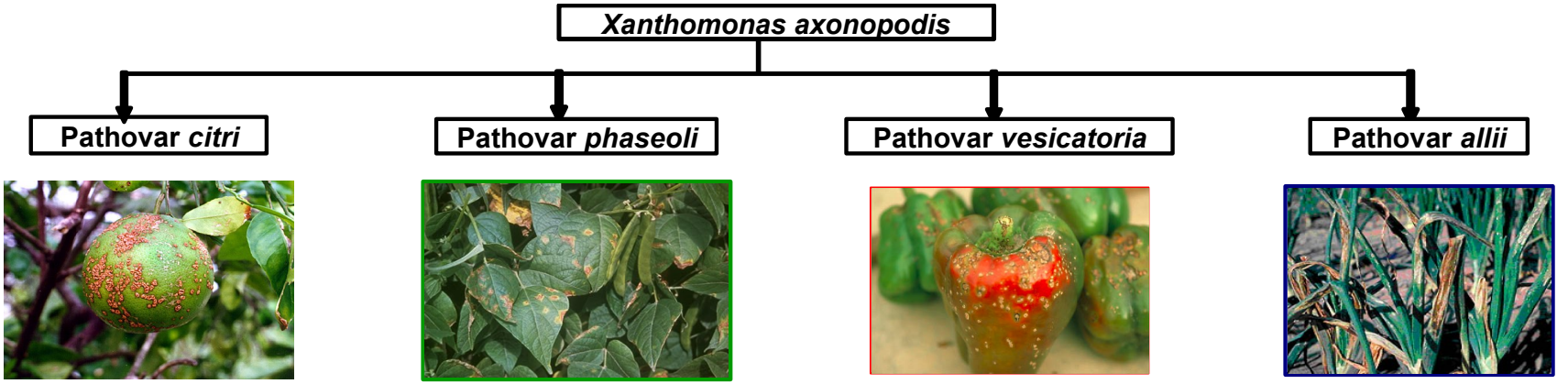
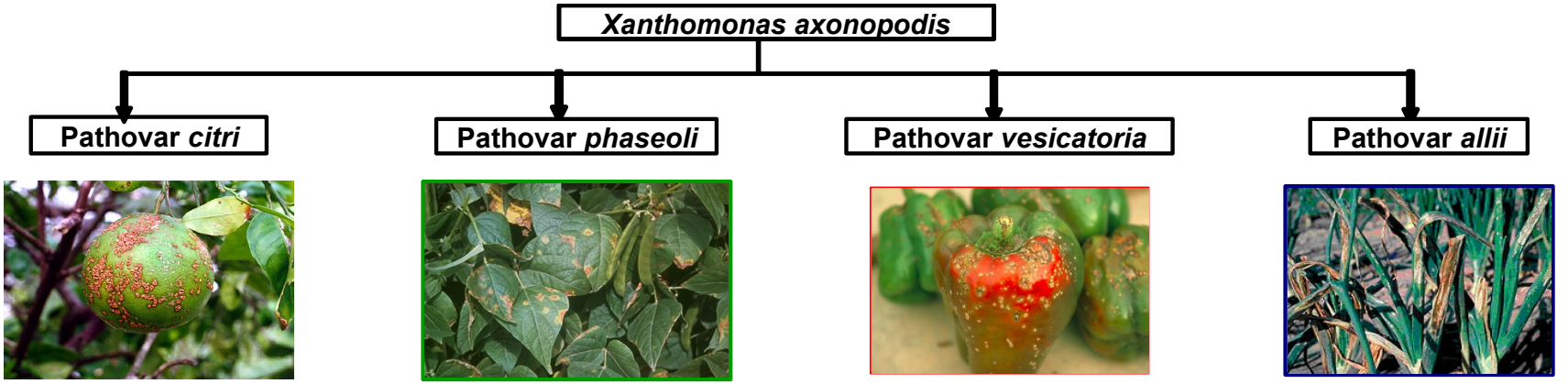
Characterization of Multiple Groups of Data

A. Chambon, T. Boureau, F. Lardeux, F. Saubion, M. Le Saux

November 11, 2015

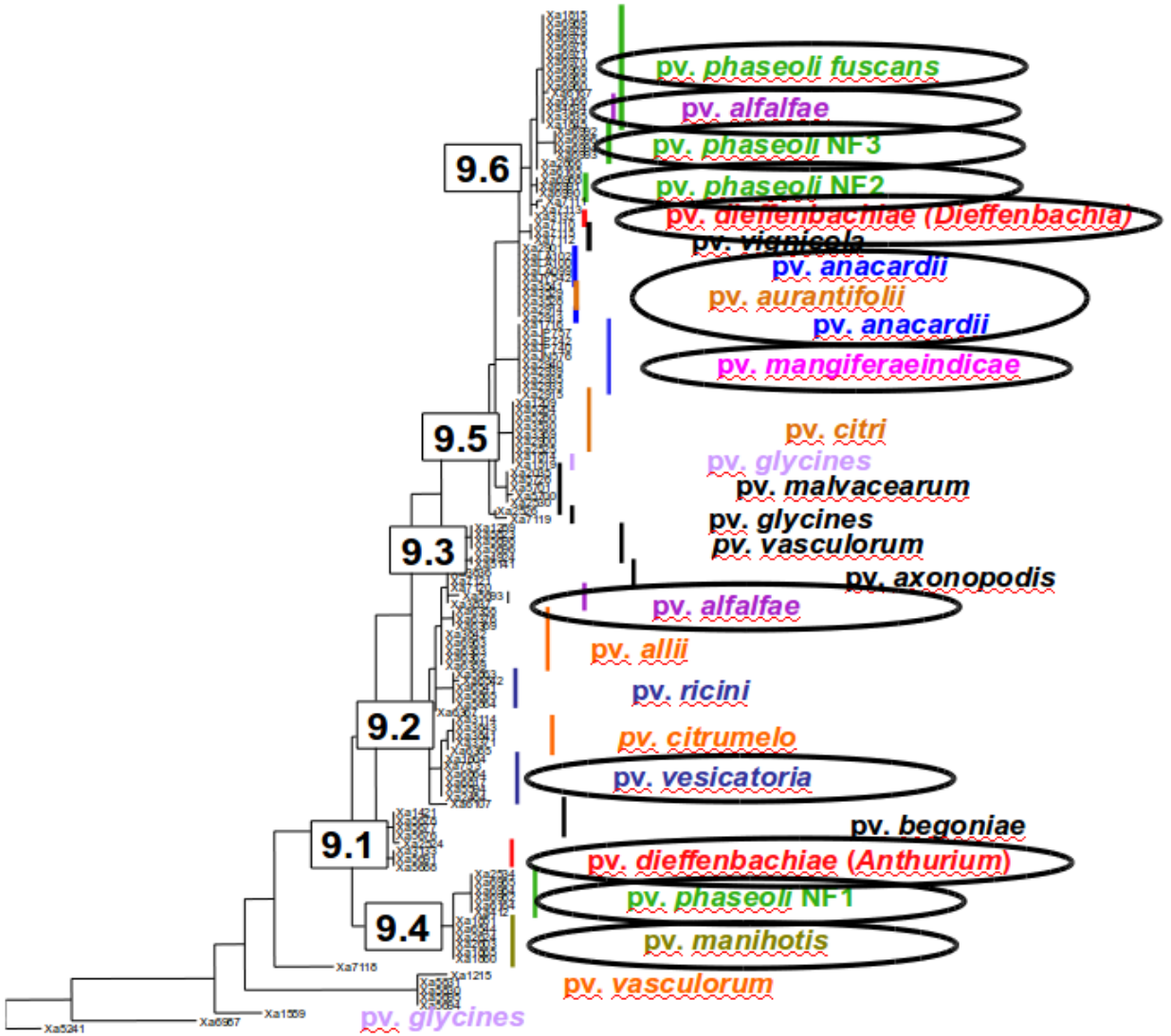


- Xanthomonas bacteria are responsible for diseases on economically important crops.

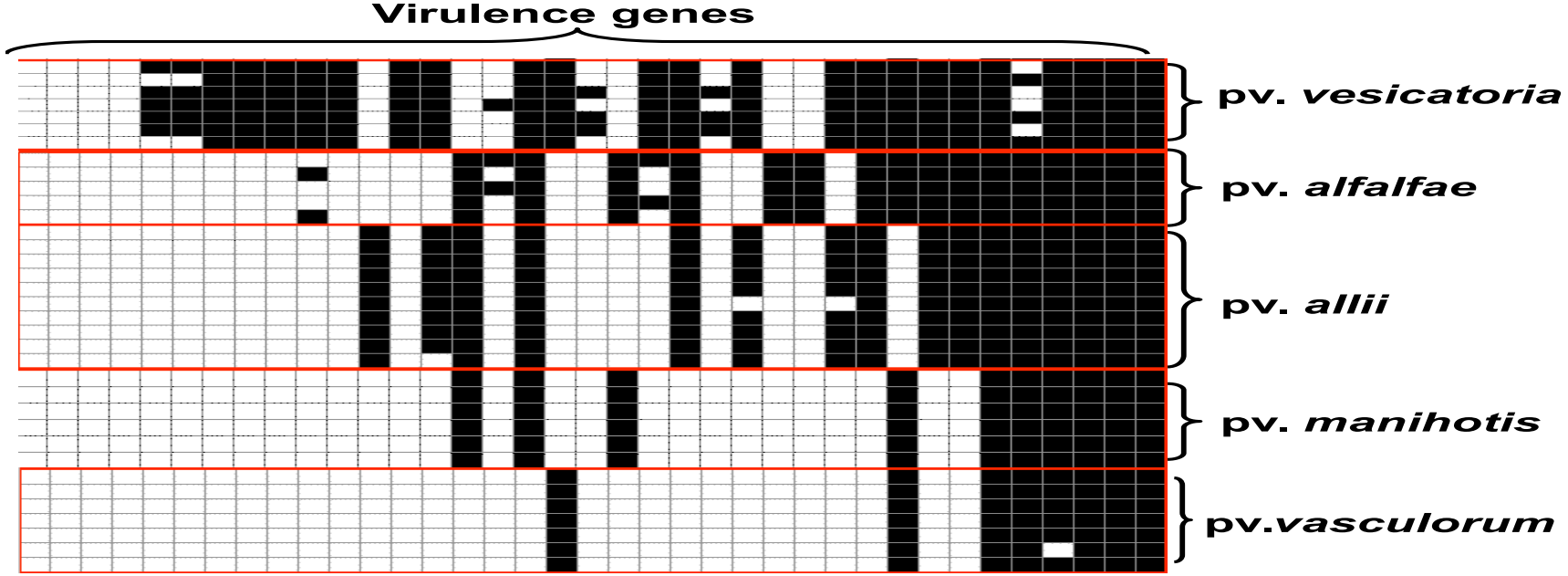


- Xanthomonas bacteria are responsible for diseases on economically important crops.
- The Xanthomonas taxonomy not yet resolved, and the delimitation of species is still under debate.

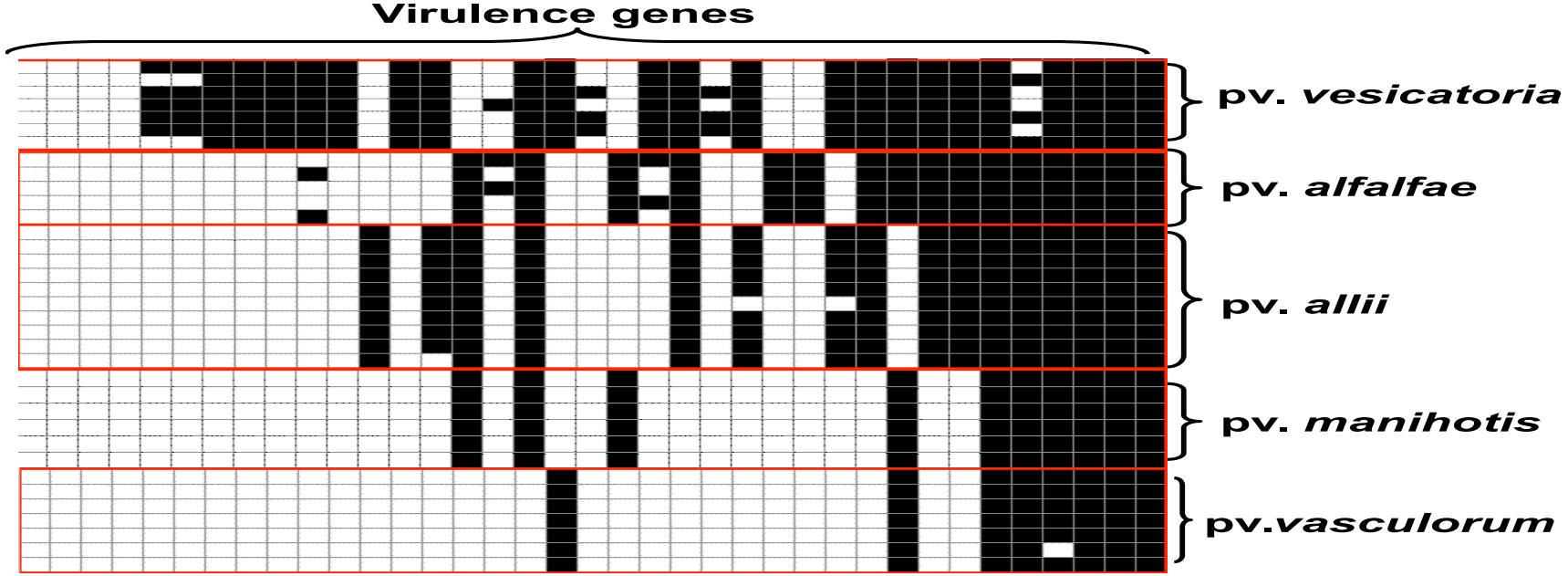
- Phylogenetic approaches aren't really applicable.



- One possible approach: use a virulence gene repository.

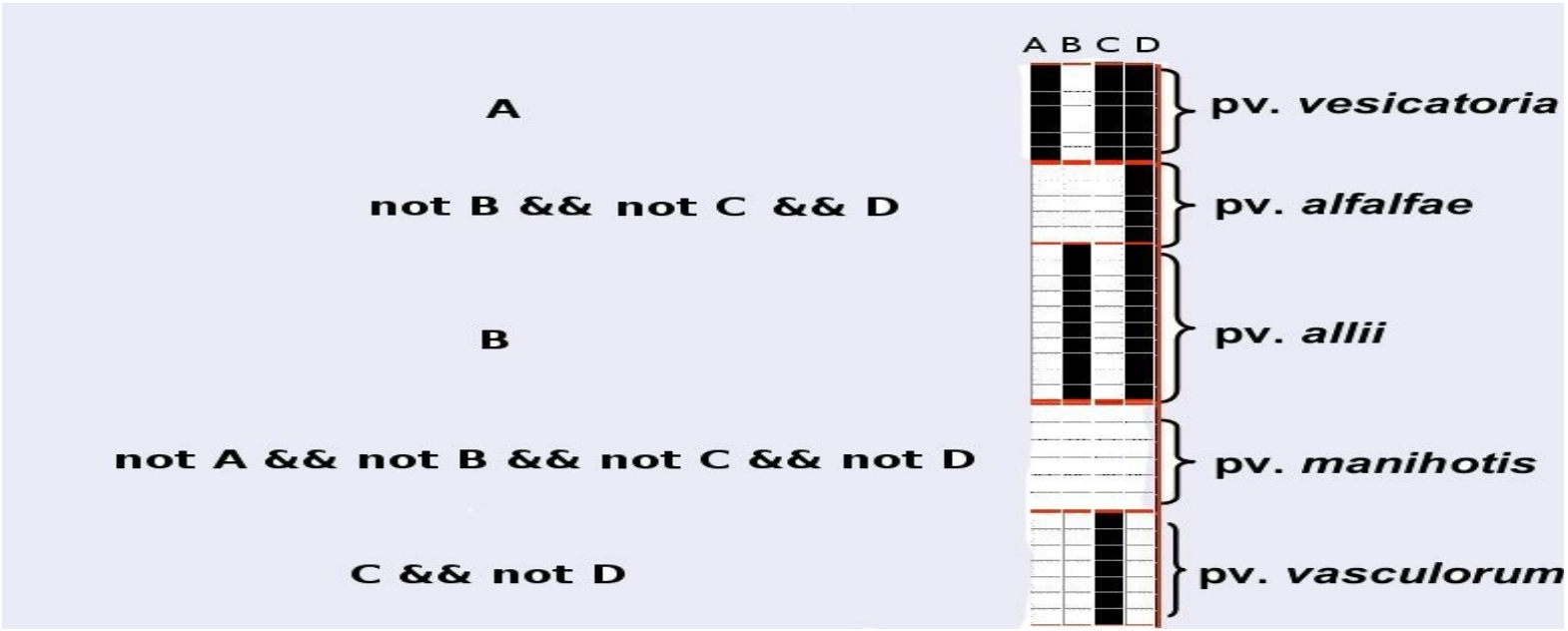


- One possible approach: use a virulence gene repository.



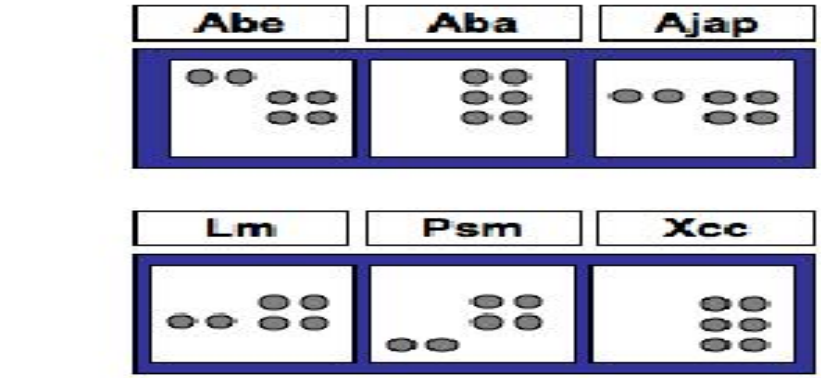
- Goal: Find the smallest combination of virulence genes specific to Xanthomonas.

- One possible approach: use a virulence gene repository.



- Goal: Find the smallest combination of virulence genes specific to Xanthomonas.

- This combination can be used to design a multiplex PCR assay for identification.



- The results show that the combination of molecular tests provides a fast technique for the identification of all pathogenic Xanthomonas strains on beans.

Mathematical representation

We represent these data in the form of a matrix:

Individuals	Characters / Groups	a	b	c	d	e
1	1	0	1	1	0	0
2	2	1	1	1	1	1
3	2	0	1	0	1	1
4	3	0	0	0	1	0
5	4	1	0	1	0	0
6	4	1	1	0	0	1

- Each row i represents an individual (bacterial strains),

Mathematical representation

We represent these data in the form of a matrix:

Individuals	Characters / Groups	a	b	c	d	e
1	1	0	1	1	0	0
2	2	1	1	1	1	1
3	2	0	1	0	1	1
4	3	0	0	0	1	0
5	4	1	0	1	0	0
6	4	1	1	0	0	1

- Each row i represents an individual (bacterial strains),
- Each column j represents a character (genes),

Mathematical representation

We represent these data in the form of a matrix:

Individuals	Characters / Groups	a	b	c	d	e
1	1	0	1	1	0	0
2	2	1	1	1	1	1
3	2	0	1	0	1	1
4	3	0	0	0	1	0
5	4	1	0	1	0	0
6	4	1	1	0	0	1

- Each row i represents an individual (bacterial strains),
- Each column j represents a character (genes),
- The value x_{ij} is 1 if the character j is present in the individual i, 0 otherwise,

Mathematical representation

We represent these data in the form of a matrix:

Individuals	Characters / Groups	a	b	c	d	e
1	1	0	1	1	0	0
2	2	1	1	1	1	1
3	2	0	1	0	1	1
4	3	0	0	0	1	0
5	4	1	0	1	0	0
6	4	1	1	0	0	1

- Each row i represents an individual (bacterial strains),
- Each column j represents a character (genes),
- The value x_{ij} is 1 if the character j is present in the individual i, 0 otherwise,
- Each individual i is associated with a group (pathovars).

- The goal is to remove as much columns as possible.

Individuals	Characters / Groups	a	b	c	d	e
1	1	0	1	1	0	0
2	2	1	1	1	1	1
3	2	0	1	0	1	1
4	3	0	0	0	1	0
5	4	1	0	1	0	0
6	4	1	1	0	0	1

- The goal is to remove as much columns as possible.

Individuals	Characters / Groups	a	b	c	d	e
1	1	0	1	1	0	0
2	2	1	1	1	1	1
3	2	0	1	0	1	1
4	3	0	0	0	1	0
5	4	1	0	1	0	0
6	4	1	1	0	0	1

- Only one rule: two individuals belonging to two different groups have to be different on at least one remaining column.

- The goal is to remove as much columns as possible.

Individuals	Groups \ Characters	a	b	c	d	e
1	1	0	1	1	0	0
2	2	1	1	1	1	1
3	2	0	1	0	1	1
4	3	0	0	0	1	0
5	4	1	0	1	0	0
6	4	1	1	0	0	1

- Only one rule: two individuals belonging to two different groups have to be different on at least one remaining column.
- A solution is therefore a subset of columns according to this rule.

- The goal is to remove as much columns as possible.

Individuals	Groups \ Characters	b	d	e
1	1	1	0	0
2	2	1	1	1
3	2	1	1	1
4	3	0	1	0
5	4	0	0	0
6	4	1	0	1

- Only one rule: two individuals belonging to two different groups have to be different on at least one remaining column.
- A solution is therefore a subset of columns according to this rule.
- A solution could be {b,d,e}.

- Note that if we find a solution, adding a column to this solution will give a new solution.
- This new solution is dominated because a smaller solution that is a subset of this new solution exists.

For example the solution {a,b,d,e} is dominated by the solution {b,d,e}:

Gr \ Char	a	b	d	e
1	0	1	0	0
2	1	1	1	1
2	0	1	1	1
3	0	0	1	0
4	1	0	0	0
4	1	1	0	1

$\prec$

Gr \ Char	b	d	e
1	1	0	0
2	1	1	1
2	1	1	1
3	0	1	0
4	0	0	0
4	1	0	1

Objectives:

- Determine the set of non-dominated solutions (MCP).
- Determine the smallest solutions (Argmin-MCP).

MCP

To solve the Multiple Characterization Problem (MCP), we create a new matrix such as:

- Each line represents two individuals from two different groups,
- Each column determines whether the character differentiates individuals or not.

For example, to compare the 1st and 3rd individual:

Ind	Gr \ Char	a	b	c	d	e
1	1	0	1	1	0	0
2	2	1	1	1	1	1
3	2	0	1	0	1	1
4	3	0	0	0	1	0
5	4	1	0	1	0	0
6	4	1	1	0	0	1

Ind	Gr \ Char	a	b	c	d	e
1	1	0	1	1	0	0

XOR

Ind	Gr \ Char	a	b	c	d	e
3	2	0	1	0	1	1

⇓

Ind \ Char	a	b	c	d	e
1-3	0	0	1	1	1

Thus, if we compare among themselves all individuals:

Comparisons \ Characters	a	b	c	d	e
1-2	1	0	0	1	1
1-3	0	0	1	1	1
1-4	0	1	1	1	0
1-5	1	1	0	0	0
1-6	1	0	1	0	1
2-4	1	1	1	0	1
2-5	0	1	0	1	1
2-6	0	0	1	1	0
3-4	0	1	0	0	1
3-5	1	1	1	1	1
3-6	1	0	0	1	0
4-5	1	0	1	1	0
4-6	1	1	0	1	1

- Finding a solution is equivalent to selecting a color for each line.
- We want to avoid getting dominated solutions.
- If a row has all the colors of another, then it is redundant.
- We can ignore these redundant lines.

After removing the redundant lines (known as dominated):

Comparisons \ Character	a	b	c	d	e
1-5	1	1	0	0	0
1-6	1	0	1	0	1
2-6	0	0	1	1	0
3-4	0	1	0	0	1
3-6	1	0	0	1	0

- This new table is the constraint matrix.
- This is the constraints of the linear program associated to the problem.

$$min : \sum_{i=1}^{m} y_i$$

$$s.t. :$$

$$CM \ . \ Y^t \geqslant \mathbb{1}^t$$

$$Y \in \{0,1\}^m = [y_1, ..., y_i, ..., y_m]$$

- Problem simplified, but still not resolved.
- Solve the problem by linear programming is still very long.

One method to reduce the number of lines is to merge the lines between them.
The idea to merge two lines is simple:

- If the first row is : $a \vee b$
- And if the second is : $a \vee c \vee e$
- Satisfying both rows requires : $(a \vee b) \wedge (a \vee c \vee e)$

Now:

$$(a \vee b) \wedge (a \vee c \vee e) = a \vee (a \wedge c) \vee (a \wedge e) \vee (a \wedge b) \vee (b \wedge c) \vee (b \wedge e)$$

Then, by applying the concept of dominance seen previously, we have:

$$a \vee \cancel{(a \wedge c)} \vee \cancel{(a \wedge e)} \vee \cancel{(a \wedge b)} \vee (b \wedge c) \vee (b \wedge e)$$

a , $(b \wedge c)$,$(b \wedge e)$ and the combinations dominated by those ones are the combinations satisfying the first two rows.

- The idea is to merge in this way each row of the constraint matrix.
- It returns all non-dominated solutions satisfying all lines.
- The problem is solved.
- The algorithm we propose also allows to avoid creating dominated solutions during these mergers.

Argmin-MCP

- Another problem is to find the smallest solutions (Argmin-MCP).
- The idea is to use the same method, removing solutions bigger than a given bound.
- If we don't find solution, we increment the bound.

Advantages:

- No need to verify dominance in this problem.
- The algorithm we propose can be parallelized and avoids identical solutions.
- Very low cost in memory.

Experiments

- For experimentation, we tested our algorithm on real biological data.
- Instance parameters are given by the table below.

Instances	ind*char	# gr.	# constr.	Kmin
ra100_phy	113*99	4	3479	7
ra100_phv	109*99	8	4146	9
rch8	132*37	21	8229	9
ra_rep1	112*155	7	4587	12
ra_rep2	112*155	7	4609	12
ra_phy	113*155	4	3479	6
920kmers	18*920	6	96	5

- We compare our algorithm for Argmin-MCP with CPLEX.
- For optimize parameters of CPLEX we use “tuning tool”.
- Our algorithm and CPLEX found all solutions.
- CPLEX is parallelized but for the experimentation, our algorithm isn’t.

Results:

Instance	Kmin	#Opt sol	Time algo argmin-MCP	CPLEX LP
ra100_phy	7	63	0.29	11.00
ra100_phv	9	119	1.23	17.00
rch8	9	16	0.11	1.66
ra_rep1	12	14	0.18	2.96
ra_rep2	12	1298	3.93	23.58
ra_phy	6	4	2.63	15.94
ra_phv	6	6	2.37	15.83
920kmers	5	1151863	90.90	-

- About our algorithm for MCP, it found all non-dominated solutions.
- CPLEX can't handle the dominance.
- We can't compare CPLEX with our algorithm.

Results:

Instance	Kmin	# Solution	Time algo
ra100_phy	7	860	10.23
ra100_phv	9	3674	76.18
rch8	9	222	0.48
ra_rep1	12	37382	113.00
ra_rep2	12	42067	132.04
ra_phy	6	4576	3719.52
ra_phv	6	1853	3760.59
920kmers	5	1151863	194.55

Conclusion and future prospects

- Previous work give us only one solution.
- We can now have all the minimum solutions in a time roughly similar.
- Many solutions means a lot of information.
- We plan to continue work assisting biologists in interpreting these results.