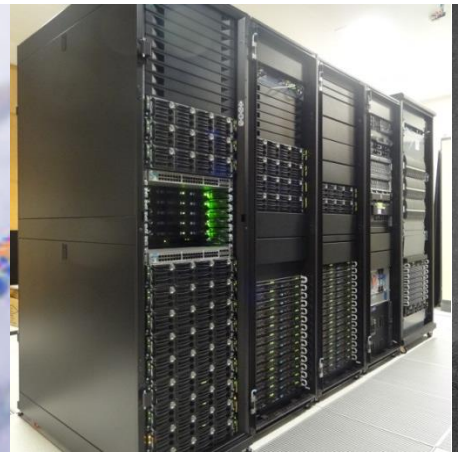
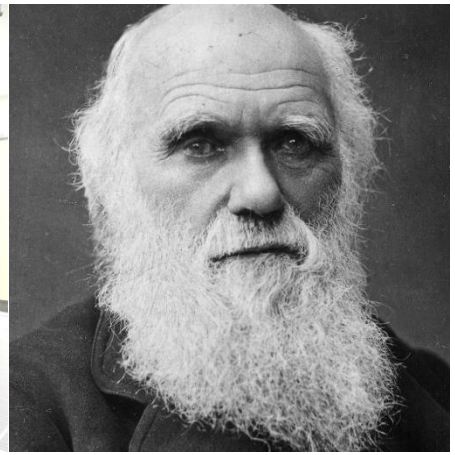
Les algorithmes génétiques peuvent-ils aider à identifier des biomarqueurs génétiques ?

SÉMINAIRE IN-THE-ART

Stéphane GAZUT (DRT/ LIST/ DM2I/ SID/ LS2D)

10 octobre 2019

PLAN

- **La sélection de variables**
  - Univariée
  - Multivariée
- **Les algorithmes génétiques**
  - Généralités
  - Application à la sélection de variables
- **Illustration avec des données génétiques**
- **Perspectives**

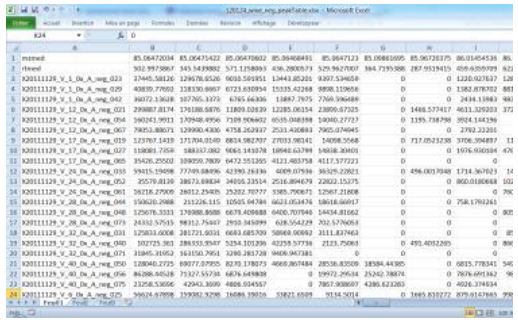
La sélection de variables

LA SÉLECTION DE VARIABLES

- **Objectif:** Parmi des **variables candidates**, **identifier** les **variables pertinentes** qui permettent d'**expliquer** et **prédire** l'étiquette liée aux données.

p
Variables, descripteurs, features …

n
Samples
Exemples
Points …



$\hat{Y}$

$\begin{pmatrix} \widehat{y_1} \\ \widehat{y_2} \\ \vdots \\ \vdots \\ \widehat{y_n} \end{pmatrix}$

Étiquette, target, status, (phénotype)

- Variables candidates ⇔ phase exploratoire
- Expliquer **et prédire** ⇔ La sélection de variables n'est jamais réalisée « seule » et est une étape préliminaire à la phase de prédiction

LA SÉLECTION DE VARIABLES

- **Objectif:** Parmi des **variables candidates**, **identifier** les **variables pertinentes** qui permettent d'**expliquer** et **prédire** l'étiquette liée aux données.

- **Pourquoi ?**
  - Réduire la dimensionnalité du problème → *réduire le nombre de variables à recueillir pour le déploiement du modèle*
  - Améliorer la connaissance du lien de « causalité » entre les descripteurs et l'état à prédire → *améliorer l'interprétation des résultats*
  - Améliorer la qualité de la prédiction → *ratio nombre de variables/nombre d'exemples plus favorable*

- **Comment ?** *Les méthodes de sélection:*
  - Expertise
  - Sélection univariée
  - Sélection multivariée

LA SÉLECTION DE VARIABLES – MÉTHODES UNIVARIÉES

- **Identification du lien statistique entre la sortie à estimer et chacune des variables prises indépendamment les unes des autres.**

Expliquer *(identification de l'information pertinente)*

- → Faible complexité algorithmique: p tests statistiques
- → Cadre théorique solide
- → Significativité du test et comparaison entre tests via la p-value
- → Ranking des variables selon la p-value

Prédire *(exploitation de l'information pertinente)*

- Dans le domaine de la santé, trop souvent, des modèles de prédictions ont été construits avec les top-k variables identifiées en univarié.
- → Cette démarche exclut *de facto*, des variables avec un faible score univarié mais en synergie (complémentarité) avec d'autres variables.

- *Méthodes classiques: Wilcoxon, Spearman, ANOVA, Information Mutuelle…*

SÉLECTION DE VARIABLES – MÉTHODES MULTIVARIÉES

- **Objectif: Trouver un (des) groupe(s) de variables pertinent(s)**

- **Recherche exhaustive de groupes**
  - Possible pour un nombre p de variables entre 15 et 20
  - **Il y a 2^p-1 combinaisons à évaluer**

- **Recherche exhaustive tronquée**
  - Recherche exhaustive jusqu'à un nombre p_{max} de variables
  - Dans ce cas, le nombre de combinaison est:

$$\sum_{i=1}^{p_{max}} C_p^i \qquad \text{où } C_n^k = \frac{n!}{k!\,(n-k)!}$$

- **Besoin:** Identifier des groupes sans avoir à évaluer l'intégralité des combinaisons

SÉLECTION DE VARIABLES – MÉTHODES MULTIVARIÉES

- **Besoin:** Identifier des groupes sans avoir à évaluer l'intégralité des combinaisons

- **Évaluer:**

En général, les méthodes multivariées utilisent un modèle pour évaluer la pertinence d'un groupe

Hypothèse: Un groupe de variables est d'intérêt si un modèle construit avec ces variables obtient de bonnes performances

- Stratégie pour **Identifier:**
  - Top-down
  - Bottom-up
  - Embedded methods
  - Optimisation combinatoire

SÉLECTION DE VARIABLES – MÉTHODES MULTIVARIÉES

- **Top-Down**
  - Regroupe un grand nombre d'algorithmes avec l'extension **RFE** (**R**ecursive **F**eatures **E**limination)
  - SVM-RFE [1], Random Forest – RFE [2]…
  - Estimation de l'importance d'une variable dans la construction du modèle en interne du processus d'apprentissage
    → exclusion, de manière itérative, des variables les moins importantes jusqu'à stabilisation du modèle
- **Bottom-up**
  - Initialisation de la sélection avec la meilleure variable
  - Ajout itératif d'une variable supplémentaire qui accroit le plus la performance du modèle
  - Inconvénient: le groupe final dépend fortement des premières sélections du groupe solution
    → On peut utiliser, l'algorithme d'échange double de Fedorov [3] qui peut échanger une sélection antérieure
- **Embedded methods**
  - La sélection fait partie de la phase de modélisation avec la mise en œuvre d'une pénalisation sur la valeur des paramètres: LASSO (norme L1), ridge (norme L2) Elastic-Net pénalisation L1 et L2.

$$\hat{\beta} = argmin_{\beta} \left\{ m(\beta, D) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 \right\}$$

SÉLECTION DE VARIABLES

Les trois grandes classes de méthodes:

- **Filter methods**
  - Une mesure indépendante de modèles qui permet d'exclure des variables non pertinentes
  - → Les méthodes univariées

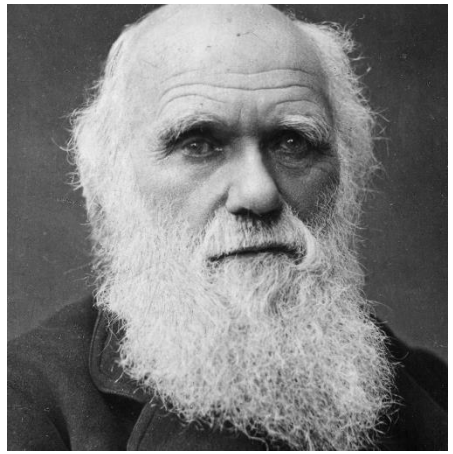
- **Embedded methods**
  - La sélection fait partie du processus de construction du modèle (pénalisation)

- **Wrapper methods**
  - Utilise un modèle prédictif pour évaluer la pertinence d'un groupe
  - Identifie un groupe de manière itérative
  - Recursive Features Elimination (RFE)
  - Optimisation combinatoire comme

  - → Les Algorithmes Génétiques [4]

Les Algorithmes Génétiques

LES ALGORITHMES GÉNÉTIQUES



- Initiés en 1975 par John Holland [5], ce sont des algorithmes d'optimisation (famille des méta-heuristiques).
- La méthode permet de résoudre des problèmes d'optimisation multicritères, optimisation combinatoire…
- Directement inspiré de la théorie de l'évolution de Darwin et de la génétique, ce sont des algorithmes itératifs qui opèrent sur des individus (solutions potentielles) codés, à partir d'une population initiale.
- Les solutions potentielles constituent une population d'individus dont les représentants les plus adaptés à leur milieu survivront et enfanteront de nouveaux individus.
- La population de solution évolue de la génération k à la génération k+1 à l'aide de trois opérateurs:
  - Opérateur de sélection – adaptation à une fonction objectif (fitness)
  - Opérateur de croisement
  - Opérateur de mutation

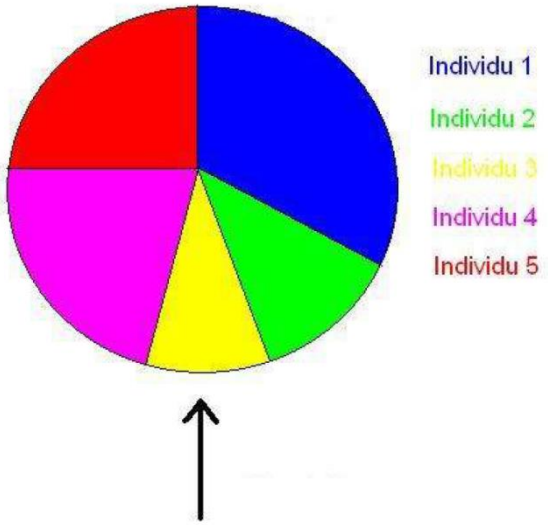
LES ALGORITHMES GÉNÉTIQUES

Opérateur de sélection → *Définit si les individus sont adaptés à leur milieu*

- Cet opérateur est chargé de définir quels individus de la génération k seront dupliqués et serviront de parents pour la génération k+1. On doit sélectionner la moitié des individus k, l'autre moitié sera obtenue par croisement. La sélection d'un individu dépend de son adaptation au problème (valeur de fitness).

Méthode de la loterie biaisée

Angle du secteur proportionnel à la fitness de l'individu



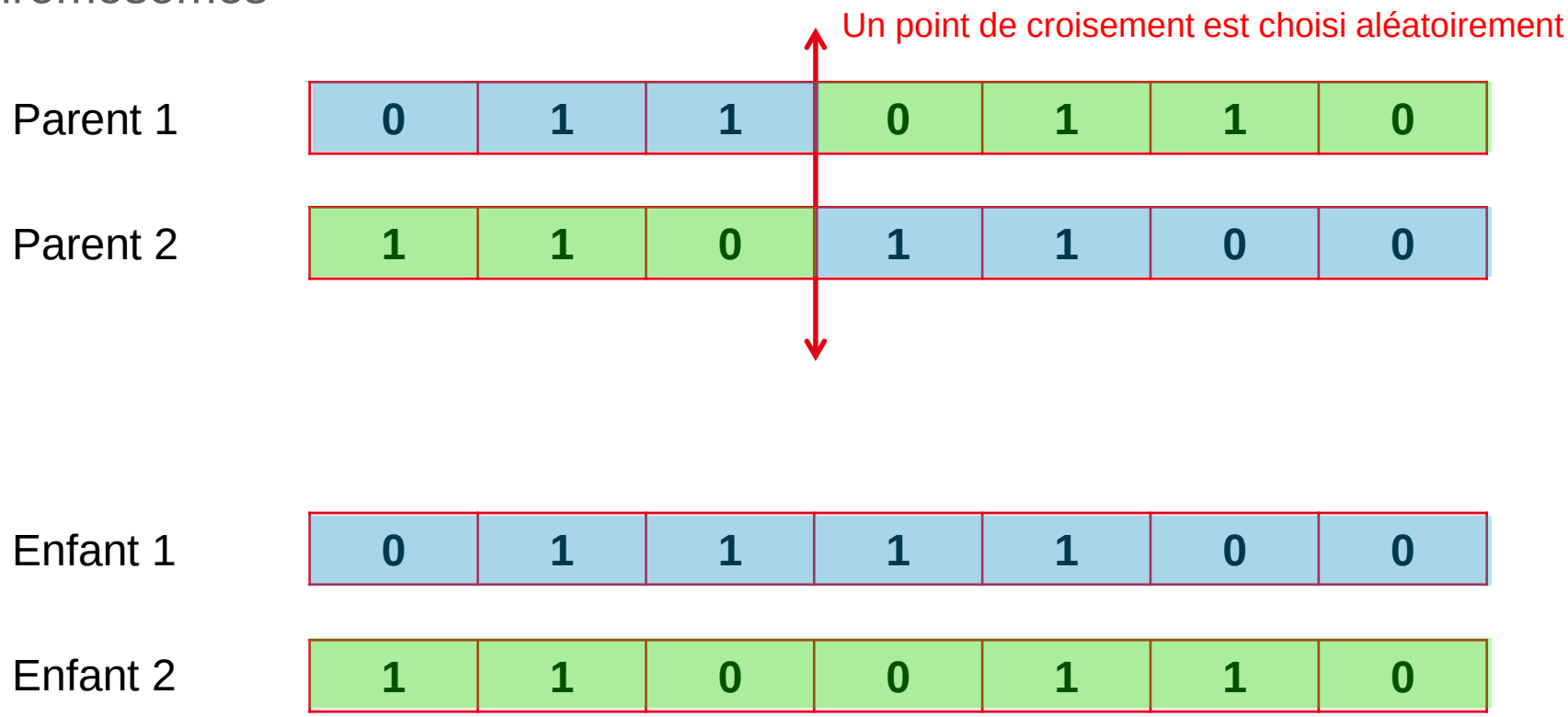
Méthode élitiste

Les individus sont classés suivant leur valeur de fitness et les x% meilleurs sont conservés pour le croisement (généralement 50%).

LES ALGORITHMES GÉNÉTIQUES

Opérateur de croisement → *Définit la façon de générer de nouveaux individus*

- Cet opérateur permet de générer deux enfants à partir de deux parents par « crossing over de leur chromosomes »

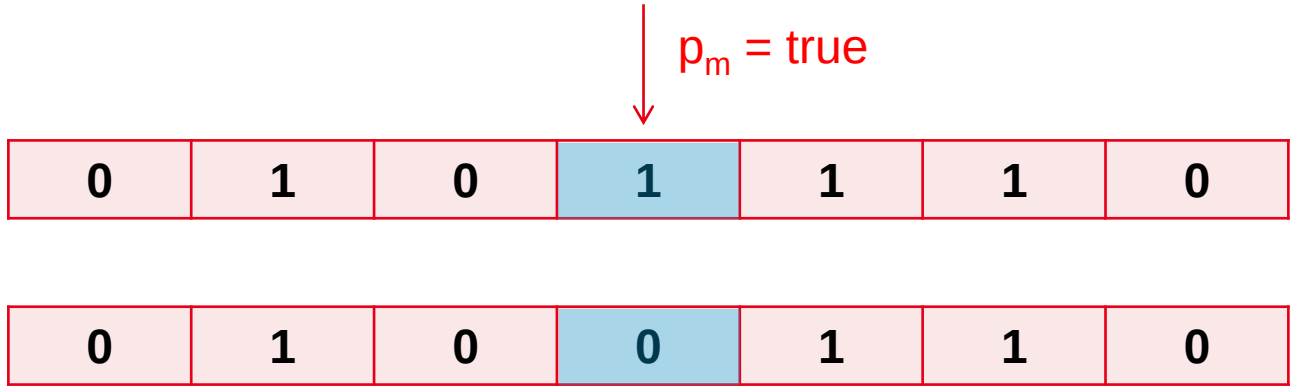


LES ALGORITHMES GÉNÉTIQUES

Opérateur de mutation → Ajoute une perturbation aléatoire (maladie) pour explorer de nouvelles solutions

- Cet opérateur consiste à changer la valeur allélique d'un gène avec une probabilité p_m généralement faible. On utilise une fonction qui renvoie true avec la probabilité p_m.

- Pour chacun des locus, si la fonction renvoie true, le locus est changé



LES ALGORITHMES GÉNÉTIQUES – EXEMPLE D'ILLUSTRATION

- Identifier le minimum d'une fonction de $\mathbb{R}^2 \rightarrow \mathbb{R}$

- Dans ce cas, la solution est une coordonnée (x_1, x_2), le point de crossover est *de facto* entre les deux composantes.

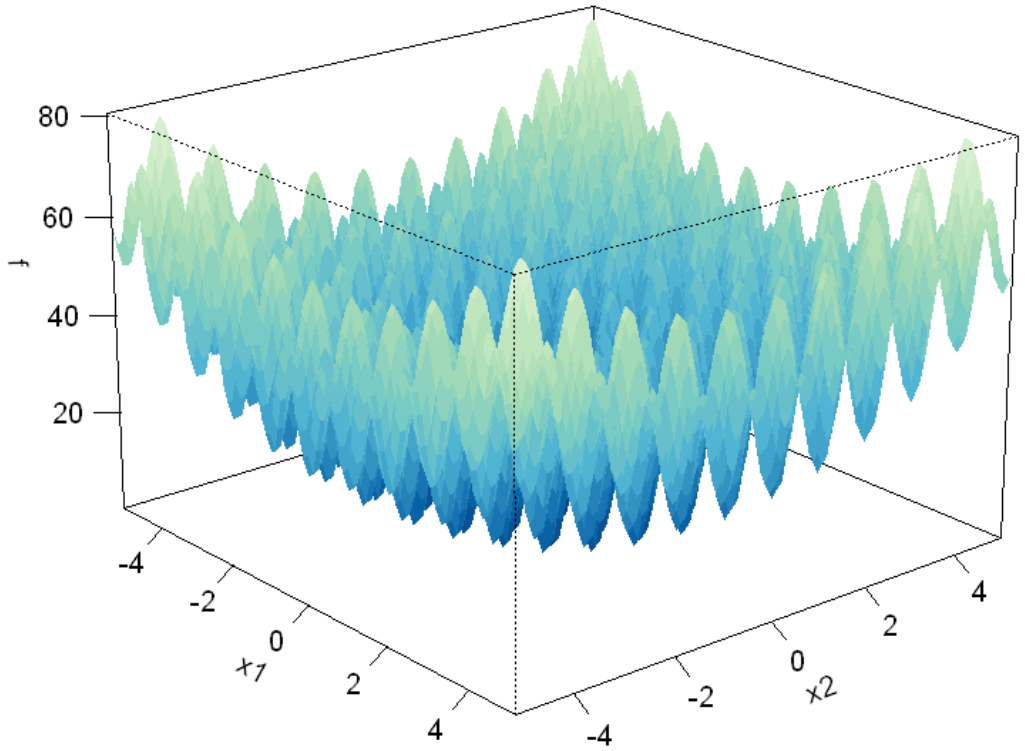
- La fonction Rastrigin est non convexe et souvent utilisée pour tester les algorithmes d'optimisation. Elle est définie par:

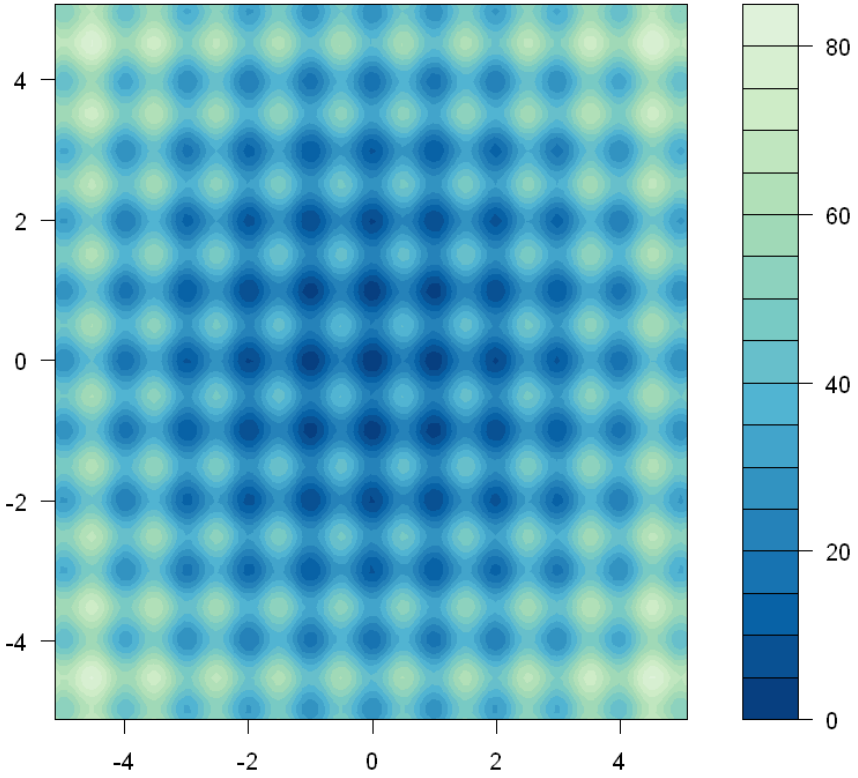
$$f(x_1, x_2) = 20 + x_1^2 + x_2^2 - 10(\cos(2\pi x_1) + \cos(2\pi x_2))$$

- Cette fonction a un minimum global en (0,0) avec f(0,0)=0

LES ALGORITHMES GÉNÉTIQUES – EXEMPLE D'ILLUSTRATION

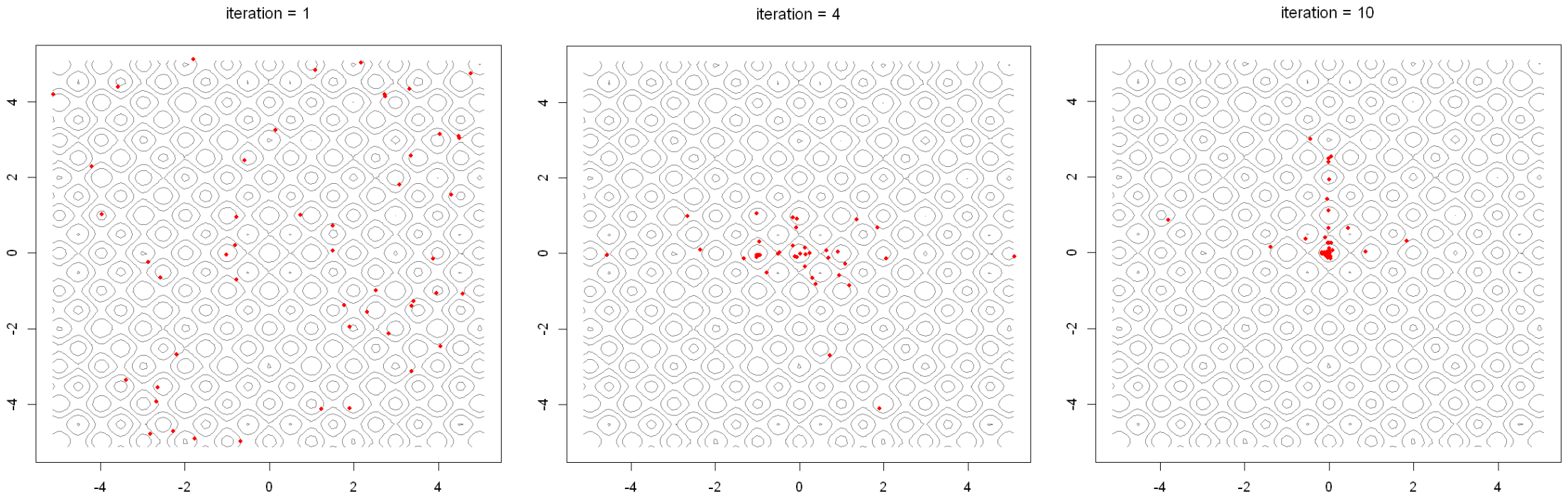
$$f(x_1, x_2) = 20 + x_1^2 + x_2^2 - 10(\cos(2\pi x_1) + \cos(2\pi x_2))$$





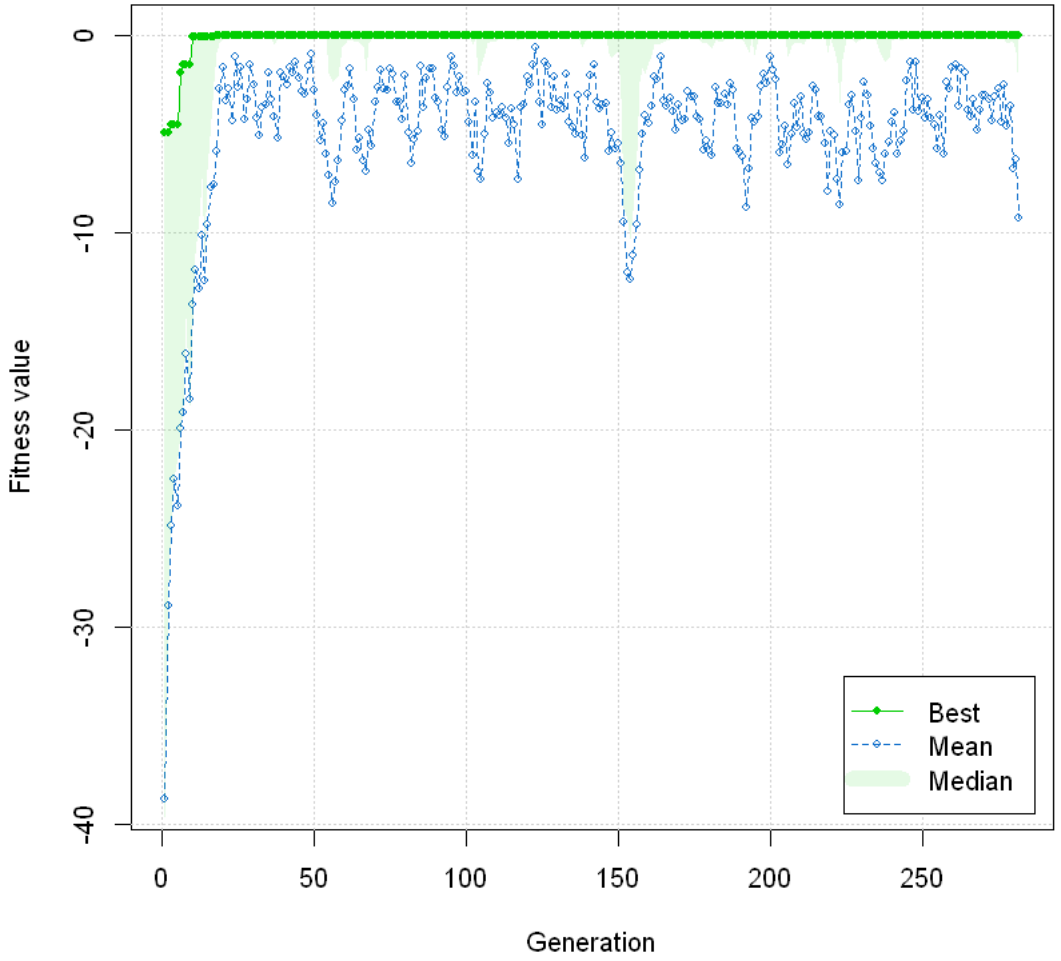
LES ALGORITHMES GÉNÉTIQUES – EXEMPLE D'ILLUSTRATION

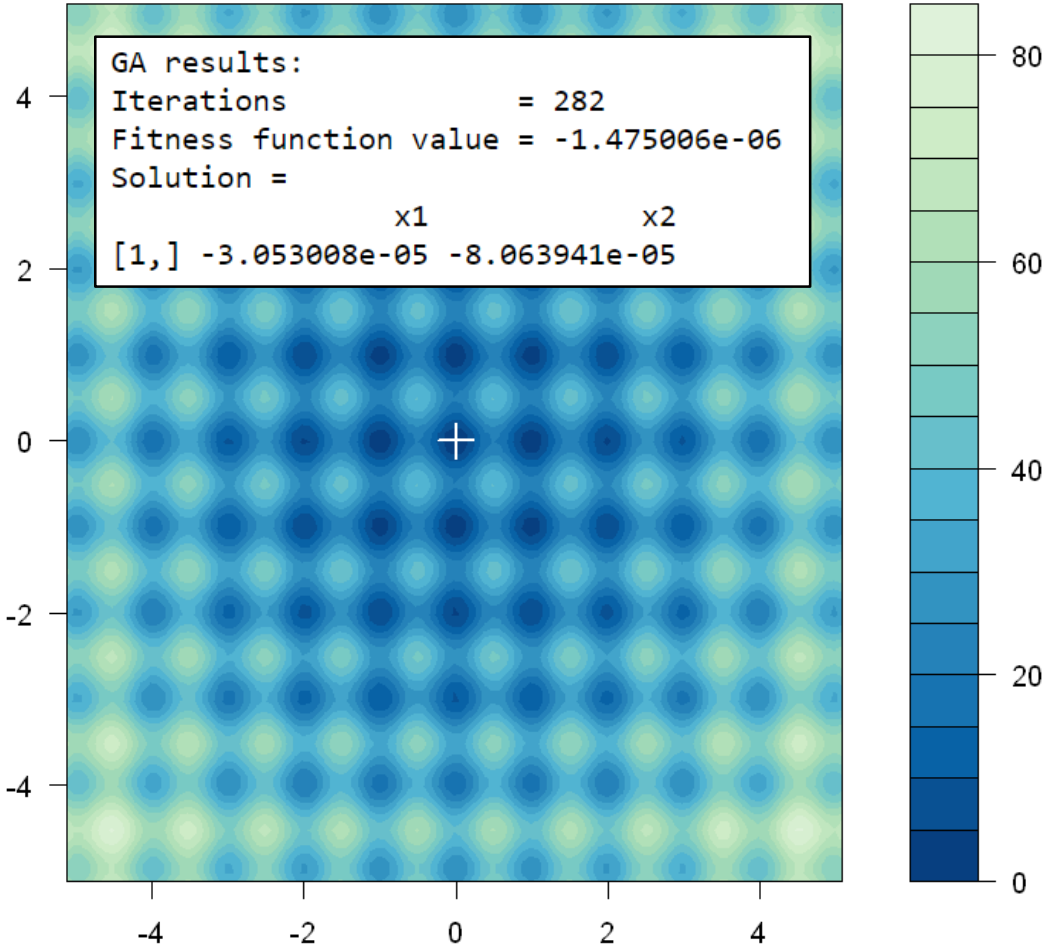
- **Disposition des points à l'initialisation**



LES ALGORITHMES GÉNÉTIQUES – EXEMPLE D'ILLUSTRATION

- Évolution de la valeur de fitness





LES ALGORITHMES GÉNÉTIQUES - SYNTHÈSE

- **Une population initiale d'individus (solutions potentielles)**
- **Une méthode de codage**
- **Une fonction objectif (fonction de fitness)**
- **Des opérateurs pour faire évoluer la population de la génération k à la génération k+1**
  - Sélection
  - Croisement
  - Mutation
- **Un critère d'arrêt**

Application à la sélection de variables

APPLICATION À LA RECHERCHE DE SIGNATURE

- **Recherche de biomarqueurs dans le cadre d'une étude GWAS (Genome Wide Association Study) → Données génétiques pour une application Cas/Control dans le cadre d'une étude sur le cancer de la prostate.**

- **Données**:
  - 27 000 variables de type SNP (BdD phase II),
  - 1 289 individus
  - 1 phénotype associé (Malade/Sain) par individu

- **Objectifs du projet:**
  - Trouver des petites signatures de quelques SNPs

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

CeRePP

Equipe du Pr Cussenot

APPLICATION À LA RECHERCHE DE SIGNATURE

Illustration de la combinatoire

Même dans le cas d'une recherche exhaustive tronquée, l'ensemble des groupes de 3 SNP pris parmi 27 000 est égal à $3{,}3e^{12}$

→ **Problème combinatoire**

- Si on évalue 1000 combinaisons par seconde il faudrait 100 ans pour toutes les tester
- Pour 4 SNP, il faudrait 700 000 ans dans les mêmes conditions

APPLICATION À LA RECHERCHE DE SIGNATURES

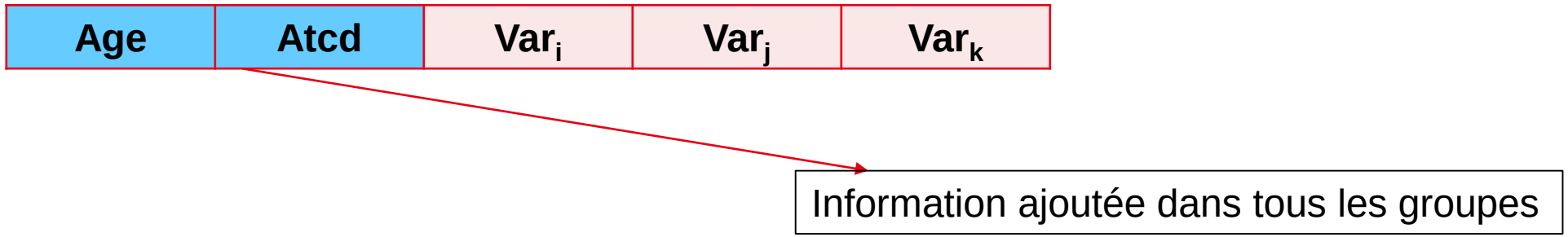
- **Définition d'un individu**

- Un individu est une solution potentielle

27 000 SNP sont disponibles dans la base de données de l'étude. Un individu est alors un groupe de variables SNP et correspond à une liste de 3 ou 4 indices parmi les 27 000 disponibles.

On se limite à 4 SNP pour assurer une certaine redondance de l'information dans les modèles qui seront construits par apprentissage au sein de la fonction de fitness.

Exemple d'individu



- Population initiale

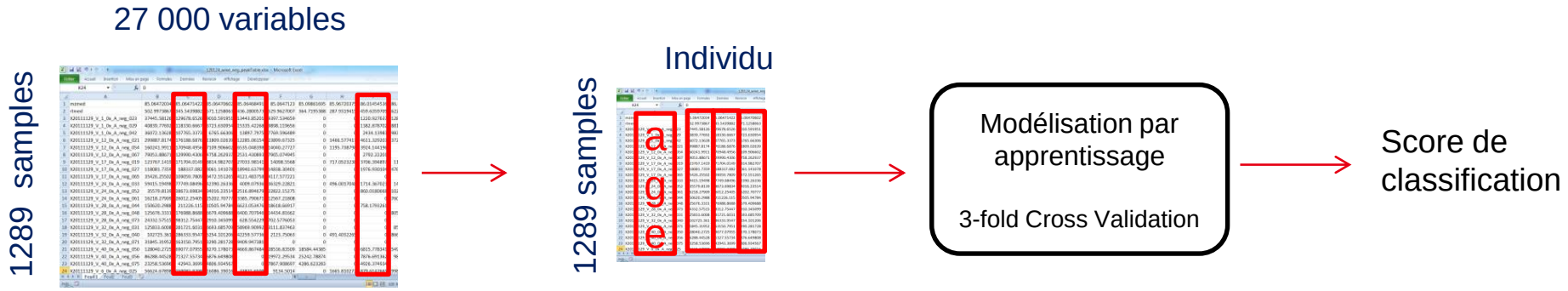
On génère 5 000 de ces individus aléatoirement.

APPLICATION À LA RECHERCHE DE SIGNATURES

Fonction de fitness

- Elle prend en argument un individu et renvoie sa valeur de fitness.

- La valeur de fitness est calculée comme étant le taux de bonne classification (accuracy) ou AUC ROC d'un modèle construit avec ces variables.
- Les modèles choisis sont des « weak learner », modèles simples à mettre en œuvre et ne nécessitant qu'une seule boucle de cross-validation (modèle à séparateur linéaire par exemple comme la régression logistique ou le SVM linéaire).

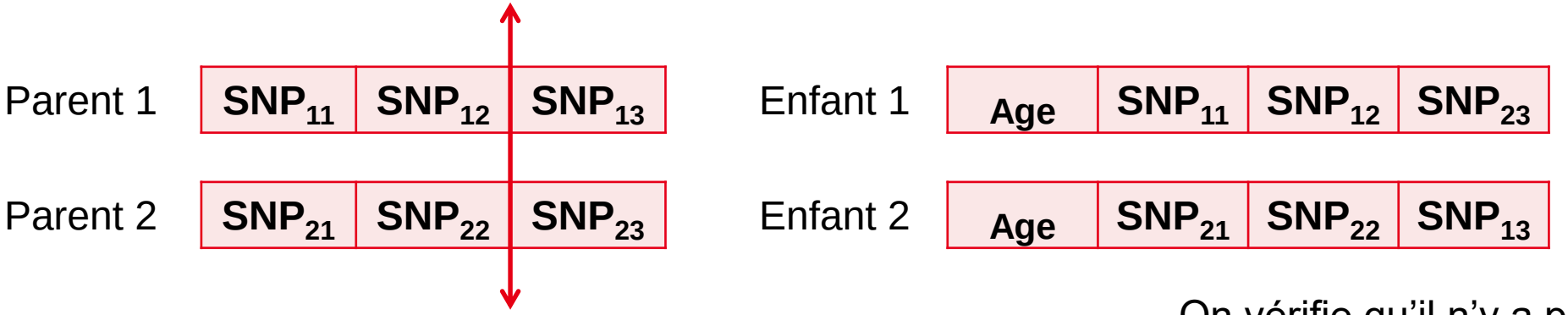


APPLICATION À LA RECHERCHE DE SIGNATURES

Sélection (méthode élitiste)

Classement des individus selon leur score de fitness. On garde les 50% meilleurs

Croisement



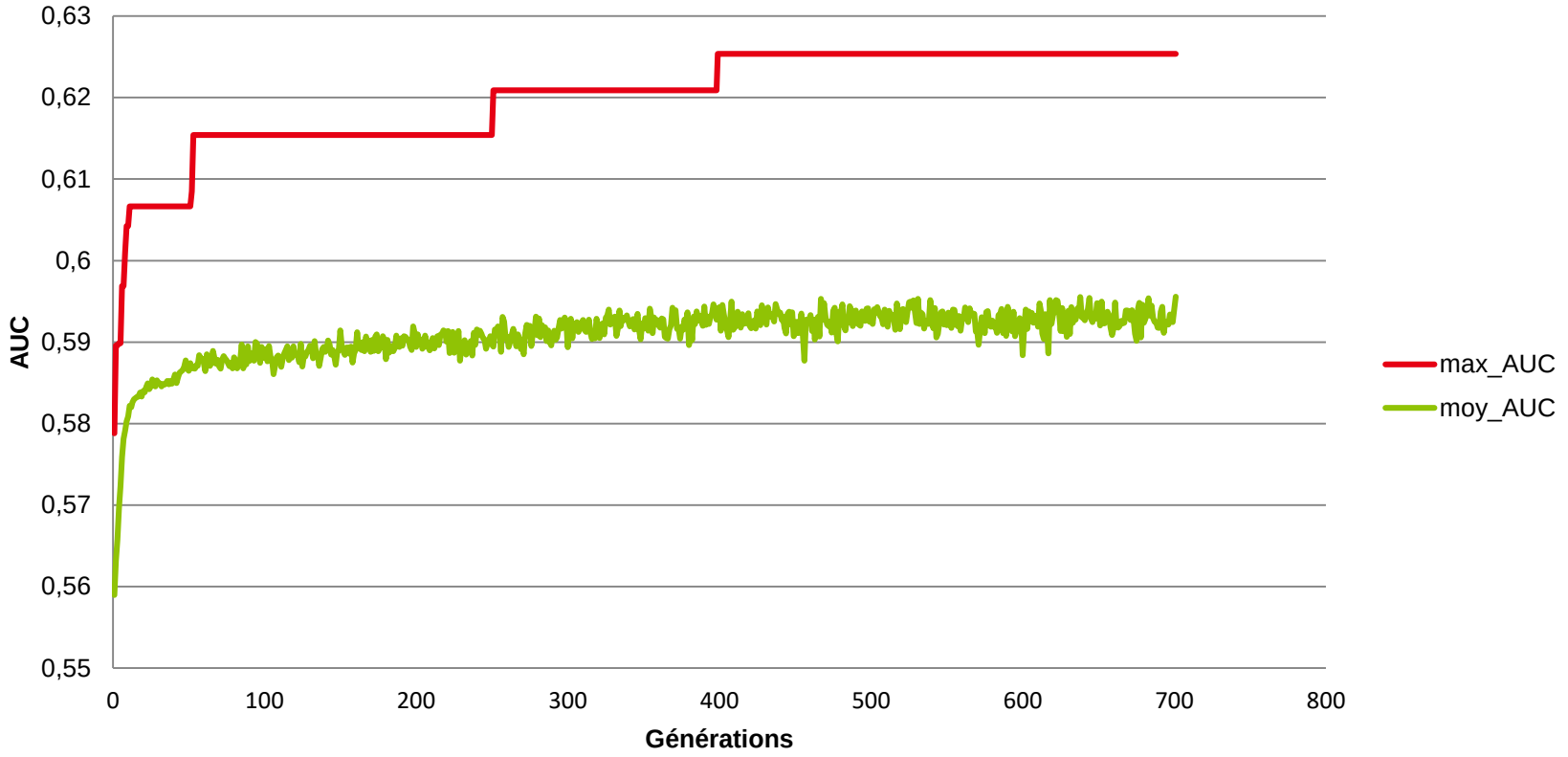
On vérifie qu'il n'y a pas de doublon

Mutation

Nous avons utilisé une probabilité assez forte au début de l'algorithme et qui décroit au cours des générations.

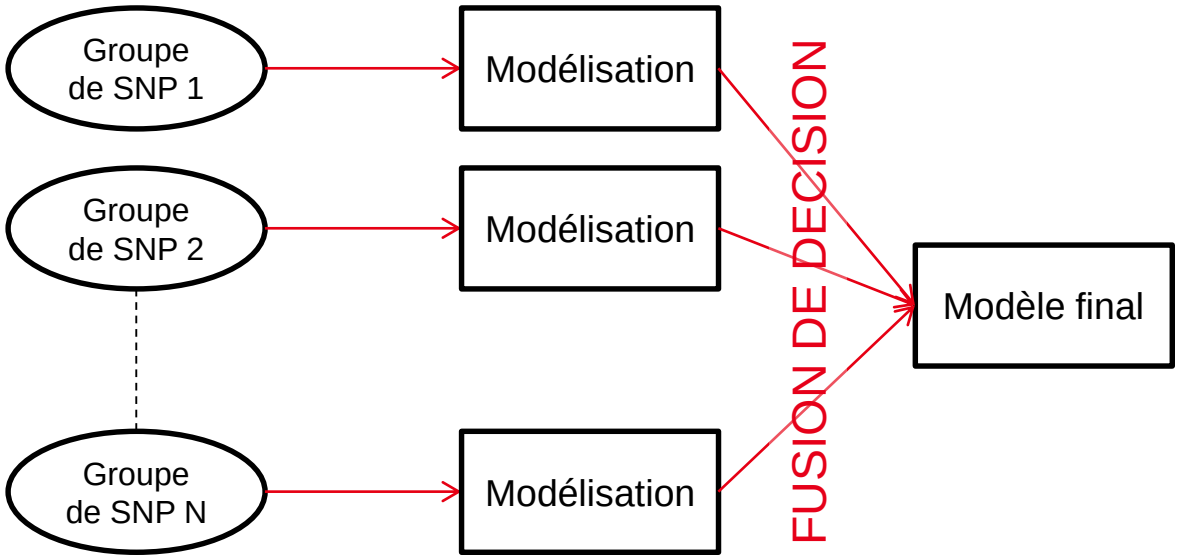
Quand une variable est choisie pour muter, on la remplace par une autre variable disponible dans le base en s'assurant qu'il n'y a pas de doublon pour l'individu obtenu.

APPLICATION À LA RECHERCHE DE SIGNATURES



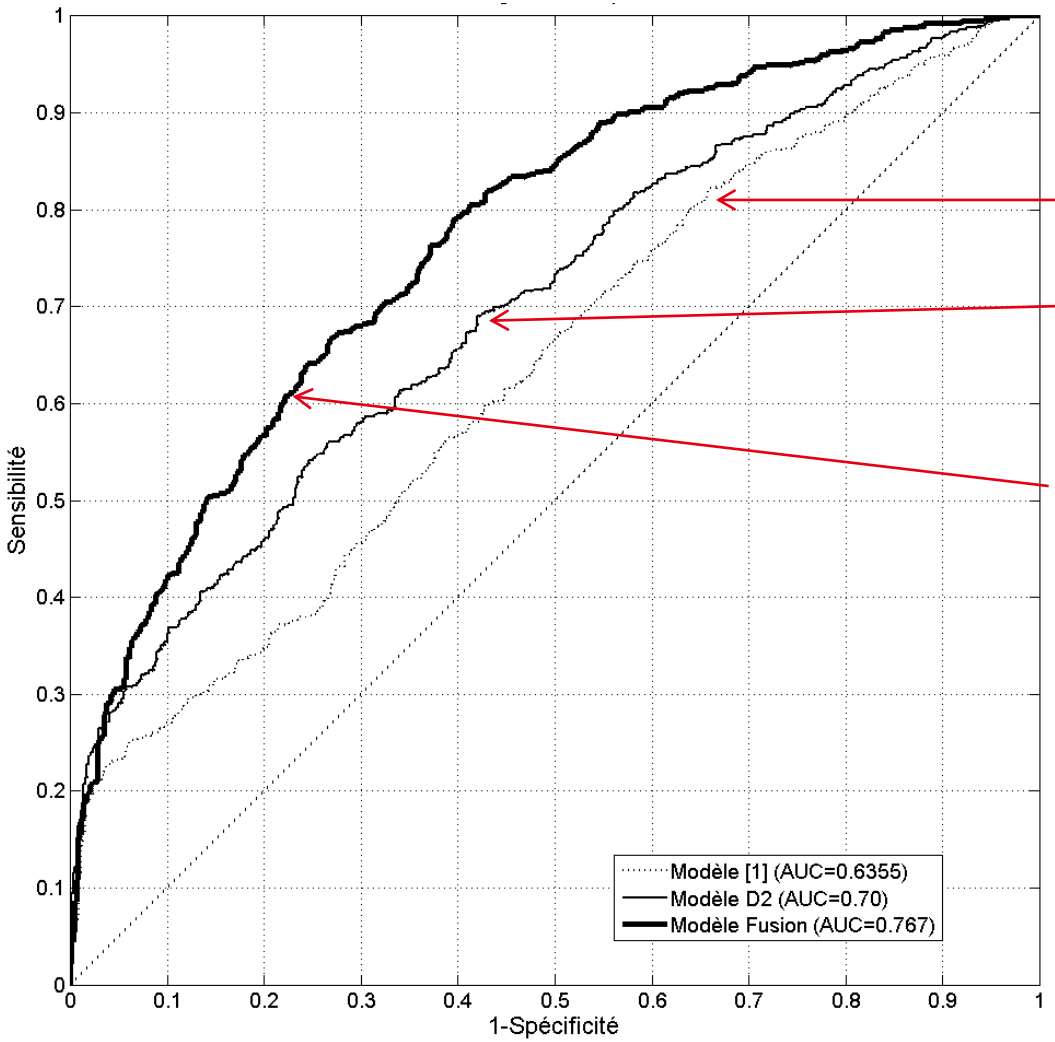
- Nous avons retrouvé des marqueurs d'intérêt connus en urologie
- Nous avons trouvé des marqueurs inédit
- Il est possible d'imposé un SNP au sein des individus d'une population pour trouver les marqueurs en synergie avec lui.

MODÉLISATION FINALE



- Création d'un modèle spécifique à chacun des groupes
- Fusion de décision afin d'exploiter les complémentarités entre modèles et créer le modèle final.

COURBES ROC



Courbes ROC:

- Modèle Antécédents sans SNP
- Modèle avec les 15 meilleurs SNP univariés
- Modèle de fusion avec 5 groupes de SNP (15 SNP)

Perspectives

PERSPECTIVES

- **Codage binaire des variables**
- Une solution est un code binaire contenant p bits

Nom des variables	Var_1	Var_2	Var_3	…	Var_i	…	Var_p
Codage binaire d'1 individus	1	0	0		1		1

- Ne limite pas la taille de la solution
- Pour favoriser les bonnes solutions utilisant peu de variables, on peut pénaliser la fonction de fitness par le nombre de variables sélectionnées ou mettre en œuvre l'aspect multicritère des algorithmes génétiques.

PERSPECTIVES

Méthode gourmande

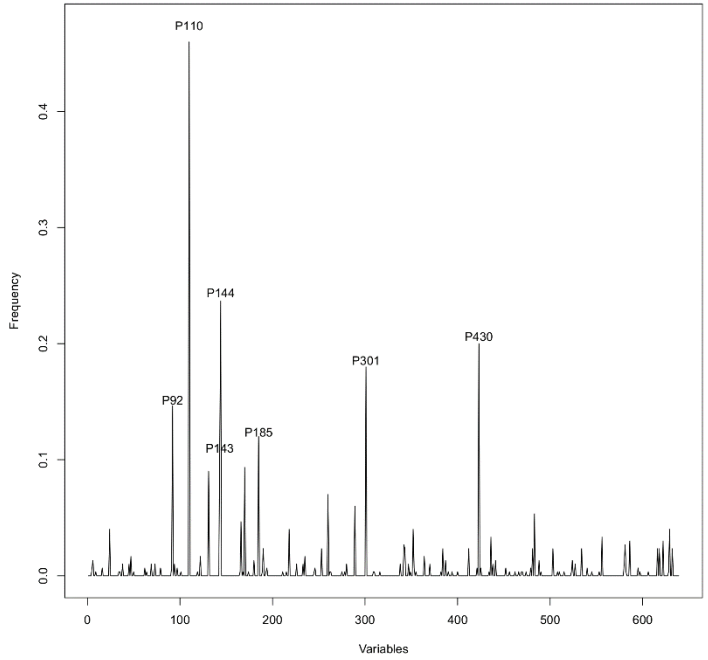
- Mais facilement parallélisable (notamment l'implémentation galgo sous R)
- Parallélisable dans les itérations, les évaluations des individus sont indépendantes
- Parallélisable à un plus haut niveau

→ **Modèle en îlots [6]**

- Consiste à créer plusieurs populations, chaque sous-population évoluant sur un processeur suivant le schéma classique auquel vient s'ajouter une étape de **migration** : chaque sous-population envoie ses meilleurs individus soit vers les populations voisines soit dans un « pool » commun.
- Chaque sous-population reçoit ensuite des individus soit envoyés par ses voisins soit péchés dans le ``pool'' central.
- Permet de faire évoluer les sous-populations avec des paramètres différents (l'un des inconvénients de la méthode des algoG)

PERSPECTIVES

- **Statistique sur les meilleures solutions**
  - On prend généralement la meilleure solution de la population finale comme étant la réponse au problème.
  - Or, les solutions classées juste en-dessous sont également d'intérêt, parfois même peuvent être considérées *ex aequo*.
  - Certaines variables peuvent apparaitre fréquemment dans les topk solutions
  - → Exploiter l'information de sélection de variables issue des topk solutions



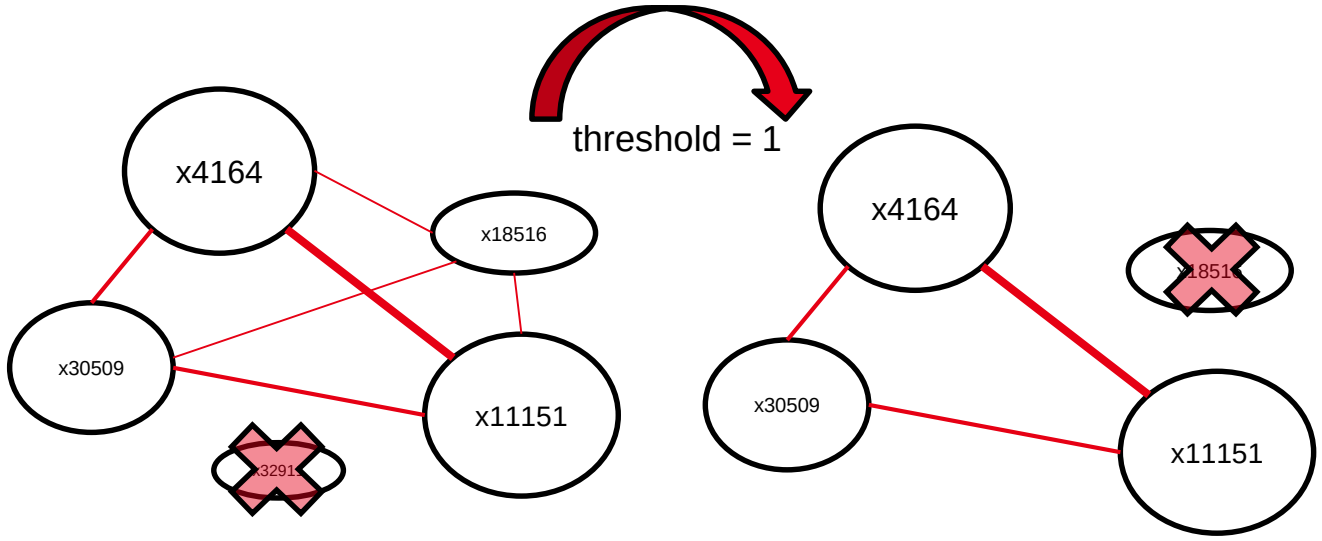
PERSPECTIVES

- **Statistique sur les meilleures solutions**

- Graphe d'interaction via une matrice d'adjacence issue des topk solutions
- Exemple de solutions

x4164	x18516	x11151	x30509	x32911
1	1	1	1	0
1	0	1	0	0
1	0	1	1	0
1	0	1	0	0

	x4164	x18516	x11151	x30509	x32911
x4164	4	1	4	2	0
x18516	1	1	1	1	0
x11151	4	1	4	2	0
x30509	2	1	2	2	0
x32911	0	0	0	0	0



PERSPECTIVES

- **D’autre algorithmes inspirés du vivant:**
  - Algorithme des colonies de fourmis [7]
  - Système immunitaire artificiel [8]

- **Autre application des algorithmes génétiques**
  - Programmation génétique [9]

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

Établissement public à caractère industriel | RCS Paris B 775 685 019

BIBLIOGRAPHIE

[1] Guyon I., Weston J., Barnhill S. and Vapnik V., Gene Selection for Cancer Classification using Suport Vector Machines, Mach.Learn 46, 389–422, (2002)

[2] Breiman L., Random Forests, Mach.Learn 45, 5–32, (2001)

[3] Algorithmes d'échange double de Fedorov: **https://hal.archives-ouvertes.fr/hal-00160194/document**

[4] Trevino V. and Falciani, F., GALGO: an R package for multivariate variable selection using genetic algorithms, Bioinformatics 22, 1154-1156, (2006)

[5] Holland J. H., Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, (1975)

[6] Modèle en îlots (parallélisme): **https://tel.archives-ouvertes.fr/tel-01293722/document** **§1.4**

[7] Colonies de fourmis: **http://www.i3s.unice.fr/~crescenz/publications/travaux_etude/colonies_fourmis-200605-rapport.pdf**

[8] Système immunitaire artificiel: **https://hal.inria.fr/inria-00347211/document**

[9] Programmation génétique: **https://tel.archives-ouvertes.fr/tel-00918968/document**

Packages

Galgo: package sous R, de nouveau maintenu (dernière version octobre 2018):
https://cran.r-project.org/web/packages/galgo/index.html

GA: package sous R
https://cran.r-project.org/web/packages/GA/index.html

Sous python: sklearn-genetic (je ne l'ai jamais testé)
https://pypi.org/project/sklearn-genetic/