Introduction
Annotation des génomes

Méthodes de prédiction: démarche générale

- Définir clairement l'objectif.
- Choisir les critères.
- Choisir le type d'approche :
  - sans système de référence,
  - avec système de référence.

2 étapes

Apprentissage → Détermine les paramètres des fonctions discriminantes

Test → Valide la généralité des fonctions établies

Mesure du pouvoir prédictif d'une méthode

4 paramètres importants :

- pourcentage de vrais positifs (VP, True positive)
- pourcentage de faux positifs (FP, False positive)
- pourcentage de vrais négatifs (VN, True negative)
- pourcentage de faux négatifs (FN, False negative)

		Réalité	
		Groupe 1	Groupe 2
prédiction	Groupe 1	% vrais positifs	% faux positifs
	Groupe 2	% faux négatifs	% vrais négatifs

Groupe 1 : exemples
Groupe 2 : contre-exemples

Mesure du pouvoir prédictif d'une méthode

Idéal: prédire le maximum d'exemples (max VP) avec un minimum d'erreurs (min FP). Mais valeurs non indépendantes donc impossible.
Solution un compromis:

- on maximise le % de VP (donc minimise le % de FN) souvent par utilisation de critères moins stricts même si cela entraîne l'augmentation du % de FP. L'élimination des FP se fait par un autre traitement informatique ultérieur. On dit que l'on privilégie la sensiblilité de la méthode

- inversement, on minimise le % de FP même si cela conduit à ne pas détecter certaines séquences d'intérêts (donc plus grand % de FN). On dit que l'on privilégie la spécificité de la méthode.

Sensibilité = VP/(VP+FN) sensibility an anglais

Spécificité = VP/(VP+FP) specificity en anglais

précision = (VP+VN)/(VP+VN+FP+FN) accuracy en anglais

Annotation d'un génome

Identification des gènes codant pour :

- les ARNr
- les ARNt
- les protéines

Identification des unités de transcription (promoteur et terminateur)

Identification des unités de traduction

Pour les gènes codant pour les protéines, prédiction fonctionnelle par recherche de similarité de séquences (Blast) et classification en grandes classes fonctionnelles (ex: biosynthèse des acides aminés, métabolisme énergétique….)

Exemple d'annotation d'un génome
Mycoplasma genitalium