

## Modèles et vecteurs

Cours 1 — Modèles en biologie

Christophe Ambroise

L3 GBI — Université d'Évry

## Génétique des populations

Novembre *et al.*, **Genes mirror geography within Europe**, *Nature*, 2008.

**3 000** individus européens

Tableau : 3 000 × 500 000.

**500 000** SNP par individu

### À l'échelle

Écrit petit : **1 m** de haut, **166 m** de large.

## Un problème concret

### Un SNP

Position du génome variable entre individus.

Deux allèles : **C** ou **T**.

Codage : nombre d'allèles **T**.

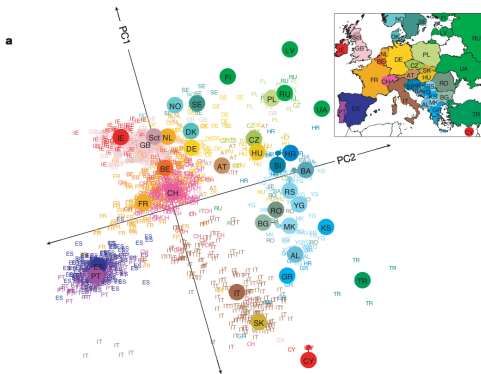
| Mère \ Père | C | T |
|-------------|---|---|
| <b>C</b>    | 0 | 1 |
| <b>T</b>    | 1 | 2 |

## Le tableau

|                | SNP 1 | SNP 2 | SNP 3 | SNP 4 | ... | SNP 500 000 |
|----------------|-------|-------|-------|-------|-----|-------------|
| Individu 1     | 0     | 2     | 1     | 0     | ... | 1           |
| Individu 2     | 1     | 2     | 0     | 2     | ... | 0           |
| Individu 3     | 0     | 1     | 1     | 0     | ... | 2           |
| ⋮              | ⋮     | ⋮     | ⋮     | ⋮     |     | ⋮           |
| Individu 3 000 | 2     | 0     | 1     | 1     | ... | 1           |

Des nombres. Rien d'autre.

## Résumé en deux dimensions



500 000 colonnes → 2. Novembre *et al.*, 2008.

## Un individu, un point

Un individu = un point de  $\mathbb{R}^{500\,000}$ .

Indessinable.

## Question

Comment **résumer** ce tableau ?

## Une carte tirée de l'ADN

Données : ADN seulement, aucune information géographique.

## Objectif

Comprendre comment générer et interpréter cette figure d'**analyse en composantes principales** (séance 16).

# Des modèles, un outil

## Un modèle en biologie

**Modèle** : représentation mathématique simplifiée d'un phénomène biologique.

Exemples : canal ionique, métabolisme des mammifères, séquences d'ADN, crabes.

### Idée du semestre

Questions différentes, même outil : l'**algèbre linéaire**.

## Quatre outils, tout le semestre

**Pivot de Gauss** : résoudre un système (réaction, état stationnaire). Séances 2, 3, 4, 6, 16

**Vecteur propre** : état reproduit par le modèle (Markov, population, ACP). Séances 4, 5, 16

**Produit scalaire, projection** : mesurer, comparer (distance, angle, corrélation).

Séances 6 à 11, 14, 16

**Variance, information** : résumer (inertie, entropie, part d'information). Séances 8, 10, 12,

14, 16

# Des vecteurs

## Vecteur de $\mathbb{R}^n$

Une liste ordonnée de  $n$  nombres réels.

$$\mathbf{x} = (x_1, x_2, \dots, x_n), \quad x_i \in \mathbb{R} \quad \Longleftrightarrow \quad \mathbf{x} \in \mathbb{R}^n$$

## Trois lectures d'un vecteur

**Profil** : masse et métabolisme d'un animal ; expression d'un gène dans  $n$  échantillons.

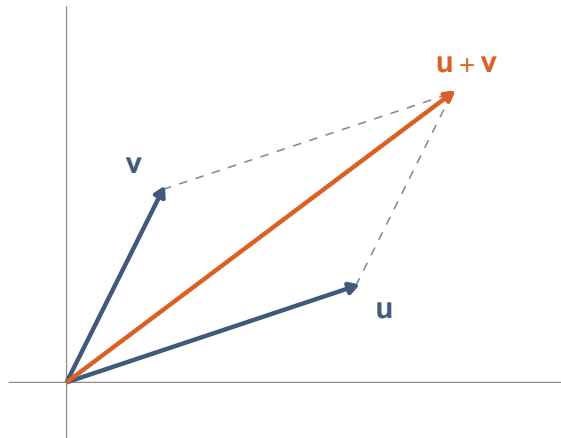
**État** : canal ionique fermé ou ouvert ; cellules quiescentes ou en division.

**Loi** : fréquences de A, C, G, T à une position de l'ADN.

### Même objet

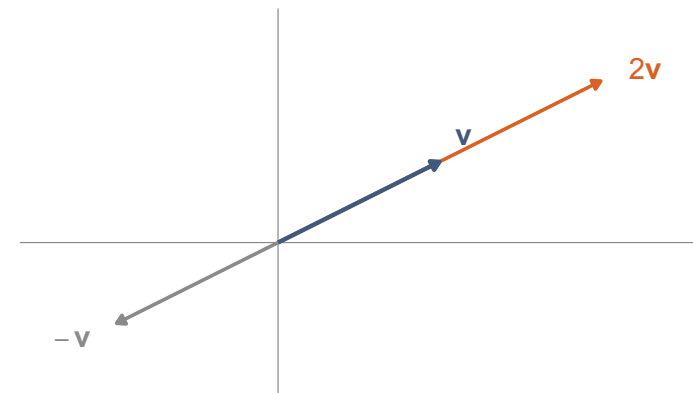
Trois lectures, mêmes opérations.

## Addition



Coordonnée par coordonnée.

## Multiplication par un scalaire

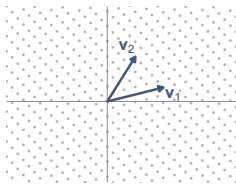


Étirement, compression, retournement. Direction conservée.

# Combinaison linéaire

Somme de vecteurs multipliés par des réels.

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \dots + \alpha_p \mathbf{v}_p, \quad \alpha_1, \dots, \alpha_p \in \mathbb{R}$$



$\alpha_1$  et  $\alpha_2$  quelconques : **tout le plan**.

17

# Organisation

19

# Combinaison linéaire, opération commune

## Opération centrale

Système, pas de temps, centre d'un groupe, axe d'ACP : des **combinaisons linéaires**.

Centre de gravité :  $\frac{1}{n} \mathbf{x}_1 + \frac{1}{n} \mathbf{x}_2 + \dots + \frac{1}{n} \mathbf{x}_n$ .

ACP : 2 axes, combinaisons des 500 000 colonnes.

18

# Le semestre

| Séances | Contenu                                                        |
|---------|----------------------------------------------------------------|
| 1 – 3   | Résoudre : systèmes linéaires, pivot de Gauss                  |
| 4 – 5   | Faire évoluer un état : chaînes de Markov                      |
| 6 – 9   | Mesurer : géométrie de $\mathbb{R}^p$ , distances, similarités |
| 10 – 11 | Regrouper : clustering                                         |
| 12 – 13 | Quantifier : information, entropie                             |
| 14 – 15 | Résumer : variance, covariance, corrélation                    |
| 16      | Résumer : analyse en composantes principales                   |
| 17 – 18 | Analyse de données                                             |

Annexe du livret de TD : résumé des notions d'algèbre linéaire.

20

## Deux livres de référence

**Algèbre linéaire** — David C. Lay, Steven R. Lay, Judi J. McDonald, *Linear Algebra and Its Applications*, 5<sup>e</sup> édition, Pearson, 2016.

Ch. 1, 2, 5, 6, 7 : systèmes, matrices, valeurs propres, moindres carrés, ACP.

**Modèles** — Scott E. Page, *The Model Thinker: What You Need to Know to Make Data Work for You*, Basic Books, 2018.

Ch. 12, 16, 17 : entropie, fonctions de Lyapunov, chaînes de Markov.