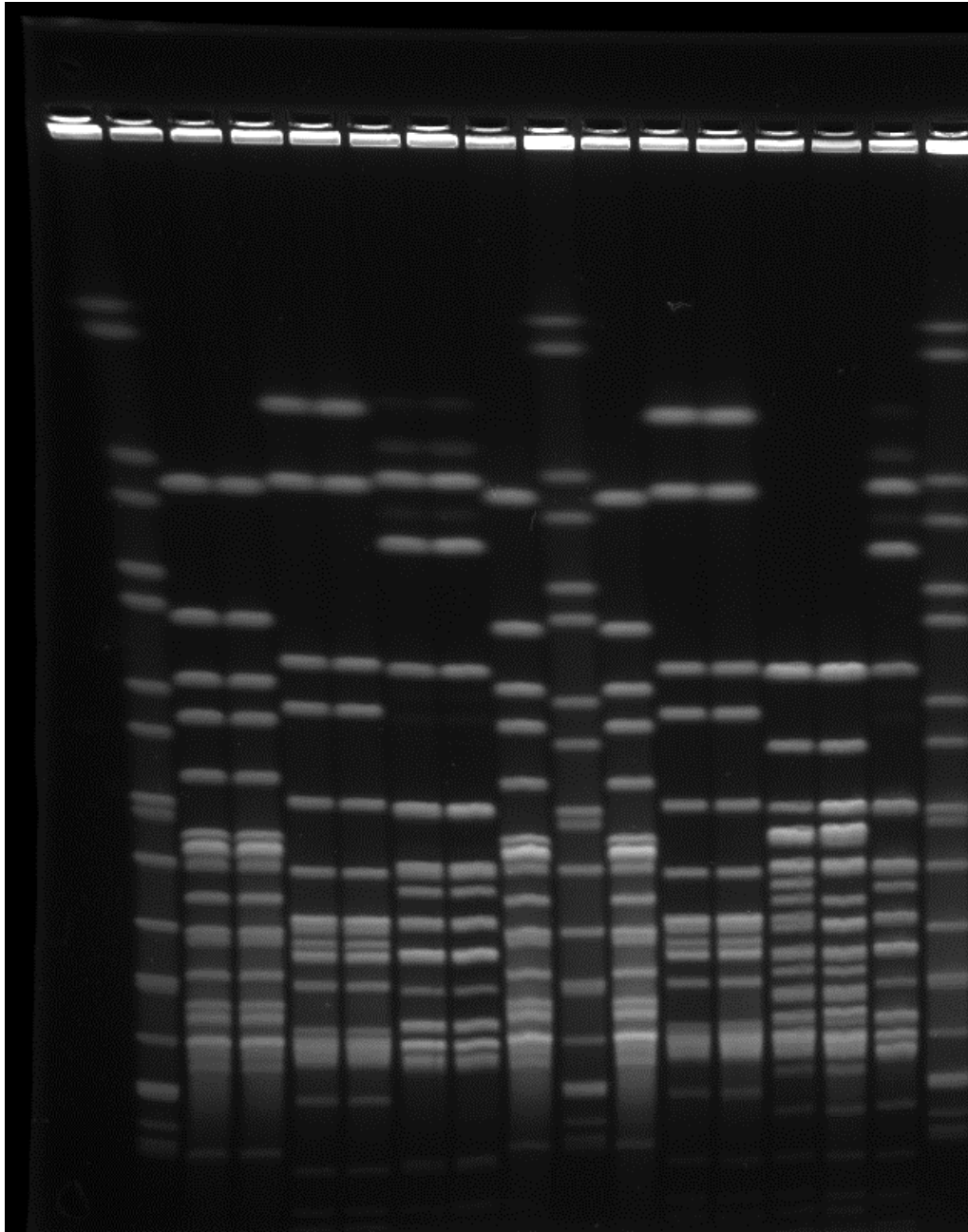
# Normalization process

**Metric =  
molecular  
size kbp**

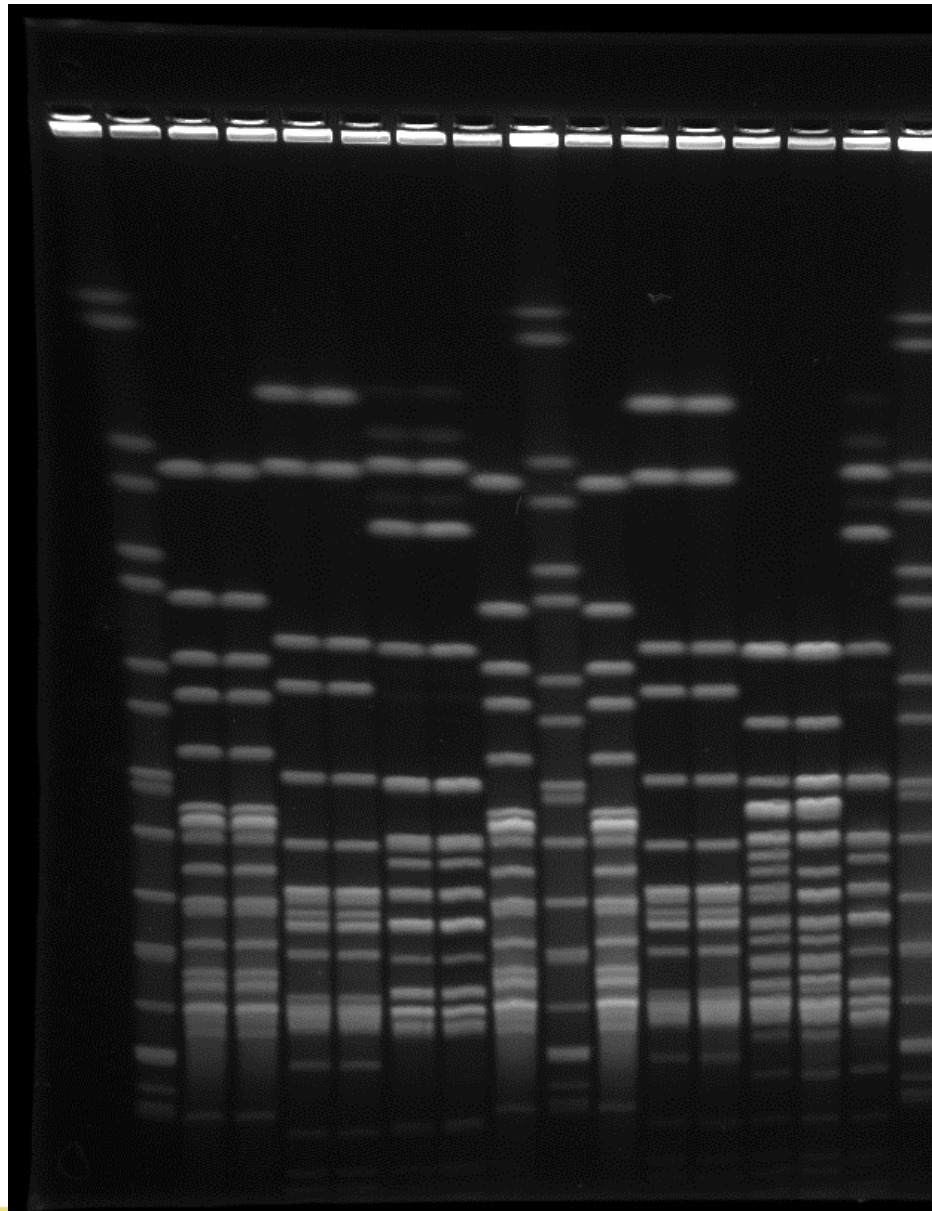


**Percentage of  
migration**

# Normalization process



# Normalization process



-ApaI'

Bandmatching Window Help

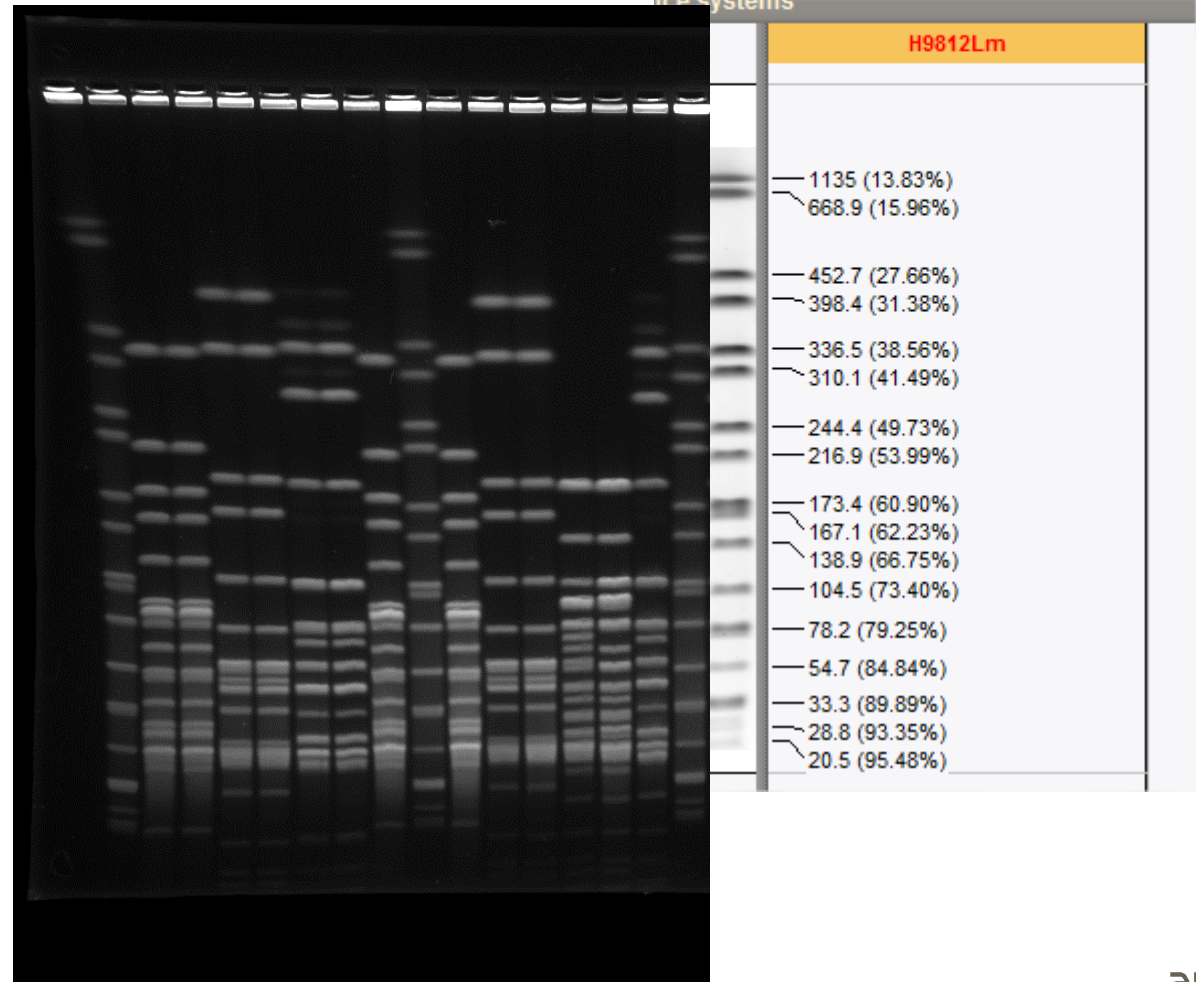
+4.3

**H9812Lm**

5	(13.83%)
9	(15.96%)
7	(27.66%)
4	(31.38%)
5	(38.56%)
1	(41.49%)
4	(49.73%)
9	(53.99%)
4	(60.90%)
1	(62.23%)
9	(66.75%)
5	(73.40%)
2	(79.25%)
7	(84.84%)
8	(89.89%)
8	(93.35%)
5	(95.48%)

# Normalization process

## Decale





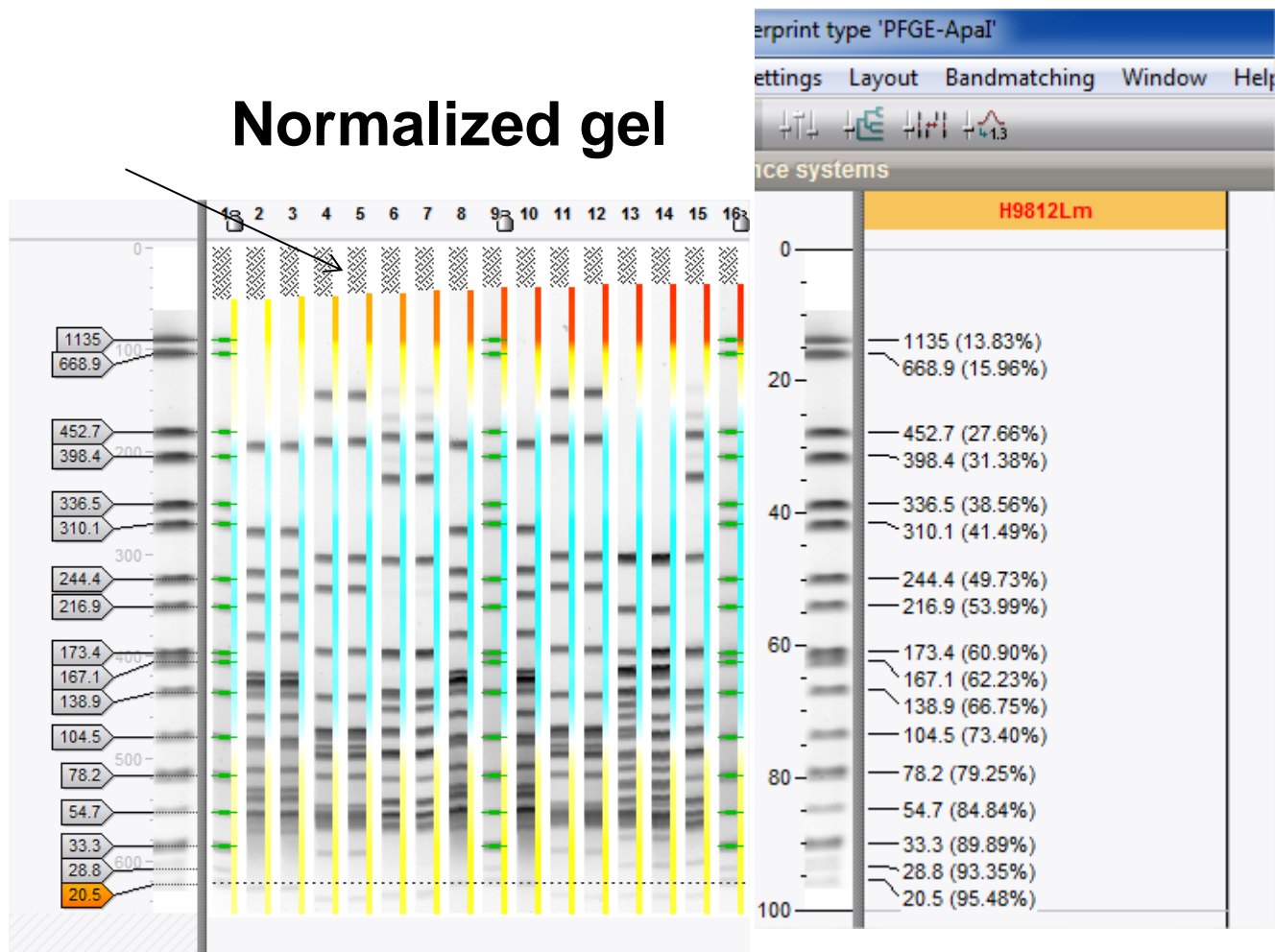
# Normalization process

## Distortion



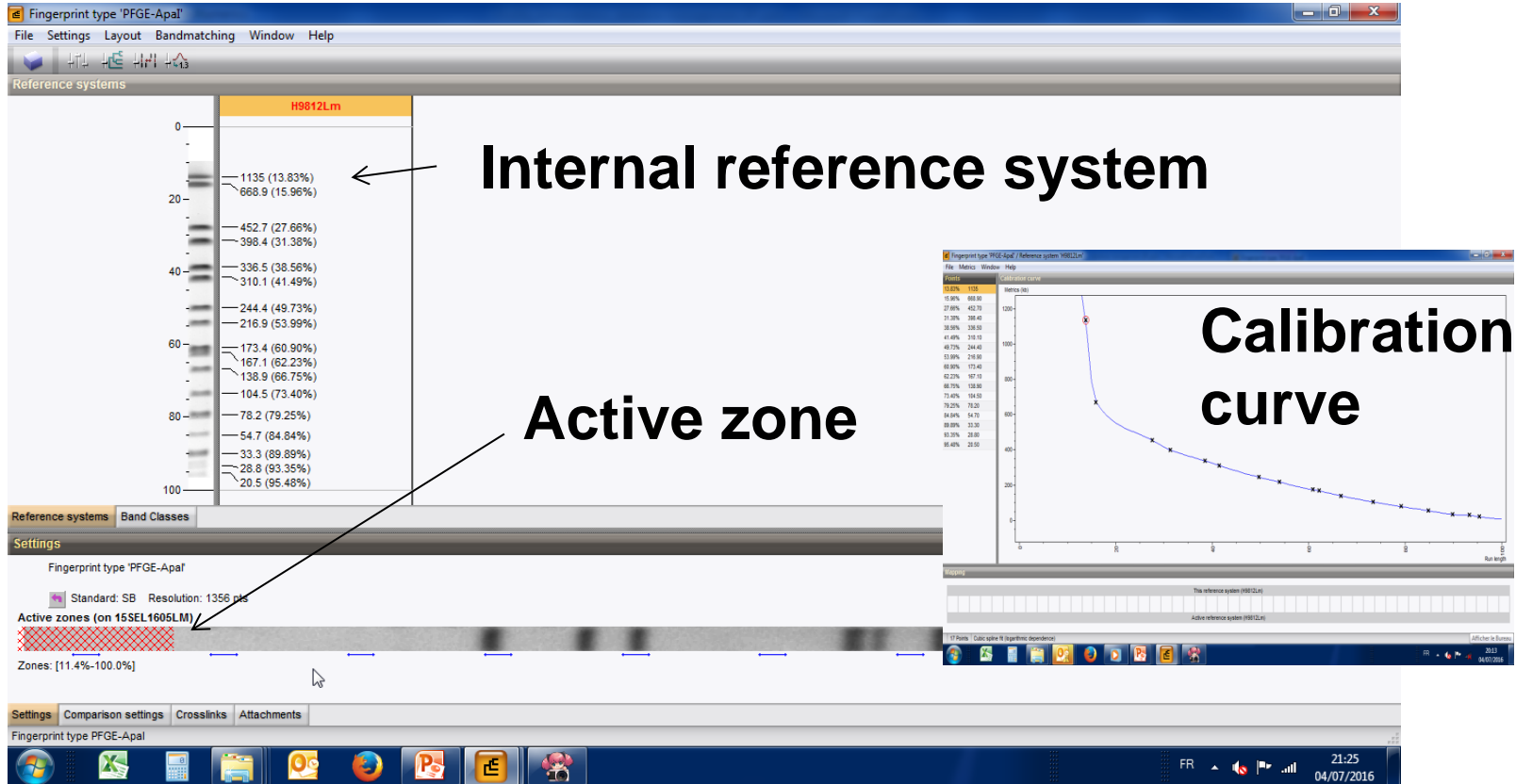
# Normalization process

## Normalized gel



# Reference system

## Database main reference parameters





---

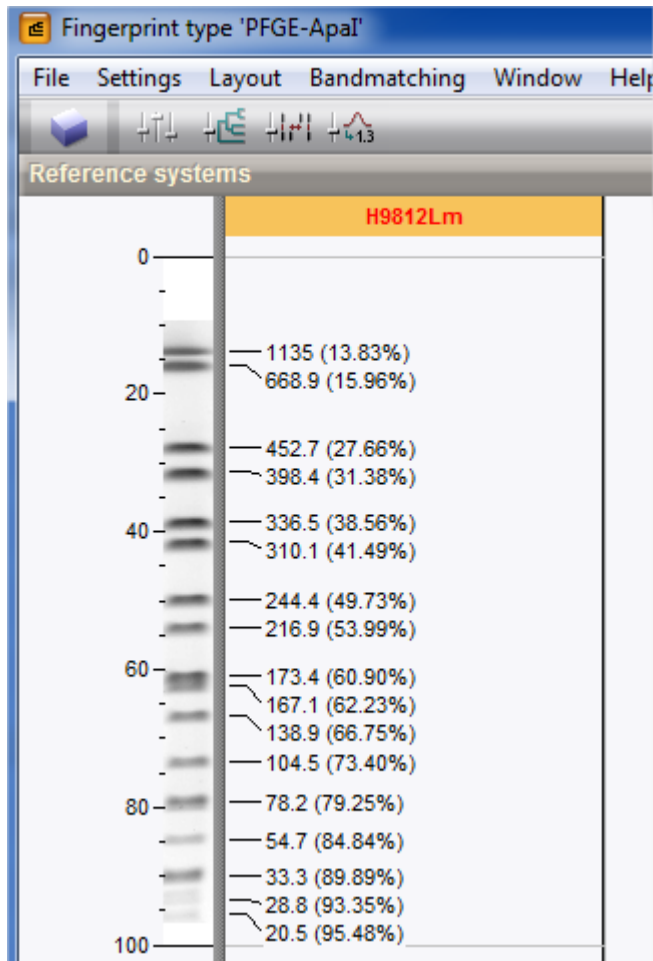
# How to proceed when the reference systems are different ?



*EURL Lm*

European Union Reference Laboratory for  
*Listeria monocytogenes*

# Re-mapping



Several reference systems can work together

# Re-mapping

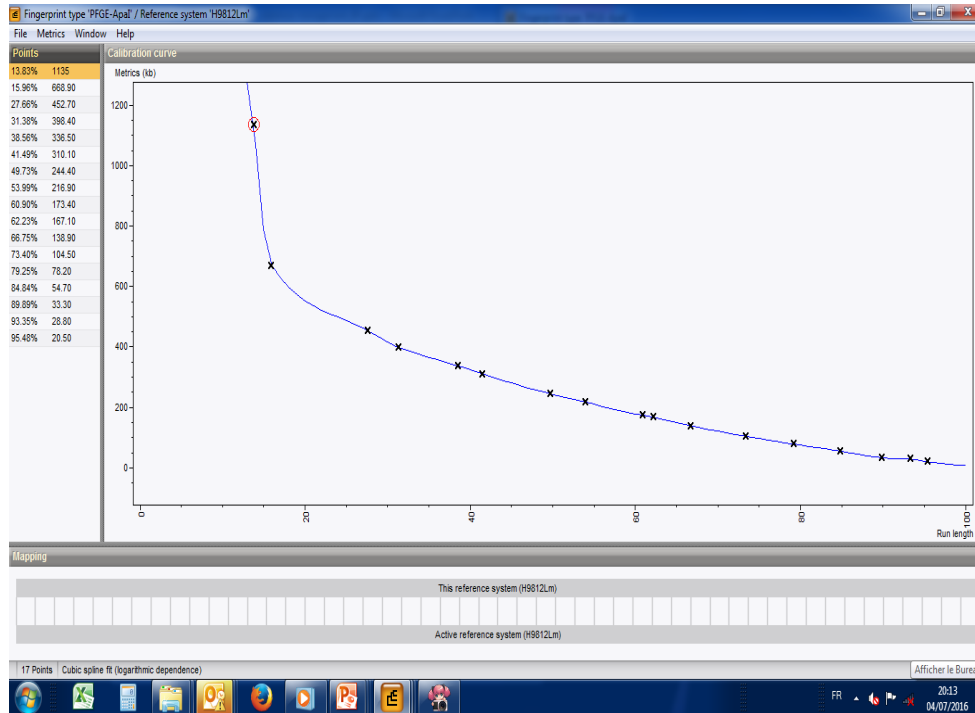
---

**The re-mapping can be used for bundles and XML import**

**Bundles need re-mapping to be set beforehand**

**XML import implement automatically parallel reference system**

# Re-mapping



The calibration curve is used to perform the re-mapping process



# Re-mapping

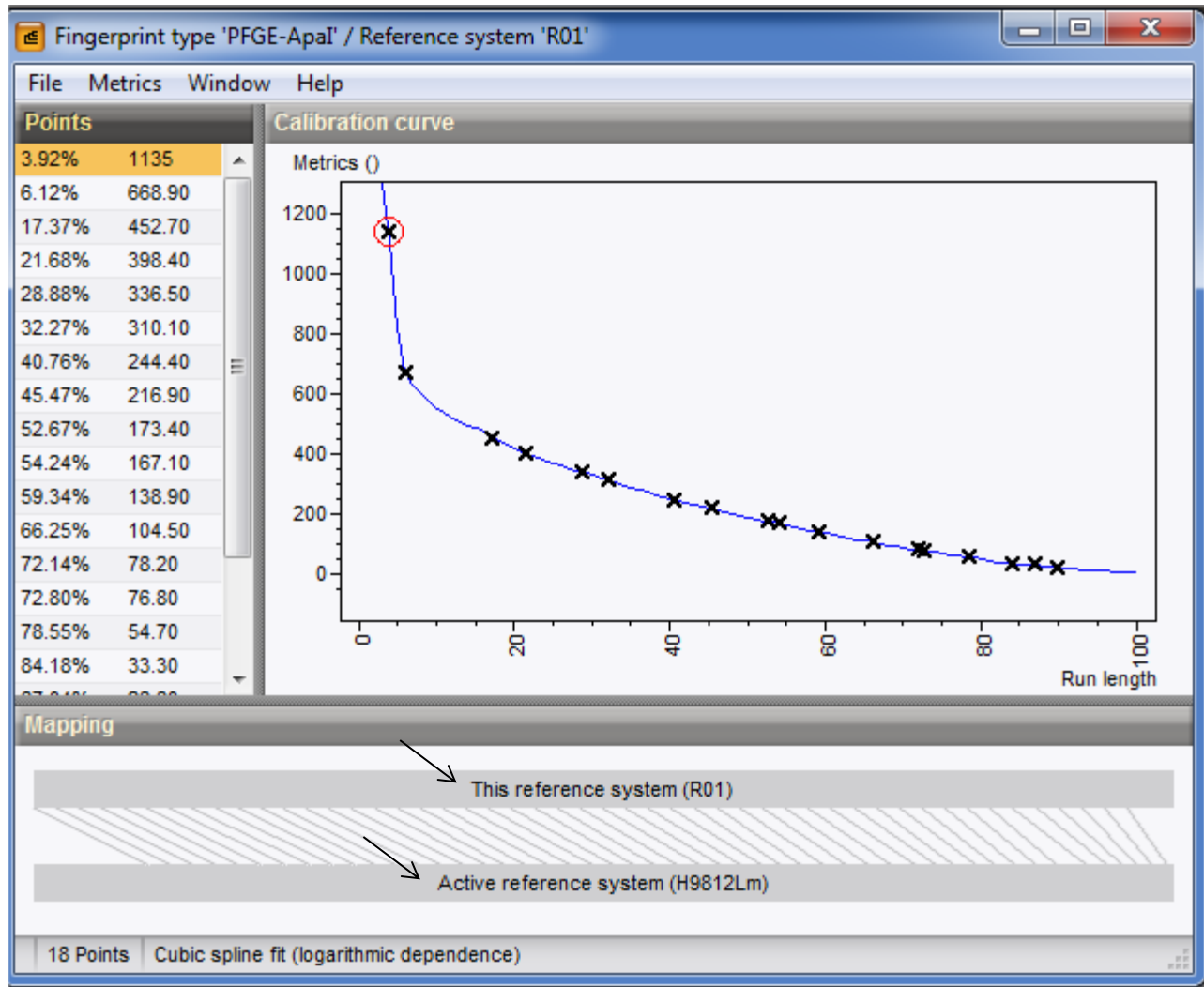
The screenshot illustrates the 'Re-mapping' process in a software application. It shows three overlapping windows:

- Settings Window:** The 'Settings' menu is open, with 'Edit reference system' selected. Other options include 'Data type...', 'General settings...', 'Comparison settings...', 'Position tolerance settings...', 'Comparative quantification...', 'Set standard...', 'Fingerprint file information fields...', 'Set as active reference system', 'New reference system (positions)...', 'New reference system (curve)...', 'Remove reference system', 'Create peak intensity profile...', 'Enable fast band matching' (checked), 'Level assignment...', and 'Summary replication settings...'.
- Metrics Window:** Displays a table of peak data for 'Fingerprint type 'PFGE-ApaI' / Reference syste...'.

Points	
13.83%	1135
15.96%	668.90
27.66%	452.70
31.38%	398.40
38.56%	336.50
41.49%	310.10
49.73%	244.40
53.99%	216.90
60.90%	173.40
62.23%	167.10
66.75%	138.90
73.40%	104.50
- Calibration curve Window:** Shows a graph titled 'Calibration curve' with 'Metrics ()' on the y-axis (0 to 1200) and 'Position' on the x-axis. A red dot marks a specific point on the curve.
- Mapping Window:** Shows a list of mapping options for 'Fingerprint type 'PFGE-ApaI' / Reference syste...'.
  - Copy markers from reference system...
  - Logarithmic dependence (checked)
  - First degree fit
  - Third degree fit
  - Cubic spline fit (checked)
  - Pole fit
  - Assign units...



# Re-mapping is implemented



# Re-normalization

- Specific script provided by Applied Maths
- Convert\_refsys.BNS
- The script can re-analysed automatically all fingerprint into another internal reference system
- However this script only to convert similar reference system e.g. SB01 into SB02
- The script works by batch of 20 gels
- Beware to backup your system before to proceed
- Beware to verify the output of the re-normalisation

