EPIDEMIOLOGICAL DATA INTEGRATION, ANALYSIS AND VISUALIZATION : FROM DATABASES TO POPULATION GENETICS

João André Carriço

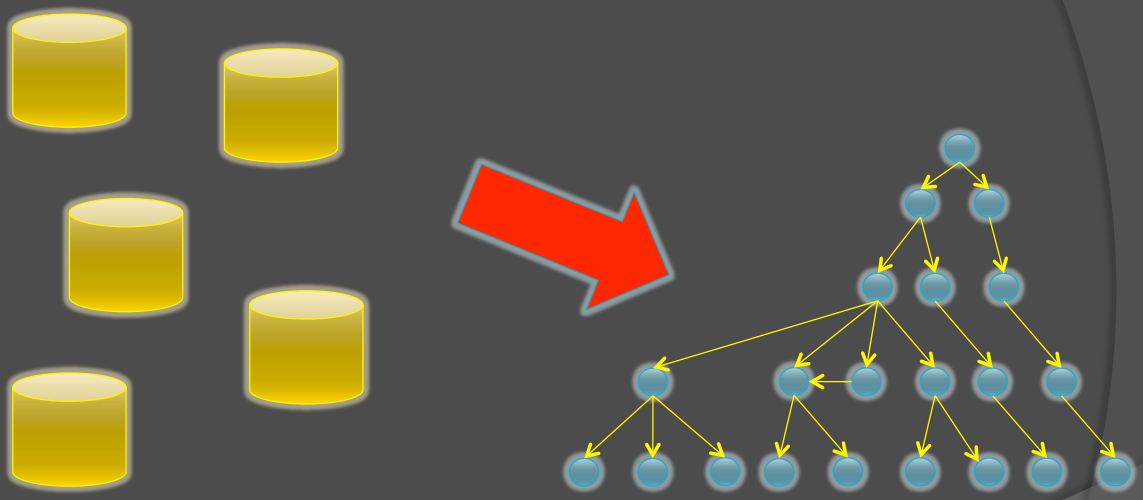
www.joaocarrico.info

UMMI

INSTITUTO DE MEDICINA MOLECULAR

FACULDADE DE MEDICINA UNIVERSIDADE DE LISBOA

From Databases to Population genetics



Microbial Typing methods

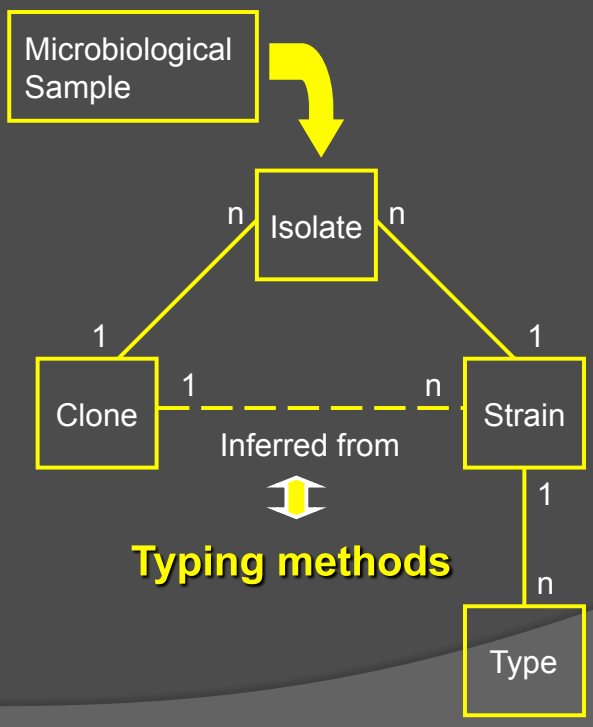
Crude classifications and false generalizations are the curse of organized life.

George Bernard Shaw (1856 - 1950)

All science is either physics or stamp collecting.

Ernest Rutherford, in J. B. Birks "Rutherford at Manchester" (1962)

Molecular Epidemiology : From isolates to clones



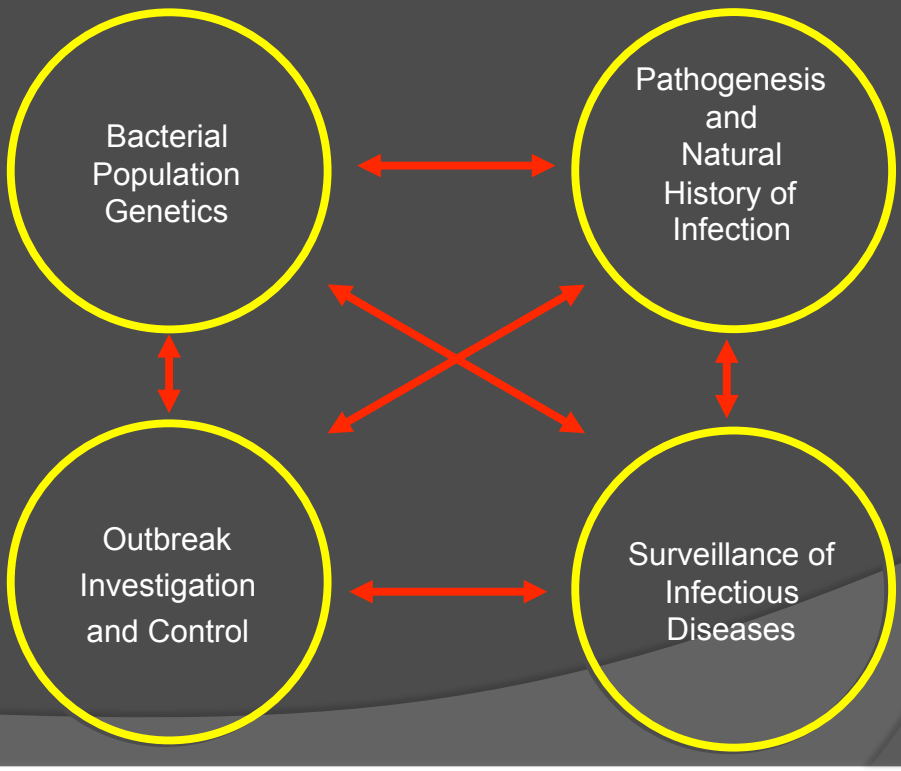
Evolution of typing methods

Phenotypic
- Growth and morphological characteristics
- Physiological characteristics: Antibiotic susceptibilty testing
- Serotyping

Genotypic (Molecular)
- Extrachromossomal: Plasmid fingerprinting
- Cromossomal: RFLP/ PFGE

Genotypic (Sequence-based)
- MLST
- MLVA
- spa Typing (S.aureus)

Applications of microbial typing



Data Integration

Multiple sources of data

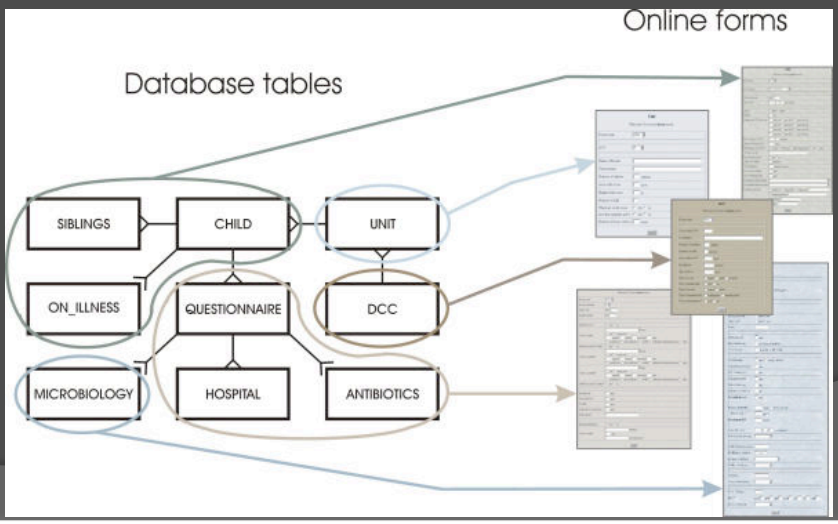
- Isolate Data :
  - Host clinical data
  - Host demographic data
  - Typing methods data
    - Sequences -> Alleles
    - Data linked to some temporal definition. Ex: MICs
    - Categorical Data
    - Etc…

Old Toys, Old Problems...
New Toys, New Solutions, New Problems...

Past and Present…

RDBMS

(Relational DataBase Management Systems)



Excel: the most (wrongly) used DB ever...

id	strain	other_name	st	country	region
1	NCTC11906-19F	NCTC11906	1	UK	
2	SAF-17244-19	17244	52	South Africa	
3	PJ23/1	PJ23	300	Sweden	
4	SP264-23F	264	81	Spain	
5	87-029044-14	87-029044	20	Slovakia	
6	SP3026-6B	3026	90	Spain	
7	UK577-23F	DN87/577	81	UK	
8	ATCC6323-23	ATCC6323	108	USA	
9	CS111-23F	CS111	37	USA	
10	GM70-19A	GM70	81	Spain	
11	GA71-19F	GA71	81	Spain	
12	GM169-19F	GM169	86	Spain	
13	GM17-6B	GM17	90	Spain	
14	KD18-23	100520	150	Kenya	
15	KD6-19		107	Kenya	
16	PN13-14	Kaguane/73	10	Papua New Guinea	
17	PN16-42	110K/70	117	Papua New Guinea	
18	PN18-19	VA1	41	USA	
19	PN20-19	6249	68	South Africa	
20	PN29-15	4865	194	Australia	
21	R6	R6	128	USA	
22	SP665-9V	SP665	156	Spain	
23	VH14-14	VH14	12	Spain	
24	M1-7F	B4-831	203	Sweden	
25	M2-4	B4-1021	205	Sweden	
26	M3-14	B4-5626	124	Sweden	
27	M4-14	B4-8660	124	Sweden	
28	M5-4	B4-8067	205	Sweden	
29	M6-4	B4-9708	205	Sweden	
30	M7-6B	B4-7681			
31	M8-9N	B4-743			
32	M9-18C	B4-8250			
33	M10-14	B4-4168			
34	M11-14	11L11B0			
35	M12-14	01L02B0			

ST	aroe	gdh_	gki_	recP	spi_	xpt_	ddl_
1	1	1	1	1	1	1	1
2	1	1	4	1	18	13	18
3	1	5	1	8	14	11	14
4	1	5	2	20	1	1	1
5	1	5	2	11	9	3	51
6	1	5	2	1	9	3	49
7	1	5	4	1	5	1	8
8	1	5	4	5	17	1	8
9	1	5	4	5	5	1	8
10	1	5	4	5	9	1	10
11	1	5	4	5	10	1	8
12	1	5	4	4	9	3	16
13	1	5	4	5	5	27	8
14	1	5	4	5	5	1	14
15	1	5	4	5	5	3	8
16	1	5	4	12	5	27	8
17	1	5	4	11	9	3	47
18	1	5	4	11	9	3	16
19	1	5	4	18	9	3	16
20	1	5	4	1	5	3	3
21	1	5	4	11	9	3	4
22	1	5	4	11	9	43	52
23	1	5	4	5	35	3	8
24	1	5	4	5	5	1	24
25	1	5	4	1	5	3	8
26	1	5	7	12	26	1	14
27	1	5	9	1	14	14	6
28	1	5	9	1	9	14	6
29	1	5	10	5	5	3	8
30	1	5	27	20	1	1	1
31	1	5	31	5	5	3	8
32	1	6	4	5	15	1	20
33	1	8	1	2	6	4	6
						4	6
						4	6
						4	6
						4	6
						4	6
						4	6
						4	6
						1	12
						4	6
						3	8

id	sequence	length
1	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
2	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
4	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
5	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
6	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
7	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
8	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
9	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
10	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
11	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
12	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
13	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
14	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCTTATAAGGAGCAGGTGATTCCTTATTT	
15	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
16	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
17	GAAGCGAGTGACCTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
18	GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
19	GAAGCGAGTGACTTGGCAGAAACAGTTGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
20	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
21	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	
22	GAAGCGAGTGACTTGGTAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTT	

Old Toys, Old Problems…
New Toys, New Solutions, New Problems…

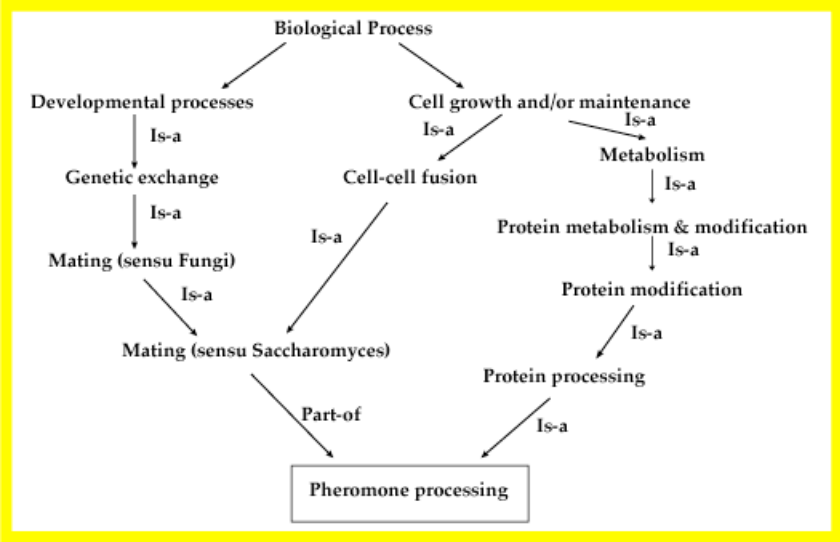
Present and Future

Semantic Web Approaches

Ontologies
+ RDF Stores

↓

Machine
readable
formats



Work in progress:
RESTful Sequence–based typing methods Databases

Typing Methods Ontology

→

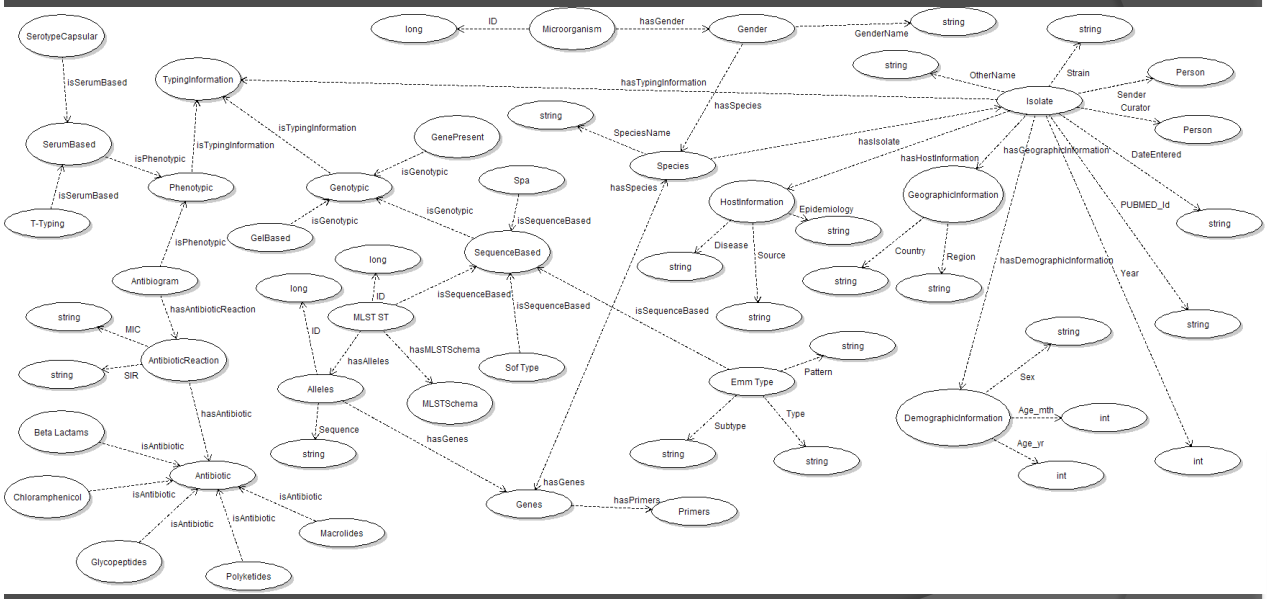
Jena – Semantic Web Framework
(http://jena.sourceforge.net/)

Programmatic environment for RDF , OWL and SPARQL and a rule-based inference engine

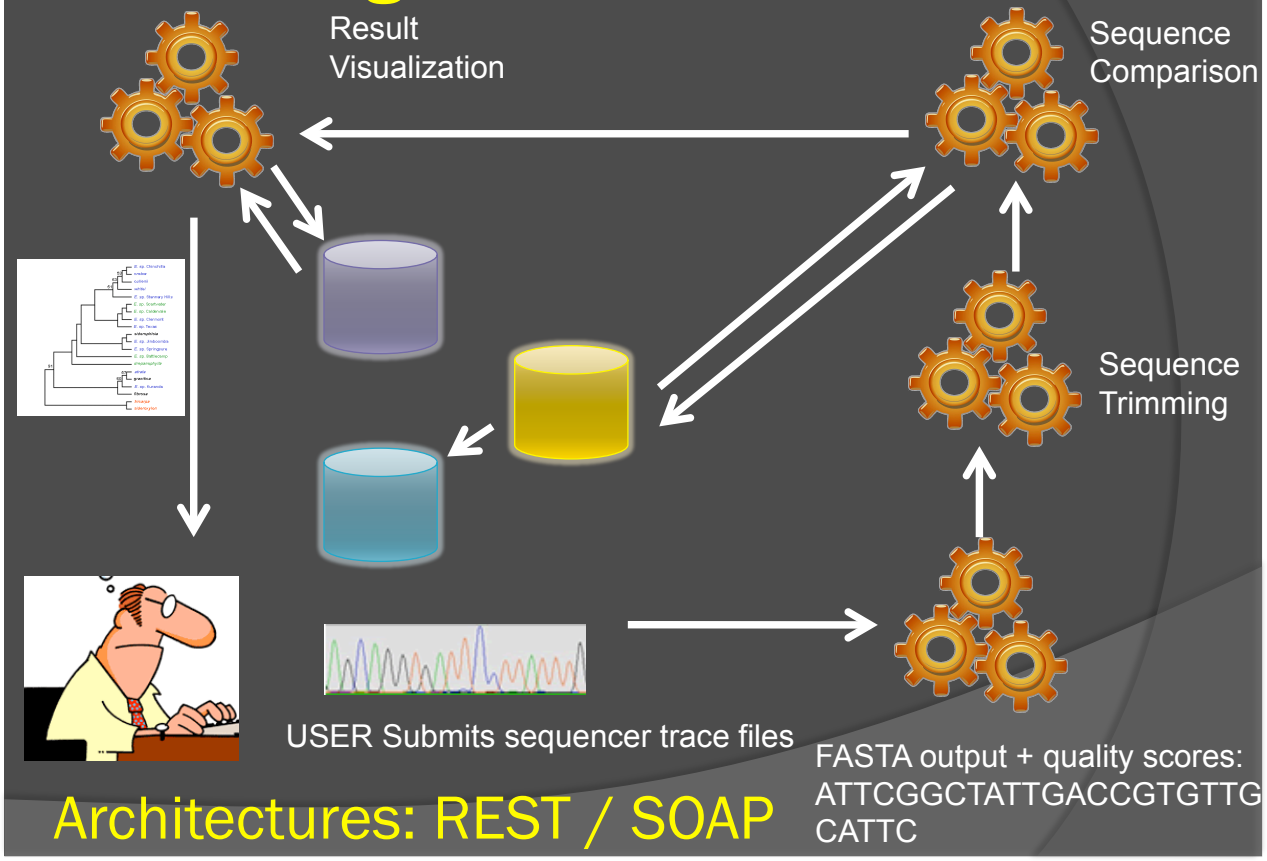
→

REST interface → Web browser interface

→ JAVA library



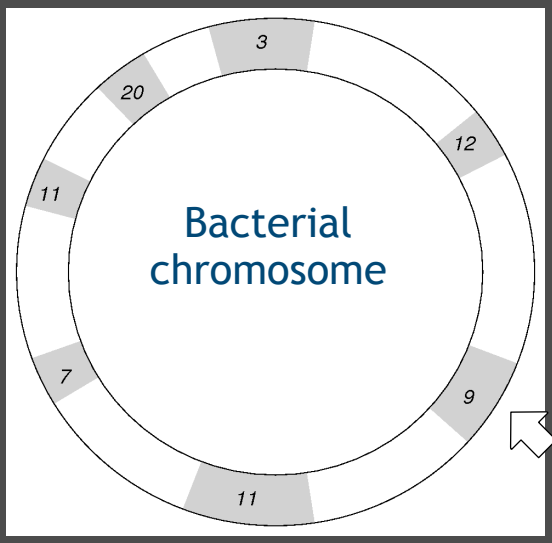
Connecting Data sources



Architectures: REST / SOAP

Data Analysis

MLST



To each unique gene sequence (allele) is attributed an integer ID, by comparison with online DBs

Allelic profile:
12 - 9 - 11 - 7 - 11 - 20 - 3

Each allelic profile, aka ST, is unequivocally identified by an integer.

housekeeping gene

Single locus variant (SLV): 12 - **10** - 11 - 7 - 11 - 20 - 3
Double locus variant (DLV): 12 - **10** - 11 - **11** - 11 - 20 - 3
Triple locus variant (TLV): 10 - **10** - 11 - **11** - 11 - **2** - 3

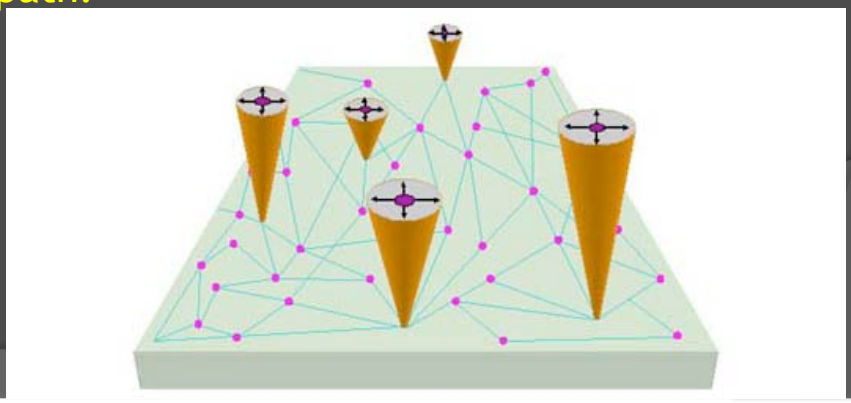
MLST Model

More similar STs should denote closely related strains from an evolutionary point of view.

STs with more SLVs can be regarded has a common ancestor.

Links between STs depict descent relations.

With these assumptions, connected STs should share an evolutionary path.



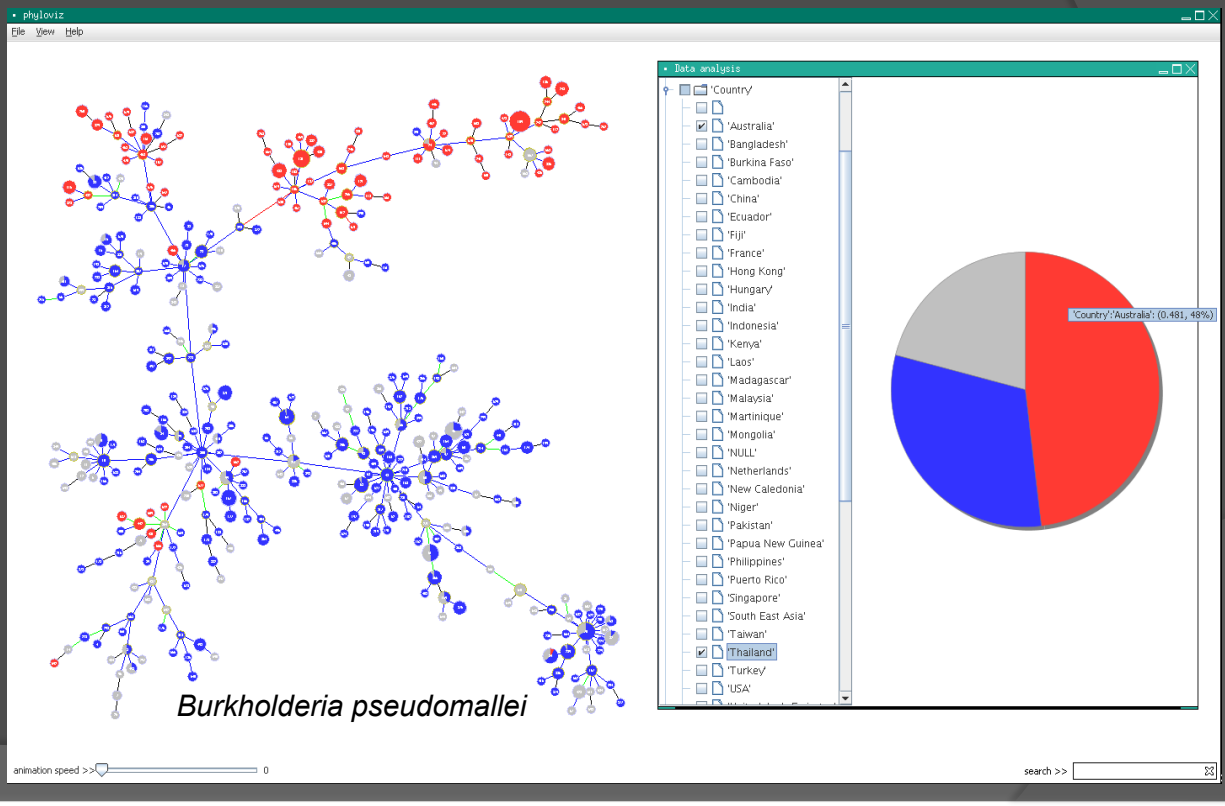
goeBURST

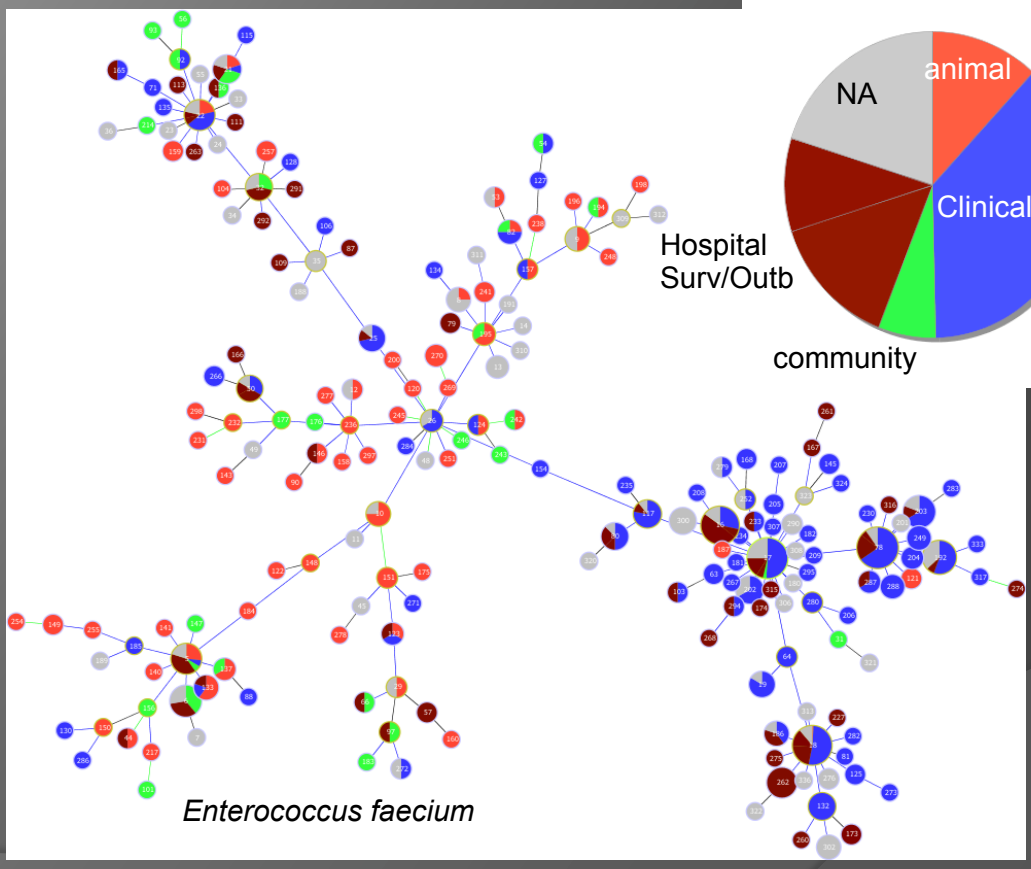
Implements the evolutionary module using the following rules:

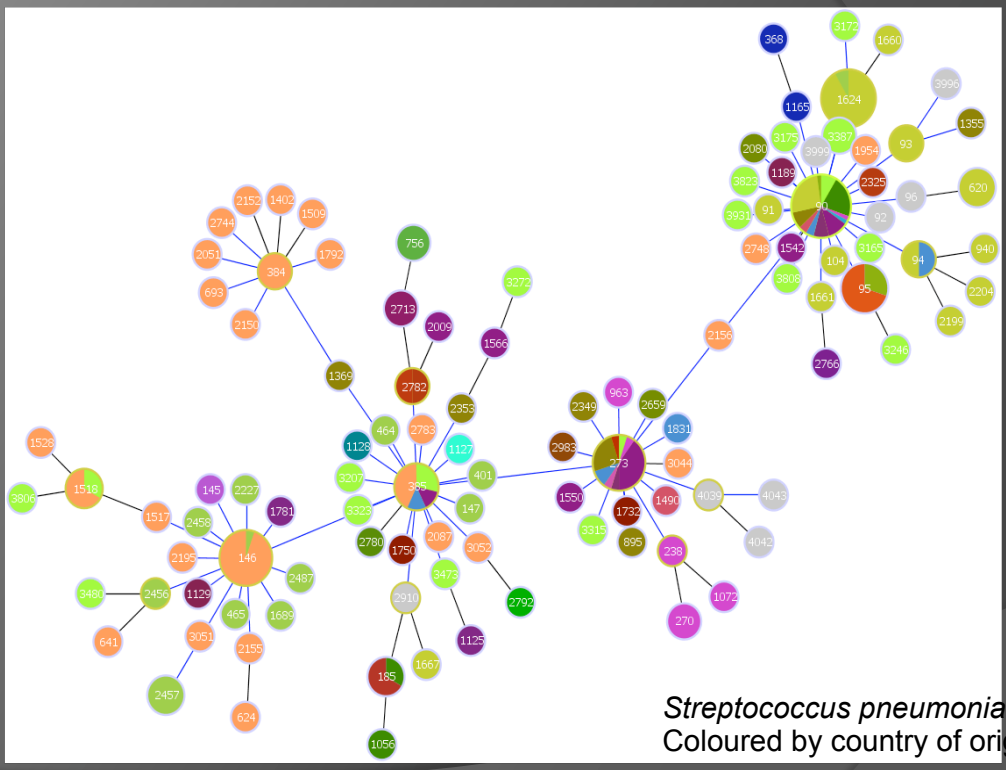
- the ST with the most SLVs connects to all its SLVs
- Repeat this procedure
  - In case of ties, use DLVs / TLVs / ST frequency and ST IDs as tie-breaker
- Proceed until all available STs have been assigned to a clonal complex or are singletons

Phyloviz Live Demo

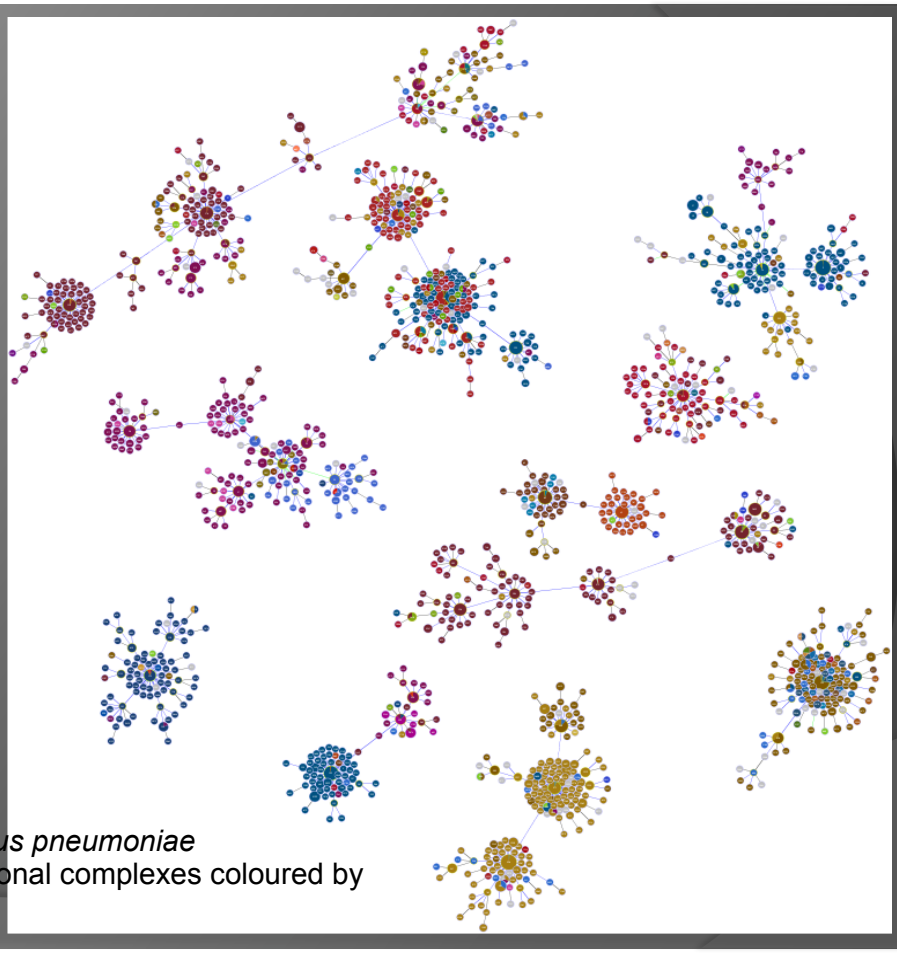
Ongoing work: Phyloviz







Streptococcus pneumoniae CC90
Coloured by country of origin



Streptococcus pneumoniae
10 largest clonal complexes coloured by serotype

The future of microbial typing

- Full genome sequencing + Comparative genomics: The solution?
  - Difficulties assembling the high-throughput reads / complete coverage
  - core genome vs. accessory genome
- Online databases : need for much better annotation and curation , and methods to submit, validate, retrieve and visualize the data

Concluding Remarks

"All models are wrong, Some are useful" – George Box

Data analysis methods can skew the view: we simply need to know how much before drawing the conclusions

Need for novel data analysis methodologies that take recombination and mutation into consideration

Large scale population simulations can lead to new conclusions on how recombination and mutation shape the bacterial populations, and can provide the validation model for new methods.