MATHÉMATIQUES VISION APPRENTISSAGE

University of Ljubljana Faculty of Electrical Engineering

NATIONAL INSTITUTE OF BIOLOGY

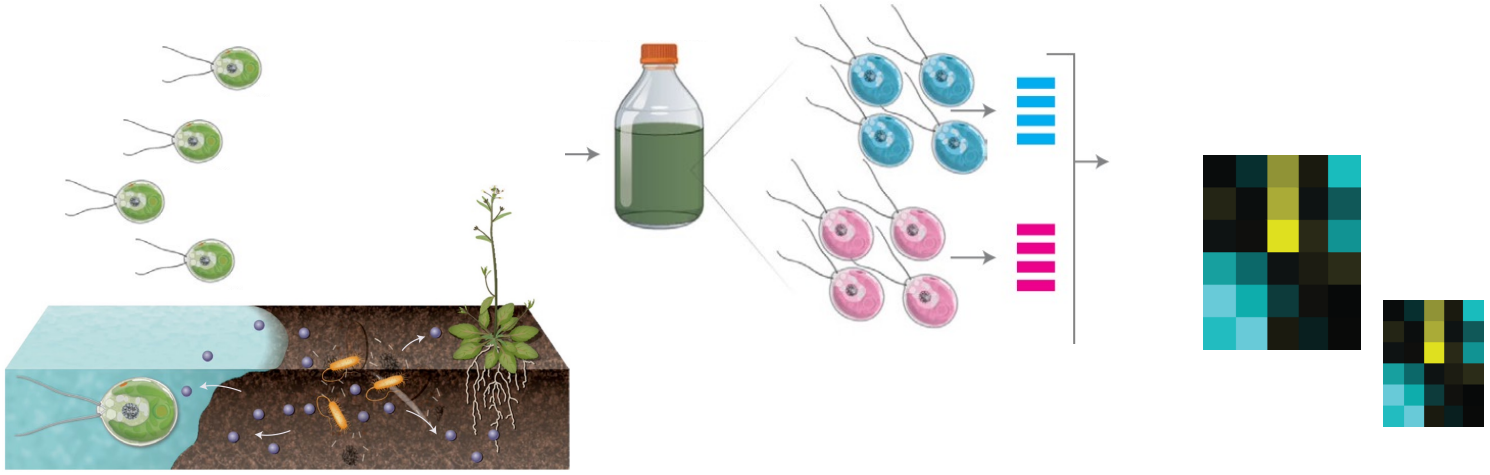
IDA Institut d'Alembert

LuMIn Lumière, Matière et Interfaces

Identifying the genetic determinants of chemical toxicity in microalga Chlamydomonas reinhardtii thanks to statistical analysis – Gaussian mixture model estimation

M2 Internship from 10/04/2023 – 10/08/2023

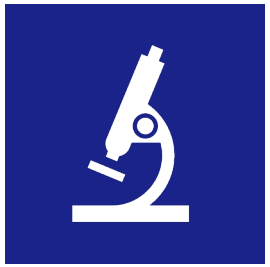
Mélanie PIETRI



école normale supérieure paris–saclay

université PARIS-SACLAY

CONTENT



CONTEXT AND STATE OF THE ART

Microalgae in academic and industrial research

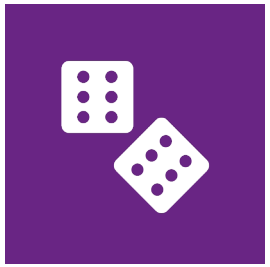
Protocol

Tools in statistical analysis

EXISTING METHOD

Description

Analysis



PROPOSED METHOD

Estimation of gaussian mixture model

Simulation

RESULTS AND DISCUSSION

Estimated data

Simulation

Perspectives

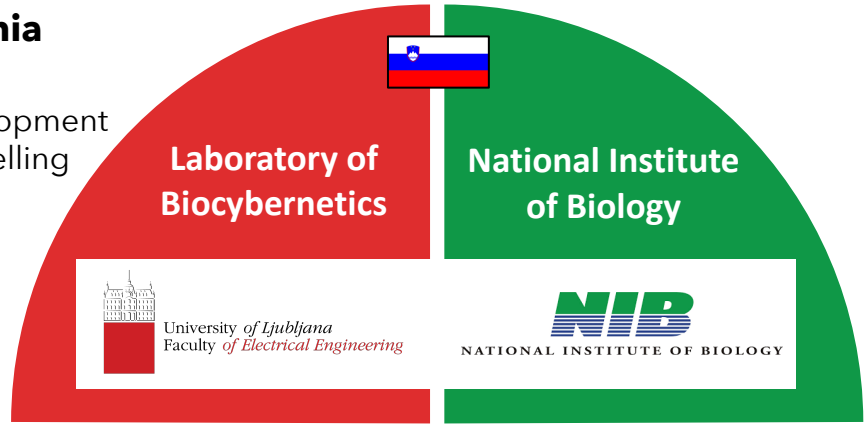
1

CONTEXT AND STATE OF THE ART

LABORATORIES IN THE PROJECT

2 months in Slovenia

Hardware development
Numerical modelling
Electroporation



Marine ecology and terrestrial biology
Cancer biology
Food security and plant protection

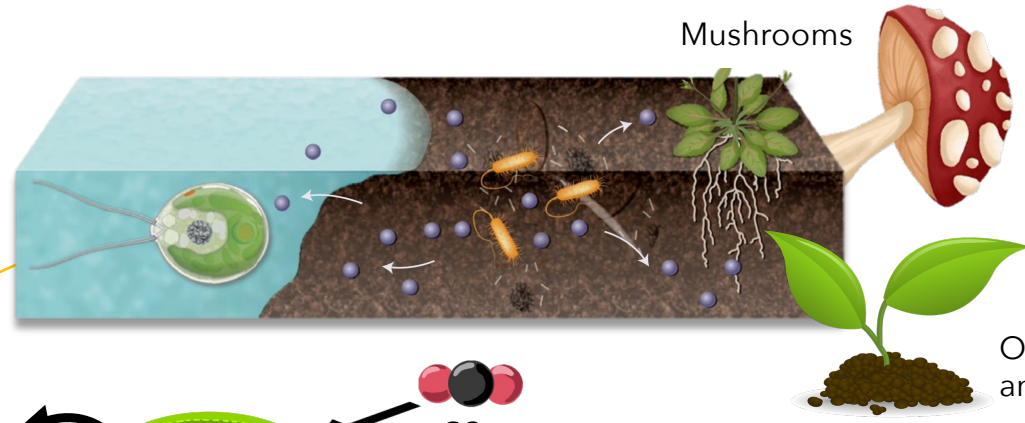
Biophotonic imaging
Microfluidic devices
Biofuels

2 months in France

Applied physics
Engineering
Applied mathematics

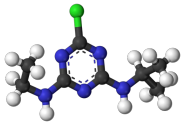
MICROALGAE AND VALUABLE COMPOUNDS

Chlamydomonas reinhardtii interacting with its natural environment (adapted from **[1]**)

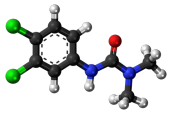


Biochemical compounds

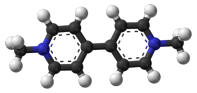
Atrazine

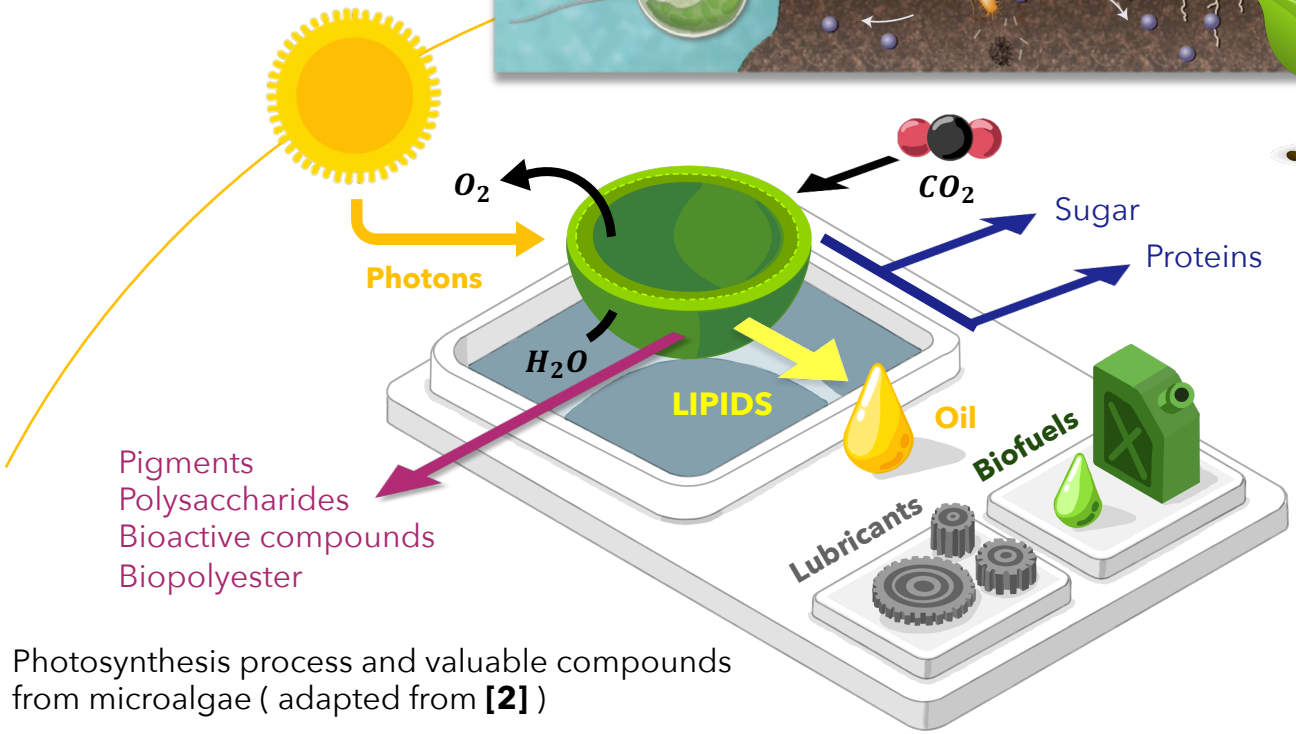


Diuron



Paraquat





Photosynthesis process and valuable compounds from microalgae (adapted from **[2]**)

Open pond for microalgae industrial culture

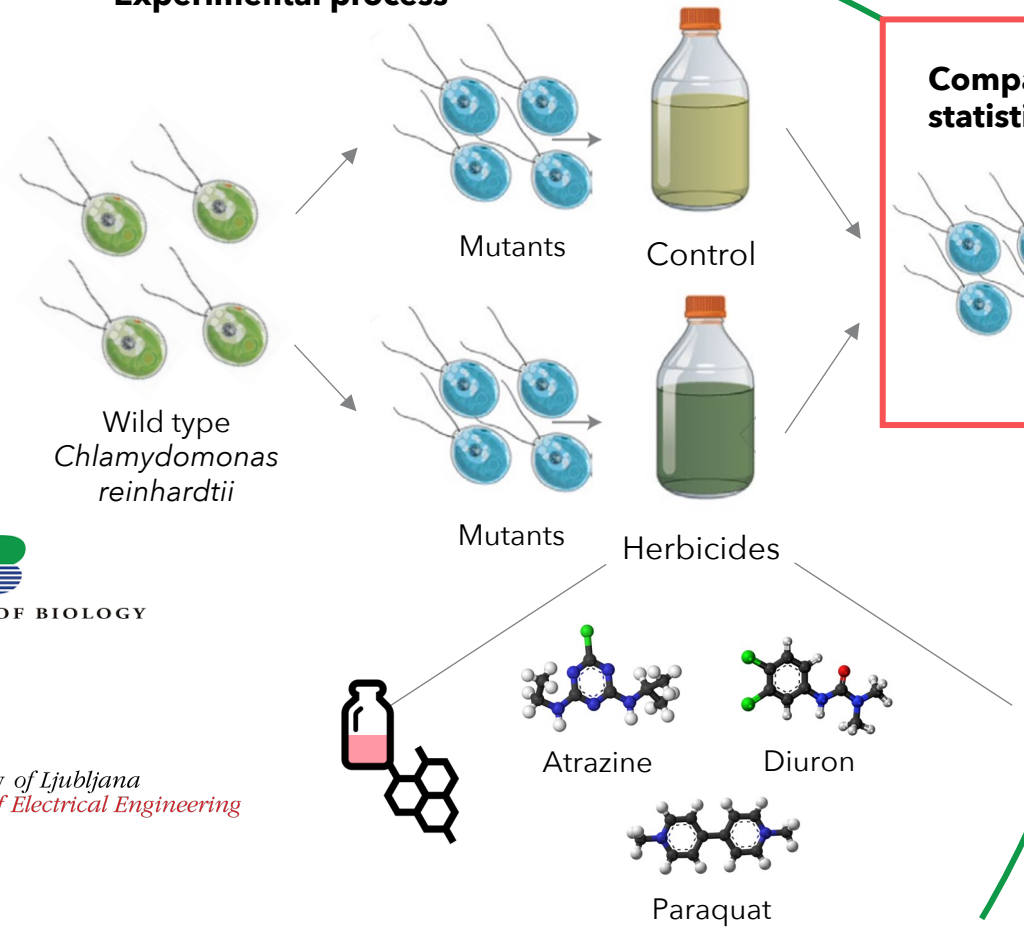
Proposal of the internship

Study of the impact of herbicides on microalga *Chlamydomonas reinhardtii*

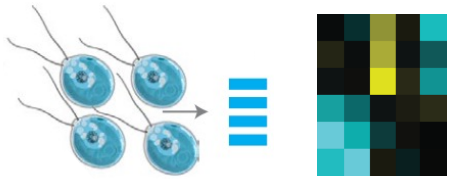
[1] Systematic characterization of gene function in photosynthetic alga Chlamydomonas reinhardtii – F. Fauser et al. (2022)
[2] *Microalgues de la recherche à l'industire – CEA Cadarache – Dossier de presse 31/01/2020*

HOW TO STUDY THE IMPACT OF HERBICIDE ON MICROALGAE ?

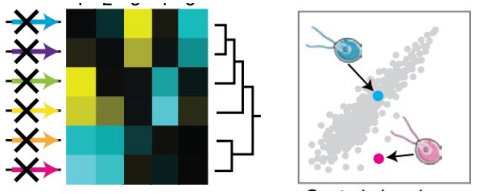
Experimental process



Comparison by statistical analysis



Determination of sensitive or tolerant genes to the herbicide



Other tests

NATIONAL INSTITUTE OF BIOLOGY

University *of Ljubljana*
Faculty *of Electrical Engineering*

IDA Institut d'Alembert

LuMIn
Lumière, Matière et Interfaces

[1] Systematic characterization of gene function in the photosynthetic alga Chlamydomonas reinhardtii – F. Fauser et al. (2022)

STATISTICAL TEST IN LITTERATURE

Hypothesis testing

$\mathcal{H}_0$

Absence of associations between data

$\mathcal{H}_1$

Alternative hypothesis

Two types of errors

Type I error

False positive (FP)

$\mathcal{H}_0$ can be true but it is rejected

Type II error

False negative (FN)

$\mathcal{H}_1$ can be true but the test does not reject $\mathcal{H}_0$

Statistical tests

Parametric tests

Hypothesis: the data follows a specific distribution

Gaussian distribution $\mathcal{N}(\mu, \sigma)$

Student's t-test (1908)

ANOVA (1925)

Non-parametric tests

$\mathcal{X}^2$ Pearson's test (1900)

Fisher's exact test (1922)

Kolmogorov-Smirnov (1933)

Wilcoxon-Mann-Whitney (1947)

FOCUS ON χ^2 AND FISHER'S EXACT TEST APPLICATION

Dataset format

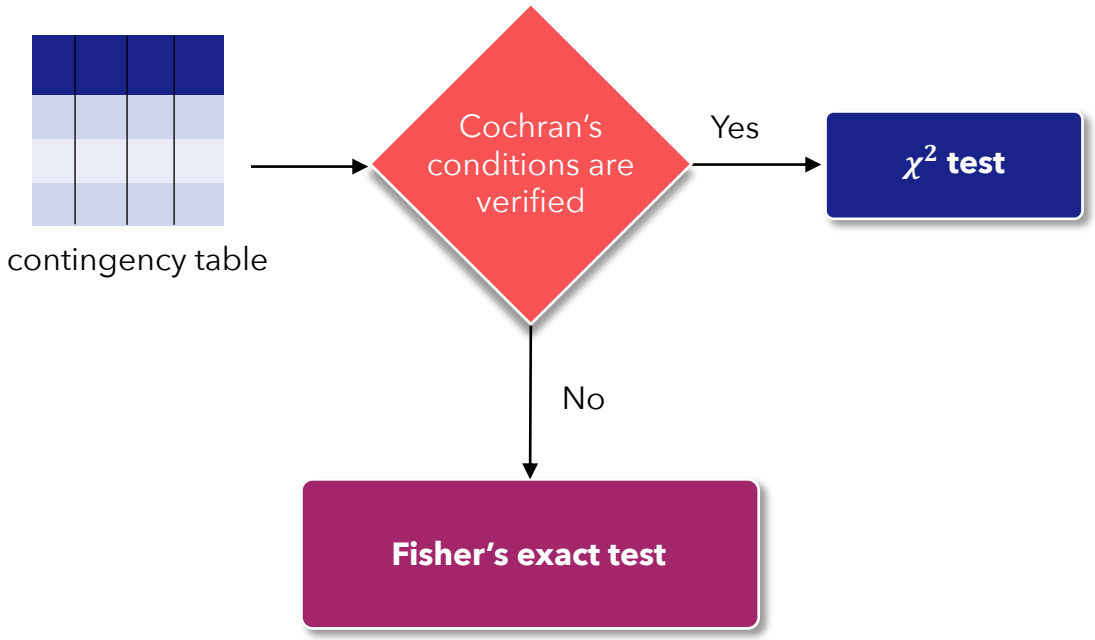
Let $\mathbf{X} \in \mathcal{M}_{r,c}(\mathbb{N})$ a contingency table. $x_{i,j} \in \mathbb{N}$ for $(i,j) \in [\![1,r]\!] \times [\![1,c]\!]$

R/C	Column 1	⋯	Column j	⋯	Column c
Row 1	$x_{1,1}$	⋯	$x_{1,j}$	⋯	$x_{1,c}$
⋯	⋯	⋯	⋯	⋯	⋯
Row i	$x_{i,1}$	⋯	x_{ij}	⋯	$x_{i,c}$
⋯	⋯	⋯	⋯	⋯	⋯
Row r	$x_{r,1}$	⋯	$x_{r,j}$	⋯	$x_{r,c}$

Cochran's conditions:

$$\forall (i,j) \in [\![1,r]\!]\times[\![1,c]\!], \mathbb{P}\left(x_{i,j} < 1\right) = 0 \qquad (1)$$

$$\frac{\mathrm{Card}\left(\left\{(i,j) \in [\![1,r]\!]\times[\![1,c]\!] \mid x_{i,j} \geq 5\right\}\right)}{r \times c} \geq 80\% \qquad (2)$$



Alternative to the χ^2 test for small sample size

$$\frac{\mathrm{Card}\left(\left\{(i,j) \in [\![1,r]\!]\times[\![1,c]\!] \mid x_{i,j} \leq 5\right\}\right)}{r \times c} \geq 20\%$$

[3] A network fisher.test: Fisher's Exact Test for Count Data – RDocumentation, stats (version 3.6.2)

FISHER'S EXACT TEST

R/C	Column 1	⋯	Column j	⋯	Column c
Row 1	$x_{1,1}$	⋯	$x_{1,j}$	⋯	$x_{1,c}$
⋯	⋯	⋯	⋯	⋯	⋯
Row i	$x_{i,1}$	⋯	x_{ij}	⋯	$x_{i,c}$
⋯	⋯	⋯	⋯	⋯	⋯
Row r	$x_{r,1}$	⋯	$x_{r,j}$	⋯	$x_{r,c}$

$\mathbf{X}$: contingency table for variables R and C for $(i,j) \in [\![1,r]\!] \times [\![1,c]\!]$

Let $\mathcal{T}$ the set of possible $r \times c$ contingency tables

$$\mathcal{T} = \left\{ \mathbf{Y} : \mathbf{Y} \in \mathcal{M}_{r,c}(\mathbb{N}), \sum_{j=1}^{c} y_{ij} = R_i \,, \sum_{i=1}^{r} y_{ij} = C_j \right\}$$

Step 1: $\mathcal{H}_0$ hypothesis of independance between R and C

The probability of observing any $\mathbf{X} \in \mathcal{T}$ can be expressed as

$$\mathbb{P}(\mathbf{X}) = \left(\prod_{j=1}^{c} \frac{C_j!}{x_{1j}!\, x_{2j}!\, \ldots x_{rj}!} \right) \Big/ \frac{T!}{R_1!\, R_2!\, \ldots R_r!} \quad \text{with} \quad T = \sum_{i=1}^{r} R_i$$

Step 2: Computation of the p value

$$\mathbf{p} = \sum_{Y \in \mathcal{S}} \mathbb{P}(\mathbf{Y}) \quad \text{where} \quad \mathcal{S} = \{ \mathbf{Y} : \mathbf{Y} \in \mathcal{T} \mid \mathbb{P}(\mathbf{Y}) \leq \mathbb{P}(\mathbf{X}) \}$$

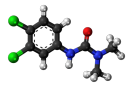
Step 3: Decision

Given a level of decision α , if $\mathbf{p} < \alpha$ then $\mathcal{H}_0$ is rejected

Usually $\alpha = 0.05$

On biological data:

Diuron 1



Genes	Effect of Diuron 1	No effect of Diuron 1	Sum
Cre01.g000200	2	4	6
All genes	3 295	34 837	38 132
Sum	3 297	34 841	38 138

$\mathcal{H}_0$: « Herbicide Diuron 1 does not have an effect on gene Cre01.g000200 »

$\mathbf{p} = 0.08865 > \alpha$ with Fisher's exact test

$\mathbf{p} = 0.154 > \alpha$ with χ^2 test

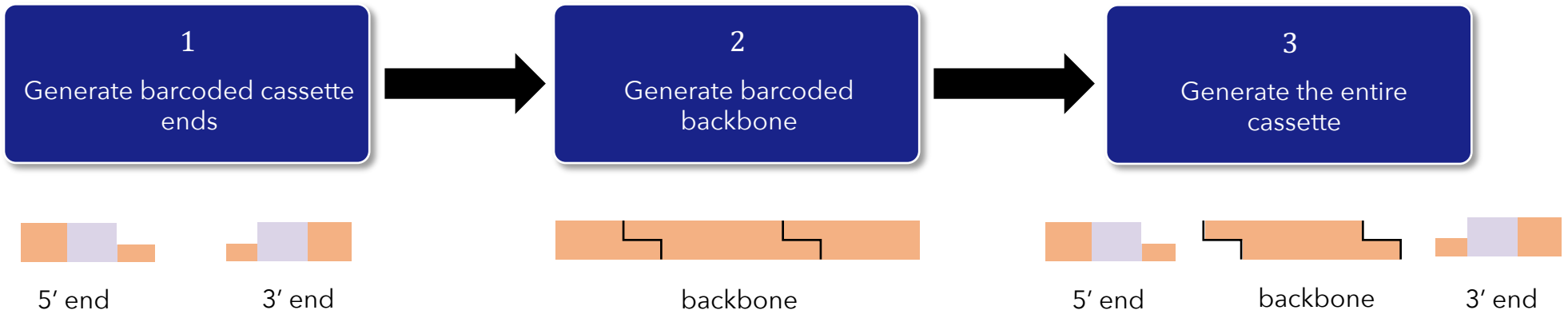
[4] A network algorithm for performing Fisher's exact test in r × c contingency tables – C. Mehta et al. (1983)

2

EXISTING METHOD

CREATION OF MUTANT LIBRARY

Genetic engineering



Characteristics	Big library	Small subset of library	Our library
Number of mutants	67 000	9 686	38 132
Number of genes	18 202	5 769	11 622
Average mutants per gene	7.8	1.68	3.28
Median mutants per gene	4	1	2

Challenge

Number of mutants per gene is relatively small

[5] A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis – X. Li et al. (2019)

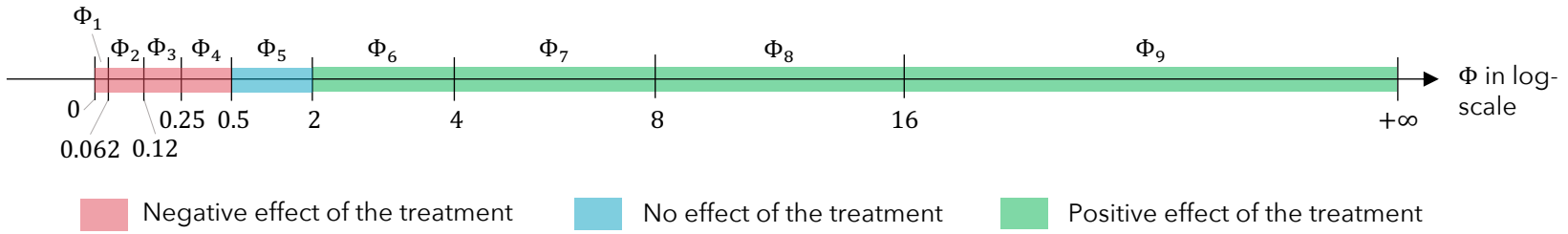
SAMPLING MUTANT PHENOTYPE AND EXISTING METHOD

Experimental measure

Mutant phenotype ratio: $\Phi = \frac{N_t}{N_c}$

N_t: abundance of mutant after growth under treatment

N_c: abundance of mutant after growth under control condition



Obtained data

Example of contingency table for Atrazine 1 treatment

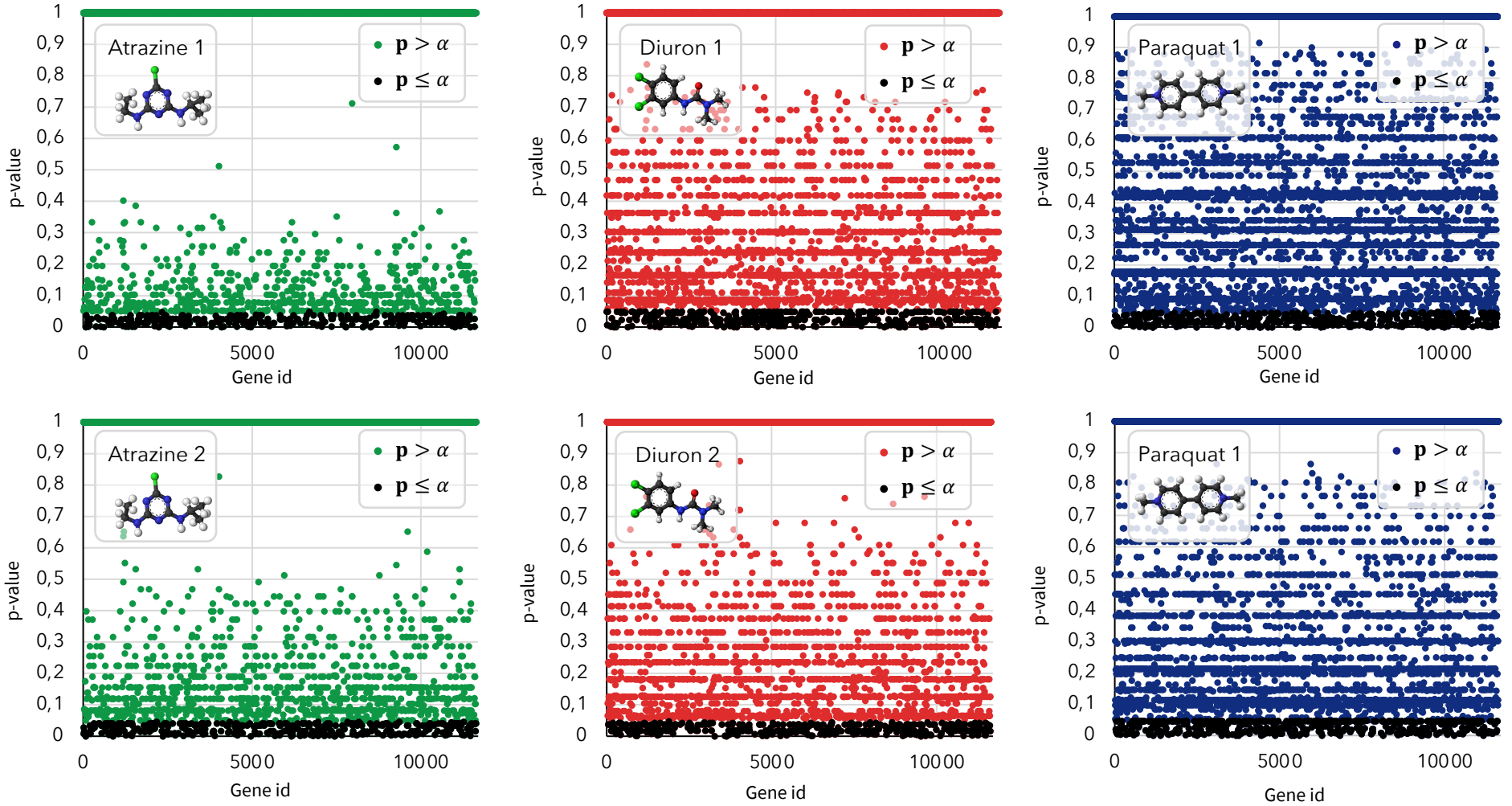
	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6	Φ_7	Φ_8	Φ_9
Cre01.g000033	0	0	0	0	1	0	0	0	0
Cre01.g000100	0	0	0	0	2	0	0	0	0
…	…	…	…	…	…	…	…	…	…
Cre03.g176325	0	1	0	2	4	0	0	0	0
…	…	…	…	…	…	…	…	…	…
Cre09.g396850	0	0	0	0	1	0	1	0	0
…	…	…	…	…	…	…	…	…	…
All genes	4	10	29	535	37114	440	1	0	0

Fisher's exact test

p-value		
1	$\mathbf{p} > \alpha = 0.05$	
1	$\mathbf{p} > \alpha = 0.05$	
…		
0.00028	$\mathbf{p} < \alpha = 0.05$	INTERESTING GENE
…		
0.00025	$\mathbf{p} < \alpha = 0.05$	INTERESTING GENE
…		

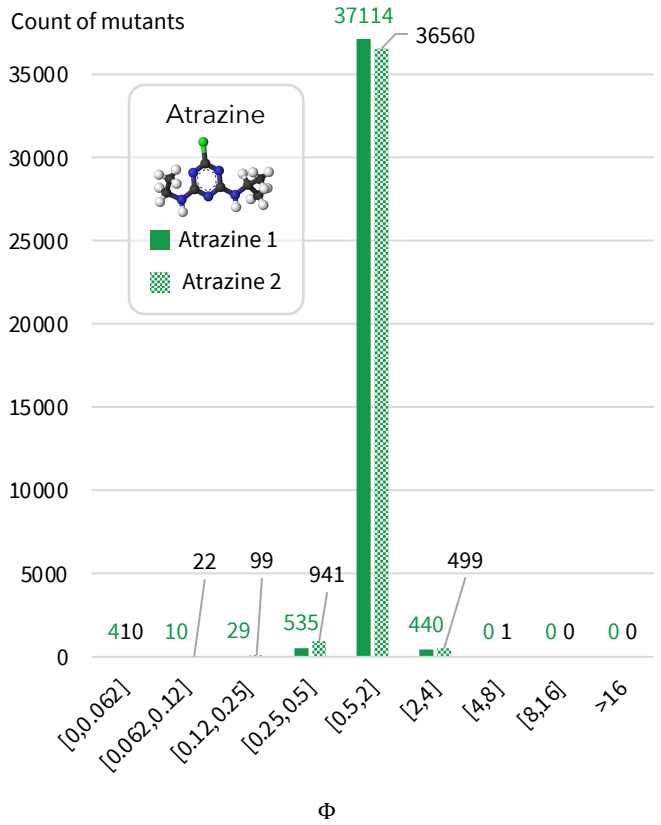
[1] Systematic characterization of gene function in photosynthetic alga Chlamydomonas reinhardtii – F. Fauser et al. (2022)

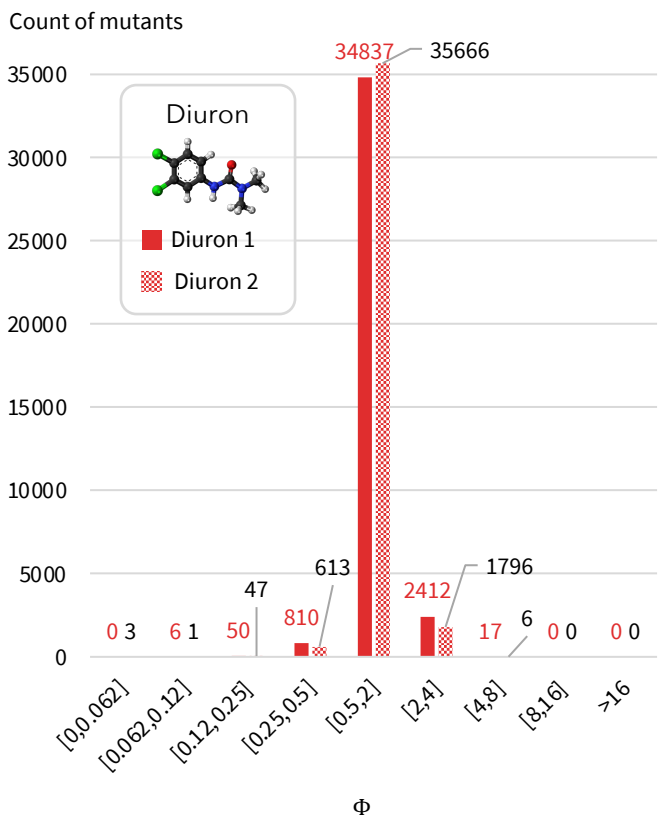
EVALUATION OF P-VALUE

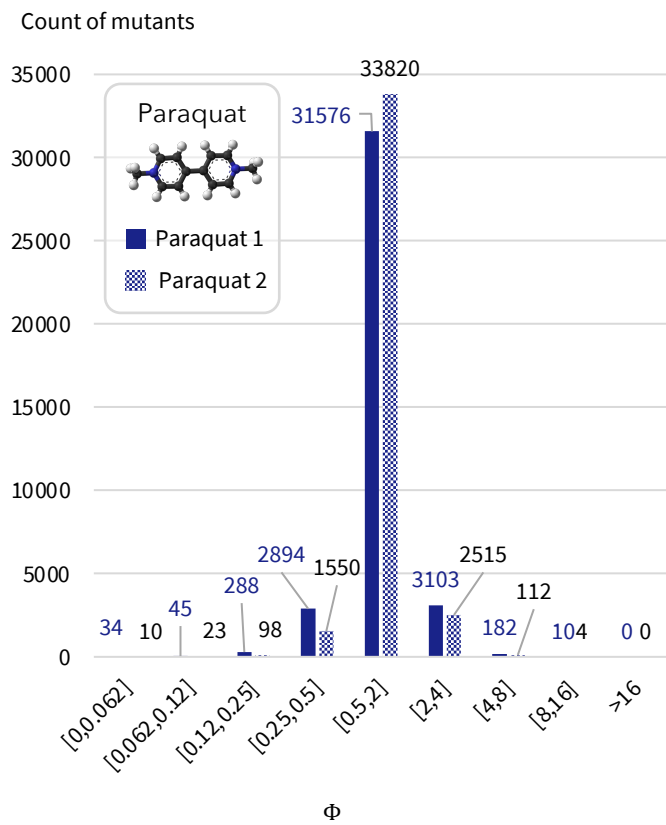


HISTOGRAMS

Cumulative histograms for each treatment







A majority of mutants are in Φ_5: herbicides do not have an impact on a large number of genes

METHOD ANALYSIS

Discrimination between positive and negative effect

With p-value we do not know if the effect of the treatment is positive or negative

Example: for Atrazine 1

Cre03.g176325	0	1	0	2	4	0	0	0	0
All genes	4	10	29	535	37114	440	1	0	0

→ $\mathbf{p} = 0.00028$

Cre09.g396850	0	0	0	0	1	0	0	0	0
All genes	4	10	29	535	37114	440	1	0	0

→ $\mathbf{p} = 0.00025$

Fisher's exact test

Genes	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6	Φ_7	Φ_8	Φ_9
Cre01.g000033	0	0	0	0	1	0	0	0	0
All genes	4	10	29	535	37114	440	1	0	0

Verified conditions:

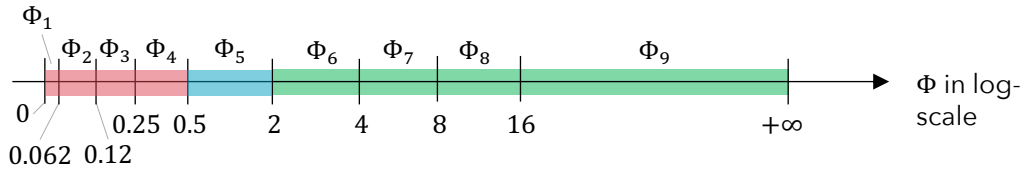
$\forall (i,j) \in [\![1,r]\!]\times[\![1,c]\!], \mathbb{P}(x_{i,j} < 1) = 0$ **BUT**

$\frac{\mathrm{Card}(\{(i,j) \in [\![1,r]\!]\times[\![1,c]\!] \mid x_{i,j} \leq 5\})}{r \times c} \geq 20\%$

Huge gap between numbers in the table

Fisher's exact test is approximated because too costly computationally speaking

Intervals partitioning



Negative effect of the treatment

Positive effect of the treatment

No effect of the treatment

- Is the sampling with Φ ratios well adapted to the experiments ?
- Sampling mutant phenotype seems to be adapted to Fisher's exact test because of low number of mutants per gene
- In particular, is it reasonable to assert « if mutants are in $\Phi_5 = [0.5, 2]$ there is no effect of the treatment » ?

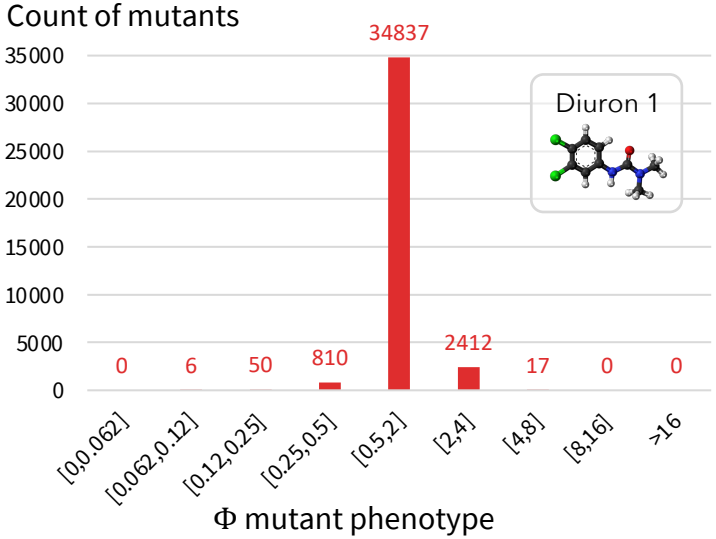
3

PROPOSED METHOD

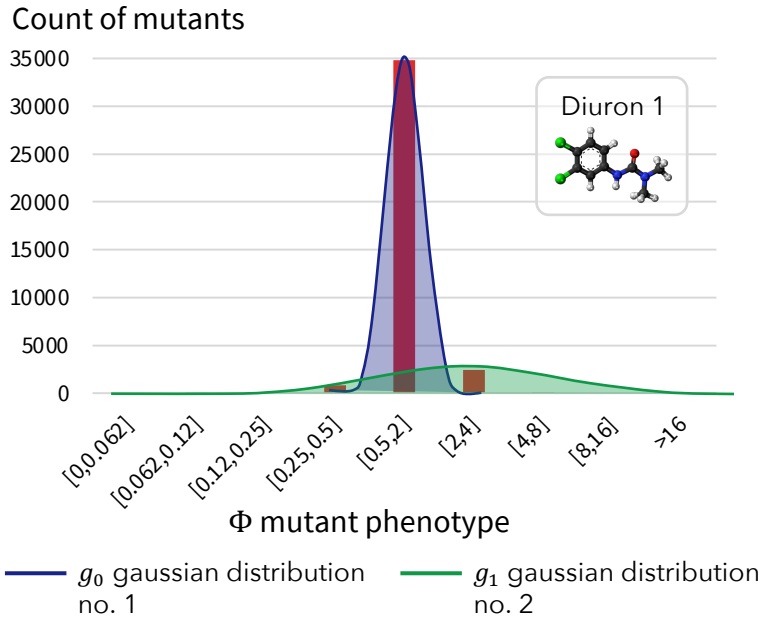
METHOD DESCRIPTION

ESTIMATION

Raw dataset for Diuron 1



Estimation with Gaussian Mixture Model (GMM)



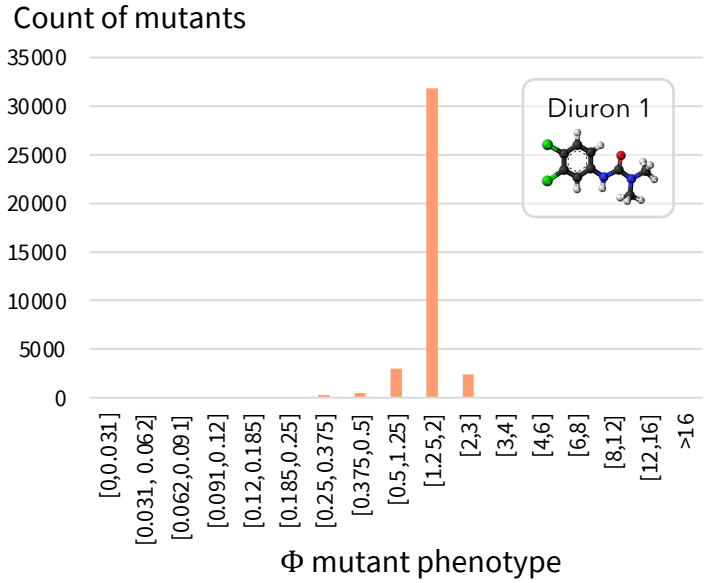
Issue

- Only a few points due to sampling of Φ to use a model
- Loss of information

Goal

- Estimate parameters of GMM

SIMULATION

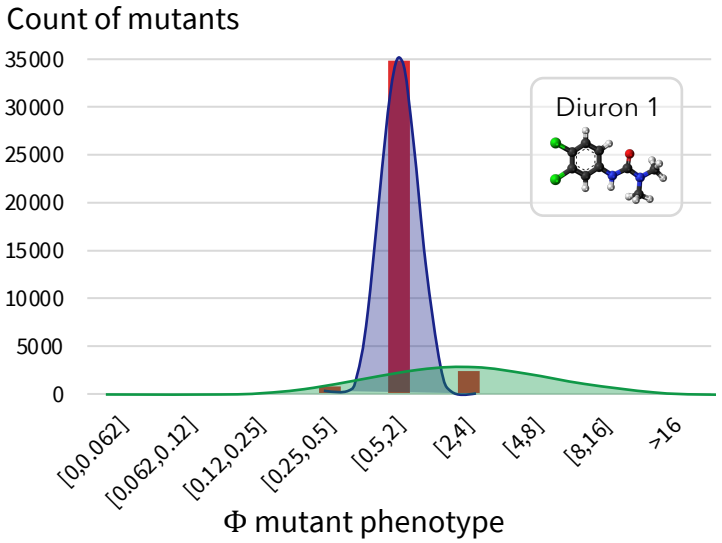


Goal

- Sample from GMM to test if the statistical analysis could be different
- Propose a new statistical method for analysis

GAUSSIAN MIXTURE MODEL AND HYPOTHESIS

Model



Gaussian mixture : $g(x) = p.g_0(x) + (1-p).g_1(x)$

with: $g_0(x) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)$

$g_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)$

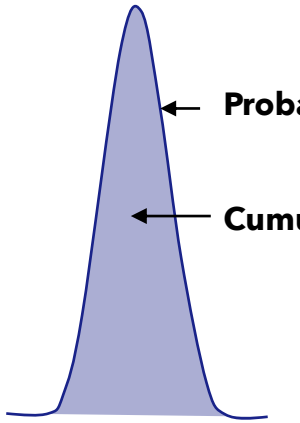
5 parameters:

p proportion of g_0

μ_i mean of g_i for $i \in \{0,1\}$

σ_i standard deviation of g_i for $i \in \{0,1\}$

g_0 gaussian distribution no. 1 — g_1 gaussian distribution no. 2



Probability density function (pdf): $g_i(x)$

Cumulative density function (cdf) :

$$G_i(x) = \int_{-\infty}^{x} g_i(t)\,\mathrm{d}t$$

Hypothesis

- (μ_0, σ_0) are fixed for g_0
- estimation is only performed to estimate (p, μ_1, σ_1)

ESTIMATION METHOD

Metropolis-Hastings (MH) algorithm

Markov Chain Monte Carlo (MCMC) method

Let $(X_i)_{i \in [\![0,N]\!]}$ a Markov chain with invariant probability $\pi \to g$

Input

y_{mes}

1) Initialization

$$x_0 = (p_0, \mu_{1,0}, \sigma_{1,0})$$
$$y'_0 = (G(\varphi_i) - G(\varphi_{i+1}))_{i \in [\![1,9]\!]}$$

2) Proposal of parameter samples

$$x'_t = (p'_t, \mu'_{1,t}, \sigma'_{1,t})$$
$$p'_t \sim \mathcal{U}(0,1)$$
$$\mu'_{1,t} \sim \mathcal{U}(-\ln 8, \ln 8)$$
$$\sigma'_{1,t} \sim \mathcal{U}(0, \ln 16)$$

3) Computation of proposed histogram

$$y'_t = (G(\varphi_i) - G(\varphi_{i+1}))_{i \in [\![1,9]\!]}$$

Repeat N times

4) Computation of the acceptation / rejection ratio

$$\alpha_{MH} = \frac{\|y_{mes} - y'_t\|^2 - \|y_{mes} - y_t\|^2}{\sigma_m^2}$$
$$u \sim \mathcal{U}(0,1)$$

5) Is x'_t accepted ?

Yes: $\alpha_{MH} > \log u$, $x_{t+1} = x'_t$

No: $\alpha_{MH} < \log u$, $x_{t+1} = x_t$

Store the x'_t

Output

6) Parameters estimation

$$\hat{p} = \frac{1}{N}\sum_{t=0}^{N-1} p_t$$
$$\widehat{\mu_1} = \frac{1}{N}\sum_{t=0}^{N-1} \mu_{1,t}$$
$$\widehat{\sigma_1} = \frac{1}{N}\sum_{t=0}^{N-1} \sigma_{1,t}$$

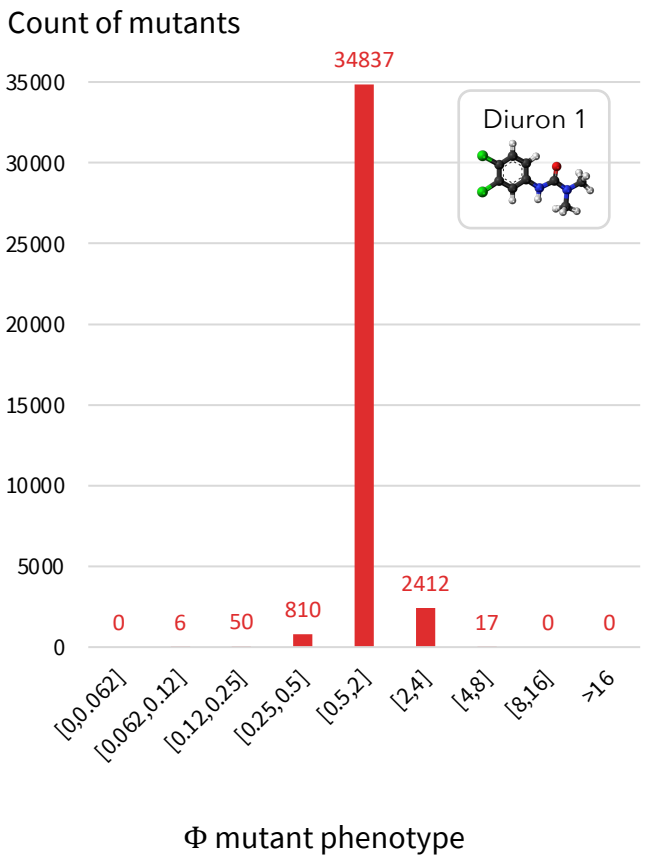
N : number of itérations

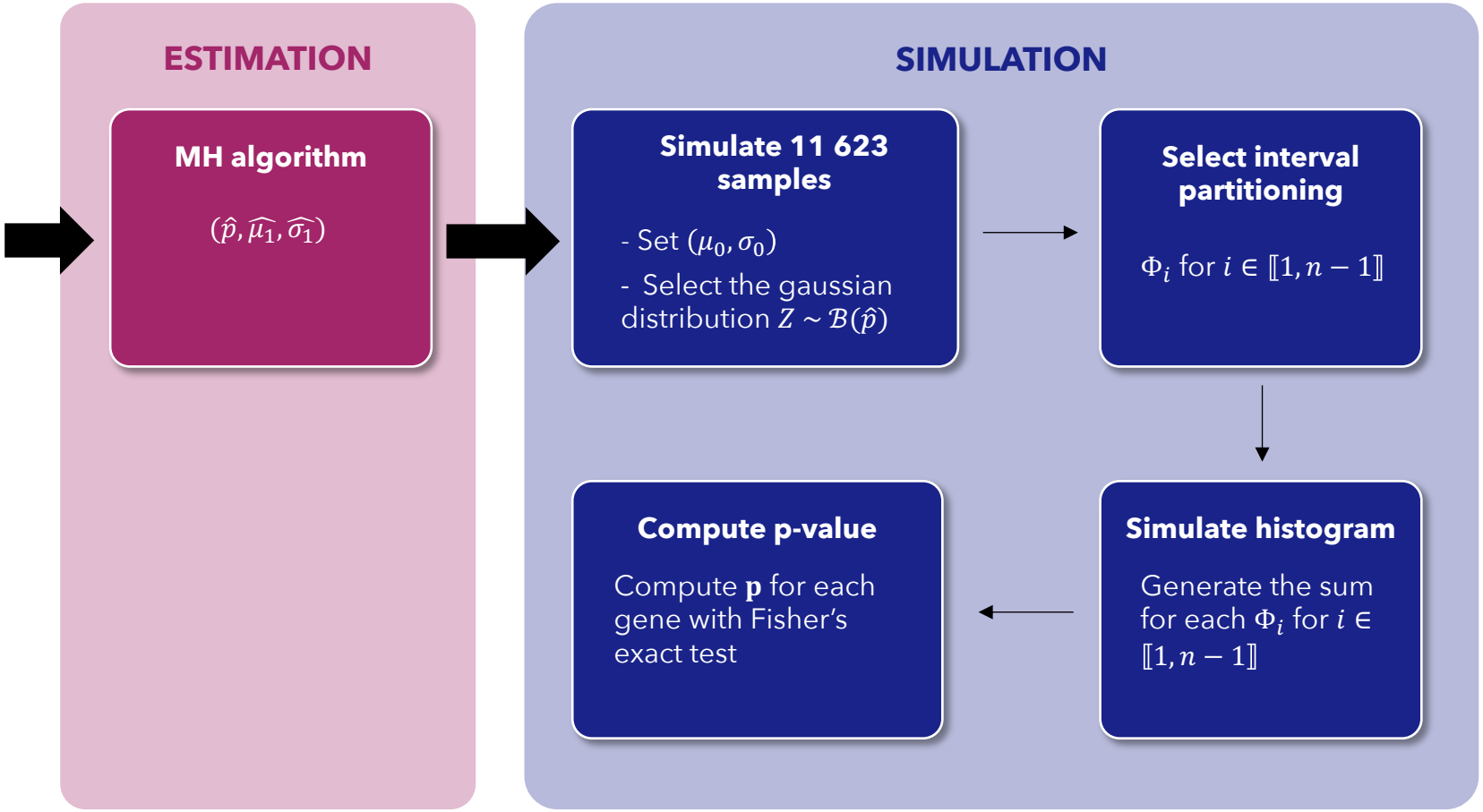
n : number of intervals of mutant phenotype Φ

σ_m : standard deviation of the GMM to the data

SIMULATION

Steps of the study





4

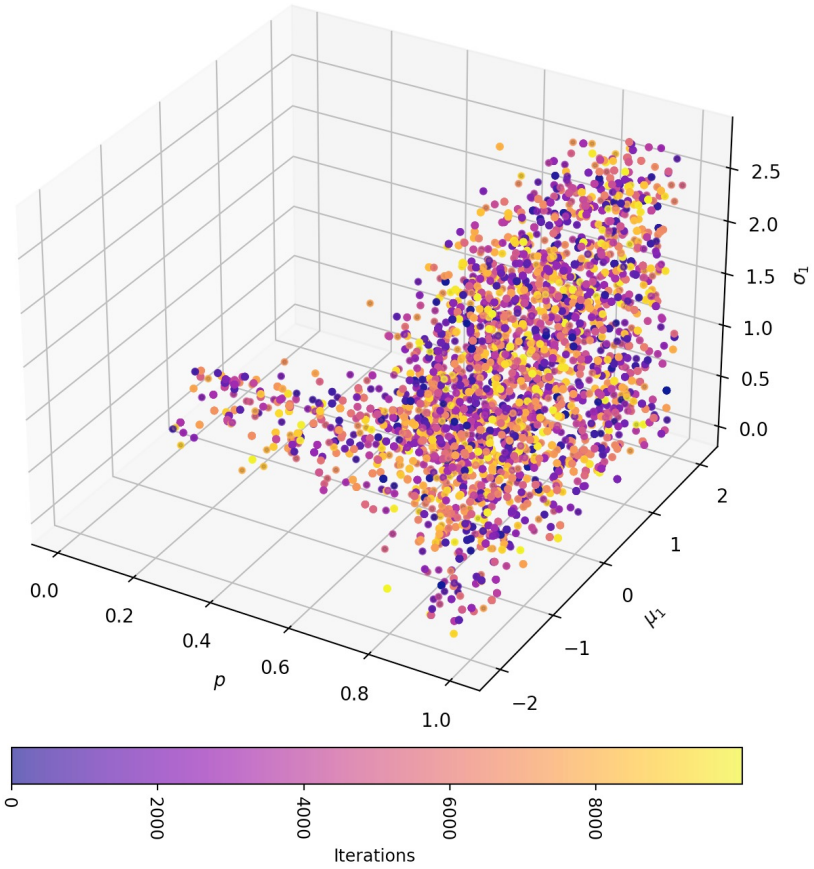
RESULTS AND DISCUSSION

ESTIMATION OF MUTANT DISTRIBUTION FOR DIURON 1

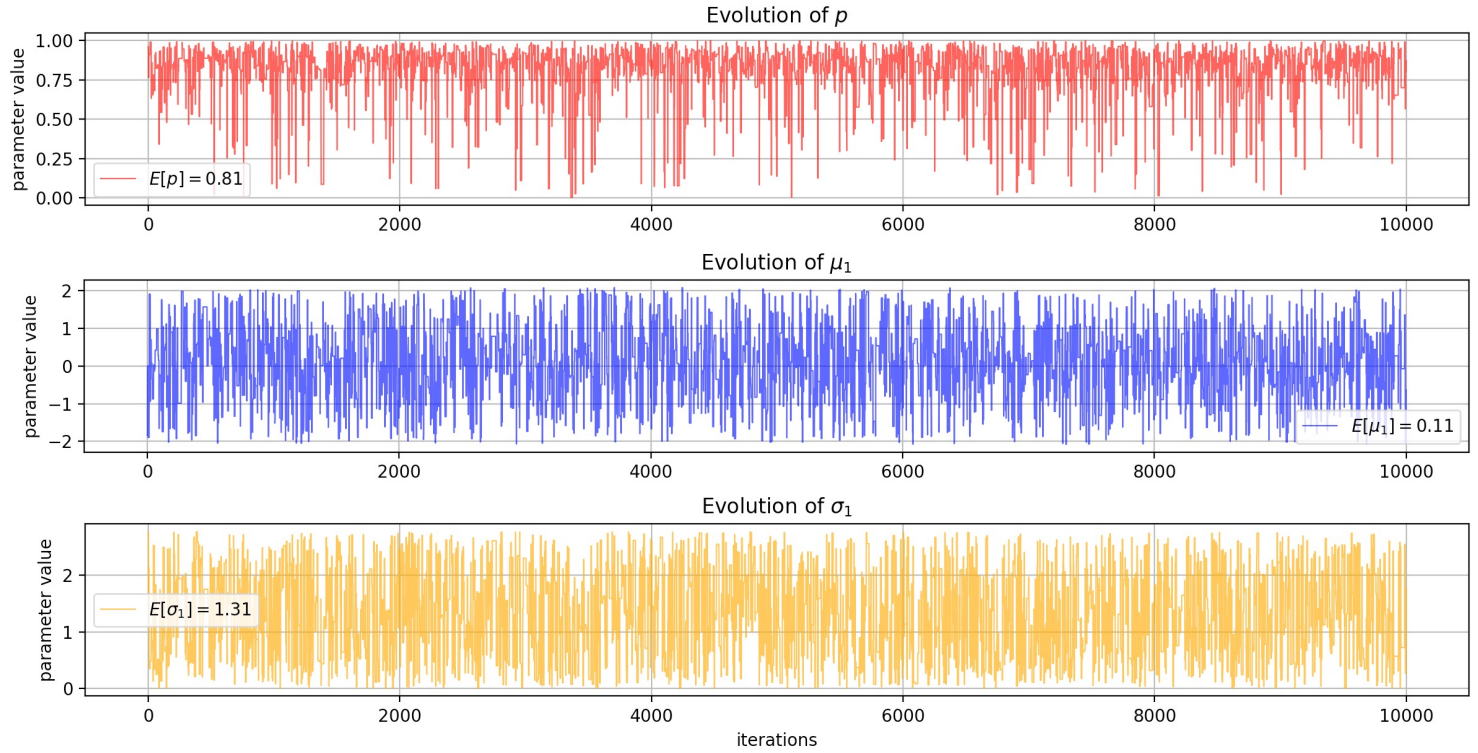
Parameter space (p, μ_1, σ_1)

Iterations: $N = 10\,000$

Distance GMM/data: $\sigma_m = 0.1$



Evolution of parameters according to iterations



Fixed parameters of GMM: $\mu_0 = 0$, $\sigma_0 = 0.01$

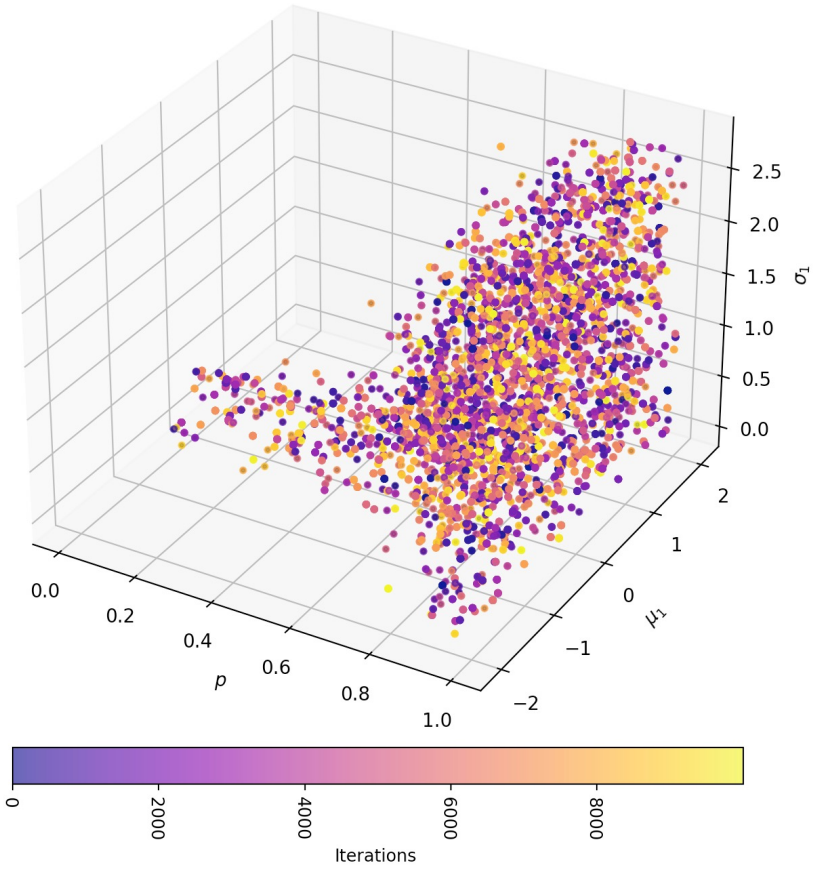
Estimated parameters: $\hat{p} = 0.81$, $\widehat{\mu_1} = 0.11$, $\widehat{\sigma_1} = 1.31$

ESTIMATION OF MUTANT DISTRIBUTION FOR DIURON 1

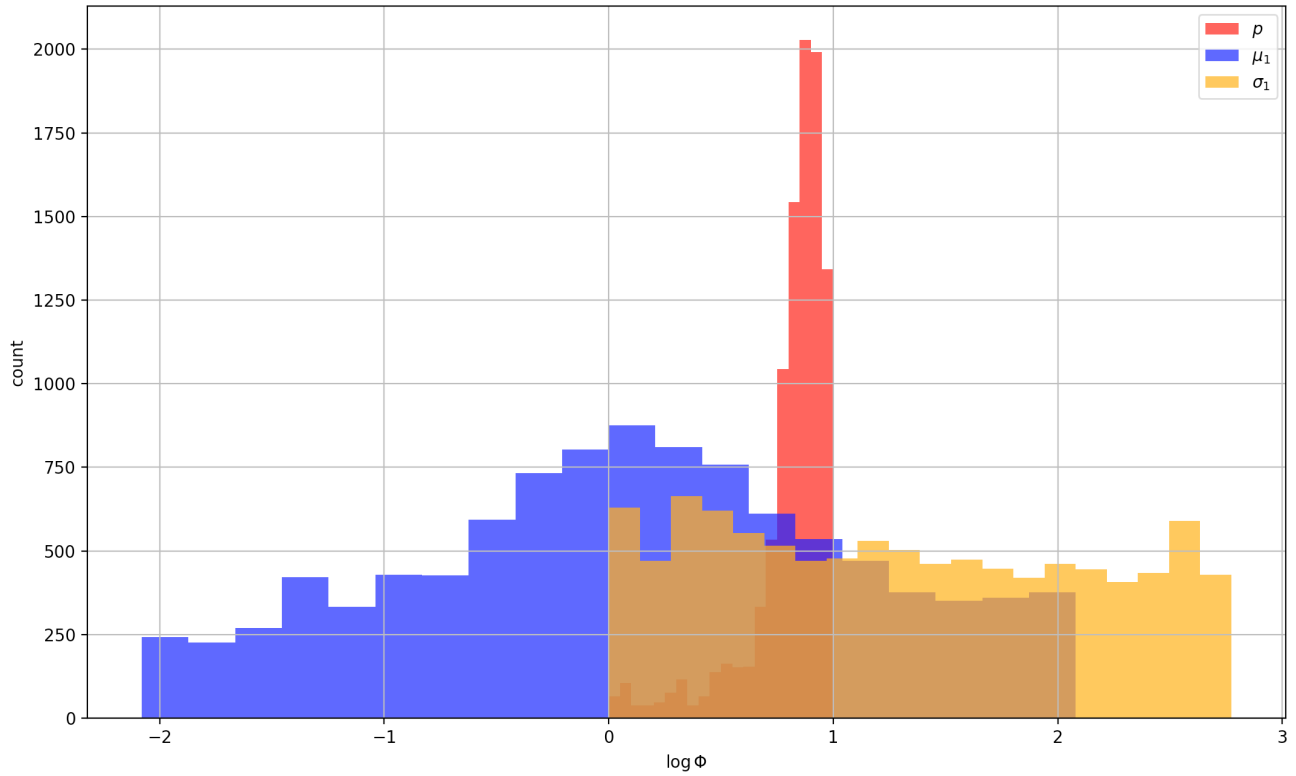
Parameter space (p, μ_1, σ_1)

Iterations: $N = 10\,000$

Distance GMM/data: $\sigma_m = 0.1$



Distributions of GMM's parameters

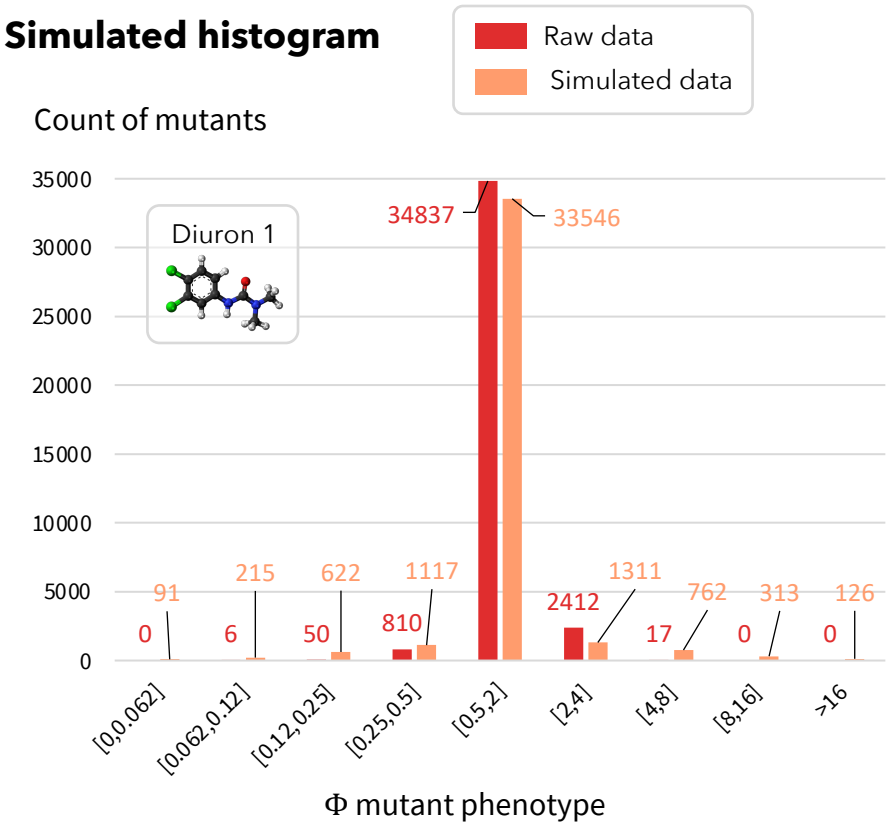


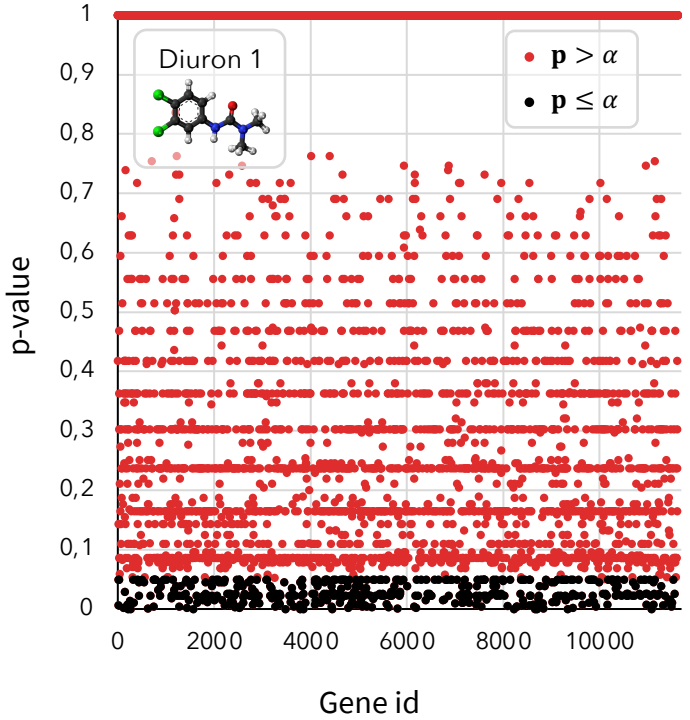
Fixed parameters of GMM: $\mu_0 = 0$, $\sigma_0 = 0.01$

Estimated parameters: $\hat{p} = 0.81$, $\widehat{\mu_1} = 0.11$, $\widehat{\sigma_1} = 1.31$

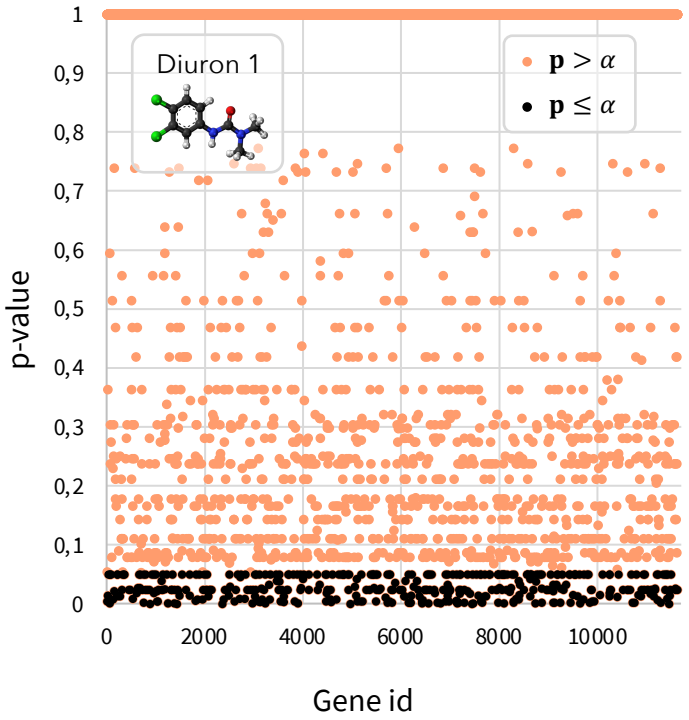
SIMULATION

Simulated histogram





36% of genes are sensistive or tolerant to Diuron 1

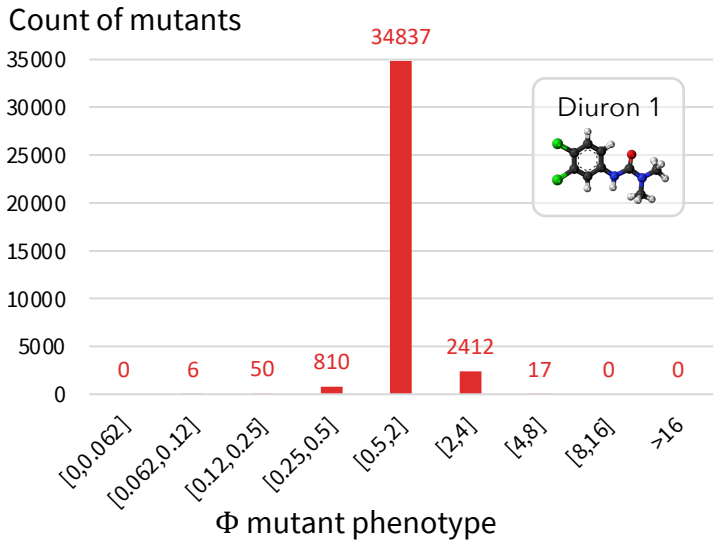


34% of genes are sensistive or tolerant to Diuron 1

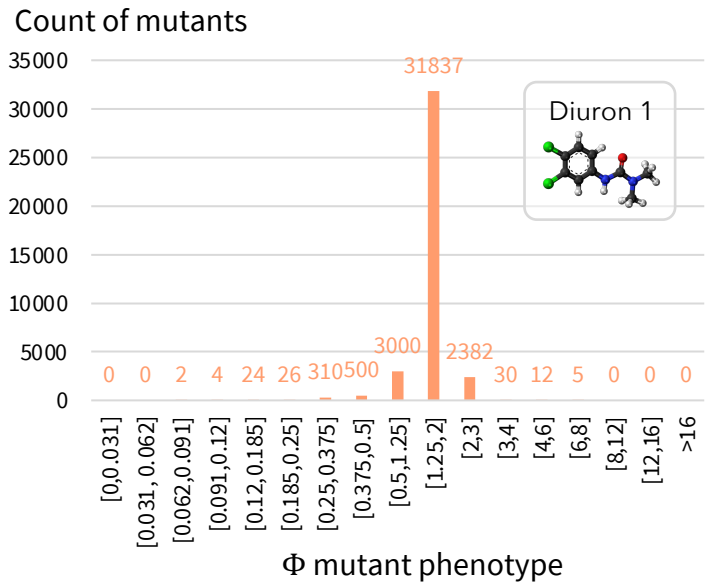
93% of genes have the same behavior in raw and simulated data

IMPACT OF INTERVAL PARTITIONING

Increase the sampling of mutant phenotype ratio



Generate
Φ_i for $i \in [\![1, n-1]\!]$



Goal

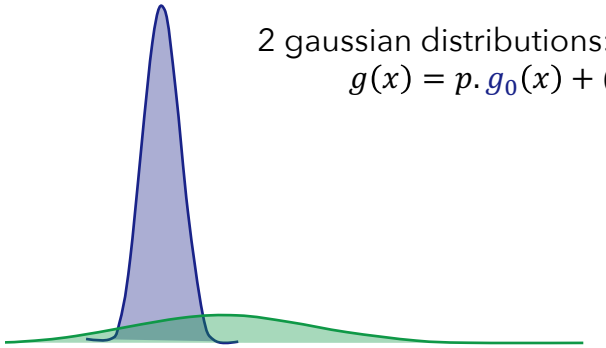
- Quantify the impact of changing n or σ_0
- Count how many « uninteresting genes » become « interesting » / Count how many « interesting genes » become « uninteresting »

PERSPECTIVES

Mix of more than 2 gaussian distributions

2 gaussian distributions:

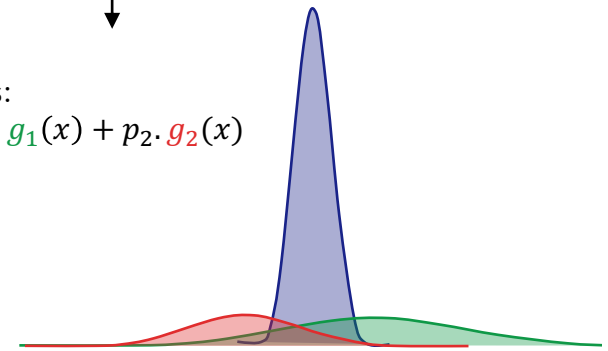
$$g(x) = p.g_0(x) + (1-p).g_1(x)$$



3 gaussian distributions:

$$g(x) = p_0.g_0(x) + p_1.g_1(x) + p_2.g_2(x)$$

$$p_0 + p_1 + p_2 = 1$$



Alternative statistical analysis based on probabilities

GMM : $g_i(x)$ for $i \in \{0,1\}$

$\mathbf{X} \mid Z \sim \mathcal{N}(\mu_i, \sigma_i)$ the probability of a sample according to the gaussian distribution is known

$Z \sim \mathcal{B}(\hat{p})$ the gaussian distribution is chosen with a probability $\hat{p}$

Bayesian approach: Compute $\mathbb{P}(Z \mid \mathbf{X})$

Fisher's exact test

- Provide p-value $\mathbf{p}$
- Indicate which genes are sensitive or tolerant the herbicide

With GMM

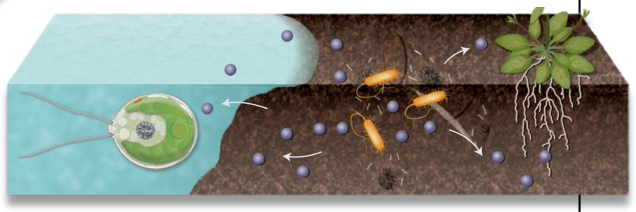
- Provide the probability to belong to g_i for $i \in \{0,1\}$
- Provide a variance on this probability

Experiment planification

In explored genes, exploit those for the variance is high

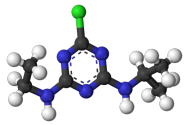
CONCLUSION

BIOLOGY

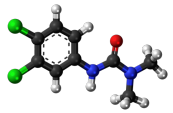


Herbicides

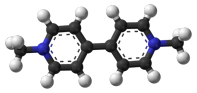
Atrazine



Diuron

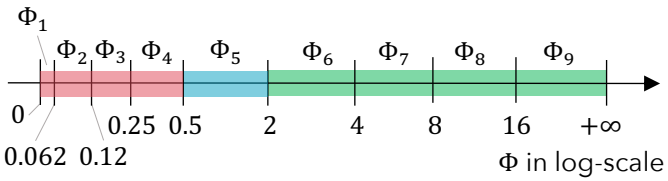


Paraquat



can have positive effect on some genes of *Chlamydomonas reinhardtii*

EXPERIMENTAL DESIGN



Negative effect of the treatment

No effect of the treatment

Positive effect of the treatment

Sampling mutant phenotype is not well adapted to fix the lack of mutants per gene because it leads to a loss of information

STATISTICAL ANALYSIS

From statistical test …

Fisher's exact test

- Provide p-value $\mathbf{p}$

… to bayesian approach and probabilistic model

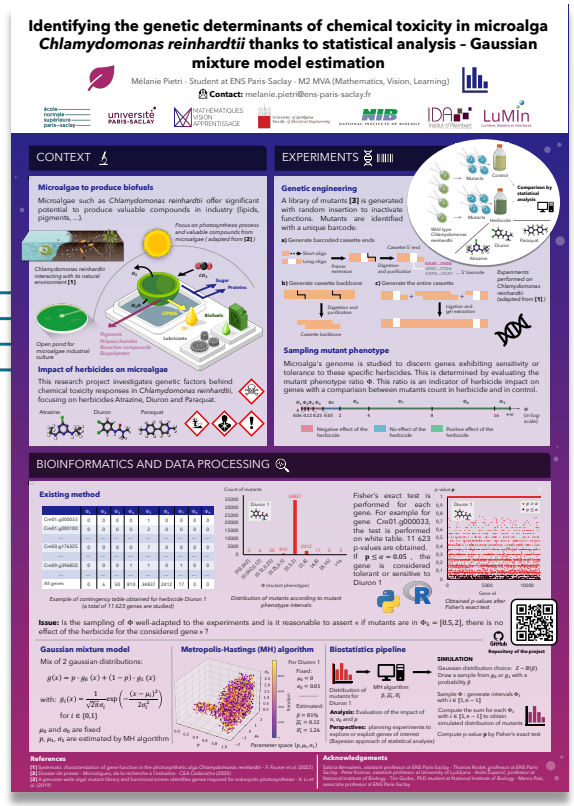
With GMM

- Provide the probability to belong to g_i for $i \in \{0,1\}$
- Provide a variance on this probability

ADDITIONAL WORK

Poster presentation

at laboratories forum of ENS Paris-Saclay
to present IDA's research activites



école normale supérieure paris–saclay

IDA Institut d'Alembert

Pluridisciplinary junior congress

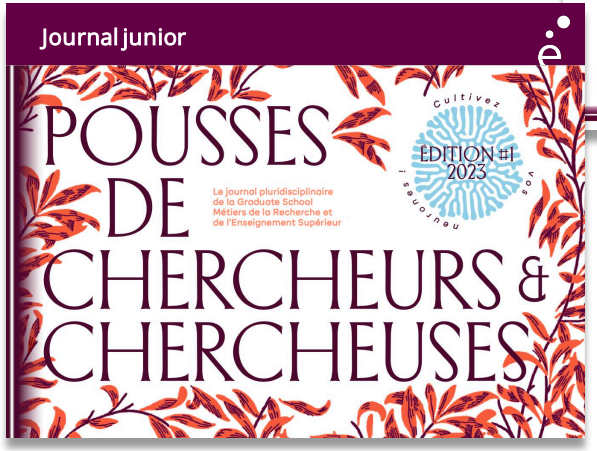
Participation and reward
Research paper in junior journal

CONGRÈS JUNIOR PLURIDISCIPLINAIRE

GRADUATE SCHOOL Métiers de la Recherche et de l'Enseignement Supérieur

université PARIS-SACLAY

PRIX DANS SA CATÉGORIE 2023 — GS MRES-UNIV. PARIS SACLAY — CONGRÈS JUNIOR PLURIDISCIPLINAIRE — J'AI PARTICIPÉ



PRODUCTION DE BIOCARBURANTS À PARTIR DE RESSOURCES RENOUVELABLES : ÉTUDE DE LA SENSIBILITÉ DES MICROALGUES À L'ÉLECTROPORATION
Réalisé dans le cadre d'un stage de M2 (10/04/23 – 10/08/23)

Présentatrice
Mélanie Pietri

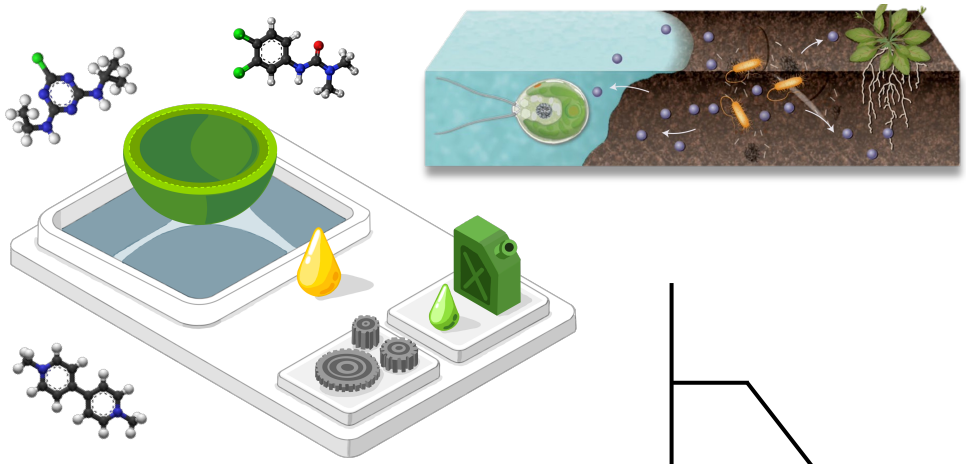
Master MATHÉMATIQUES VISION APPRENTISSAGE

Encadré par Peter Kramar, Anže Županič, Sakina Bensalem

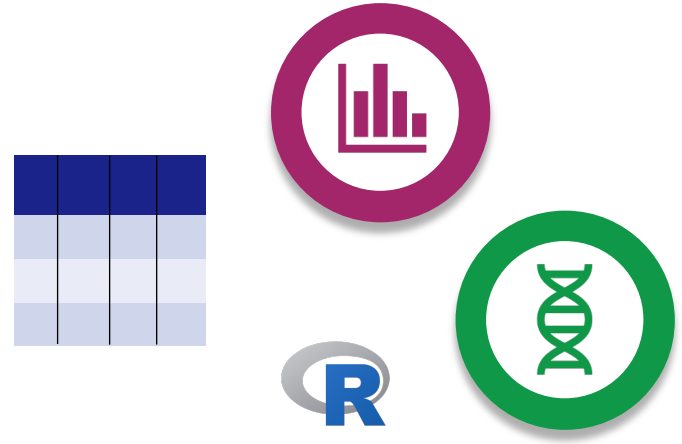
université PARIS-SACLAY · AgroParisTech · CentraleSupélec · école normale supérieure paris–saclay · Institut d'optique Graduate School ParisTech · Université d'Évry · Université de Versailles St-Quentin-en-Yvelines

Possible workshop on this topic

Preparation of a workshop for student during my PhD



… about context of the study ?

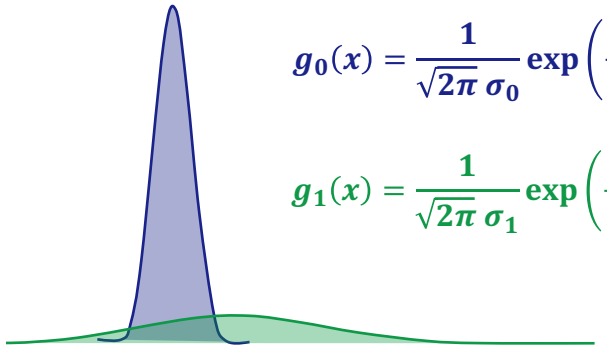


… about existing method ?

THANK YOU FOR YOUR ATTENTION !

Do you have any questions …

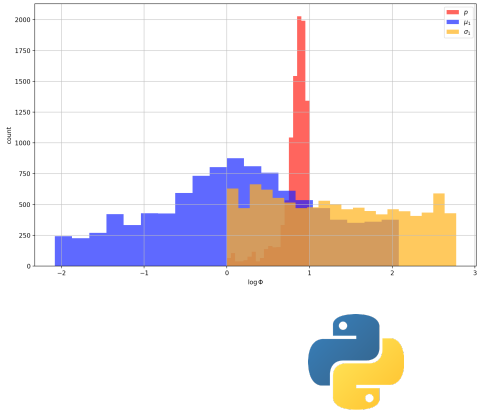
$$g(x) = p.g_0(x) + (1-p).g_1(x)$$

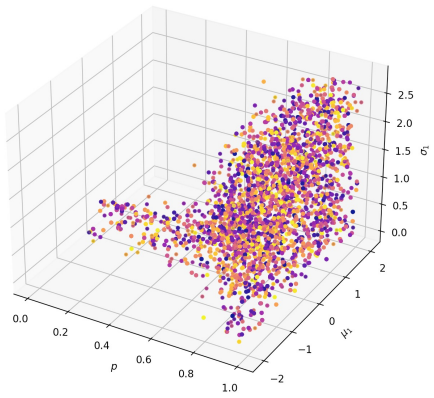


$$g_0(x) = \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)$$

$$g_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)$$

… about proposed method ?





… about results ?