Apprentissage à partir de **données bio et médicales**

Antoine Cornuéjols et **Jean-Daniel Zucker**

MMIP, AgroParisTech, Paris

IRD, Paris

éEGC - 5 février 2010

Plan

Plan

Masses de données

Dans tous les secteurs

Nouveauté radicale

- Masses de
- ... données brutes

Masses de données

Dans tous les secteurs

Nouveauté radicale

- Masses de
- ... données brutes
- ... disponibles (??)

Masses de données

Dans tous les secteurs

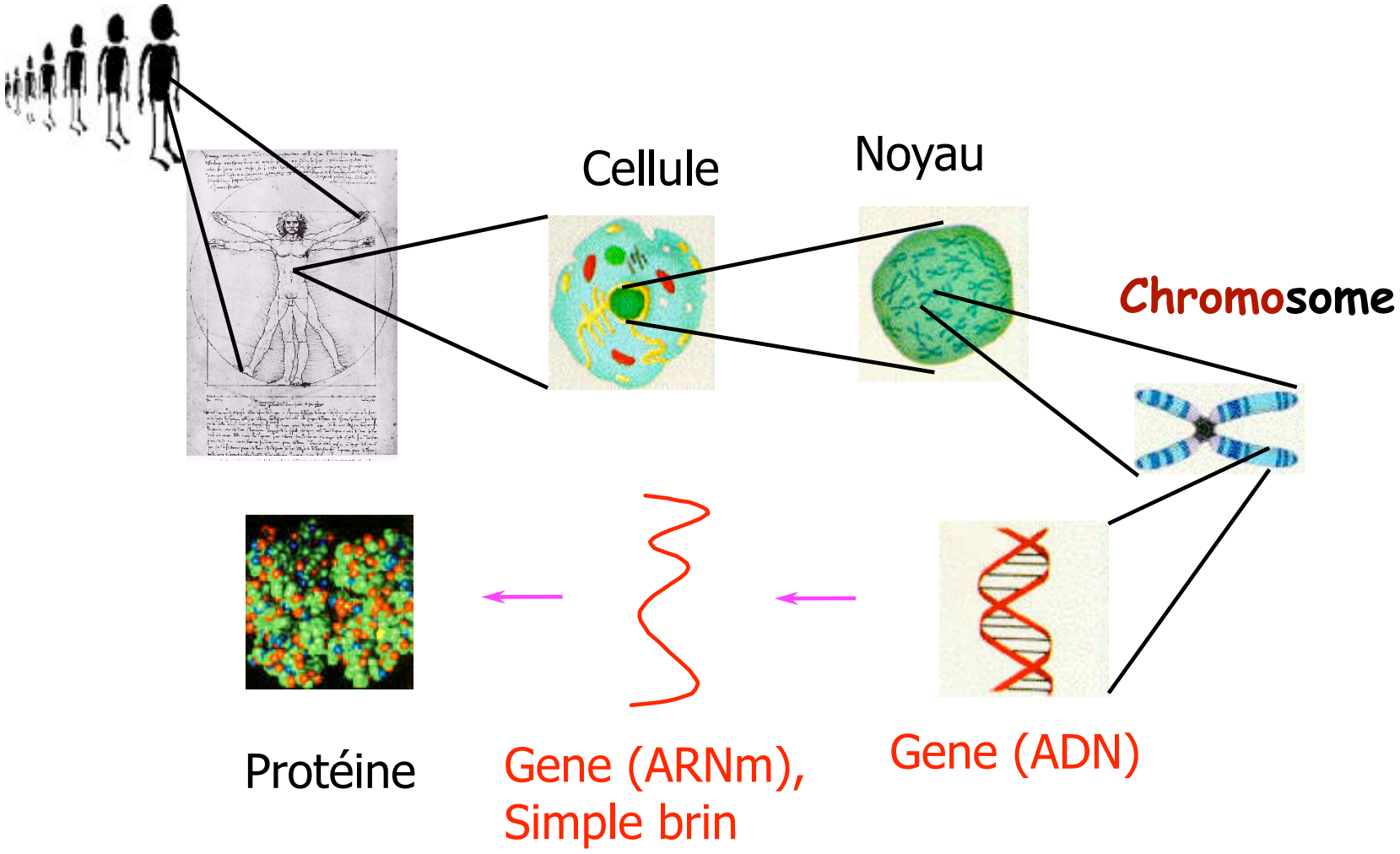
Nouveauté radicale

- Masses de
- ... données brutes
- ... disponibles (??)
- ... dans tous les secteurs du vivant

- Environnement
- Milieux naturels
- Médecine / Santé humaine
- Cellule
- Génome

Masses de données

À toutes les échelles



Masses de données

Tous types de données

Du génome au vivant

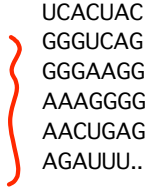
- **Nucléotides** : ADN, ARN, ...
- **Génome** : Séquences, chromosomes, gènes exprimés, ...
- **Protéines** : Séquences, structure 3-D, interactions, ...
- **Systèmes** : réseaux génomiques, réseaux protéiniques, facteurs de transcription, ...
- **Médecine** : symptômes (e.g. prédisposition à l'AVC), effet des médicaments, ...
- **Autres données** : spectrométrie de masse, micropuces, images, ontologies, journaux, ...

But : { **comprendre** / **prédire** } comment le système fonctionne

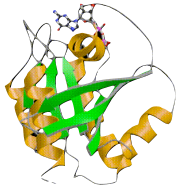
Masses de données

Données ... *omes*

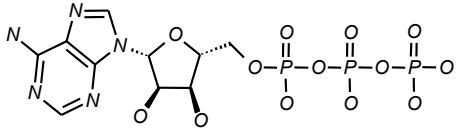
Génome



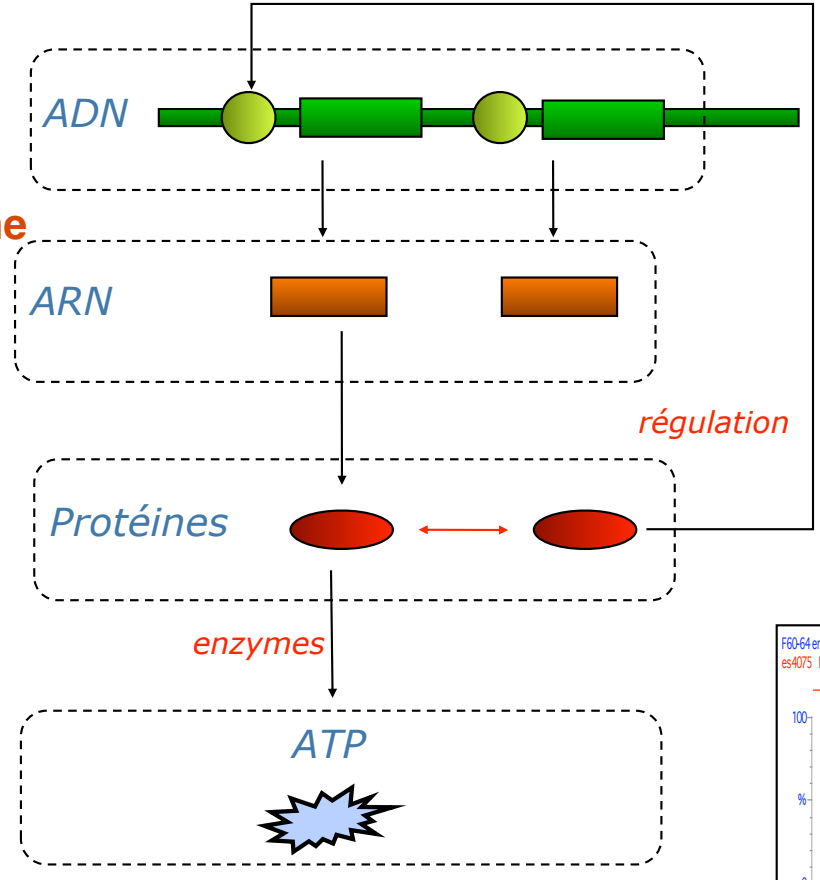
Transcriptome

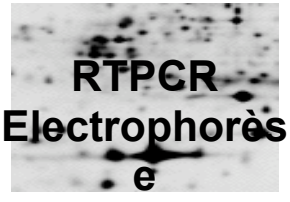


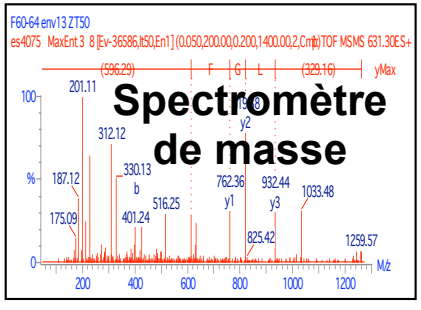
Protéome



Métabolome







Masses de données

Types d'information "omes"

1. **Génome**
   (l'ensemble du matériel génétique d'un individu ou d'une espèce.)

2. **Transcriptome**
   (l'ensemble des ARN messagers transcrits à partir du génome)

3. **Protéome**
   (l'ensemble des protéines exprimés à partir du génome)

4. **Métabolome**
   (l'ensemble des composés organiques (sucres, lipides, amino-acides, ...)

5. **Intéractome**
   (l'ensemble des interactions protéine-protéine)...

Masses de données

Très nombreuses Bases de Données

- Base de Données ADN
  - GenBank, DDBJ, EMBL,...
- Base de Données Protéines
  - PIR, Swiss-Prot, PRF, GenPept, TrEMBL, PDB,...
- Base de Données EST
  - dbEST, DOTS, UniGene, GIs, STACK,...
- Base de Données Structure
  - MMDB, PDB, Swiss-3DIMAGE,...
- Base de Données voies métabol.
  - KEGG, BRITE, TRANSPATH,...
- Base de Données intégrées
  - SRS

Base de Données de Motifs
- Prosite, Pfam, BLOCKS, TransFac, PRINTS, URLs,...

Base de Données sur les maladies
- GeneCards, OMIM, OMIA,...

Base de Données taxonomique

Base de données littérature scient.
- PubMed, Medline,...

Base de données de brevets
- Apipa, CA-STN, IPN, USPTO, EPO, Beilstein,...

Autres...
- RNA databases, QTL...

Masses de données

... en croissance exponentielle

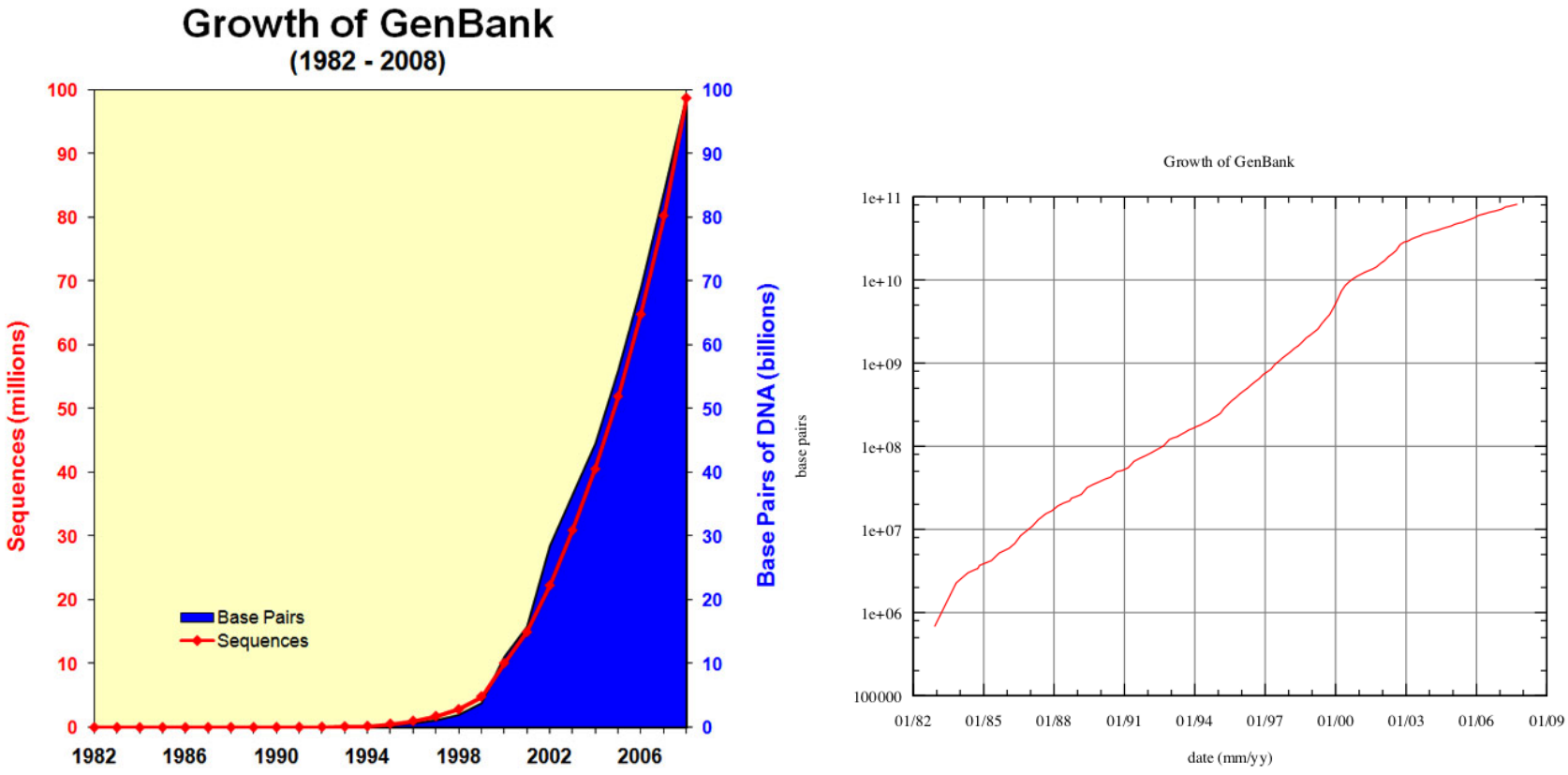


Figure: Croissance des données dans GenBank

Caractéristiques des données et des applications

Grands programmes / Nouveaux paradigmes

De grands programmes

- Human Genome Project
- Human Macrobiome Project
- ...

De nouveaux paradigmes

- “Evidence-Based medicine”

Masses de données

Types d'information "omes"

1. Explosion de la **quantité de données** (ADN 73 Gb, arrivée des données biopuces, voies métaboliques, . . .)
2. **Croissance exponentielle** des données (11-15% tous les 3 mois), plus traitable localement
3. **Données hétérogènes** dans leur structure et leur sémantique
4. **Systèmes d'information hétérogènes**
5. Beaucoup de **connaissances cachées**, privées ou inconnues.
6. ...

Masses de données
Problèmes algorithmiques

Génome

- **Identifier, prédire** les gènes dans une séquence (HMM)
- **Aligner** et comparer des **séquences** (ex: BLAST (Basic Local Alignment Search Tool))

Transcriptome

- **Analyser l'expression** des gènes différentiel. exprimés / conditions
- **Regrouper** des gènes co-exprimés, **Réseaux** de régulation des gènes
- **Identifier** la fonction de gènes.

Protéome

- **Prédire** la **structure** secondaire, la fonction des protéines, ...
- **Analyser, mesurer l'expression** en fonction des organes

Caractéristiques des données et des applications

Questions en BioInformatique

Analyse de séquence

- Alignement de séquences
- Prédiction de structures et de fonctions
- Recherche de gènes

Analyse de structure

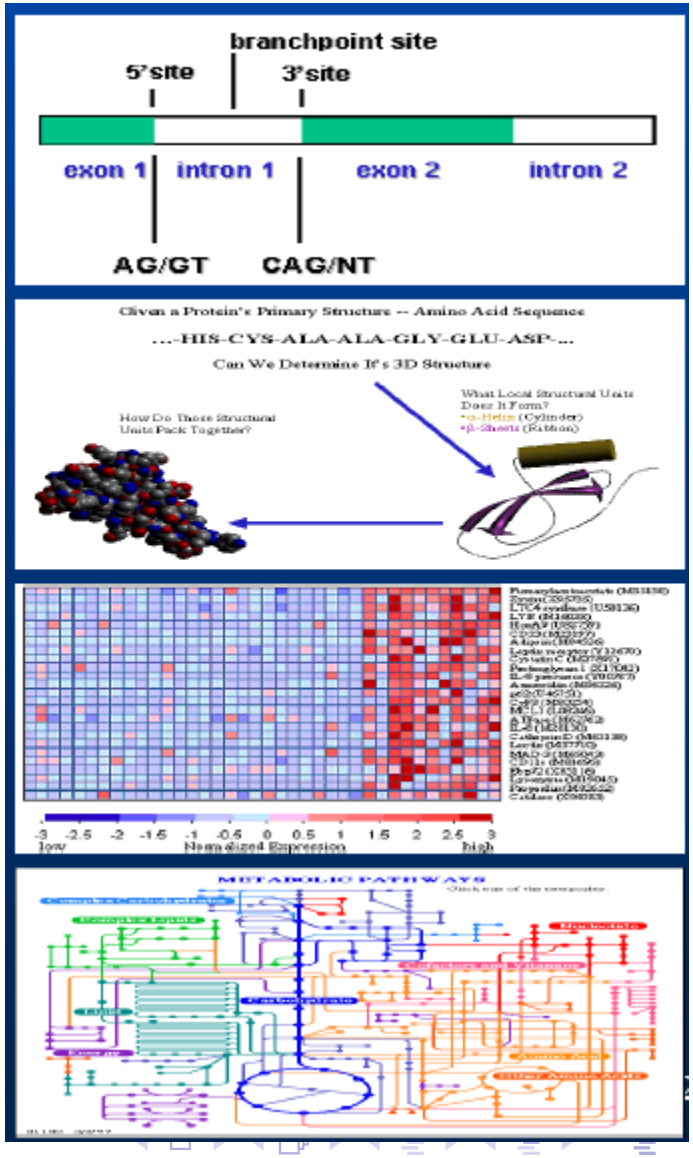
- Comparaison de structures de protéines
- Prédiction de structures de protéines
- Modélisation de la structure de l'ARN

Analyse d'expression

- Analyse de l'expression des gènes
- Clustering de gènes

Analyse d'interactions

- Voies métaboliques
- Réseau de régulation



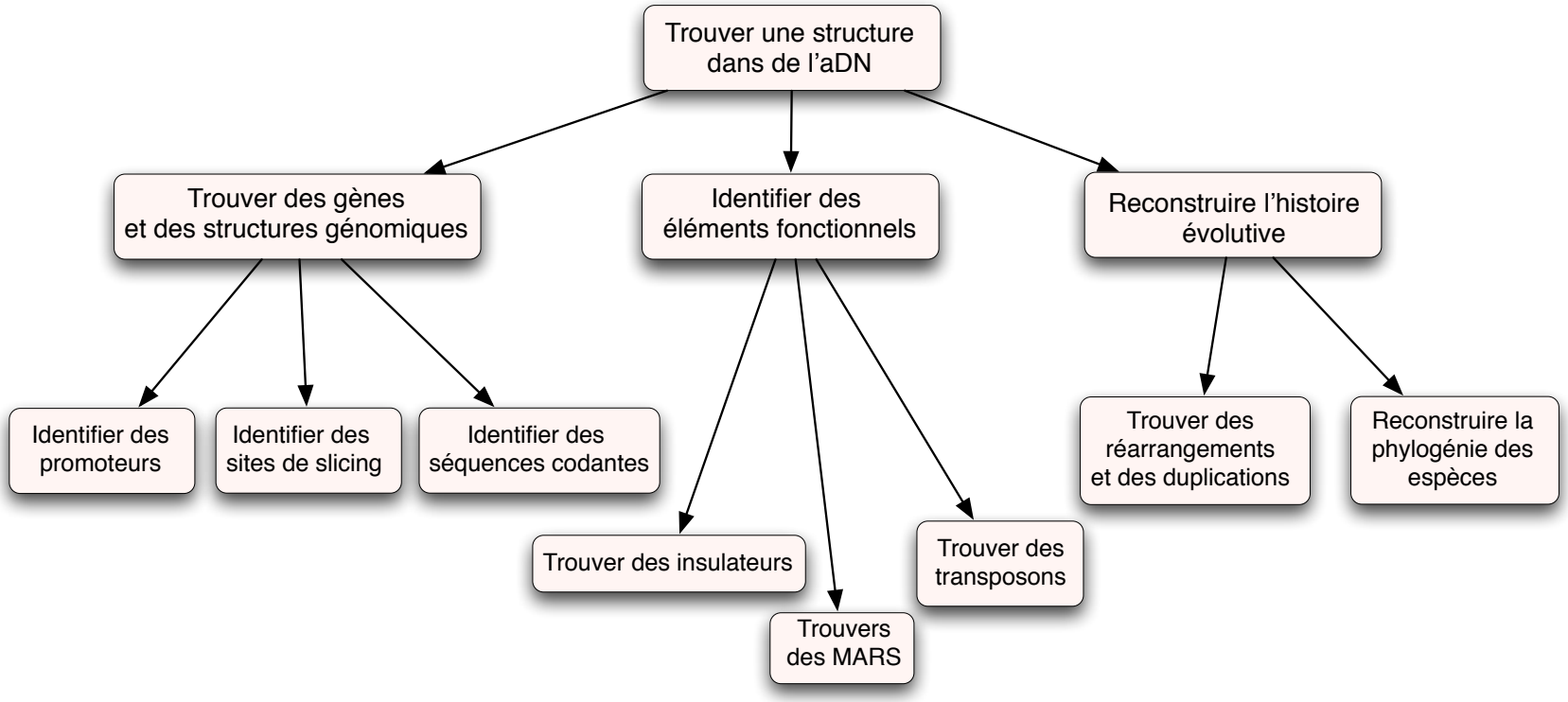
Les grands problèmes
Point de vue de l'Apprentissage Artificiel

- **Décoder** l'information contenue dans les séquences d'ADN et de protéines
  - Trouver les gènes
  - Différencier entre introns et exons
  - Analyser les répétitions dans l'ADN
  - Identifier les sites des facteurs de transcription
  - Étudier l'évolution des génomes
- **Génomique Comparative**
  - Construire les relations de parenté entre organismes
- **Génomique fonctionnelle**
  - Étudier l'expression des gènes
  - Étudier la régulation des gènes
  - Déterminer les réseaux d'interaction entre les protéines
- **Génomique structurale**
  - Modéliser les structures 3D des protéines et des ARN structurels
  - Déterminer la relation entre structure et fonction
- **Pharmacogénomique**

Caractéristiques des données et des applications

Questions en BioInformatique

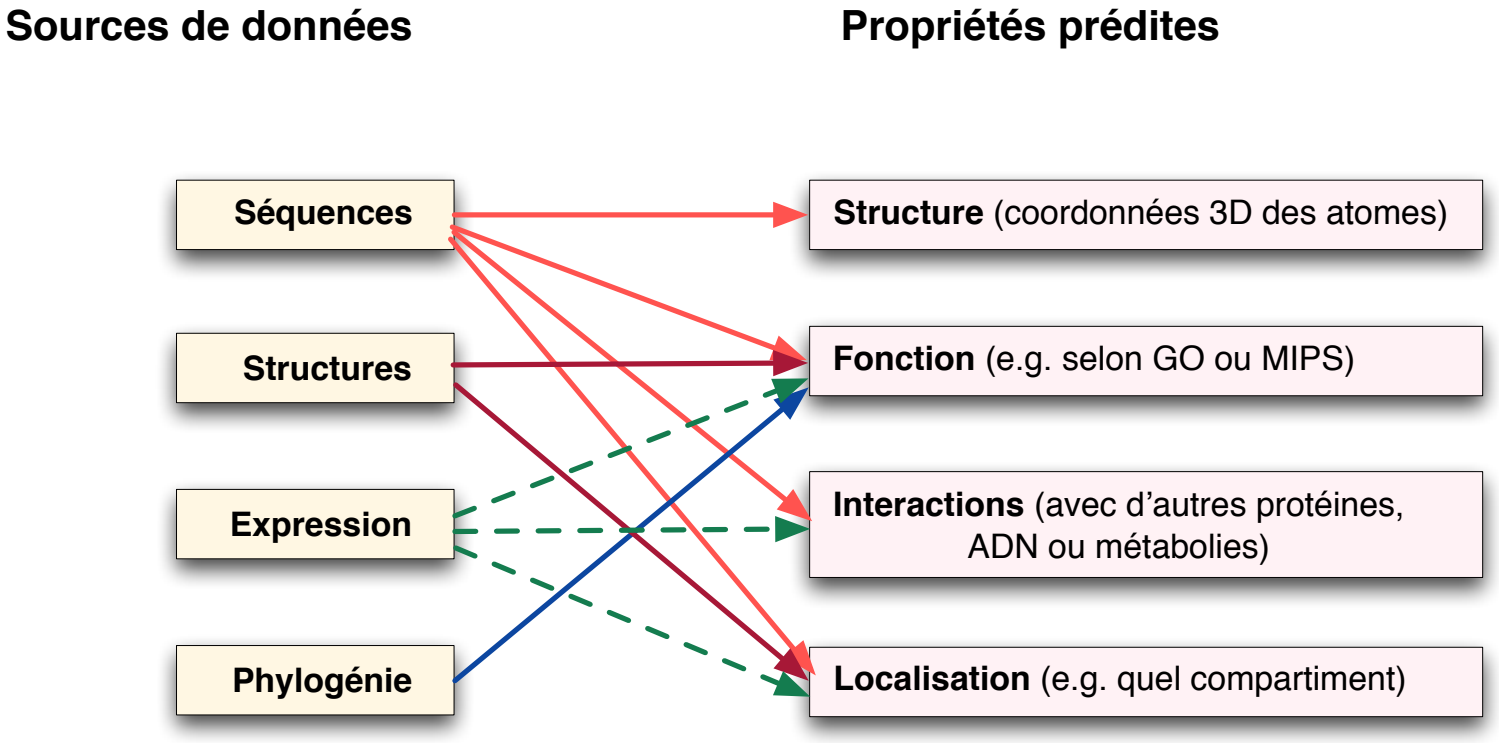
Trouver des structures dans les données génomiques



Caractéristiques des données et des applications

Questions en BioInformatique

Problèmes de prédiction et sources de données



Caractéristiques des données et des applications

Questions plus récentes en BioInformatique

- Étude du métabolome : ensemble des petites molécules se trouvant dans un organisme biologique (aspect dynamique)
- Étude de biotopes complets (mares, tract gastro-intestinal, ...)

=> Apprentissage artificiel

Caractéristiques des données et des applications

Questions en Médecine

“Evidence-Based Medecine”

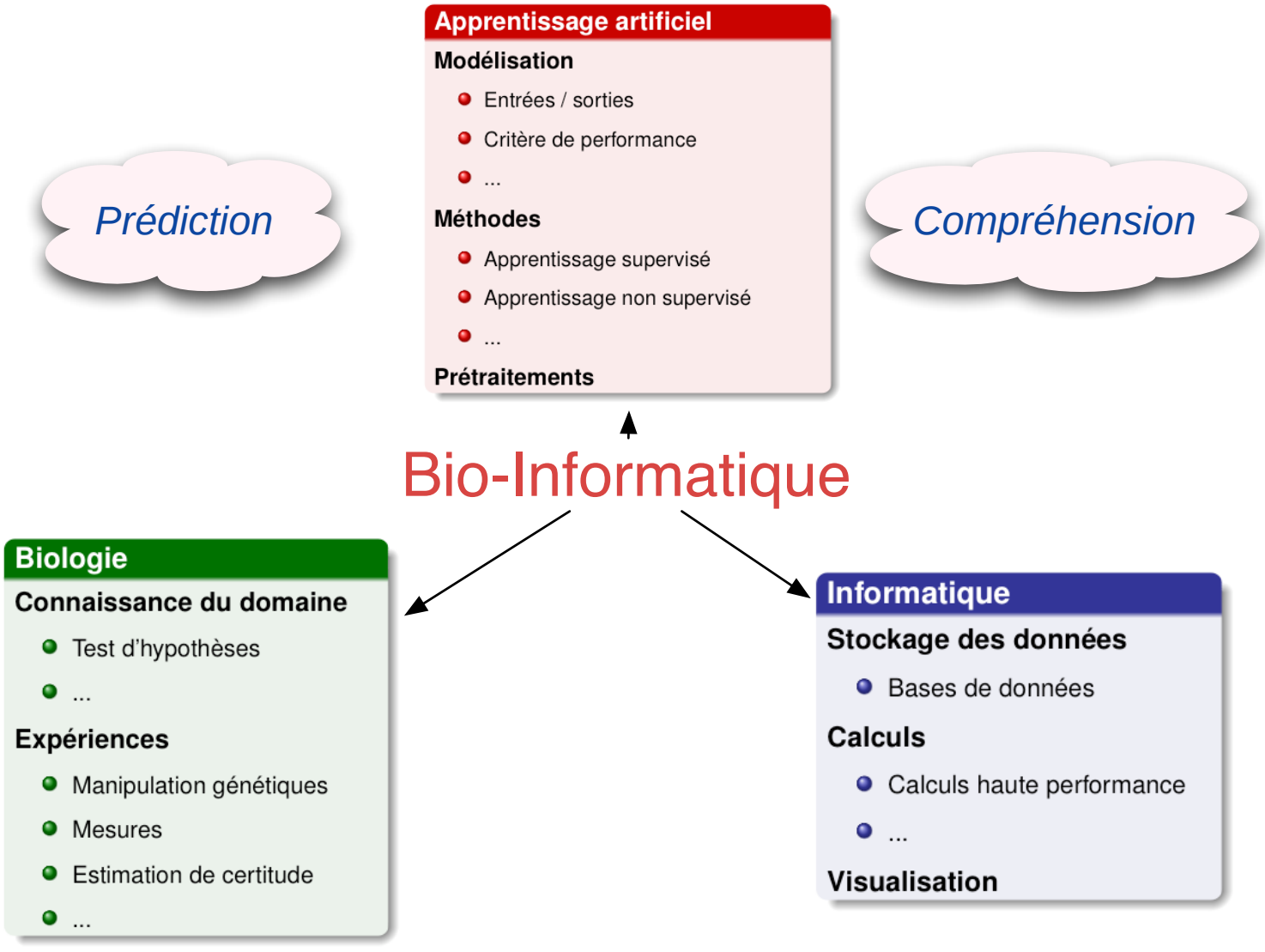
- consiste à baser les décisions cliniques, non seulement sur les connaissances théoriques, le jugement et l’expérience qui sont les principales composantes de la médecine traditionnelle, mais également sur des "preuves" scientifiques
- **données** provenant d’études cliniques systématiques, telles que des essais contrôlés randomisés, des méta-analyses, éventuellement des études transversales ou de suivi bien construites.

=> Recherche d’information et apprentissage artificiel

Plan

Caractéristiques des données et des applications

Une science systémique



Les défis

1. Beaucoup de données :

Les défis

1. Beaucoup de données : la réalité

Les défis

1. Beaucoup de données : la réalité
2. La description est un problème en soi

Les défis

1. Beaucoup de données : la réalité
2. La description est un problème en soi
3. Qualité des données

Les défis

1. Beaucoup de données : la réalité
2. La description est un problème en soi
3. Qualité des données très aléatoire
   - Tous les défauts évoqués par Laure sont présents
   - Sources, sémantiques, formats **hétérogènes**

Les défis

1. Beaucoup de données : la réalité
2. La description est un problème en soi
3. Qualité des données très aléatoire
   - Tous les défauts évoqués par Laure sont présents
   - Sources, sémantiques, formats **hétérogènes**
4. Types de données
   - Séquences
   - En grande dimension
   - Relationnelles (structures, évolution temporelle, ...)
   - Exemples positifs seuls

Les défis

1. Beaucoup de données : la réalité
2. La description est un problème en soi
3. Qualité des données très aléatoire
   - Tous les défauts évoqués par Laure sont présents
   - Sources, sémantiques, formats **hétérogènes**
4. Types de données
   - Séquences
   - En grande dimension
   - Relationnelles (structures, évolution temporelle, ...)
   - Exemples positifs seuls
5. Analyse exploratoire
   - Non supervisé
   - Interactif $\longrightarrow$ interprétable
   - Rôle important d'hypothèses nulles

Les défis

1. Beaucoup de données : la réalité
2. La description est un problème en soi
3. Qualité des données très aléatoire
   - Tous les défauts évoqués par Laure sont présents
   - Sources, sémantiques, formats **hétérogènes**
4. Types de données
   - Séquences
   - En grande dimension
   - Relationnelles (structures, évolution temporelle, ...)
   - Exemples positifs seuls
5. Analyse exploratoire
   - Non supervisé
   - Interactif $\longrightarrow$ interprétable
   - Rôle important d'hypothèses nulles
6. En médecine : calibration

Cas d'école

Le cas INDANA

Différents types

- Numérique
- Symbolique

Différentes sources

- Les descripteurs n'ont pas la même signification
- Biais en fonction de la source

Étiquettes

- ... semi cachées
- biaisées

Cas d'école

Le cas des protéines

Différents types

- Numérique
- Symbolique
- Relationnel (structure, géométrie)

Incertitudes ; Imprécisions

- Protéines en mouvement
- Cristallographie (image statique)
- Difficultés expérimentales

- Seulement des exemples positifs

Caractéristiques des données

Données hétérogènes

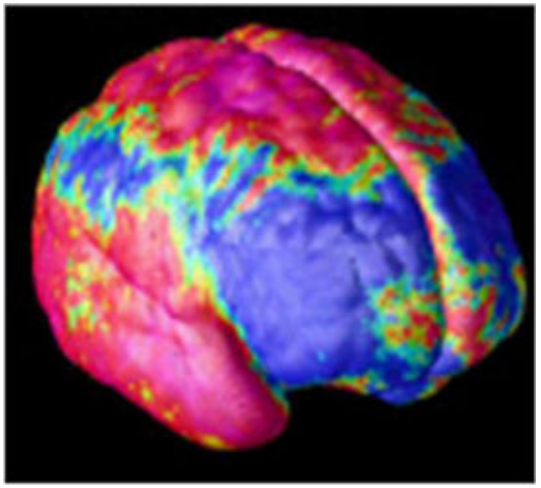
Le cas de la magnetoencéphalographie (EMG)

Différents types

- Signal temporel
- Données spatiales

Imprécisions

- Faible reproductivité pour un sujet
- Sujets différents



Caractéristiques des données

Problème de la description des données

Importance des prétraitements

Caractéristiques des données

Données en grande dimension

Exemples

- Puces à ADN
- Séquences biologiques
- Description de protéines
- ...

Les problèmes

- Stockage
- Traitements (en mémoire centrale)
- Complexité
- Recherche des descripteurs pertinents (validité statistique, ...)
- ...

Les grands problèmes

Problèmes et méthodes

	Sélection d'attributs	Clustering	Apprentissage supervisé	HMM	Réseaux bayésiens
Analyse d'expression	√				
Alignement de séquences		√		√	
Prédiction de structure et de fonctions	√	√	√		
Clustering de molécules		√	√		
Classification de molécules	√		√		
Analyse de dépendances		√	√		√

Plan

Apprentissage de séquences
Les HMMs

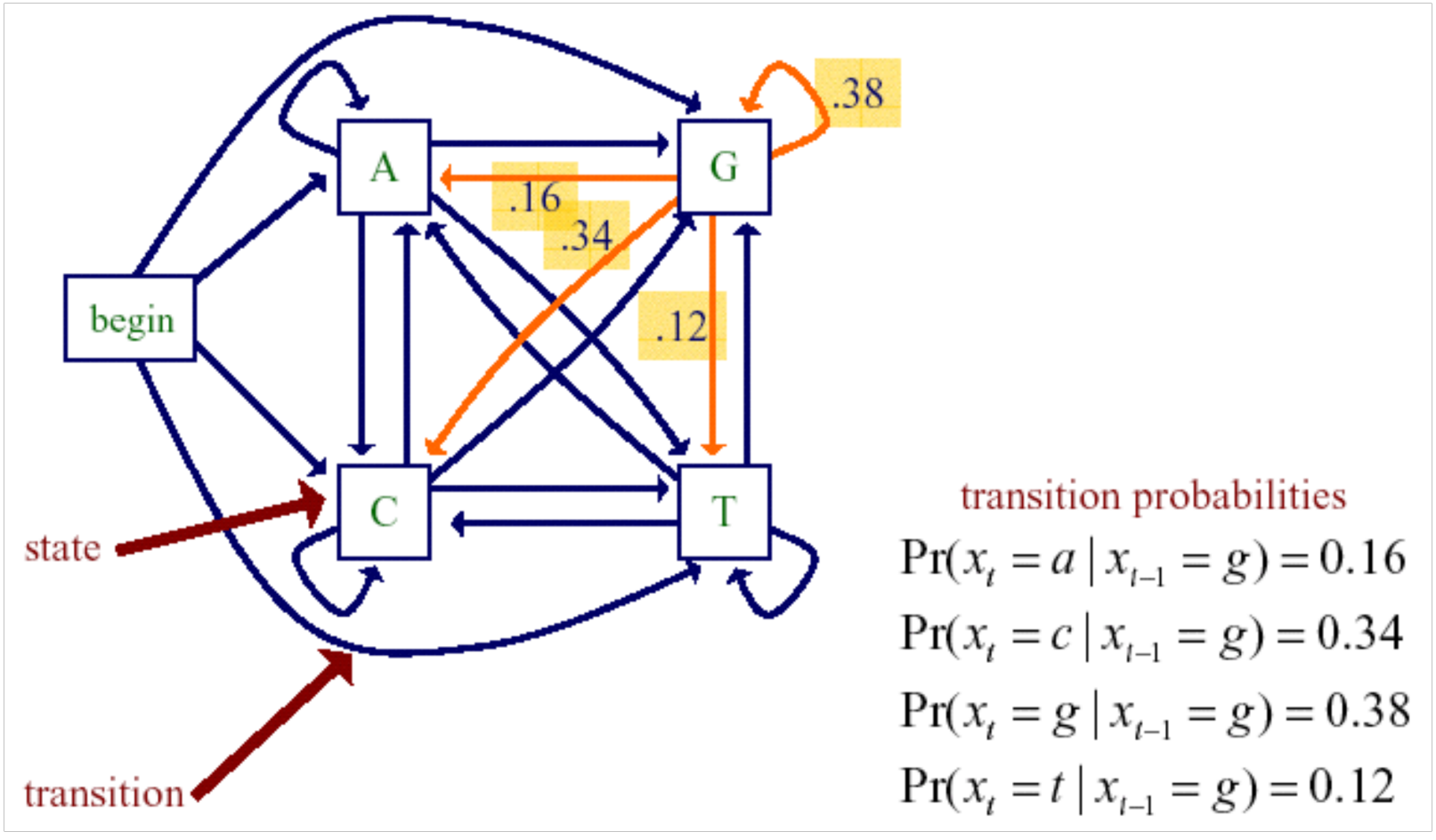
Modèles probabilistes généraux pour les séquences de symboles

Types de questions

- Cette séquence appartient-elle à une famille donnée ?
- Si cette séquence provient de cette famille, quelle est sa structure interne ?
  - *Exemple* : identifier des hélices alpha ou des feuillets beta dans une séquence de protéine
- Meilleur appariement de séquences
  - *Exemple* : Séquençage ou étude de phylogénie

Apprentissage de séquences

Les HMMs



Apprentissage de séquences

Les HMMs

Un **processus stochastique** est un phénomène temporel où intervient le hasard, i.e. une variable aléatoire X(t) évoluant en fonction du temps.

Ex: Une suite de lancers de dés 1,3,2,5,3,6,2,4

Un processus stochastique est dit **markovien** si son évolution ne dépend pas de son passé, mais uniquement de son état présent.

Si à l'instant t_i, on a observé la réalisation x_i de $X(t)$ alors :

$$Pr\big(X(t_n) \leq x_n | X(t_{n-1}), ..., X(t_1)\big) \;=\; Pr\big(X(t_n) \leq x_n | X(t_{n_1})\big)$$

Apprentissage de séquences
Les HMMs

On parle de chaîne de Markov lorsque le processus $X(t)$ est discret.

$$Pr\big(X(t_n) = x_n | X(t_{n-1}) = x_{n-1}, ..., X(1) = x_1\big) \;=\; Pr\big(X(t_n) = x_n | X(t_{n_1} = x_{n-1})\big)$$

est noté par abus de notation :

$$Pr\big(X(t_n) = x_n | X(t_{n-1}) = x_{n-1}\big) \;=\; Pr(x_n | x_{n-1})$$

Exemple : Alphabet $\Sigma = \{$`'A','C','G','T'`$\}$

Séquence : `ACGCCTAGGCTAGCTTATCG`

Apprentissage de séquences

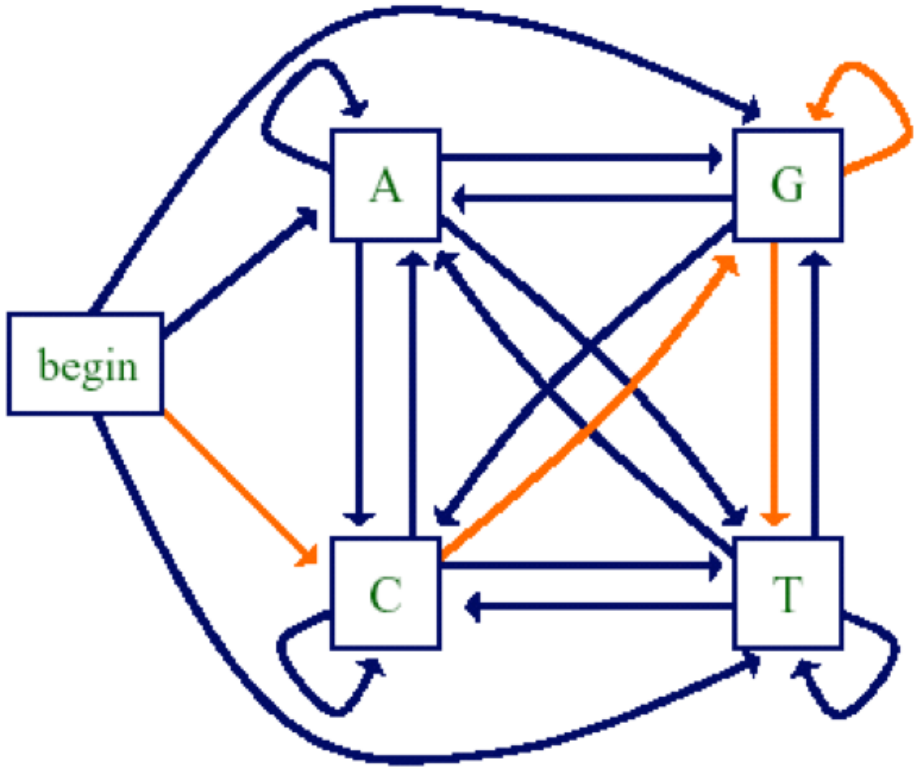
Les HMMs

Probabilité d'une séquence $\mathcal{O} = (x_1, \ldots, x_n)$

$$\mathbf{P}(\mathcal{O}) = \mathbf{P}(x_n, x_{n-1}, \ldots, x_1) = \mathbf{P}(x_1) \times \Pi_{t=2}^{n} \mathbf{P}(x_n | x_{n-1})$$

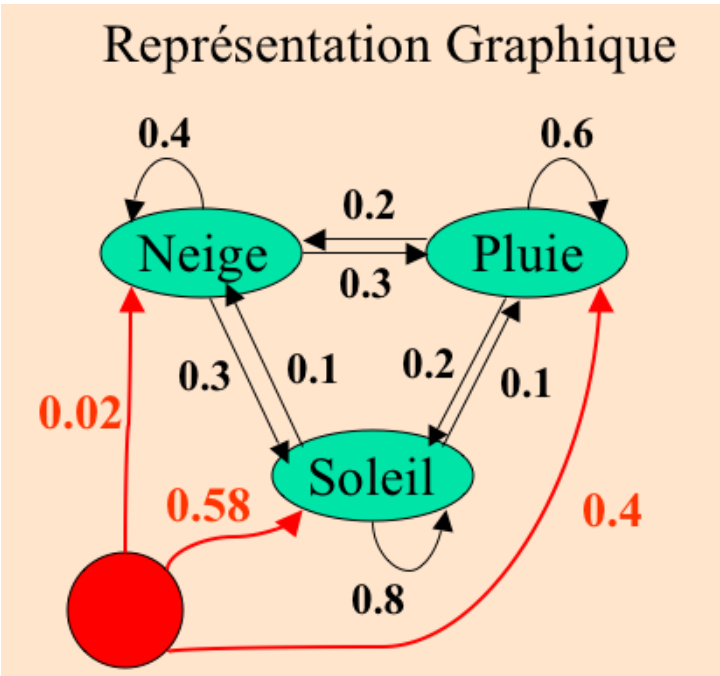
Apprentissage de séquences

Les HMMs



$$\mathbf{P}(cggt) = \mathbf{P}(c) \cdot \mathbf{P}(g|c) \cdot \mathbf{P}(g|g) \cdot \mathbf{P}(t|g)$$

Apprentissage de séquences
Les HMMs



Quelle est la probabilité qu'il fasse '*soleil*' 5 jours de suite ?

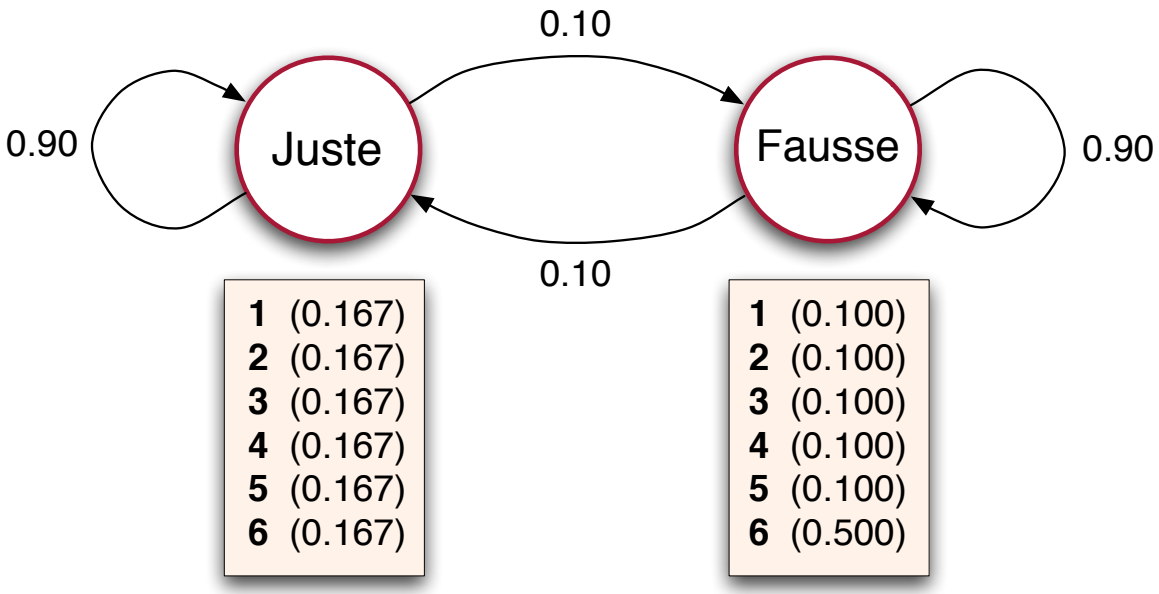
$$\mathbf{P}(Soleil, Soleil, Soleil, Soleil, Soleil) = 0.58 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = 0.2375$$

$$\mathbf{P}(Neige, Neige, Pluie, Pluie, Soleil) = 0.02 \times 0.4 \times 0.3 \times 0.6 \times 0.2 = 0.000288$$

Apprentissage de séquences
Les HMMs

Distinction entre symbole émis x_i (observé) et état caché s_k

Plusieurs états peuvent désormais être responsables du symbole observé.



Caché : **s** = 11111111111111111111122221111111222222222

Visible : **x** = 45536531633635551333626651321416366516666

Apprentissage de séquences
Les HMMs

Définition d'un HMM

- Alphabet $S = \{s_1, \ldots, s_L\}$ des **états** de la chaîne de Markov (L états distincts)
- **Matrice de transition** $\mathbf{A} = \{a_{i,j} = \mathbf{P}(s_i|s_j)\}$
- Probabilités de départ $\Pi = \{\pi_i = \mathbf{P}(s_i)\}$
- Alphabet $\Sigma = \{x_1, \ldots, x_M\}$ des symboles émis par les s_i pour un HMM (M symboles distincts)
- Probabilité d'émission $E = \{e_j(x_k) = \mathbf{P}(x_k|s_i)\}$

Apprentissage de séquences

Que peut-on faire avec les HMMs

Problème 1 : Étant donnés un HMM $H = \langle \Sigma, S, \Pi, A, E \rangle$ et une séquence observée $\mathcal{O} = x_1, \ldots, x_T$

quelle est la vraisemblance $\mathbf{P}(\mathcal{O}|H)$ de $\mathcal{O}$ suivant H ?

Problème 2 : Étant donnés un HMM $H = \langle \Sigma, S, \Pi, A, E \rangle$ et une séquence observée $\mathcal{O} = x_1, \ldots, x_T$

quelle est la séquence $\mathcal{Q}$ des états qui a la probabilité maximale d'avoir engendré $\mathcal{O}$?

Problème 3 : À partir d'un ensemble d'observations $\mathbf{x} = \{x_1, \ldots, x_n\}$

ajuster les paramètres du HMM H afin de maximiser la vraisemblance de l'ensemble d'apprentissage $\mathbf{P}(\mathcal{O}|H)$?

Apprentissage de séquences

Les HMMs : **tâches d'apprentissage et de prédiction**

1. **Classification**
   - Étant donnés un ensemble de modèles représentant différentes classes de séquences et une séquence test
   - **Trouver quel modèle** explique le mieux la séquence
2. **Segmentation**
   - Étant donnés un modèle représentant différentes classes de séquences et une séquence test
   - **Segmenter** la séquence en sous-séquences, en **prédisant la classe de chaque sous-séquence**
3. **Apprentissage**
   - Étant donnés un modèle et des séquences
   - **Trouver les paramètres d'un modèle** permettant de rendre compte avec une grande probabilité des séquences d'apprentissage (le but étant de généraliser à d'autres séquences)

Apprentissage de séquences

Les HMMs : **tâches d'apprentissage et de prédiction**

1. **Classification** : probabilité d'une séquence d'observables donnée
   - Modèle de Markov simple
     => Calcul de probabilité de séquence le long de chemins simples pour chaque modèle
   - Modèle de Markov à état caché (HMM)
     => **Algorithme Forward** pour calculer la probabilité de séquence le long de tous les chemins pour chaque modèle

2. **Segmentation** : calcul de la séquence d'états la plus probable
   - Modèle de Markov à état caché (HMM)
     => **Algorithme de Viterbi**

3. **Apprentissage** : des paramètres du modèle
   - Modèle de Markov simple
     => Méthode de **maximum de vraisemblance** ou d'**estimation bayésienne**
   - Modèle de Markov à état caché (HMM)
     => **Algorithme Forward-Backward + ML : *Baum-Welch***

Apprentissage et séquences

Les HMMs : **Exemples d'applications en biologie : la segmentation**

Apprentissage et **segmentation**

Chaque nucléotide d'une séquence d'ADN appartient soit à une région normale (**N**) soit à une région riche en GC (**R**).

On suppose de plus que les catégories (**N**) et (**R**) ont tendance à former des îots.

Exemple :
NNNNNNNNNRRRRRNNNNNNNNNNNNNNNNNNRRRRRRRNNNN

Ces catégories sont non observées. Ce qui est mesuré peut être :

TTACTTGACGCCAGAAATCTATATTTGGTAAACCCGACGCTAA

Supposons qu'une analyse ait permis d'obtenir les estimations :

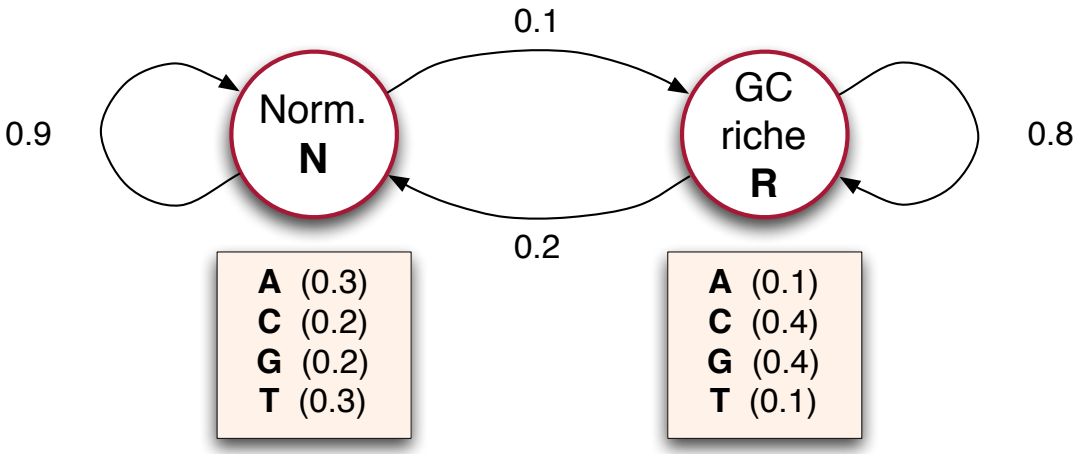
	A	**T**	**G**	**V**	**longueur moyenne**
Normal (N)	0.3	0.3	0.2	0.2	10
G-C rich (R)	0.1	0.1	0.4	0.4	5

Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

Apprentissage

	A	T	G	V	longueur moyenne
Normal (N)	0.3	0.3	0.2	0.2	10
G-C rich (R)	0.1	0.1	0.4	0.4	5



Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

Apprentissage et **Segmentation**

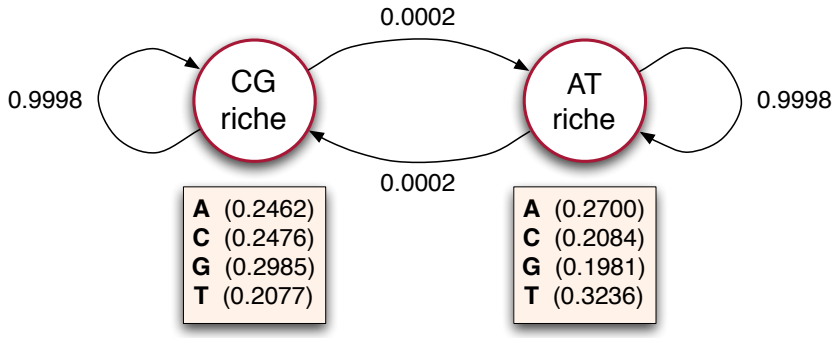


Figure: Obtenu par algorithme EM en partant de matrices de transition et d'émission aléatoires.

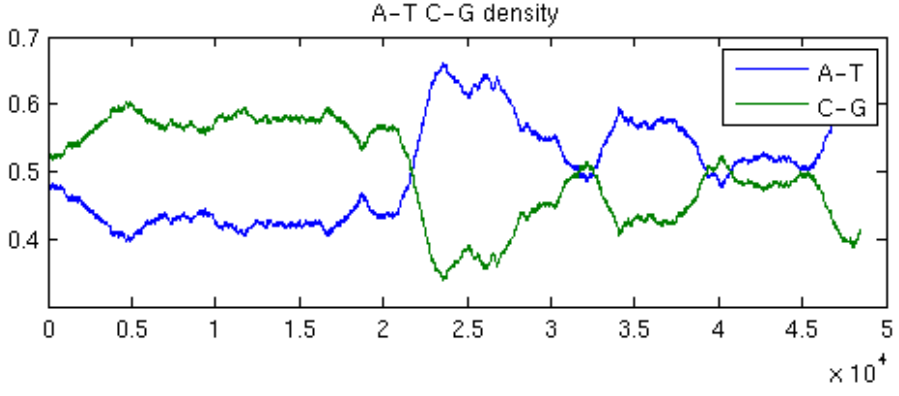


Figure: Segmentation trouvée

Apprentissage de séquences
Exemples d'applications en biologie

Alignements de séquences : Algorithme de Viterbi

- Substitution de T en T → +2

Score de ACGGCTA / ??? / ACTGTA + 2

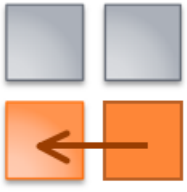
- Délétion de T → -2

Score de ACGGCTA / ??? / ACTGTAT - 2

- Insertion de T→ -2

Score de ACGGCTAT / ??? / ACTGTA - 2

Insertion



Délétion

Substitution
Ou Identité

Apprentissage de séquences
Exemples d'applications en biologie

Alignements de séquences : Algorithme de Viterbi

	-	A	C	G	G	C	T	A	T
-	0	-2	-4	-6	-8	-10	-12	-14	-16
A	-2	2	0	-2	-4	-6	-8	-10	-12
C	-4	0	4	2	0	-2	-4	-6	-8
T	-6	-2	2	3	1	-1	0	-2	-4
G	-8	-4	0	4	5	3	1	-1	-3
T	-10	-6	-2	2	3	4	5	3	1
A	-12	-8	-4	0	1	2	3	7	5
T	-14	-10	-6	-2	-1	0	4	5	9

```
ACGGCTAT
|||| |||
ACTG-TAT
```

Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

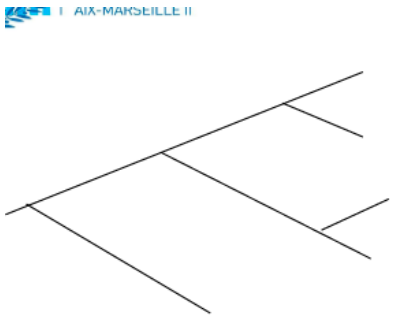
Alignements multiples

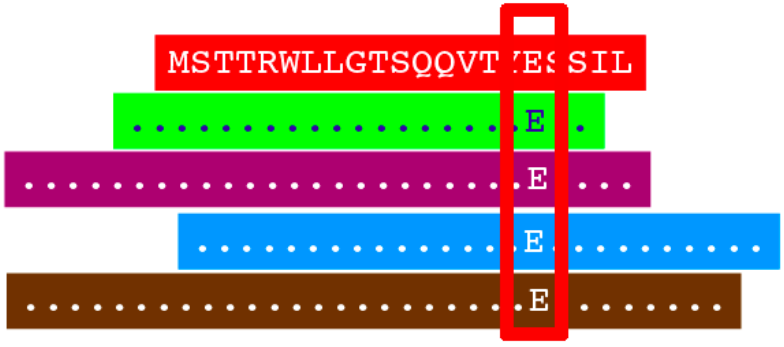
```
INS_HUMAN          1 MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY     50
                     ||||:|.|||||||...||..:.|..||||||||||||||||||||||||
INS_CHICK          1 MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFY     50

INS_HUMAN         51 TPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIC    100
                     :||.||:.|...|.....|   .||.|.....|....|||||||||.:.|
INS_CHICK         51 SPKARRDVEQPLVSSPLRG---EAGVLPFQQEEYEKVKRGIVEQCCHNTC     97

INS_HUMAN        101 SLYQLENYCN    110
                     ||||||||||
INS_CHICK         98 SLYQLENYCN    107
```

Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

Alignements multiples





alignement multiple:permet d'identifier les AA/nucléotides invariants dans des séquences **homologues**

"pression évolutive"

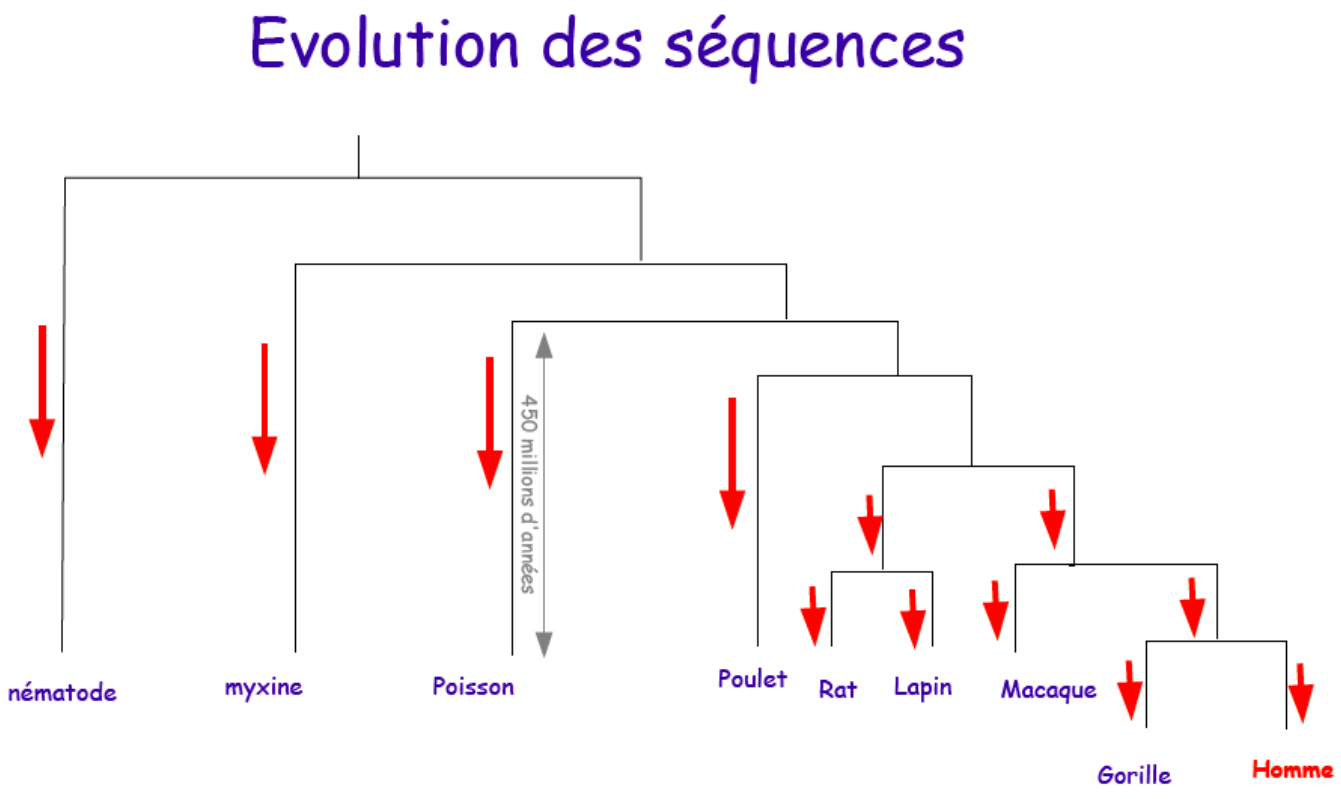
fonction ?

("pourquoi le glutamate est-il conservé dans toutes les séquences ???")

Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

Alignements multiples



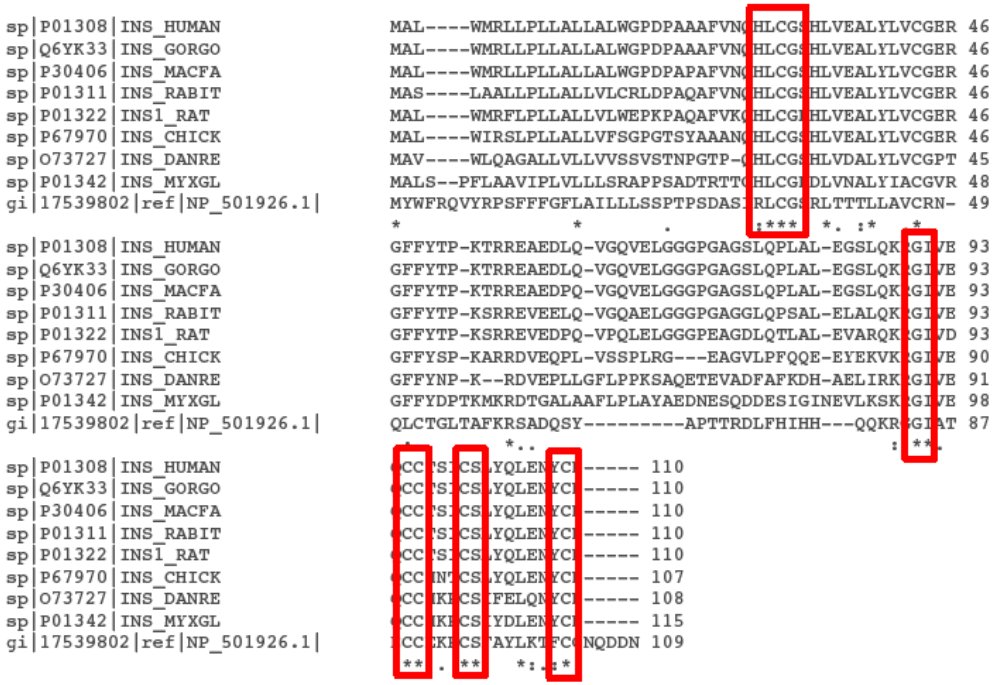
Evolution **indépendante** des différentes séquences

- mutations
- insertions/délétions

Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

Alignements multiples



homme/gorille/macaque
rat/poulet/poisson/
myxine/nématode

Certains acides aminés n'ont jamais subi de mutations
en 500 millions d'années d'évolution: ils ont été **préservés**

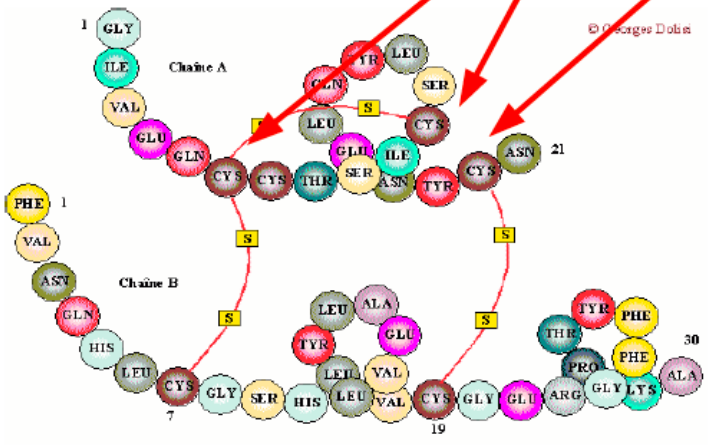
Pourquoi ?

Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

Alignements multiples

```
sp|P01308|INS_HUMAN          QCCTSICSLYQLENYCN----- 110
sp|Q6YK33|INS_GORGO          QCCTSICSLYQLENYCN----- 110
sp|P30406|INS_MACFA          QCCTSICSLYQLENYCN----- 110
sp|P01311|INS_RABIT          QCCTSICSLYQLENYCN----- 110
sp|P01322|INS1_RAT           QCCTSICSLYQLENYCN----- 110
sp|P67970|INS_CHICK          QCCHNTCSLYQLENYCN----- 107
sp|O73727|INS_DANRE          QCCHKPCSIFELQNYCN----- 108
sp|P01342|INS_MYXGL          QCCHKLCSIYDLENYCN----- 115
gi|17539802|ref|NP_501926.1| RCCEKRCSFAYLKTPCCNQDDN 109
                             :** . *:  *:. *
```



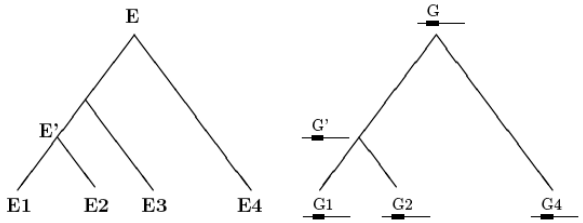
Certains acides aminés ont une fonction **essentielle** dans la protéine: toute mutation rend la protéine non-fonctionnelle **pression évolutive** pour préserver ces acides aminés

Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

Alignements multiples

```
V I V A L A S V E G A S
V I V A D A — V I — — S
V I V A D A L L — — A S
```



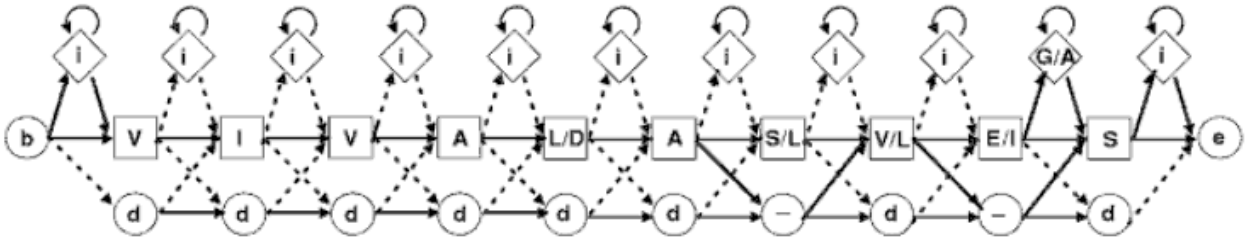


Figure: HMM pair pour l'alignement multiple ci-dessus.

Apprentissage de séquences

Les HMMs : **Exemples d'applications en biologie**

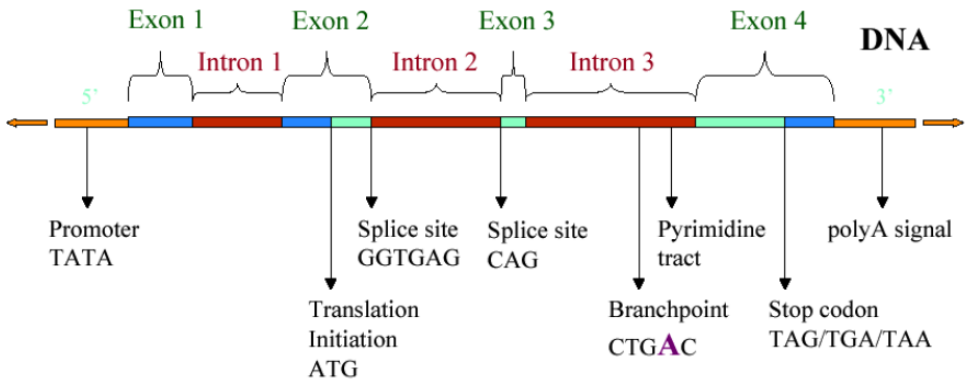
Autres applications

Recherche de gènes

Détection introns / exons

Annotation de génome

...



Sélection d'attributs

Analyse du transcriptome

But identifier des familles et des réseaux fonctionnels de gènes mis en jeu sous l'effet du stimulus étudié

Sous-but identifier et quantifier la sur- ou sous-expression d'un ensemble de gènes dans une situation biologique donnée

Puces à ADN ou microarrays

Données :

- des milliers de **gènes** (*attributs*)
- des dizaines de **lames** (*exemples étiquetés*)

=> **Méthodes de sélection d'attributs**

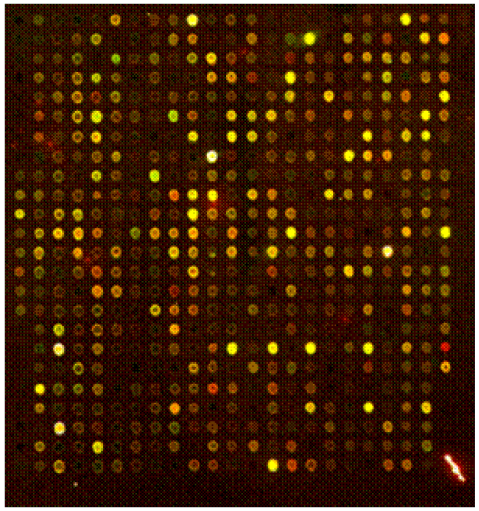
Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

Étude de l'effet des très faibles radiations sur le génome (2001-2004)

Expériences et données

- 6 cultures de *S. Cerevisiae* ($\approx$ 6400 gènes) irradiées pendant 20h à [15, 30] mGy/h
- 12 cultures non irradiées
- Mesure effectuées sur puce Corning où l'hybridation a été faite avec double marquage fluorescent (Cy3 pour les cADN contrôles et Cy5 pour les cADN étudiés).



Questions

1. **Combien** de gènes impliqués dans la réponse ?
2. Identification de **groupes de gènes** ?
3. **Prédiction** d'irradiation d'une culture ?
4. **Généralisable** à d'autres organismes ?

Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

Le **problème de la sélection d'attributs**

- Problème NP-difficile
- Mais *a priori* plus simple que la classification

Exemple : 3 attributs binaires et des fonctions booléennes

a_1	a_2	a_3	XOR
0	0	0	$-$
0	0	1	$+$
0	1	0	$+$
0	1	1	$-$
1	0	0	$-$
1	0	1	$+$
1	1	0	$+$
1	1	1	$-$

$2^{2^3} = 2^8 = 256$ fonctions possibles

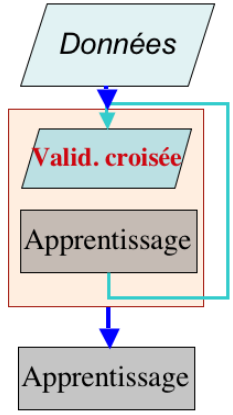
Mais seulement 10 tris possibles sur les 3 attributs (e.g. $a_1 \prec a_2 \prec a_3$)

Sélection d'attributs

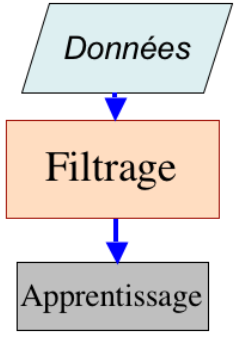
Analyse du transcriptome : **Étude de cas**

Les **grandes approches**

1. **Approche "embarquée"** (*embedded*)
   - Directement dans le système apprenant (e.g. *arbre de décision*)



2. **Approche "symbiose"** (*wrapper*)
   - Utilisent la performance en test en apprentissage pour sélectionner les attributs en entrée
   - *forward selection* : par ajouts successifs d'attributs
   - *backward selection* : par retraits successifs d'attributs



3. **Approche "filtre"** (*filter*)
   - Indépendante des traitements aval

Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

Les **méthodes de filtre**

Mesurer la **corrélation entre chaque attribut et la classe**

Hypothèse de linéarité

- Corrélation de Pearson : $\mathcal{R}_i = \frac{cov(X_i,Y)}{\sqrt{var(X_i)var(Y)}}$
- Information mutuelle : $\mathcal{I}_i = \int_{x_i}\int_y p(x_i,y)\log\frac{p(x_i,y)}{p(x_i)\cdot p(y)}dxdy$
- SAM : $\mathcal{R}_i = \frac{x_{\mathsf{I}}(i)-x_{\mathsf{NI}}(i)}{\sigma(i)+\sigma_0}$ (σ_0 : constante > 0)
- RELIEF :

$$\mathcal{R}_{\text{gène}} = \frac{1}{m}\sum_{L=1}^{m}\left\{\left[\text{expr}_{\text{gène}}(L) - \text{expr}_{\text{gène}}(M)\right] - \left[\text{expr}_{\text{gène}}(L) - \text{expr}_{\text{gène}}(H)\right]\right\}$$

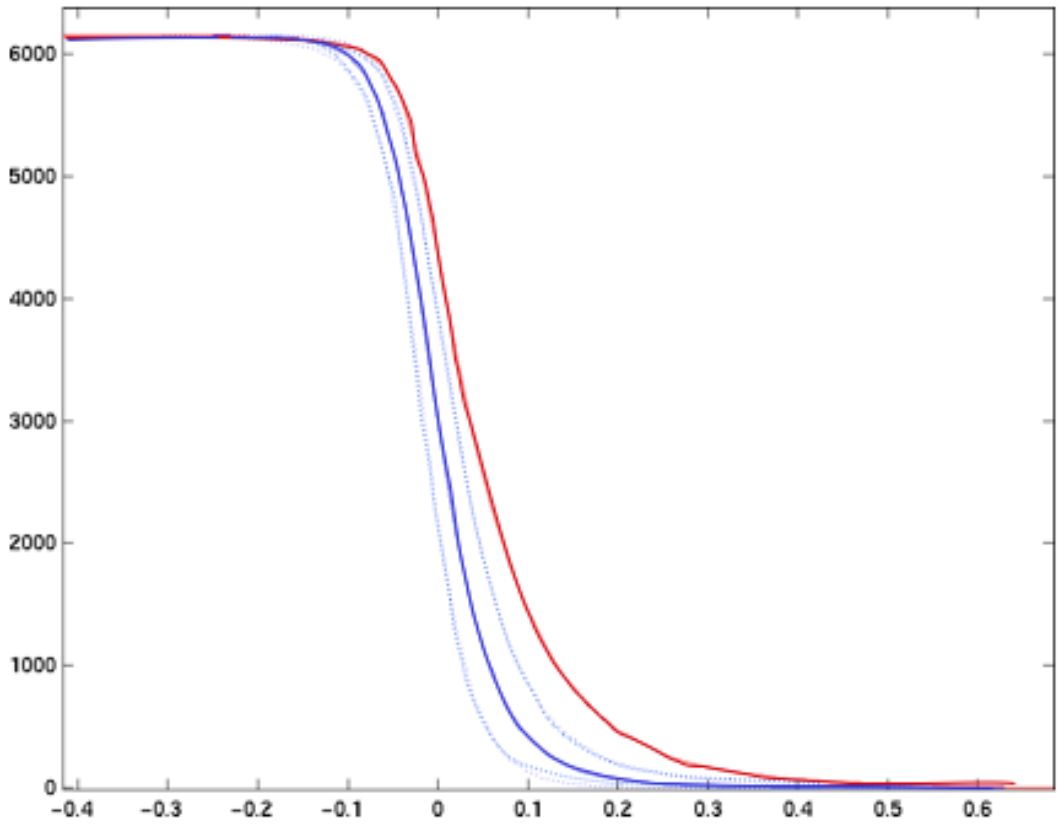
où $\text{expr}_{\text{gène}}(x)$: projection sur *gène* du point x, et m nb total de gènes.

- ...

Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

Test d'hypothèses multiples



Nombre de gènes dont le poids dépasse la valeur repérée en abscisse

rouge : Avec les classes réelles ;

bleu : Courbe moyenne obtenue avec des classes aléatoires

Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

Test d'hypothèses multiples : problème du seuil

Compromis **rappel** vs. **précision** (*faux positifs vs. faux négatifs*)

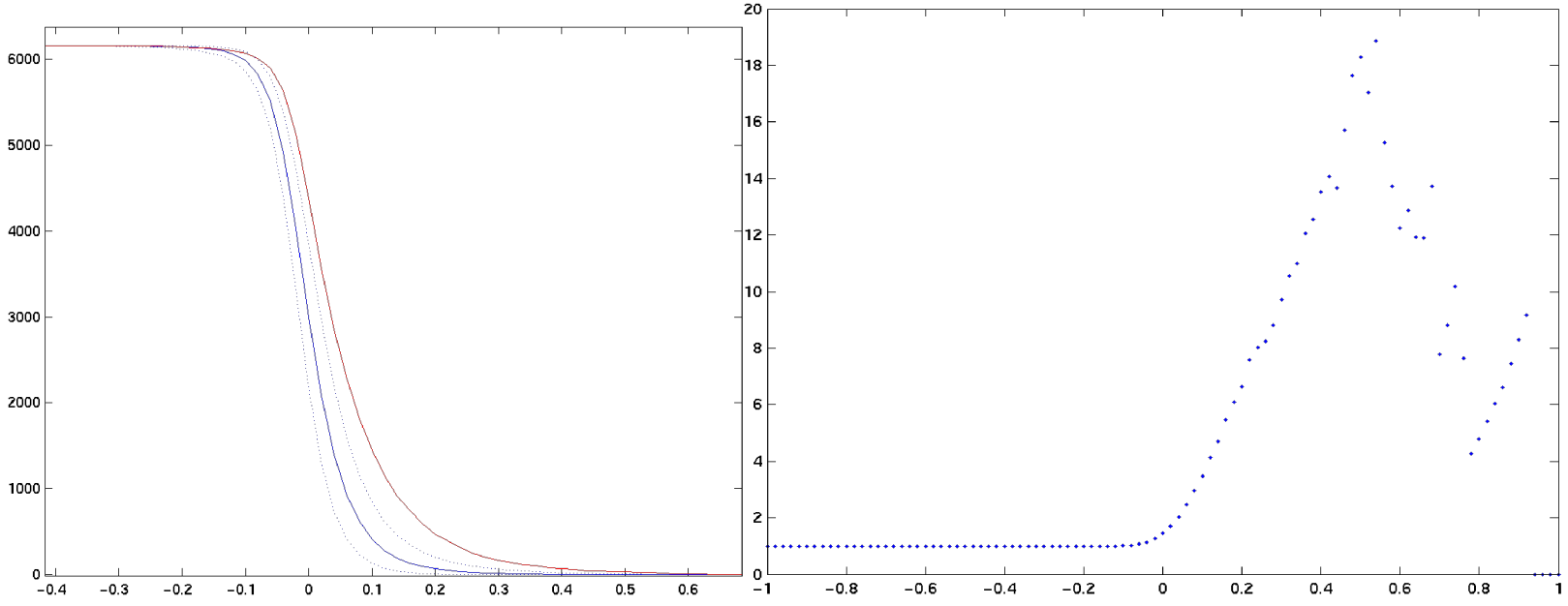
Classe estimée \ *Classe réelle*	+	−
+	**Vrais positifs**	Faux positifs
−	Faux négatifs	**Vrais négatifs**

- Le *rappel* : $\frac{VP}{VP+FN}$
- Le *précision* : $\frac{VP}{VP+FP}$

Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

Estimation d'un seuil de signification



=> 171 gènes

Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

Résultat

Cytochrome bc1	Cytochrome c oxidase	ATP synthase
Cyt1	COX5A	ATP3
QCR7	COX6	ATP5
QCR10	COX4	ATP16
	COX 13	ATP15
	COX12	ATP7
	COX7	ATP17
	COX8	ATP18
	COX20	ATP19
		ATP20
		TIM11

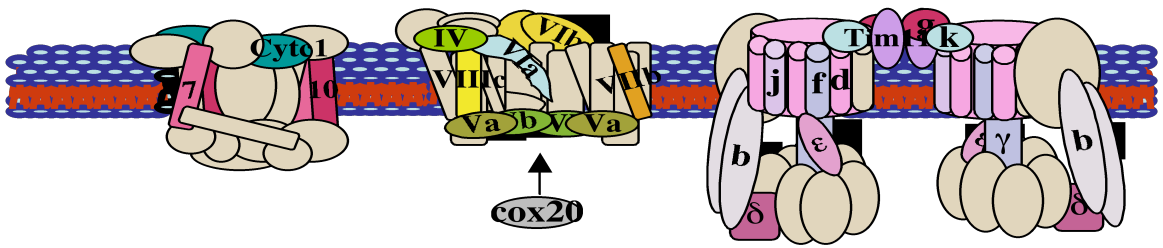


Figure: Fonctions de la membranes de la mitochondrie

Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

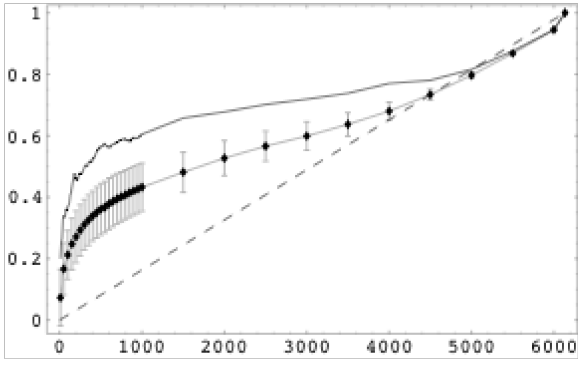
Autres méthodes de choix de seuil

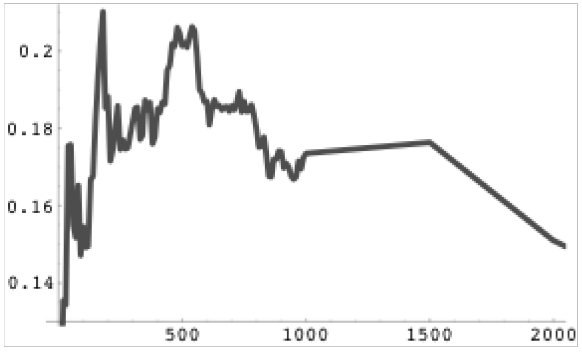
- Insérer des **gènes artificiels non corrélés** à la classe
- Examen de la **courbe des intersections des top–n** pour deux méthodes d'évaluation des gènes (e.g. RELIEF et ANOVA)
- ...

Sélection d'attributs

Analyse du transcriptome : **Étude de cas**

Méthode par combinaison de filtres





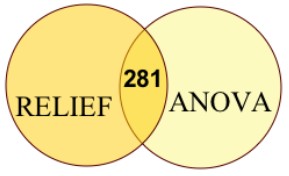


Figure: (*À gauche*) Valeur de n en abscisse et taille k de l'intersection en proportion de n en ordonnée (e.g. k = 0.6 n). Courbe du haut : k, du milieu : mH0 (corrélation a priori), du bas : taille de l'intersection due au hasard.
(*Au centre*) Différence relative entre k et mH0.

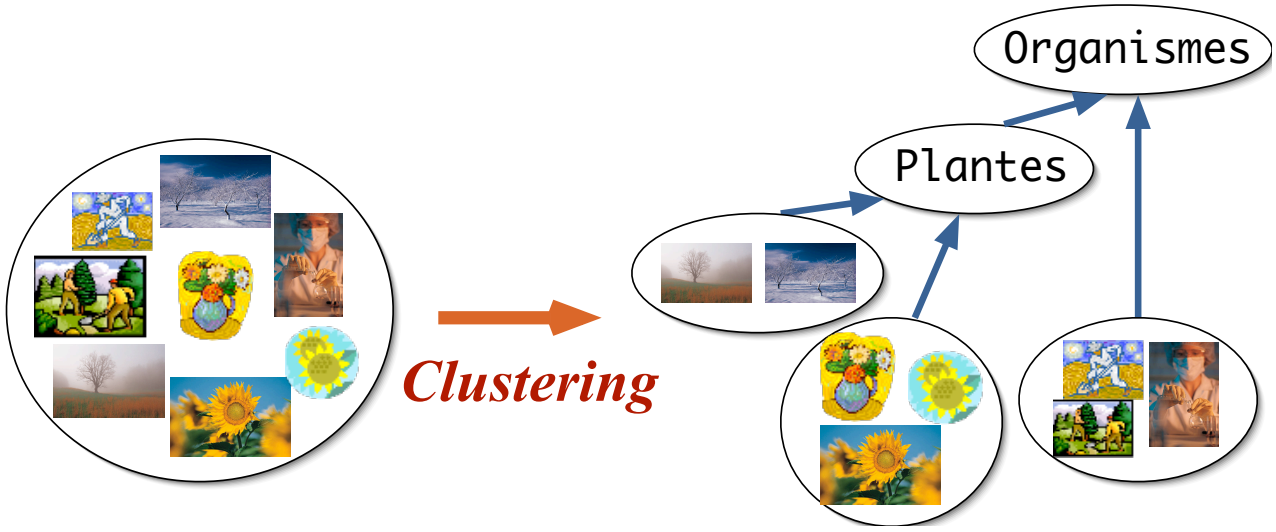
L'information apportée par les données est maximale pour $n \approx 180$ et $n \approx 540$.

Apprentissage non supervisé
Clustering

- Organiser un ensemble de *formes* en **groupes contrastés**

En vue de :

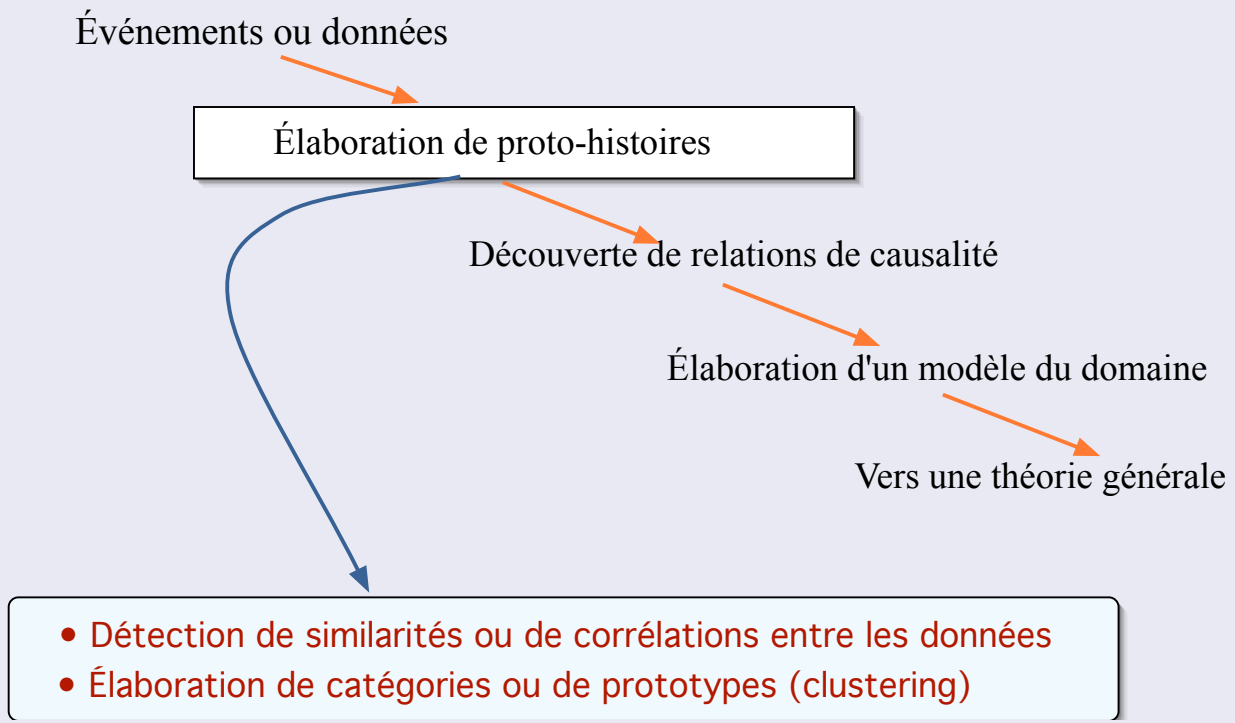
- **Comprimer** et **structurer** les données pour **permettre des prédictions**



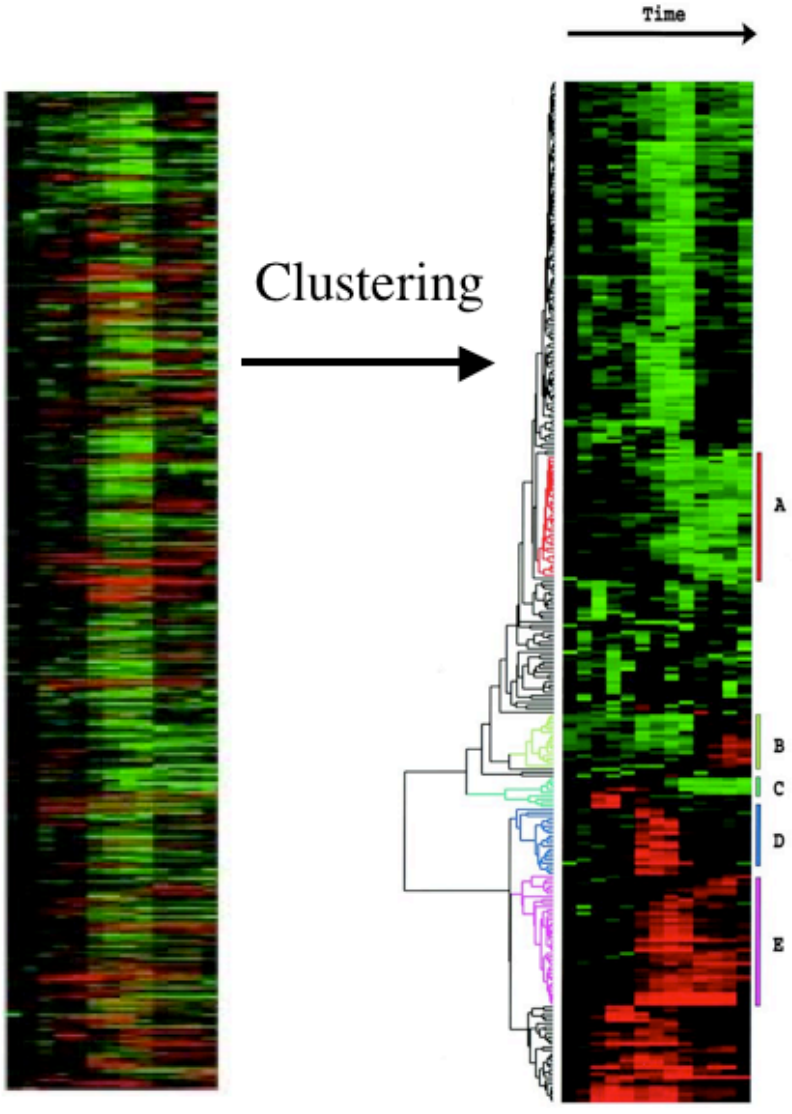
Apprentissage non supervisé
Clustering

Exemple. *Classification chinoise* [Borges, 1966] :
(1) Ce qui appartient à l'empereur ; (2) Embaumé ; (3) Fabuleux ;
(4) Inclue dans cette classification ; (5) Domestique
(6) Ce qui ressemble à un moustique vu de loin ; (7) Autre

Étape essentielle dans le processus de découverte scientifique [Forbus & Gentner, 1986]



Apprentissage non supervisé
Clustering



Much easier to look at large blocks of similarly expressed genes

Dendogram helps show how 'closely related' expression patterns are

A. Cholesterol syn.
B. Cell cycle
C. Immediate-early response
D. Signaling
E. Tissue remodeling

Clustering
Les approches

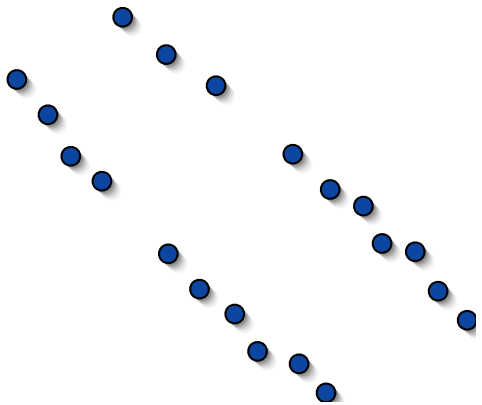


Figure: Données.

Clustering
Les approches

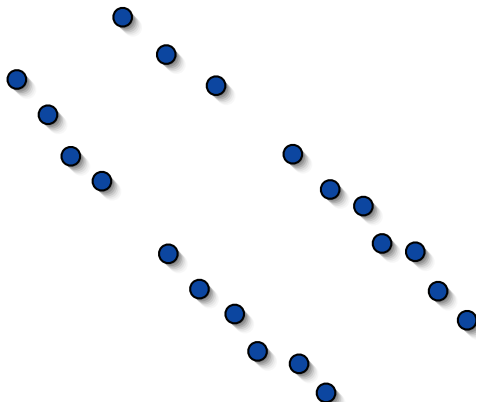


Figure: Données.

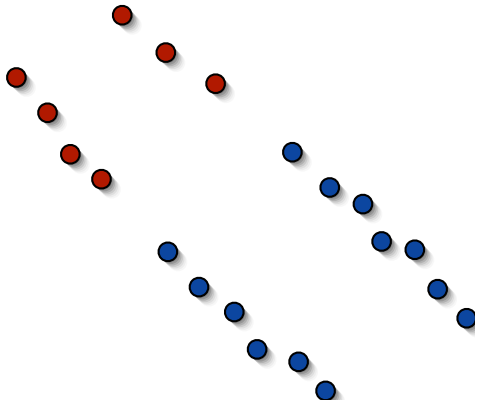


Figure: Groupes.

Clustering
Les approches

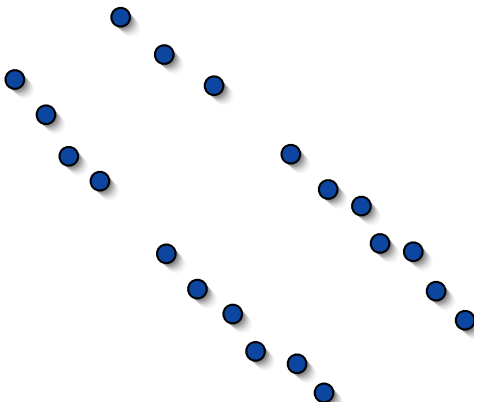


Figure: Données.

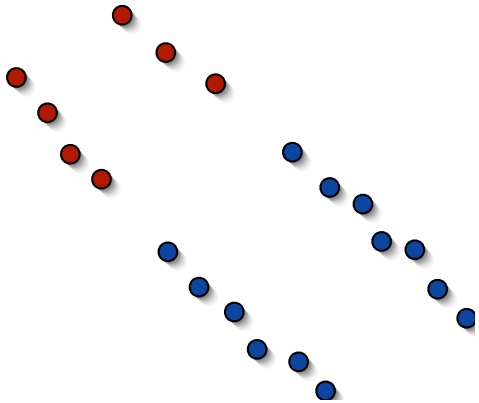


Figure: Groupes.

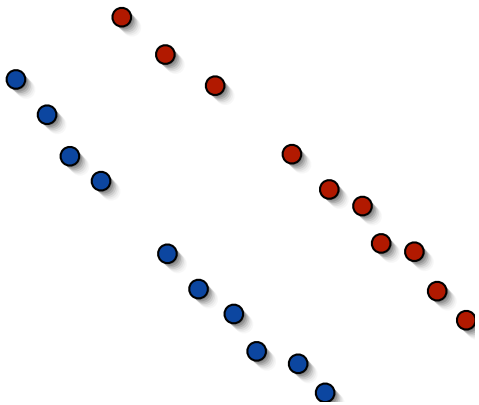


Figure: Linérarité.

Clustering
Les problèmes

1. **Sélection d'attributs**
2. Choix de la **mesure de proximité**
3. Choix du **critère de clustering**
4. Choix de l'**algorithme de clustering**
5. **Validation** des résultats
6. **Interprétation** des résultats

Clustering
Les approches

Méthodes fondées sur les distances

Méthodes fondées sur les modèles (probabilistes)

Clustering
Les approches

Méthodes fondées sur les distances

- Partition "plate" : k-moyenne, k-médoides, ...
- Catégorisation hiérarchique : ascendante ; descendante

Méthodes fondées sur les modèles (probabilistes)

Clustering
Les approches

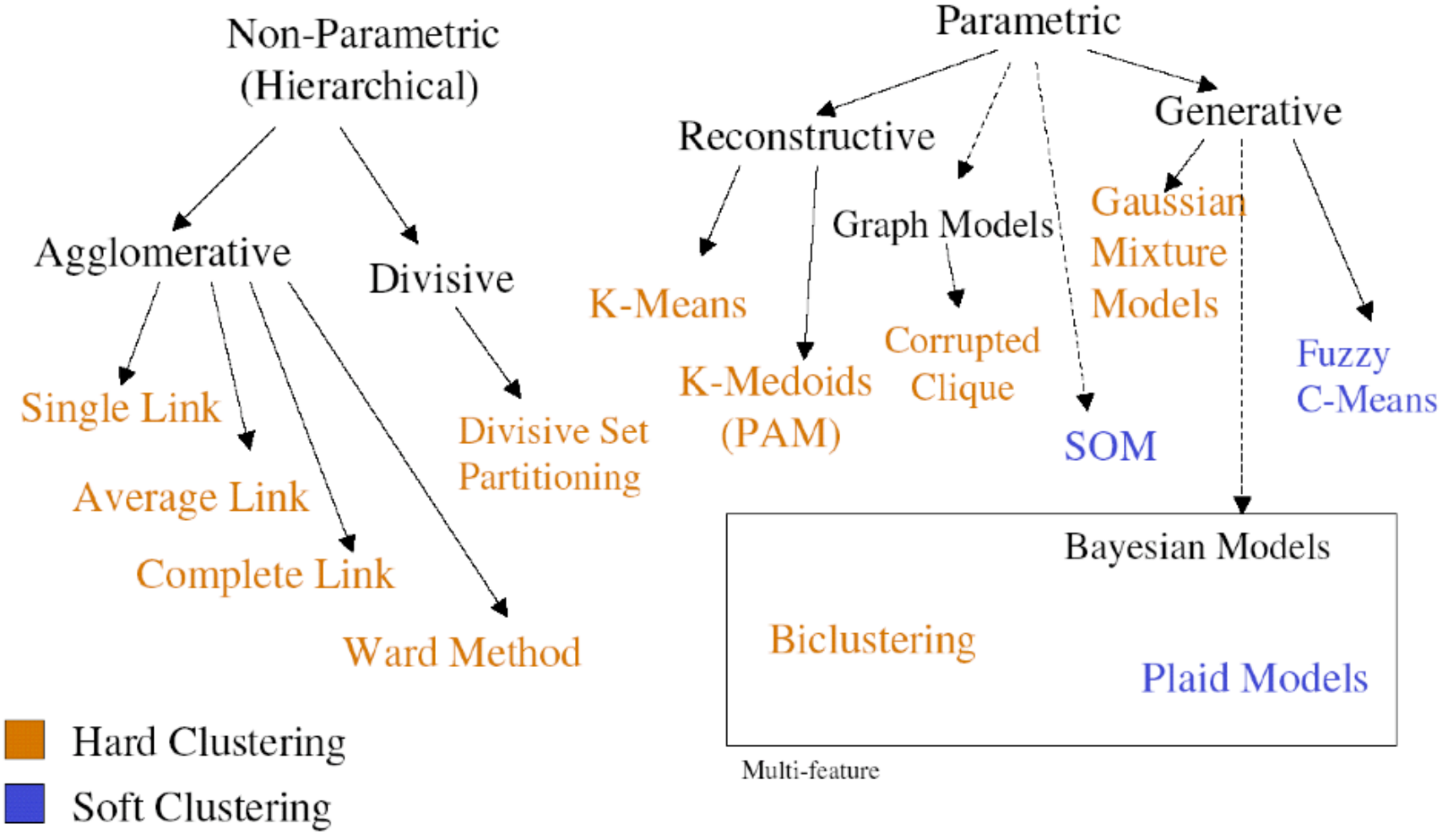
Méthodes fondées sur les distances

- Partition "plate" : k-moyenne, k-médoides, ...
- Catégorisation hiérarchique : ascendante ; descendante

Méthodes fondées sur les modèles (probabilistes)

- Mélanges de modèles
- Hiérarchie de modèles

Clustering
Taxonomie des méthodes



Clustering
La notion de similarité

Choisir de manière avisée
la mesure de similarité / dissimilarité appropriée

Clustering
Algorithme des k-moyennes

1. ***Démarrer en plaçant K centroïdes aléatoirement***

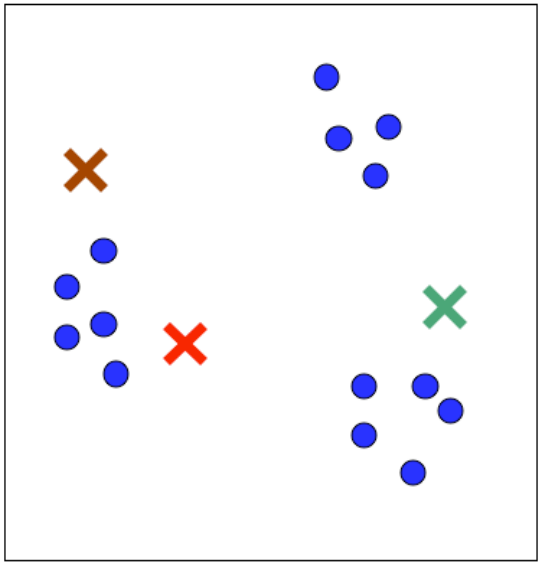


Figure: Itération = 1

Clustering
Algorithme des k-moyennes

1. ***Démarrer en plaçant K centroïdes aléatoirement***
2. ***Assigner les points aux centroïdes (en fonction de la distance)***

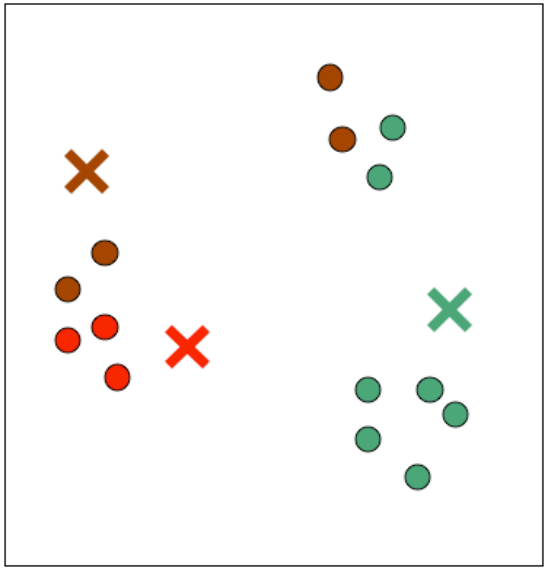


Figure: Itération = 1

Clustering
Algorithme des k-moyennes

1. ***Démarrer en plaçant K centroïdes aléatoirement***
2. ***Assigner les points aux centroïdes (en fonction de la distance)***
3. ***Déplacer les centroïdes au barycentre des points assignés***

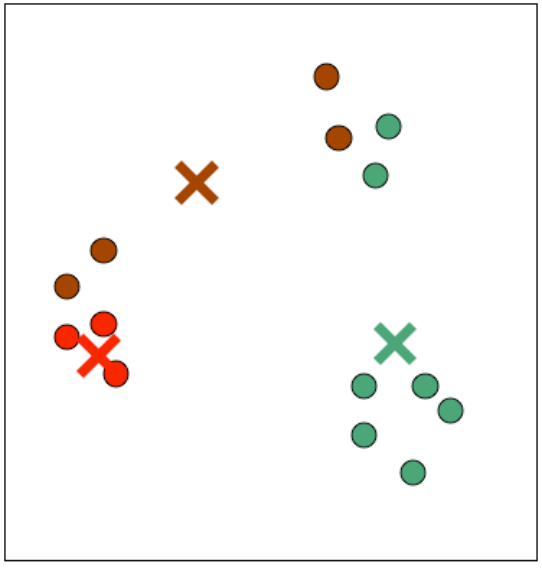


Figure: Itération = 1

Clustering
Algorithme des k-moyennes

1. ***Démarrer en plaçant K centroïdes aléatoirement***
2. ***Assigner les points aux centroïdes (en fonction de la distance)***
3. ***Déplacer les centroïdes au barycentre des points assignés***
4. ***Itérer jusqu'à déplacements sous un seuil donné***

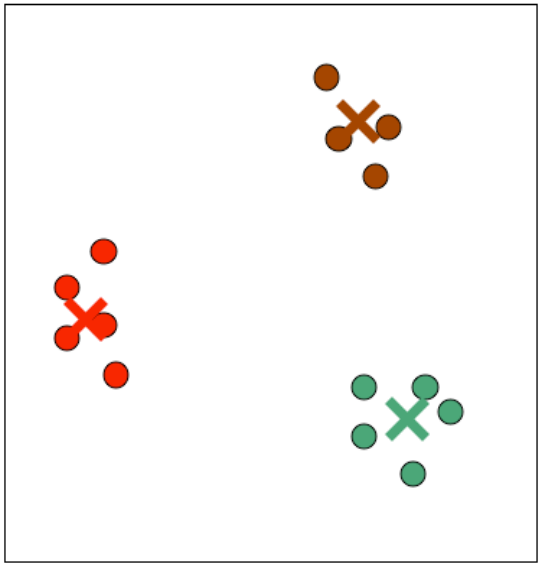


Figure: Itération = 3

Clustering
Algorithme des k-moyennes

Le nombre de groupes K est choisi *apriori*.

Initialisation

(En général) choix de K points tirés aléatoirement parmi $\mathbf{x}_1 \dots \mathbf{x}_G$. Ils deviennent les "moyennes" $\mathbf{m}_1^0 \dots \mathbf{m}_K^0$ des K groupes

Étape n :

$a)$ Affecter chaque élément $\mathbf{x}$ au groupe le plus proche $\mathcal{C}_k^n$:

$$d(\mathbf{x}_x, \mathbf{m}_k^n) = \min_{k'} d(\mathbf{x}_x, \mathbf{m}_{k'}^n)$$

$b)$ Mise à jour de la moyenne de chaque groupe :

$$\mathbf{m}_k^{n+1} = \frac{1}{|\mathcal{C}_k|} \sum_{g \in \mathcal{C}_k} \mathbf{x}_g$$

Clustering
Algorithme des k-moyennes

Propriété :

L'inertie intra-groupe décroît à chaque étape

Soit I_n cette inertie à l'étape n : $I_n = \sum_{k=1}^{K} \sum_{g \in \mathcal{C}_k^n} d\left(\mathbf{x}_g - \mathbf{m}_k^n\right)^2$

$a)$ Si l'élément $\mathbf{x}$ passe du groupe $\mathcal{C}_k$ au groupe $\mathcal{C}_{k'}$ alors

$$d\left(\mathbf{x}_g - \mathbf{m}_{k'}^n\right)^2 \leq d\left(\mathbf{x}_g - \mathbf{m}_k^n\right)^2$$

$b)$ Puisque $\mathbf{m}^{n+1}$ est la moyenne du groupe $\mathcal{C}_k^{n+1}$

$$\sum_{g \in \mathcal{C}_k^{n+1}} d\left(\mathbf{x}_g - \mathbf{m}_k^{n+1}\right)^2 \leq \sum_{g \in \mathcal{C}_k^{n+1}} d\left(\mathbf{x}_g - \mathbf{m}_k^n\right)^2$$

Clustering
Algorithme des k-moyennes

Propriété :

L'inertie intra-groupe décroît à chaque étape

Soit I_n cette inertie à l'étape n : $I_n = \sum_{k=1}^{K} \sum_{g \in \mathcal{C}_k^n} d\left(\mathbf{x}_g - \mathbf{m}_k^n\right)^2$

$a)$ Si l'élément $\mathbf{x}$ passe du groupe $\mathcal{C}_k$ au groupe $\mathcal{C}_{k'}$ alors

$$d\left(\mathbf{x}_g - \mathbf{m}_{k'}^n\right)^2 \leq d\left(\mathbf{x}_g - \mathbf{m}_k^n\right)^2$$

$b)$ Puisque $\mathbf{m}^{n+1}$ est la moyenne du groupe $\mathcal{C}_k^{n+1}$

$$\sum_{g \in \mathcal{C}_k^{n+1}} d\left(\mathbf{x}_g - \mathbf{m}_k^{n+1}\right)^2 \leq \sum_{g \in \mathcal{C}_k^{n+1}} d\left(\mathbf{x}_g - \mathbf{m}_k^n\right)^2$$

L'algorithme des K-moyennes converge mais des groupes peuvent finir vides.

Clustering
Algorithme des k-moyennes

Avantages :

- Algorithme simple et efficace en temps et en mémoire
- Utilisable avec de grandes bases de données (e.g. milliers d'objets)

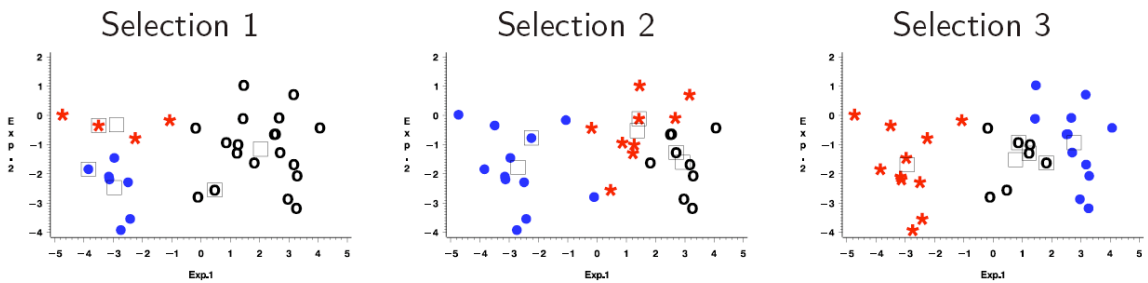
Clustering
Algorithme des k-moyennes

Avantages :

- Algorithme simple et efficace en temps et en mémoire
- Utilisable avec de grandes bases de données (e.g. milliers d'objets)

Limites :

- Sensible au choix des moyennes initiales



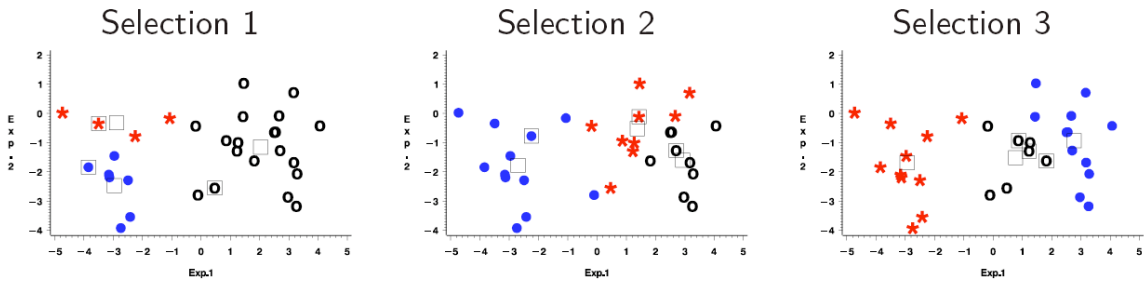
Clustering
Algorithme des k-moyennes

Avantages :

- Algorithme simple et efficace en temps et en mémoire
- Utilisable avec de grandes bases de données (e.g. milliers d'objets)

Limites :

- Sensible au choix des moyennes initiales



A utiliser quand :

- Il existe des informations *a priori* sur le choix de $\mathbf{m}_1^0 \ldots \mathbf{m}_K^0$
  - Essayer plusieurs initialisations différentes
  - En post-traitement d'un clustering hiérarchique

Clustering
Clustering hiérarchique

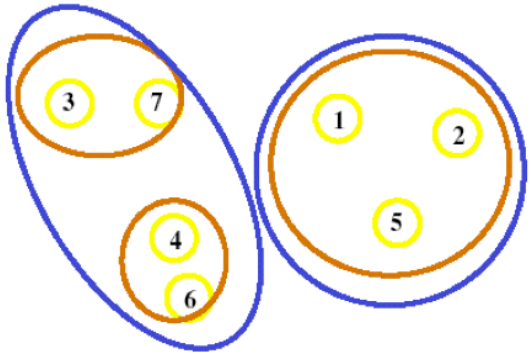
Entrées :

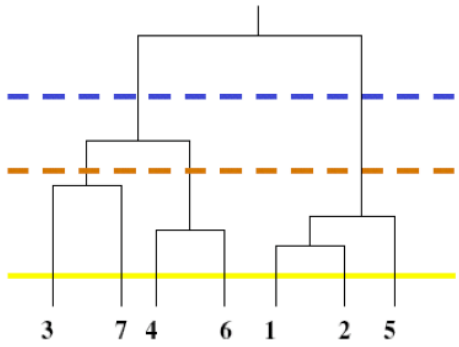
Données $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$

Sortie :

Un **arbre** :

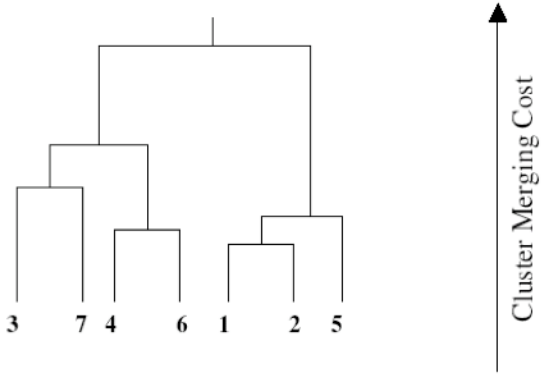
- dont les feuilles sont les données
- les branchements indiquent une similarité entre sous-groupes
- des coupures horizontales dans l'arbre induisent des groupements





Clustering
Clustering hiérarchique

1. Construire un cluster pour chaque donnée $\mathbf{x}_i$
2. Calcul du **coût de fusion** pour chaque paire de clusters
3. Identifier les deux clusters $\mathcal{C}_i$ et $\mathcal{C}_j$ de coût de fusion minimal
4. **Fusionner** $\mathcal{C}_i$ et $\mathcal{C}_j$ en un nouveau cluster $\mathcal{C}_{ij}$ (qui sera le parent de $\mathcal{C}_i$ et $\mathcal{C}_j$ dans l'arbre résultant)
5. Ré-itérer en (2) jusqu'à ce qu'il n'y ait pus qu'un seul cluster



Nombre maximal d'itérations = $N - 1$

Clustering

Clustering hiérarchique : **distances**

1. Une **distance entre formes** : $d(\mathbf{x}, \mathbf{x}')$
2. Une **distance entre groupes** : $d_{\mathcal{C}}(\mathcal{C}, \mathcal{C}')$

Clustering

Clustering hiérarchique : **distances**

Le ***coût d'agglomération*** est généralement = à la distance entre groupes $d_{\mathcal{C}}(\mathcal{C}, \mathcal{C}')$
La distance utilisée donne le nom de l'algorithme résultant :

Single linkage : $d_{\mathcal{C}}(\mathcal{C}, \mathcal{C}') = \min_{\mathbf{x} \in \mathcal{C}, \mathbf{x}' \in \mathcal{C}'} d(\mathbf{x}, \mathbf{x}')$

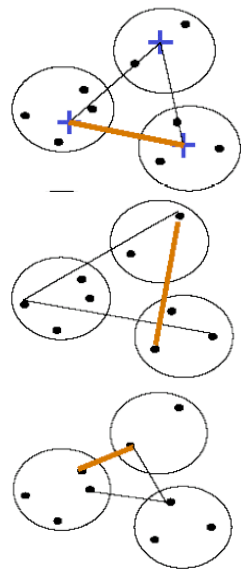
Average linkage : $d_{\mathcal{C}}(\mathcal{C}, \mathcal{C}') = \frac{1}{|\mathcal{C}||\mathcal{C}'|} \sum_{\mathbf{x} \in \mathcal{C}} \sum_{\mathbf{x}' \in \mathcal{C}'} d(\mathbf{x}, \mathbf{x}')$

Complete linkage : $d_{\mathcal{C}}(\mathcal{C}, \mathcal{C}') = \max_{x \in \mathcal{C}, \mathbf{x}' \in \mathcal{C}'} d(\mathbf{x}, \mathbf{x}')$

Centroid : $d_{\mathcal{C}}(\mathcal{C}, \mathcal{C}') = d(\overline{\mathbf{x}}, \overline{\mathbf{x}}')$

Ward : $d^2_{\mathcal{C}}(\mathcal{C}, \mathcal{C}') = \frac{|\mathcal{C}||\mathcal{C}'|}{|\mathcal{C}| + |\mathcal{C}'|} d^2(\overline{\mathbf{x}}, \overline{\mathbf{x}}')$

$= $ intra-variance dans les groupes



Clustering

Clustering hiérarchique : **distances**

UPGMA : Average linkage

UPGMC : $$d(\mathcal{C}, \mathcal{C}'\mathcal{C}'') = \frac{|\mathcal{C}'|d(\mathcal{C}, \mathcal{C}') + |\mathcal{C}''|d(\mathcal{C}, \mathcal{C}'')}{|\mathcal{C}'| + |\mathcal{C}''|} - \frac{|\mathcal{C}'||\mathcal{C}''|d(\mathcal{C}', \mathcal{C}'')}{(|\mathcal{C}'| + |\mathcal{C}''|)^2}$$

WPGMA : $$d(\mathcal{C}, \mathcal{C}'\mathcal{C}'') = \frac{d(\mathcal{C}, \mathcal{C}') + d(\mathcal{C}, \mathcal{C}'')}{2}$$

WPGMC : $$d(\mathcal{C}, \mathcal{C}'\mathcal{C}'') = \frac{d(\mathcal{C}, \mathcal{C}') + d(\mathcal{C}, \mathcal{C}'')}{2} - \frac{d(\mathcal{C}', \mathcal{C}'')}{4}$$

$$\left.\begin{array}{r}\text{Unweighted}\\ \text{Weighted}\end{array}\right\} \text{Pairwise Group Method} \left\{\begin{array}{l}\text{Average}\\ \text{Centroid}\end{array}\right.$$

Sokal & Sneath (63)

Clustering
Clustering hiérarchique

$d(g,g')$ $p(g)$	a (1)	b (1)	c (1)	d (1)	e (1)	f (?)
a		3	7	3	4	?
b	3		4	4	1	?
c	7	4		2	6	?
d	3	4	2		.5	$*$
e	4	1	6	.5		$*$
f	?	?	?	$*$	$*$	

À la première étape de l'algorithme, les "groupes" d et e sont rassemblés en un nouveau groupe f.

- Quelles sont les distances $d(\cdot, f)$?
- Quelles sont les poids f ?

Bouroche & Saporta (98)

Clustering
Clustering hiérarchique

$d(g,g')$ / $p(g)$	a (1)	b (1)	c (1)	d (1)	e (1)	$f(single)$ (2)	$f(aver.)$ (2)	$f(compl.)$ (2)
a		3	7	3	4	3	7/2	4
b	3		4	4	1	1	5/2	4
c	7	4		2	6	2	4	6
d	3	4	2		.5	$*$	$*$	$*$
e	4	1	6	.5		$*$	$*$	$*$
f	?	?	?	$*$	$*$			

À la première étape de l'algorithme, les "groupes" d et e sont rassemblés en un nouveau groupe f.

- Quelles sont les distances $d(\cdot, f)$?
- Quelles sont les poids f ?

Bouroche & Saporta (98)

Clustering
Clustering hiérarchique

Attention : Différentes méthodes produisent des arbres différents.

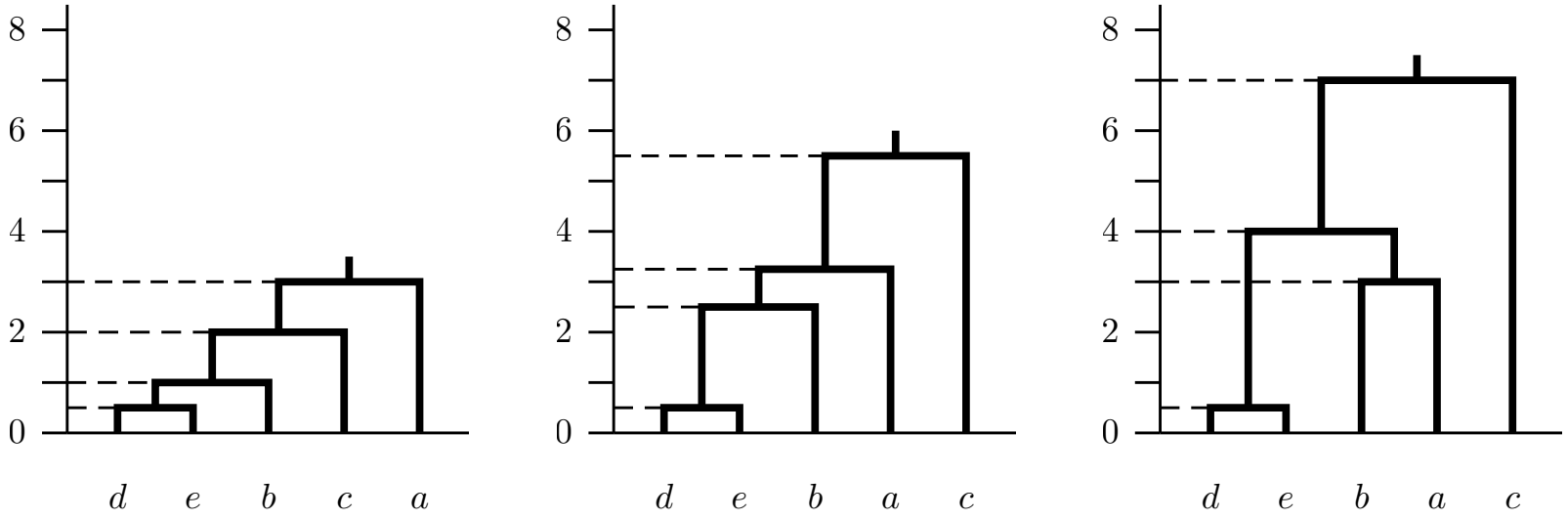


Figure: **Single linkage**, **Average linkage**, **Complete linkage**

Apprentissage non supervisé
Clustering

Algorithme de **clustering hiérarchique**

1. **Entrée** : n gènes
2. Calculer la matrice diagonale supérieure des similarités entre gènes
3. Identifier la paire de gènes la plus corrélée
4. La remplacer par un gène “moyen”
5. Recommencer (2) jusqu’à ce qu’il n’y ait plus qu’un “gène”
6. **Sortie** : dendrogramme

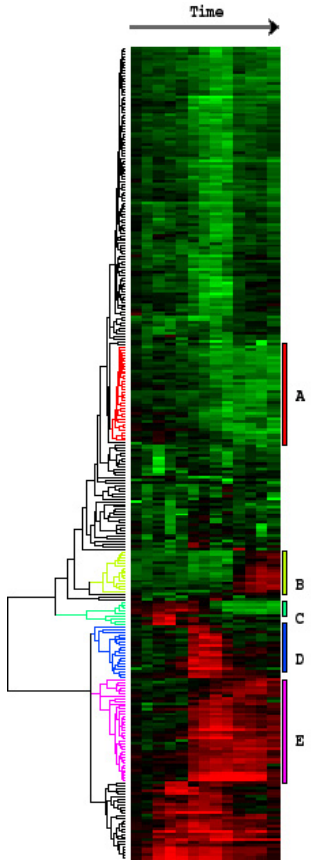
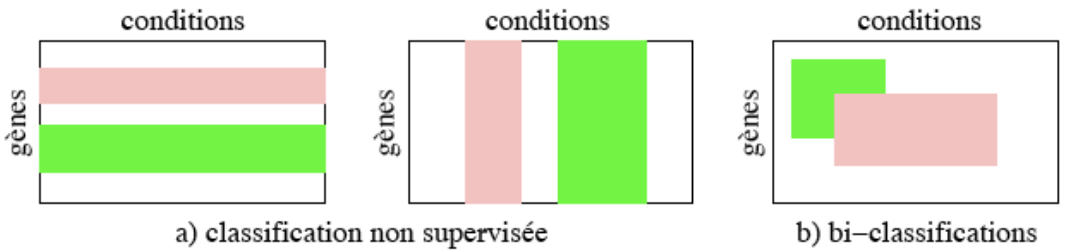


Figure: From [ESBB98]

Apprentissage non supervisé
Bi-Clustering

But : identifier des ensembles de gènes ayant le même motif d'expression (lignes) dans un sous-ensemble de conditions biologiques (colonnes).

⇒ Gènes co-exprimés dans un même sous-ensemble de conditions.

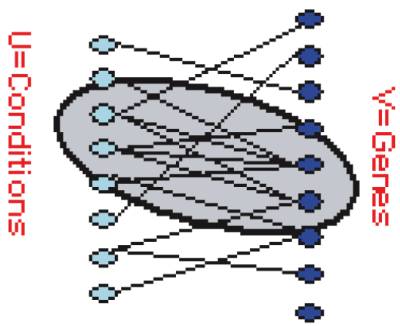


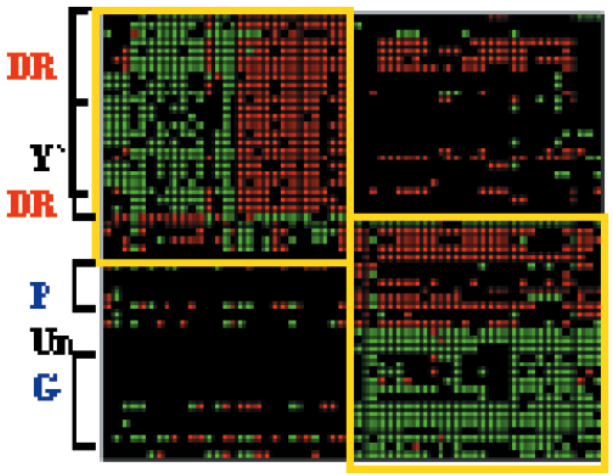
Problème NP-complet ⇒ recherche heuristique
Classiquement : on recherche des « blocs » en permutant lignes et colonnes

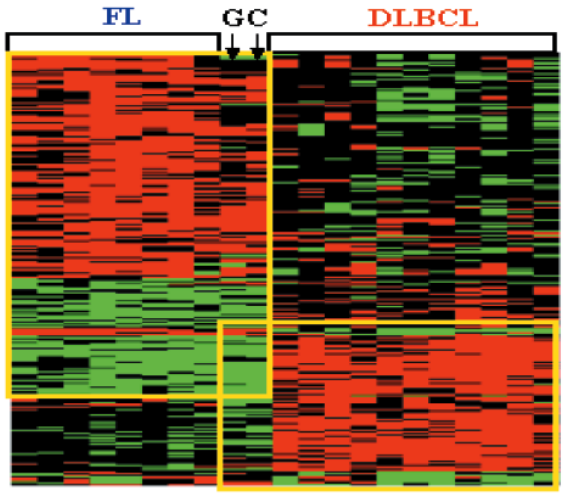
1. Méthodes statistiques
2. Méthodes symboliques

Apprentissage non supervisé
Bi-Clustering

Algorithme de **biclustering**







Graph bipartie des *conditions* et des *gènes*. Un arc (u, v) indique la réponse du gène v dans la condition u. L'ellipse indique un bi-cluster significatif.

Apprentissage non supervisé

Bi-Clustering hiérarchique

Clustering genes /

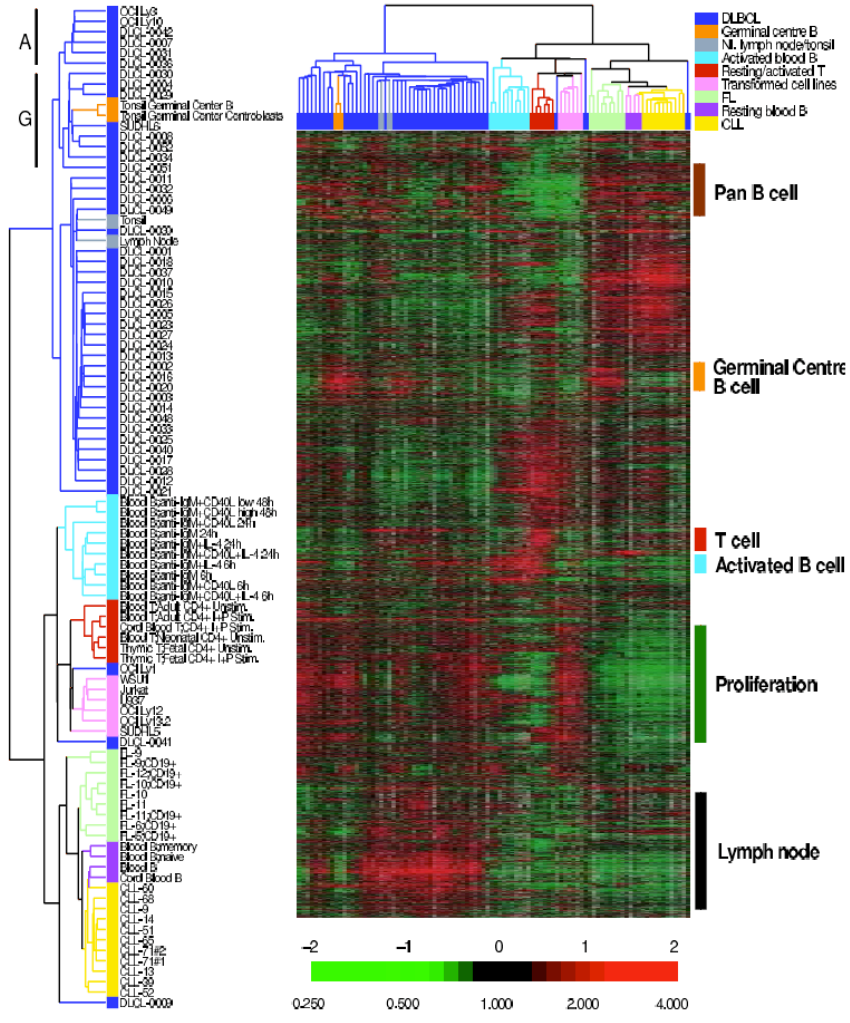
clustering patients:

lignes = ***gènes***

colonnes = ***patients***

(blue = BLBCL)

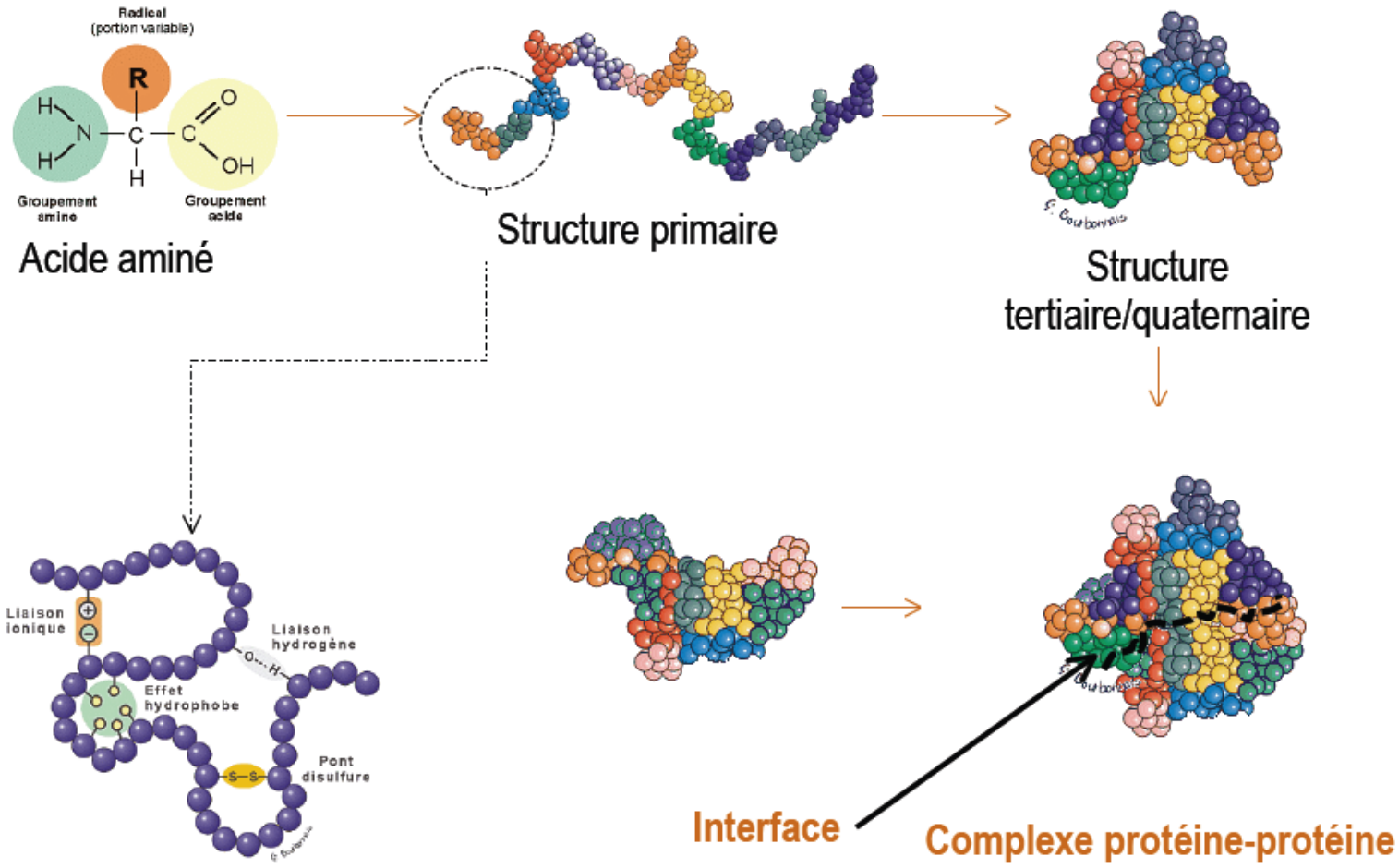
Alizadeh & al. (00)



Amarrage de protéines

Les protéines

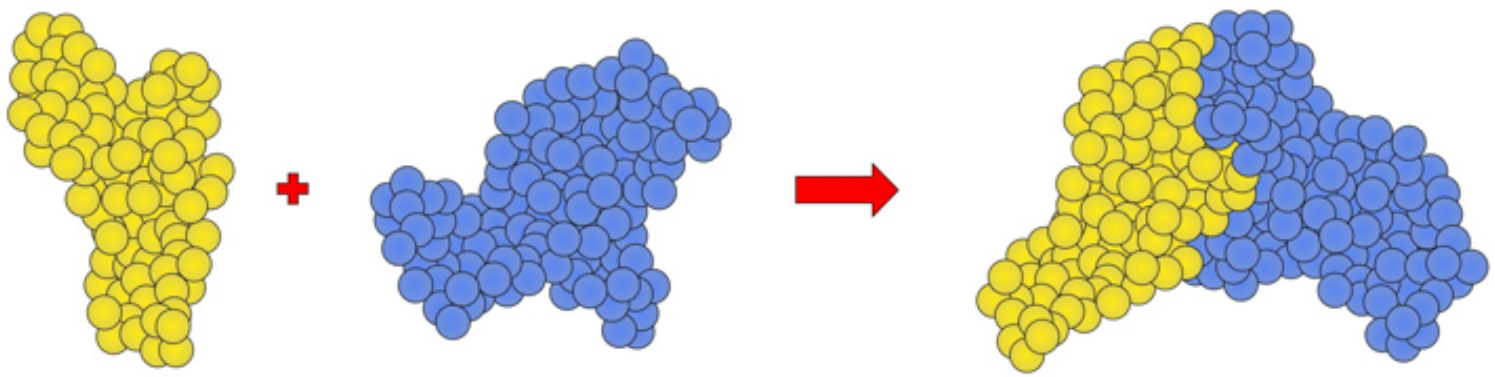
Prédiction de l'amarrage de protèines



Amarrage de protéines

L'amarrage

Prédiction de l'amarrage de protèines



Molécules de départ

Complexe moléculaire

Amarrage de protéines

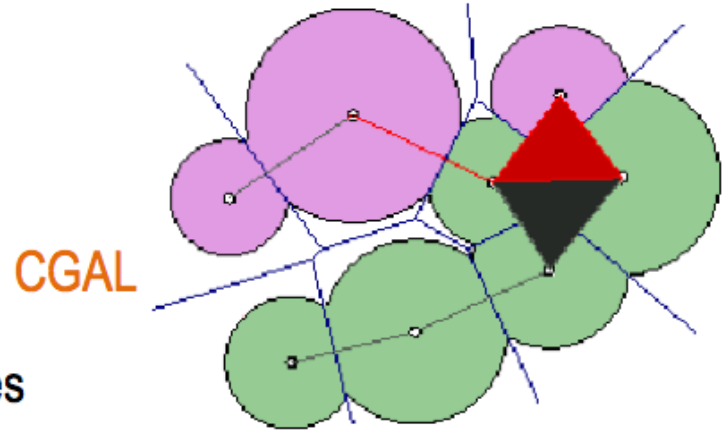
Les problèmes

- Description à construire
- Exemples positifs seuls

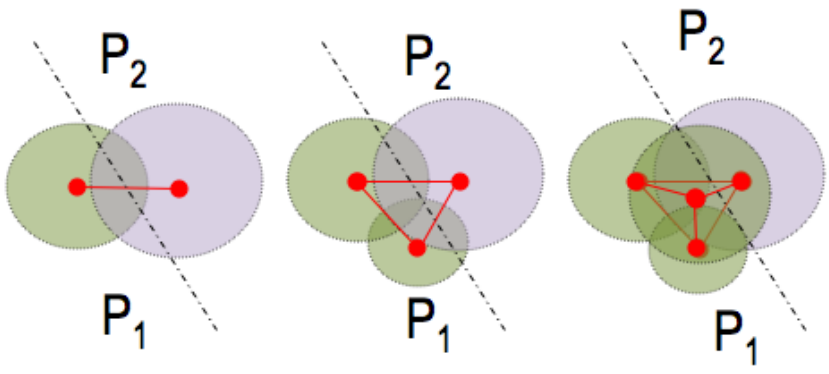
- Techniques de fouille de données
- Rôle de $\mathcal{H}_0$
- Construction de descripteurs

Amarrage de protéines
Description des exemples

- 1 sphère par acide aminé
- Représentation par l'α-complexe d'une interface
  - Inclus dans la triangulation régulière



  - Un ensemble d'éléments géométriques
    - arêtes,
    - triangles,
    - tétraèdres



Amarrage de protéines
Description des exemples

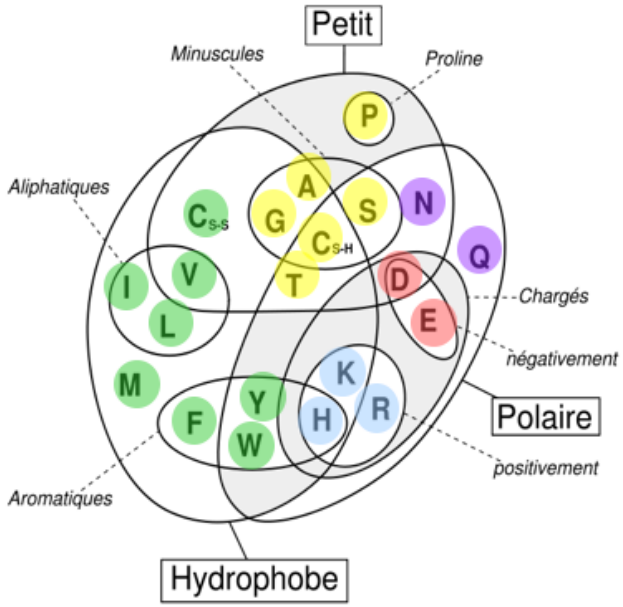
- 20 acides aminés distincts → Combinaisons avec remises
  - 210 arêtes
  - 1540 triangles
  - 8855 tétraèdres

  10 605 Attributs !!! → trop pour le nombre d'exemples (459)

- Groupement en 5 classes
  - petits (S), hydrophobes (H), polaires (P), positifs (+) and négatifs (-)
    - 15 arêtes
    - 35 triangles
    - 70 tétraèdres

  120 Attributs

Attributs : {SS, SH, ..., SHP,..., HH+-, ...}
Exemple de motif : {P+,SSP}



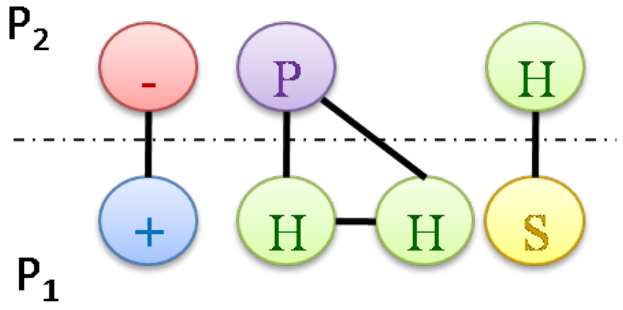
Amarrage de protéines
Fouille de données

Prédiction de l'amarrage de protèines

Problème

- Étant donnée la composition de surface de deux protéines
  - Prédire si les protéines peuvent s'amarrer

Données : Composition de surface de 459 complexes de protéines (Base PDB : Dockground) (120 attributs entiers)



Amarrage de protéines
Fouille de données

La méthode aCID

Propriétés des **motifs caractéristiques** :

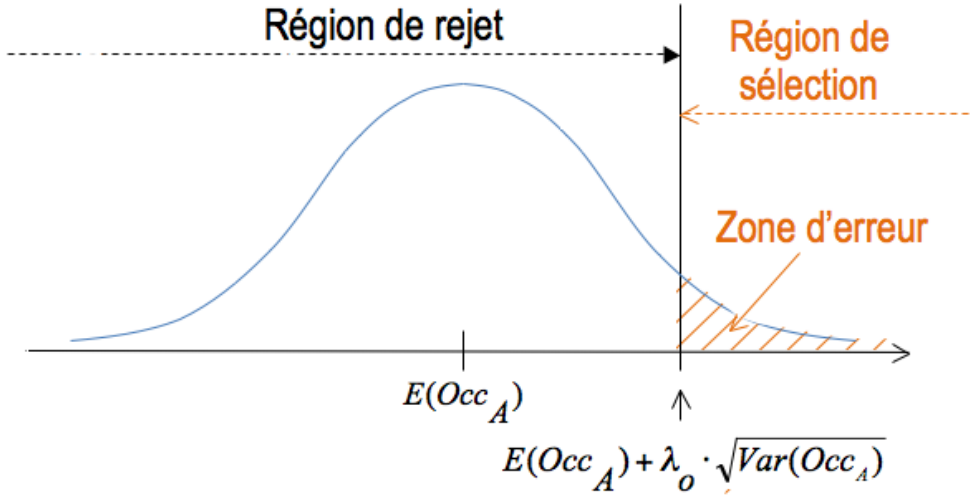
- apparaissant **suffisamment souvent** dans les exemples positifs (*taux de couverture*)
- apparaissant dans les exemples positifs **plus que ce que l'on attendrait** pour l'ensemble de la population

Combine :

- Méthode d'extraction de motifs fréquents
- Tests statistiques (Hypothèses nulles, p-value)

Amarrage de protéines

Description des exemples



- Critères de sélection:
  - $Couv(A) \geq E(Couv_A) + \lambda_c \sqrt{Var(Couv_A)}$
  - $Occ(A) \geq E(Occ_A) + \lambda_o \sqrt{Var(Occ_A)}$
  - $Couv(A) \geq c_{\min}$ → Motifs fréquents

Motifs caractéristiques

Amarrage de protéines
Fouille de données

Prédiction de l'amarrage de protèines

Méthode : ACID

- Chercher des **itemset fréquents** (combinaison d'acides aminés sur un même spot)
- **caractéristiques** (différence avec $\mathcal{H}_0$)

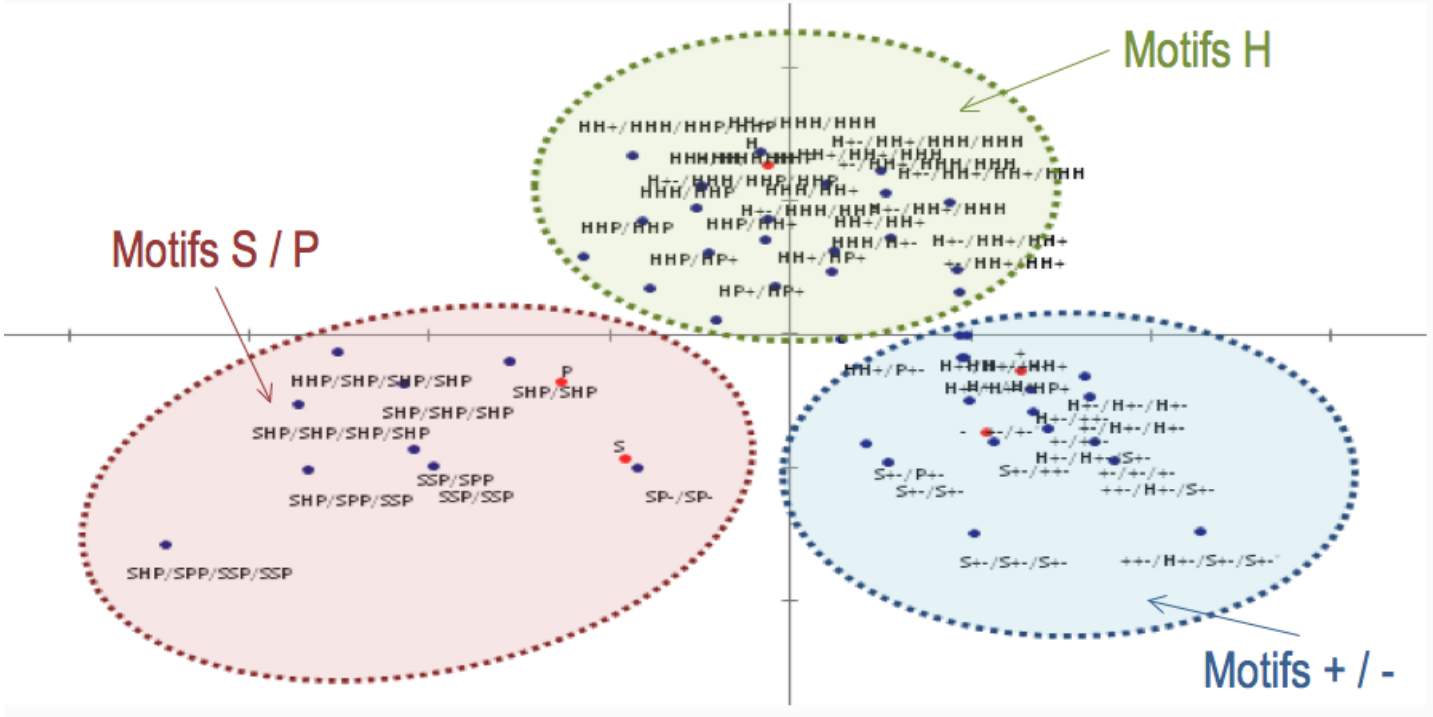
Doublets	***SSP/SSP***, ***SSP/SPP***, ***SP−/SP−***, ***S+−/S+−***, ***S+−/++−***, ***HHH/HHH***, ***HHH/HHP***, ***HHH/HH+***, ***HHH/H+−***, ***HHP/HHP***, ***HHP/HP+***, ***HH+/HH+***, ***HP+/HP+***, ***H+−/H+−***, ***H+−/++−***, *+−/+−*, *+−/++−*, *SHP/SHP*, *S+−/P+−*, *HHP/HH+*, *HH+/HP+*, *HH+/P+−*, *HH+/++−*
Triplets	{***S+−/S+−/S+−***, ***++−/H+−/S+−***, ***HH+/HHH/HHH***, ***HH+/HH+/HHH***, ***H+−/H+−/H+−***, {*+−/+−/+−*, *+−/HH+/HH+*, *+−/H+−/H+−*, *SHP/SPP/SSP*, *SHP/SHP/SHP*, *H+−/H+−/S+−*, *HH+/HHH/HHP*, *H+−/HHH/HHP*, *H+−/HH+/HHH*, *H+−/H+−/HH+*, *H+−/H+−/HP+*, *H+−/HH+/HH+*}
Quadruplets	{***H+−/HH+/HHH/HHH***, *+−/HH+/HHH/HHH*, *SHP/SPP/SSP/SSP*, *SHP/SHP/SHP/SHP*, *++−/H+−/S+−/S+−*, *H+−/HH+/HH+/HHH*, *HHP/SHP/SHP/SHP*, *HH+/HHH/HHP/HHP*, *H+−/HHH/HHP/HHP*}

Items that are over-represented (C=2 (bold), and C=1) and cover at least 5% of the interfaces. Results for C=1 are a superset of the results for C=2.

Amarrage de protéines

Description des exemples

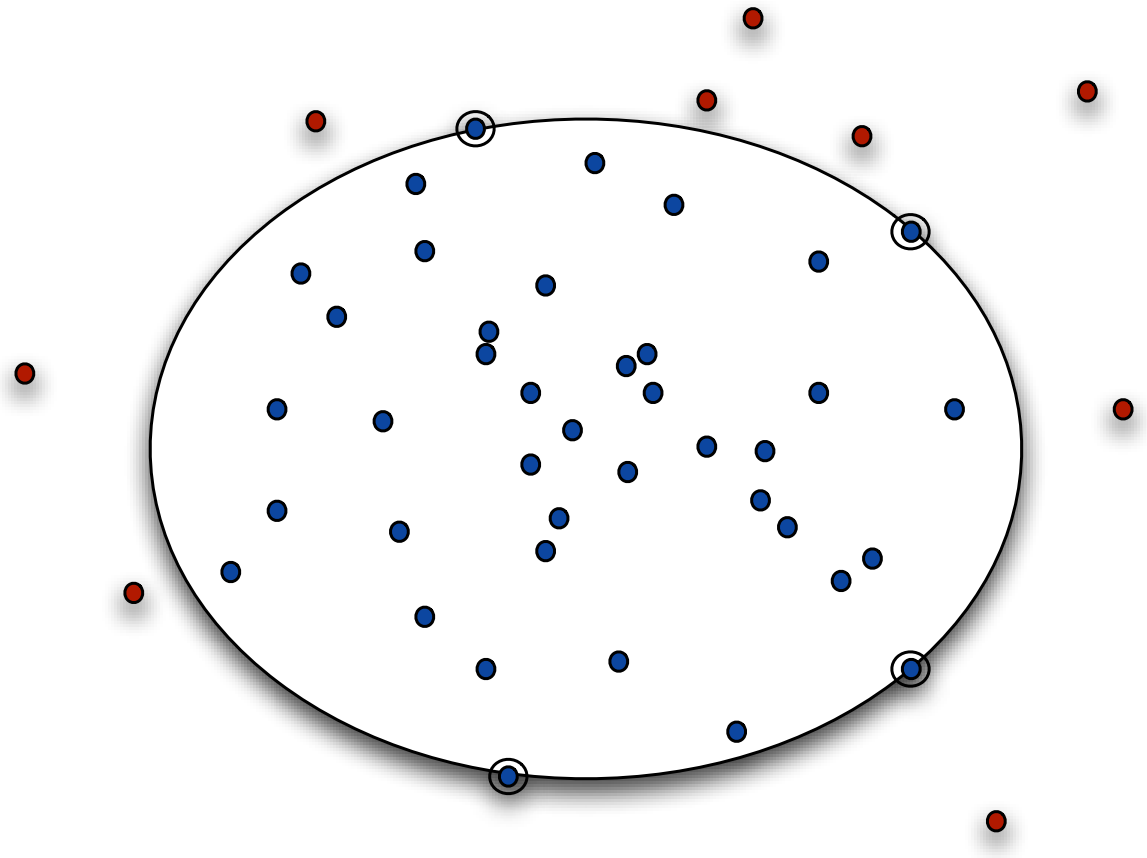
- Une tendance stable (86% sur 2 axes)
- Résultats connus sur la composition des interfaces protéine-protéine



- Résultats cohérents avec ce qui est connu
- Robustes
  - (aux choix de valeurs de paramètres et aux affectations des acides aminés dans les groupes)

Apprentissage à partir d'exemples positifs seuls

Recherche d'éléments surprenants : « One-class SVM »



Prédiction du risque cardio-vasculaire

Les problèmes

- Données hétérogènes
- Suffisamment informatives ?

- Apprentissage non supervisé
- Mélange de gaussiennes et algorithme EM

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

Étude du risque cardiovasculaire

- 26000 sujets ; 10 pays
- 2230 sujets retenus (suivis sur 6 ans)
- 21 descripteurs
  (10 discrets ; 11 numériques)
- Étiquette “semi-sûre” ($+$ ok ; $-$ non)
- 4.8% de ‘$+$’ ; 95.2% de ‘$-$’

Caractéristiques des données

Sexe	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT Female: 57.6 % / 55.1 % / 57.7 % / 4.6 % / 95.4 % Male: 42.4 % / 44.9 % / 42.3 % / 5.1 % / 94.9 %
Age	continu	parmi population / parmi EVENT / parmi NO EVENT 60.00 .. 71.47 .. 96.00 / 60.00 .. 74.69 .. 96.00 / 60.00 .. 71.31 .. 92.00
Taille	continu	parmi population / parmi EVENT / parmi NO EVENT 121.92 .. 163.15 .. 200.66 / 137.16 .. 162.82 .. 187.96 / 121.92 .. 163.17 .. 200.66
Poids	continu	parmi population / parmi EVENT / parmi NO EVENT 36.74 .. 73.46 .. 136.08 / 43.09 .. 71.13 .. 128.37 / 36.74 .. 73.58 .. 136.08
Diabète	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT YES: 10.4 % / 22.4 % / 9.8 % / 10.3 % / 89.7 % NO: 89.6 % / 77.6 % / 90.2 % / 4.2 % / 95.8 %
Fréquence cardiaque	continu	parmi population / parmi EVENT / parmi NO EVENT 38.00 .. 71.33 .. 124.00 / 48.00 .. 71.91 .. 106.00 / 38.00 .. 71.30 .. 124.00
Pression artérielle systolique	continu	parmi population / parmi EVENT / parmi NO EVENT 158.00 .. 170.09 .. 213.00 / 160.00 .. 172.80 .. 200.00 / 158.00 .. 169.95 .. 213.00
Pression artérielle diastolique	continu	parmi population / parmi EVENT / parmi NO EVENT 20.00 .. 76.70 .. 90.00 / 44.00 .. 75.56 .. 88.00 / 20.00 .. 76.76 .. 90.00
Hypertrophie ventriculaire (ECG)	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT c0: 86.1 % / 77.6 % / 86.5 % / 4.3 % / 95.7 % c1: 4.7 % / 12.1 % / 4.3 % / 12.5 % / 87.5 % c2: 9.2 % / 10.3 % / 9.2 % / 5.3 % / 94.7 %
ST depression (ECG)	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT c0: 90.5 % / 76.6 % / 91.2 % / 4.1 % / 95.9 % c1: 0.4 % / 1.9 % / 0.3 % / 22.2 % / 77.8 % c2: 3.8 % / 3.7 % / 3.8 % / 4.8 % / 95.2 % c3: 4.2 % / 13.1 % / 3.7 % / 15.1 % / 84.9 % c4: 1.2 % / 4.7 % / 1.0 % / 19.2 % / 80.8 %
T wave inversion (ECG)	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT c0: 68.2 % / 53.3 % / 68.9 % / 3.8 % / 96.2 % c1: 0.7 % / 0.9 % / 0.7 % / 6.7 % / 93.3 % c2: 21.4 % / 20.6 % / 21.4 % / 4.6 % / 95.4 % c3: 9.3 % / 24.3 % / 8.6 % / 12.5 % / 87.5 % c4: 0.4 % / 0.9 % / 0.4 % / 10.0 % / 90.0 %
Cholestérol total	continu	parmi population / parmi EVENT / parmi NO EVENT 2.53 .. 6.10 .. 11.20 / 3.67 .. 6.08 .. 8.79 / 2.53 .. 6.10 .. 11.20
Infarctus du myocarde	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT YES: 4.0 % / 10.3 % / 3.7 % / 12.2 % / 87.8 % NO: 96.0 % / 89.7 % / 96.3 % / 4.5 % / 95.5 %
Accident vasculaire cérébral	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT YES: 1.6 % / 3.7 % / 1.5 % / 11.4 % / 88.6 % NO: 98.4 % / 96.3 % / 98.5 % / 4.7 % / 95.3 %
Tabagisme	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT CURRENT SMOKER: 13.2 % / 19.6 % / 12.9 % / 7.1 % / 92.9 % NOT SMOKER: 86.8 % / 80.4 % / 87.1 % / 4.4 % / 95.6 %
Index pondéral	continu	parmi population / parmi EVENT / parmi NO EVENT 15.73 .. 27.55 .. 53.71 / 18.60 .. 26.71 .. 46.10 / 15.73 .. 27.59 .. 53.71
BL_LVH2	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT YES: 5.5 % / 9.3 % / 5.3 % / 8.2 % / 91.8 % NO: 94.5 % / 90.7 % / 94.7 % / 4.6 % / 95.4 %
BL_LVH1	discret	parmi population / parmi EVENT / parmi NO EVENT / implique EVENT / implique NO EVENT YES: 13.9 % / 22.4 % / 13.5 % / 7.7 % / 92.3 % NO: 86.1 % / 77.6 % / 86.5 % / 4.3 % / 95.7 %
Classe	discrete	parmi population EVENT: 4.8 % (107) NO EVENT: 95.2 %(2123)

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

classé dans		
Event	No-Event	classe réelle
31	133	Event
804	4445	No-Event

Arbre de décision J48 (82.69%)

classé dans		
Event	No-Event	classe réelle
68	96	Event
1206	4043	No-Event

SVM (75.95%)

classé dans		
Event	No-Event	classe réelle
10	154	Event
406	4843	No-Event

Boosting J48 (89.65%)

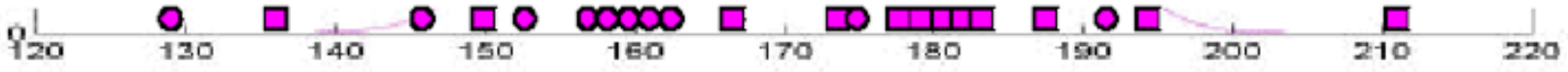
classé dans		
Event	No-Event	classe réelle
79	85	Event
1445	3804	No-Event

Boosting SVM (71.73%)

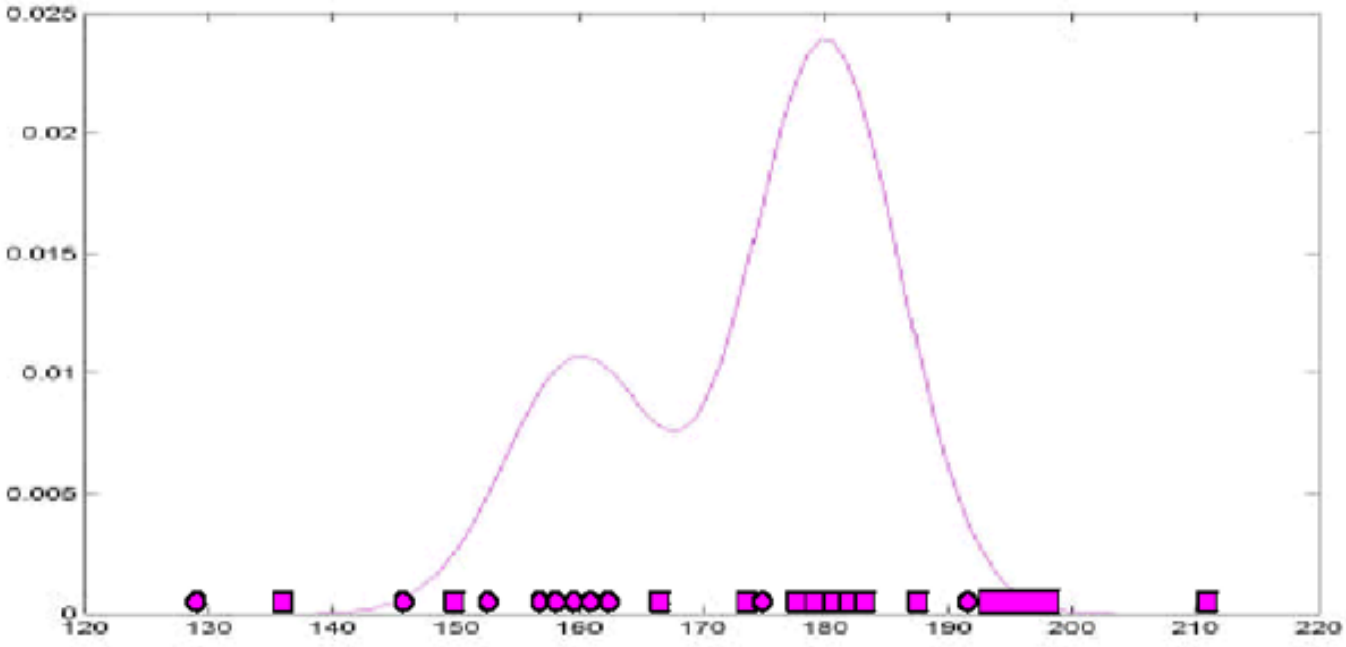
=> Très mauvais

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives



- **Soit le relevé des tailles d'un échantillon de personnes**



- **S'explique-t-il par un mélange de gaussiennes ?**

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

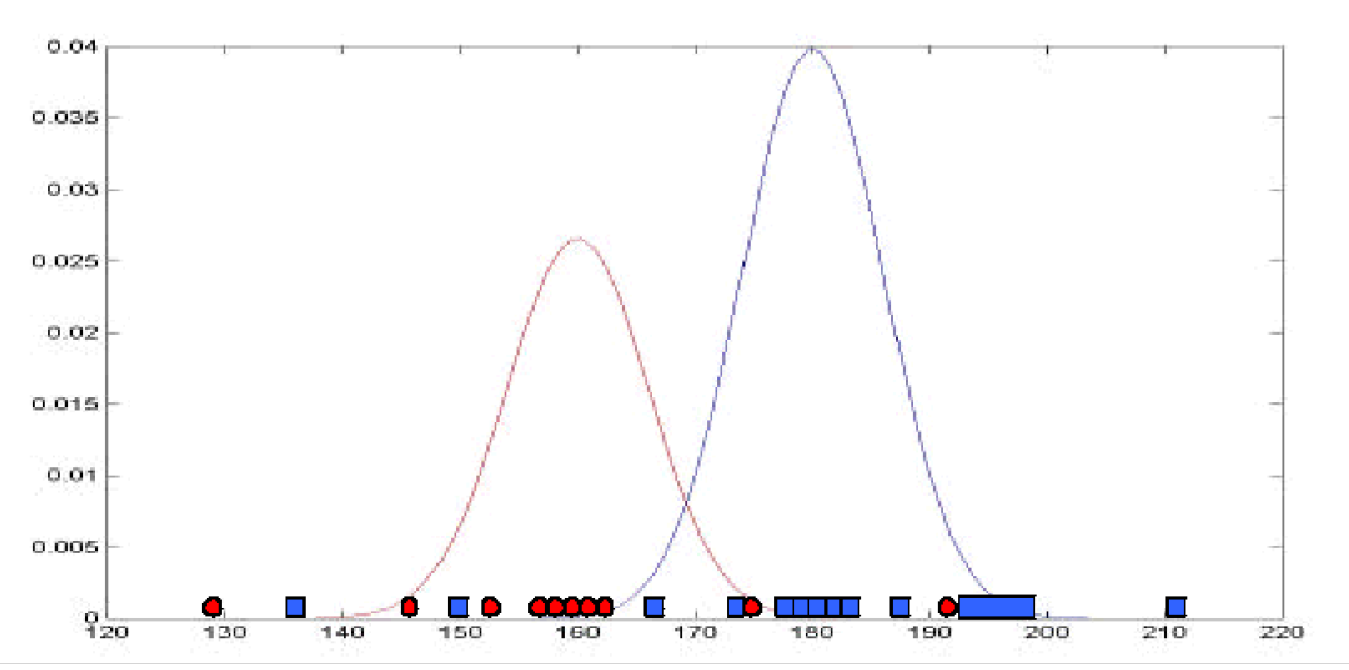


Figure: Résultat de EM après 10 itérations

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

Application de EM aux données INDANA

Problèmes sur les données

1. Données en dimension > 2
   - Malédiction de la dimensionnalité
2. Attributs numériques et symboliques
   - on traite les attributs symboliques comme des attributs numériques
3. Certaines probabilités très faibles
   - Organiser les calculs

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

Application de EM aux données INDANA

Problèmes sur la mise en œuvre

1. Initialisation des gaussiennes
2. Choix du nombre de gaussiennes
3. Mesure de la qualité de la distribution générative obtenue
   - Mesure de précision des gaussiennes
   - Mesure de proximité entre les gaussiennes
4. Répéter plusieurs fois

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

	C1	C2
Sex	0.400769	0.470284
Age	71.606773	71.228608
Taille	162.714471	163.967867
Poids	73.851848	72.748055
AnteDiabète	0.095598	0.120375
FreqCard	71.272287	71.443167
PresCardSy	169.521769	171.132822
PresCardDi	76.651002	76.793103
ECG3	**-0.000730**	**0.671186**
ECG4	**0.001666**	**0.713560**
ECG5	**0.464403**	**1.228933**
Cholestérol	6.127467	6.056233
AnteInfarMyo	-0.000066	0.109638
AnteAVC	0.001889	0.040733
Tabagisme	**-0.001972**	**0.375628**
DuréeDeVie	4.433309	4.090839
MortCardio	**0.002880**	**0.136196**
Test	1.000096	0.998852
IndexPonde	27.846327	27.006821
PI	0.648879	0.351121

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

	C1	C2	C3
Sex	0.402577	0.490435	0.473543
Age	71.720945	70.422680	71.085106
Taille	162.981711	162.625463	163.948893
Poids	73.025545	**84.802156**	70.246233
AnteDiabète	0.097973	0.122754	0.117753
FreqCard	71.194764	70.865979	71.982979
PresCardSy	169.795019	169.814433	171.174468
PresCardDi	76.994891	**72.994845**	77.251064
ECG3	0.184697	0.179129	**0.426587**
ECG4	**0.001782**	**0.535823**	**0.966781**
ECG5	**0.508375**	**1.054911**	**1.347823**
Cholestérol	6.131862	6.035777	6.032000
AnteInfarMyo	-0.002153	0.407016	0.021620
AnteAVC	0.000097	0.158685	0.007852
Tabagisme	**-0.000748**	**0.127645**	**0.569514**
DuréeDeVie	4.423003	4.424127	3.900896
MortCardio	0.003273	0.007893	**0.221601**
Test	0.999917	0.989657	1.002930
IndexPonde	27.417748	**32.311629**	26.032610
PI	0.702242	0.085202	0.212556

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

	C1	C2	C3	C4	C5
Sex	0.635777	1.006265	0.370937	-0.002173	0.280018
Age	69.022152	71.149123	72.078370	72.071864	72.828431
Taille	167.848987	171.089052	160.991410	157.585189	159.509510
Poids	74.459301	81.184055	75.146874	67.531150	71.599871
AnteDiabète	0.121072	0.111483	-0.005290	0.085876	0.303739
FreqCard	72.411392	69.696491	69.717868	72.401949	72.450980
PresCardSy	169.180380	168.428070	172.310345	170.196102	172.215686
PresCardDi	78.784810	78.140351	76.272727	75.987820	72.990196
ECG3	0.307111	-0.005471	1.072427	0.004335	0.416133
ECG4	0.293057	-0.001244	0.917058	0.003976	0.850126
ECG5	0.822984	0.489603	1.525769	0.434965	1.231773
Cholestérol	5.964612	5.826077	6.187988	6.319205	6.082155
AnteInfarMyo	0.195134	-0.003429	-0.013093	0.003355	0.134632
AnteAVC	0.084035	0.004356	0.002823	-0.001870	0.030509
Tabagisme	0.684516	-0.007599	-0.001494	0.001130	0.386453
DuréeDeVie	4.272555	4.350402	4.399177	4.479694	3.466186
MortCardio	-0.002026	-0.001107	0.001453	0.006354	0.521566
Test	0.998168	0.992083	0.997809	1.005380	1.003013
IndexPonde	26.344907	27.687979	29.107563	27.175360	28.120365
PI	0.138565	0.255157	0.141256	0.369955	0.095067

Prédiction du risque cardio-vasculaire

Apprentissage non supervisé : Méthodes génératives

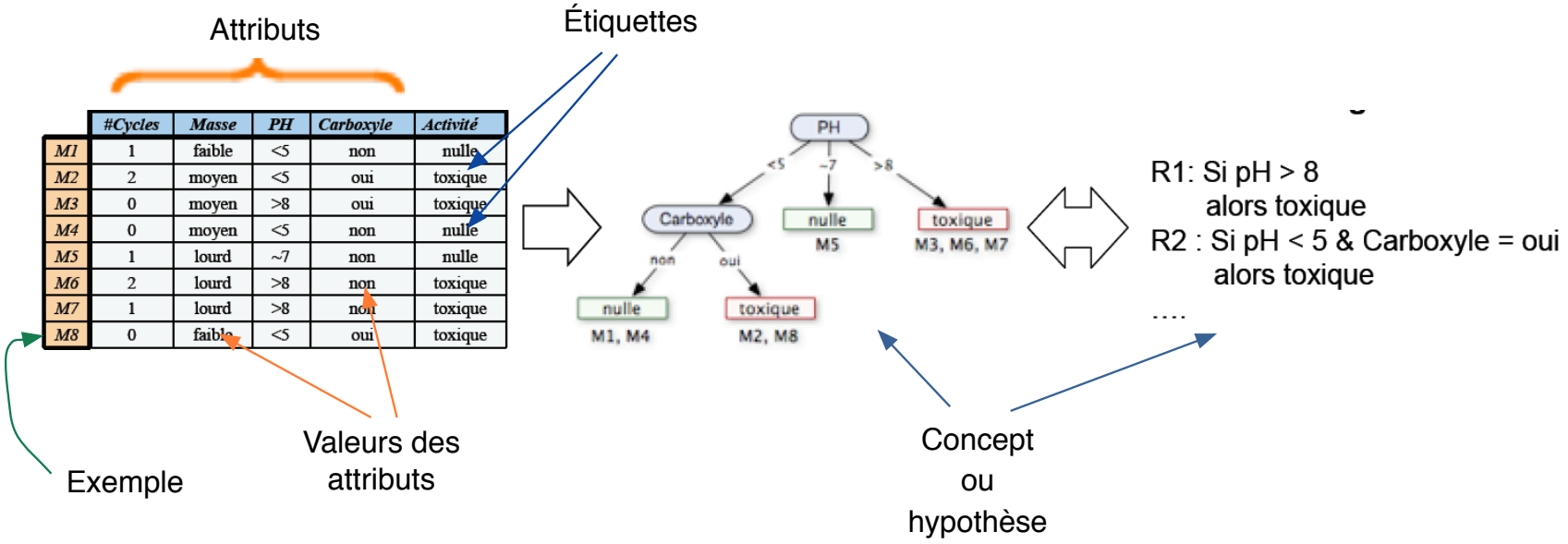
	C1	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*	*C10*
Sex	0.275320	0.314712	1.006361	0.273841	0.000018	0.516653	-0.000793	0.667932	0.580374	0.328783
Age	70.072464	73.814286	70.795645	71.585106	71.976705	69.852830	71.105042	74.754386	72.160714	73.491228
Taille	159.210144	155.992285	171.205360	160.614467	157.433511	166.633584	158.515210	165.233685	166.506071	160.242806
Poids	75.192417	73.424880	80.322151	69.135396	66.864867	78.803158	67.858941	72.957979	77.727600	67.077095
AnteDiabète	0.396307	-0.000245	0.005262	0.581352	-0.003536	0.409263	-0.000376	0.190324	0.102993	0.330896
FreqCard	72.608696	67.657143	70.289782	71.021277	72.735441	71.577358	71.806723	70.561404	72.125000	71.543860
PresCardSy	173.724638	175.778571	168.229481	167.468085	170.259567	169.743396	170.134454	173.350877	169.232143	173.491228
PresCardDi	69.826087	74.042857	78.514238	76.159574	76.031614	77.773585	76.886555	75.438596	77.267857	74.894737
ECG3	0.605651	0.979926	-0.006506	0.007533	0.002252	0.997865	0.013376	0.887922	0.215892	0.086252
ECG4	1.457110	1.204218	-0.000031	1.145312	0.006149	-0.001327	-0.005722	2.991792	0.015498	0.177305
ECG5	1.811443	1.877092	0.485615	1.410199	0.001409	0.663039	1.512070	2.854661	0.693466	0.827776
Cholestérol	6.323332	5.915452	5.829199	6.193195	6.340963	6.000691	6.348895	6.004057	6.015192	6.205946
AnteInfarMyo	0.015293	0.025705	-0.000824	-0.006045	0.001378	-0.018315	0.005667	0.340093	0.602781	-0.036198
AnteAVC	0.017484	-0.003429	0.003746	-0.000665	-0.003150	-0.000104	-0.004503	0.100204	0.012443	0.484453
Tabagisme	0.553261	-0.005677	0.143526	0.004757	-0.000934	0.159551	0.362135	0.211041	0.113917	0.264127
DuréeDeVie	4.373657	4.414569	4.341177	4.283863	4.495191	4.388671	4.399128	3.259724	3.623179	3.521693
MortCardio	-0.020055	-0.013741	-0.003954	-0.003346	0.003141	0.008787	0.012828	0.427185	0.474588	0.561736
Test	1.015702	0.993923	0.993874	1.004672	1.006162	1.000616	1.001697	0.986555	0.999047	0.979460
IndexPonde	29.812904	30.453723	27.359682	26.668951	26.962509	28.332166	27.029597	26.550892	28.000128	26.031612
PI	0.030942	0.062780	0.267713	0.042152	0.269507	0.118834	0.106726	0.025561	0.050224	0.025561

Prédiction du risque cardio-vasculaire

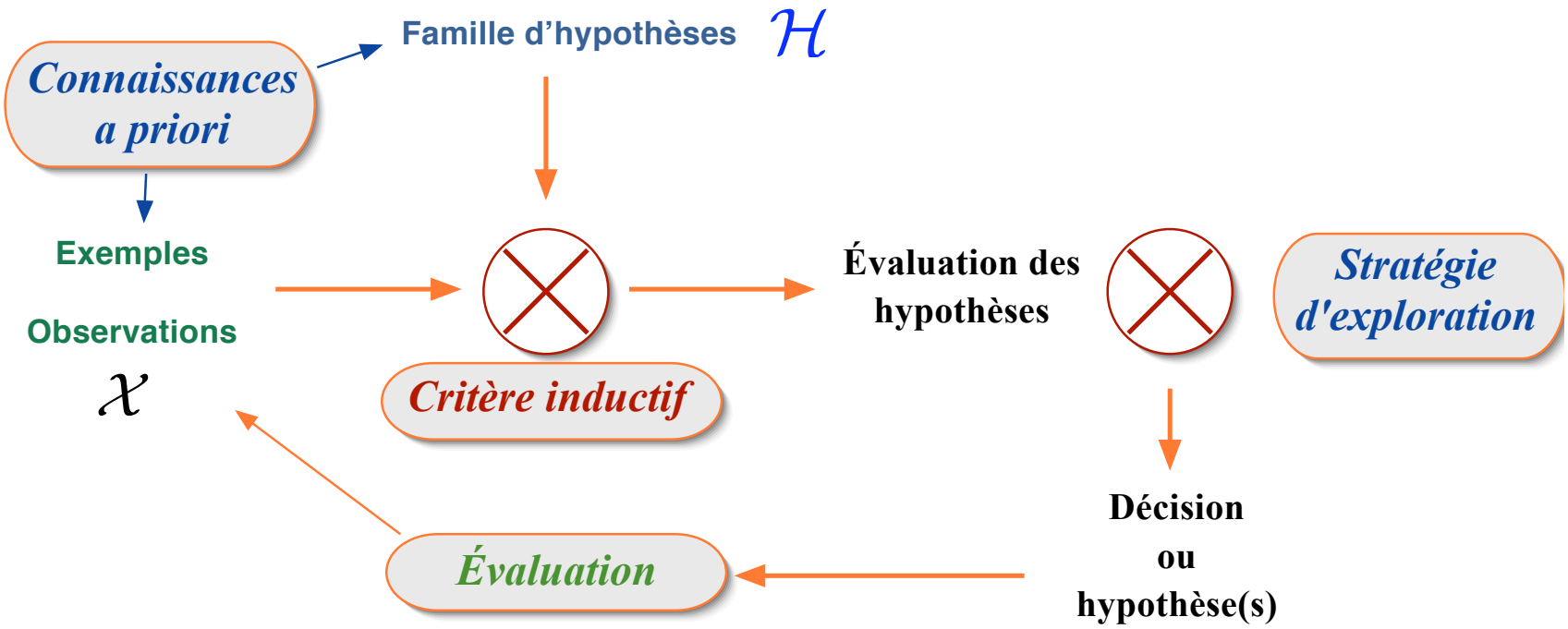
Apprentissage supervisé

Et la calibration ... ?

Apprentissage supervisé
Méthodes



Apprentissage supervisé
Méthodes



Apprentissage supervisé
Méthodes

Apprentissage supervisé
Méthodes

Identifier une dépendance cible :

- $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$
- *Fonction cible* $f : \mathcal{X} \to \mathcal{Y}$

Apprentissage supervisé
Méthodes

Identifier une dépendance cible :

- $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$
- *Fonction cible* $f : \mathcal{X} \to \mathcal{Y}$

Prédire correctement

$$\begin{aligned} & \mathcal{X} \to \mathcal{Y} \\ & x \mapsto \texttt{decision bayésienne}(x) \\ \text{ou} \quad & x \mapsto \texttt{decision}(x) = f(x) \end{aligned}$$

Apprentissage supervisé
Méthodes

Identifier une dépendance cible :

- $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$
- *Fonction cible* $f : \mathcal{X} \to \mathcal{Y}$

Prédire correctement

$$\begin{aligned} & \mathcal{X} \to \mathcal{Y} \\ & x \mapsto \texttt{decision bayésienne}(x) \\ \text{ou} \quad & x \mapsto \texttt{decision}(x) = f(x) \end{aligned}$$

Expliquer le monde

$$h(\cdot) = f(\cdot)$$

Apprentissage supervisé
Méthodes

Identifier une dépendance cible :

- $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$
- *Fonction cible* $f : \mathcal{X} \to \mathcal{Y}$

Prédire correctement

$$\begin{aligned} & \mathcal{X} \to \mathcal{Y} \\ & x \mapsto \texttt{decision bayésienne}(x) \\ \text{ou} \quad & x \mapsto \texttt{decision}(x) = f(x) \end{aligned}$$

Expliquer le monde

$$h(\cdot) = f(\cdot)$$

Indécidable et illusoire

Apprentissage supervisé
Méthodes

Fonction de perte :

$$\begin{aligned} \ell(h) : \mathcal{X} \times \mathcal{Y} &\rightarrow \mathbb{R}^+ \\ (\boldsymbol{x}, y) &\mapsto \ell(h(\boldsymbol{x}), y) \end{aligned}$$

Apprentissage supervisé
Méthodes

Fonction de perte :

$$\begin{aligned} \ell(h) : \mathcal{X} \times \mathcal{Y} &\rightarrow \mathbb{R}^+ \\ (\boldsymbol{x}, y) &\mapsto \ell(h(\boldsymbol{x}), y) \end{aligned}$$

Risque réel : espérance de perte

$$R(h) = \mathbb{E}[\ell(h(\boldsymbol{x}), y)] = \int_{\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}} \ell(h(\boldsymbol{x}), y)\, \mathbf{P}_{\mathcal{X}\mathcal{Y}}\, d(\boldsymbol{x}, y)$$

Apprentissage supervisé
Méthodes

Mais, on ne connaît pas $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$

Échantillon d'apprentissage supposé représentatif

$$\mathcal{S}_m = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$$

Apprentissage supervisé
Méthodes

Mais, on ne connaît pas $\mathbf{P}_{\mathcal{X}\mathcal{Y}}$

Échantillon d'apprentissage supposé représentatif

$$\mathcal{S}_m = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$$

Minimisation du Risque Empirique

$$R_m(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(\boldsymbol{x}_i), y_i)$$

Apprentissage supervisé

Méthodes : critère de minimisation du risque empirique

MRE

Choisir l'hypothèse $\hat{h}$ telle que : $\hat{h} = \text{ArgMin}_{h \in \mathcal{H}} \left[R_{Emp}(h) \right]$

$$R_{\text{Emp}}(h) = \frac{1}{m} \sum_{(\boldsymbol{x}_i, u_i) \in S} \ell(h(\boldsymbol{x}_i), u_i)$$

Apprentissage supervisé

Méthodes : autres critères inductifs

Compression maximale d'**information**

Choisir l'hypothèse $\hat{h}$ telle que : $\hat{h} = \text{ArgMin}_{h \in \mathcal{H}} \left[L(\mathcal{S}_m) \right]$

$$L(\mathcal{S}_m) = L(h) + L(\mathcal{S}_m | h))$$

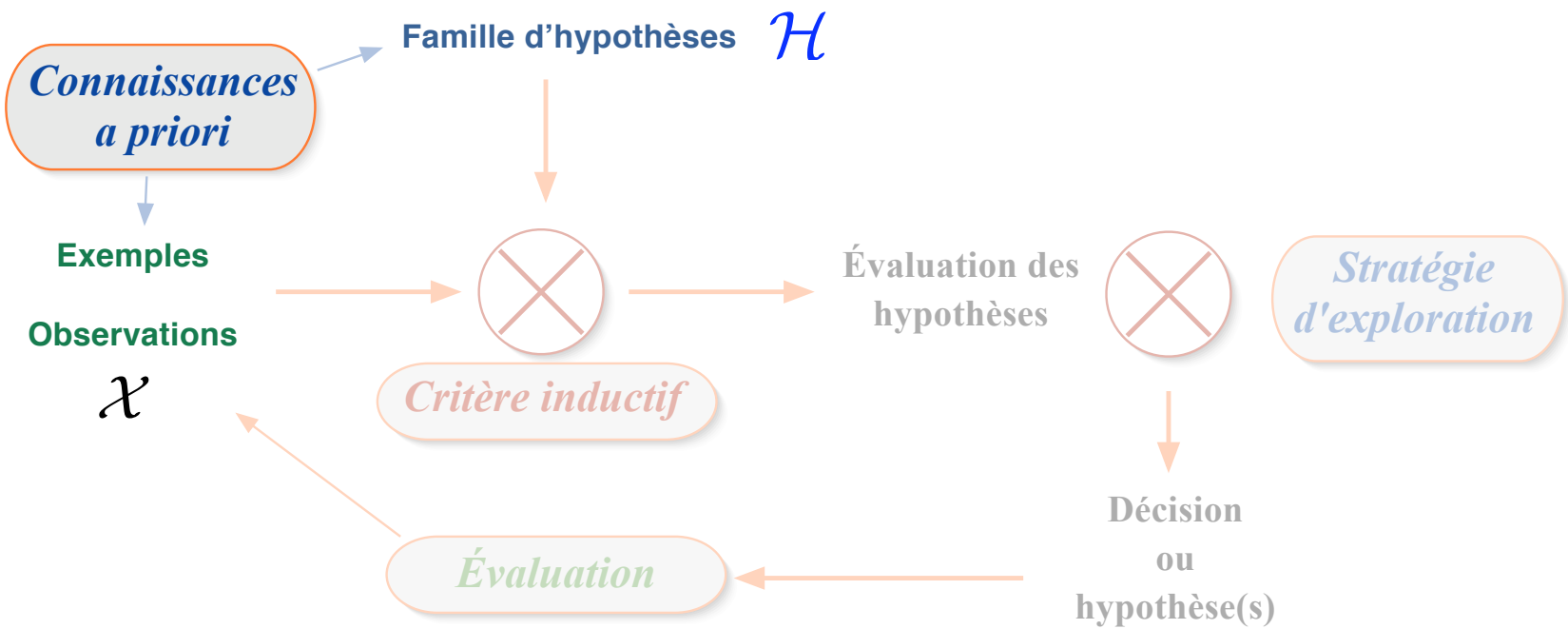
MLE et MAP

Choisir l'hypothèse $\hat{h}$ telle que :

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{ArgMax}}\, l(h) = \underset{h \in \mathcal{H}}{\text{ArgMax}}\, \ln \left[p(\mathcal{S}_m | h) \right] \quad \text{(MLE)}$$

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{ArgMax}}\, p(\mathcal{S}_m | h)\, p(h) \quad \text{(MAP)}$$

Apprentissage supervisé
Méthodes



Apprentissage supervisé
Méthodes

- “simple” plus proche(s) voisin(s)
- SVM

- Modèles linéaires
- Modèles bayésiens

- Réseaux de neurones ; Modèles de Markov à états cachés (HMM)

- Arbres de décision
- Règles (ILP : Induction of Logic Programs)
- Grammaires

Apprentissage supervisé
Méthodes

Apprentissage supervisé
Méthodes

Pas de $\mathcal{H}$

Plus-proches-voisins

Apprentissage supervisé
Méthodes

Pas de $\mathcal{H}$

Plus-proches-voisins

$\mathcal{H}$ muni d'une distance

Réseaux de neurones ; régression logistique ; modèles bayésiens ; HMM ; ...

- Optimisation directe (e.g. pseudo-inverse)
- Adaptation itérative = descente de gradient

Apprentissage supervisé
Méthodes

Pas de $\mathcal{H}$

Plus-proches-voisins

$\mathcal{H}$ muni d'une distance

Réseaux de neurones ; régression logistique ; modèles bayésiens ; HMM ; ...

- Optimisation directe (e.g. pseudo-inverse)
- Adaptation itérative = descente de gradient

$\mathcal{H}$ muni d'une relation de généralité

Inférence grammaticale ; Induction de règles ; Apprentissage relationnel

- Apprentissage symbolique
- Bruit

Apprentissage supervisé
Méthodes

Apprentissage supervisé
Méthodes

Modèles génératifs

Demandent $p(\boldsymbol{x}|\mathcal{C}_k)$ et $p(\mathcal{C}_k)$

Apprentissage supervisé
Méthodes

Modèles génératifs

Demandent $p(\boldsymbol{x}|\mathcal{C}_k)$ et $p(\mathcal{C}_k)$

Fonctions de décision

- Fonctions composées de fonctions de base
- Méthodes à Noyaux (*Kernel methods*)

Apprentissage supervisé
Méthodes

Modèles génératifs

Demandent $p(\boldsymbol{x}|\mathcal{C}_k)$ et $p(\mathcal{C}_k)$

Fonctions de décision

- Fonctions composées de fonctions de base
- Méthodes à Noyaux (*Kernel methods*)

Approches constructives

- Spécialisation dans l'espace $\mathcal{X}$
- Modèles hiérarchiques ou "profonds" (*Deep models*)

Apprentissage supervisé

Méthodes : Fonctions de décision à base de dictionnaire

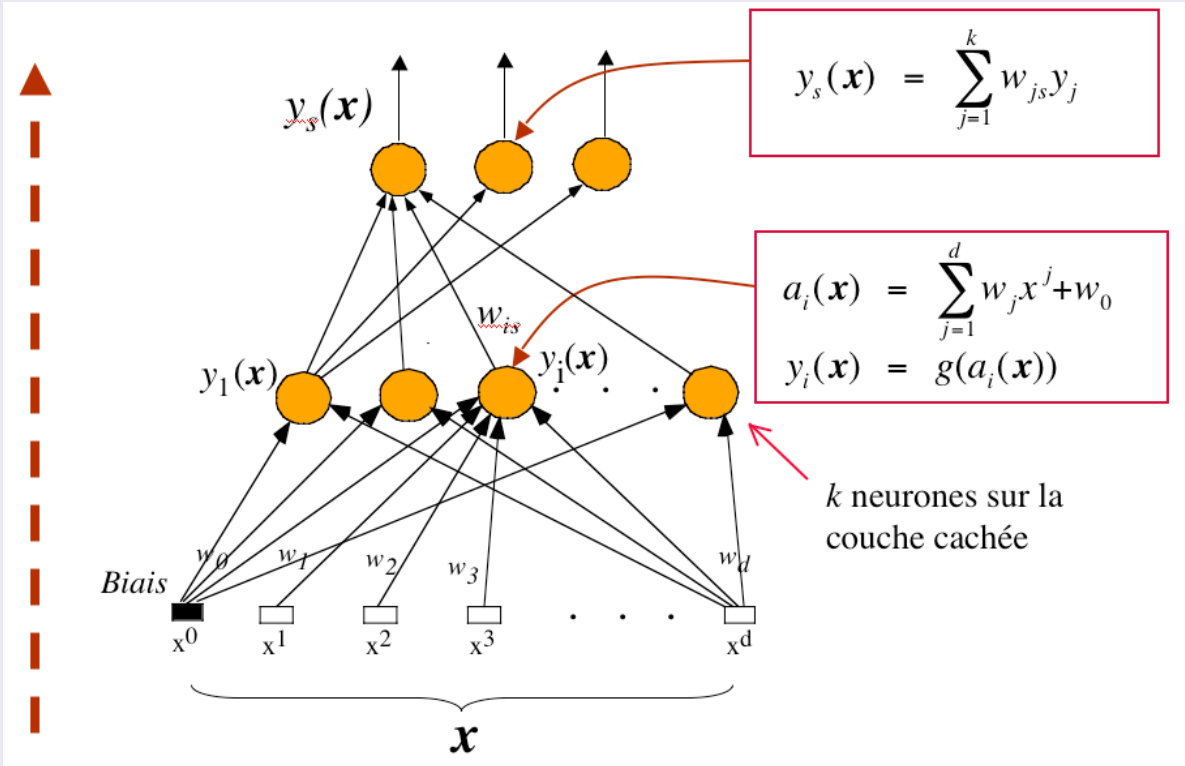
- $h(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{n} w_i \, g_i(\boldsymbol{x}) + w_0$
- où les $g_i(\boldsymbol{x})$ sont des ***fonctions de base***

Apprentissage supervisé

Méthodes : Fonctions de décision à base de dictionnaire

- $h(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{n} w_i \, g_i(\boldsymbol{x}) + w_0$
- où les $g_i(\boldsymbol{x})$ sont des ***fonctions de base***

Exemple : Perceptron multi-couches



Apprentissage supervisé

Méthodes : Fonctions de décision par noyaux

- $h(\boldsymbol{x}) = \sum_{i \text{ ``critiques''}} \alpha_i \, y_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) + \alpha_0$
- où les $K_i(\cdot, \cdot)$ sont des ***fonctions noyaux***

Apprentissage supervisé

Méthodes : Fonctions de décision par noyaux

- $h(\boldsymbol{x}) = \sum_{i \text{ ``critiques''}} \alpha_i \, y_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) + \alpha_0$
- où les $K_i(\cdot, \cdot)$ sont des ***fonctions noyaux***

$$K_G(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\frac{||\boldsymbol{x}, \boldsymbol{x}_i||^2}{2\,\sigma^2}\right)$$

$$K_L(\boldsymbol{x}, \boldsymbol{x}_i) = \boldsymbol{x}^\top \boldsymbol{x}_i$$

$$K_{Poly1}(\boldsymbol{x}, \boldsymbol{x}_i) = (\boldsymbol{x}^\top \boldsymbol{x}_i)^d$$

$$K_{Poly2}(\boldsymbol{x}, \boldsymbol{x}_i) = (\boldsymbol{x}^\top \boldsymbol{x}_i + c)^d$$

$$K_{sig}(\boldsymbol{x}, \boldsymbol{x}_i) = \tanh(\kappa \boldsymbol{x}^\top \boldsymbol{x}_i + \theta)$$

Apprentissage supervisé

Méthodes : Fonctions de décision par noyaux

- $h(\boldsymbol{x}) = \sum_{i \text{ ``critiques''}} \alpha_i \, y_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) + \alpha_0$
- où les $K_i(\cdot,\cdot)$ sont des ***fonctions noyaux***

$$K_G(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\frac{||\boldsymbol{x}, \boldsymbol{x}_i||^2}{2\,\sigma^2}\right)$$

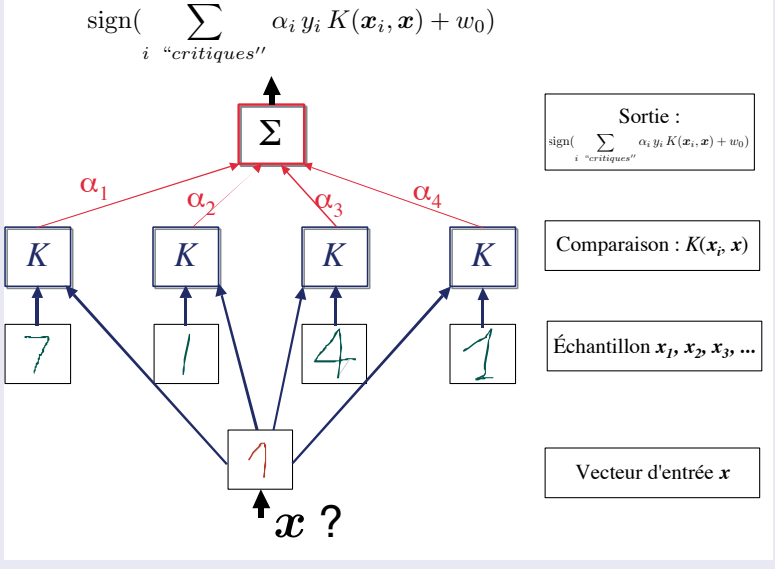
$$K_L(\boldsymbol{x}, \boldsymbol{x}_i) = \boldsymbol{x}^\top \boldsymbol{x}_i$$

$$K_{Poly1}(\boldsymbol{x}, \boldsymbol{x}_i) = (\boldsymbol{x}^\top \boldsymbol{x}_i)^d$$

$$K_{Poly2}(\boldsymbol{x}, \boldsymbol{x}_i) = (\boldsymbol{x}^\top \boldsymbol{x}_i + c)^d$$

$$K_{sig}(\boldsymbol{x}, \boldsymbol{x}_i) = \tanh(\kappa \boldsymbol{x}^\top \boldsymbol{x}_i + \theta)$$

Exemple : Séparateurs à Vastes Marges (SVM)

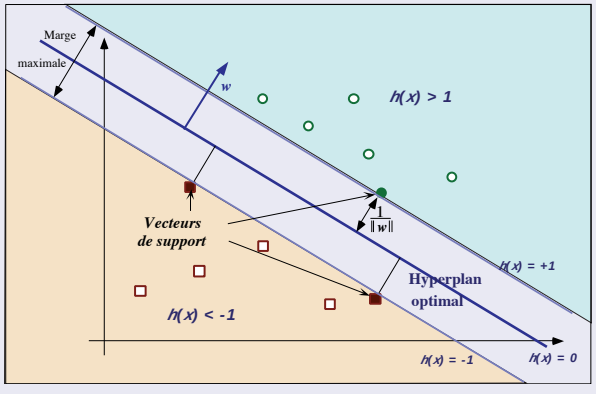


Apprentissage supervisé

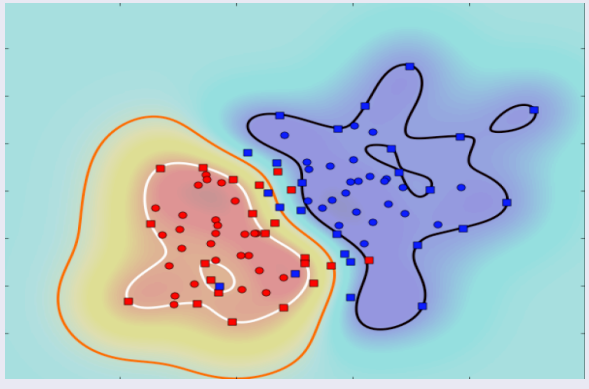
Méthodes. Fonctions de décision par noyaux : Les SVM

$$h^*(\mathbf{x}) \;=\; (\boldsymbol{w}^\star\ \mathbf{x}) \;+\; w_0^\star \;=\; \sum_{i=1}^{m} \alpha_i^\star\, u_i\, .\, \langle \phi(\mathbf{x}_i)\,,\, \phi(\mathbf{x}) \rangle \;+\; w_0^\star$$

SVM : espace des redescripteurs $\Phi(\mathcal{X})$



SVM : espace initial $\mathcal{X}$

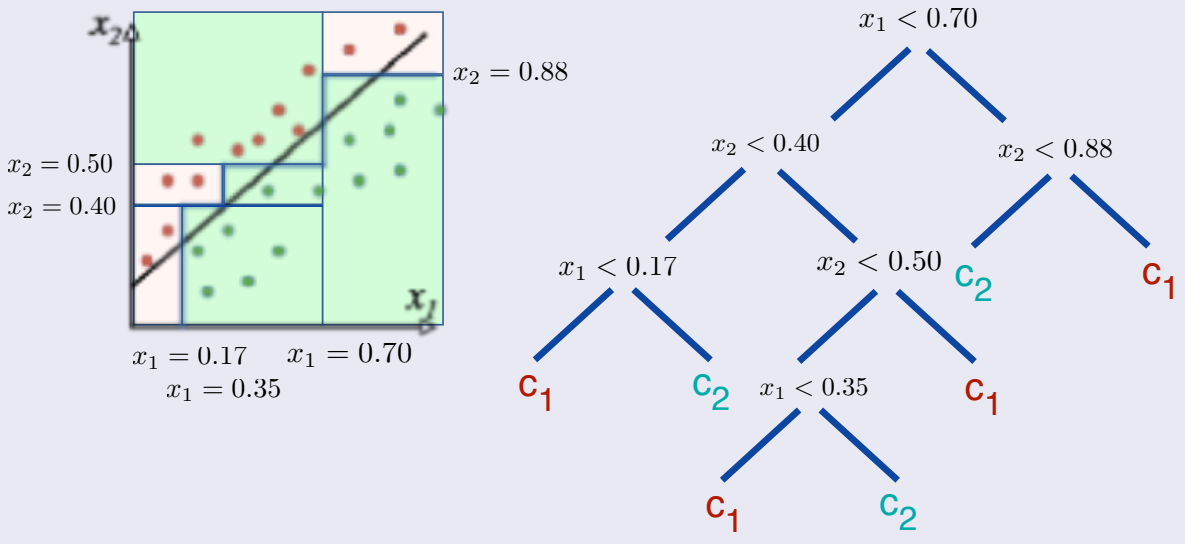


Apprentissage supervisé

Méthodes : Fonctions de décision par spécialisation

$$h_{\mathcal{X}} = h_{\mathcal{X}_1} \times h_{\mathcal{X}_2} \times \ldots \times h_{\mathcal{X}_n}$$
$$\mathcal{X} = h_{\mathcal{X}_1} \cup h_{\mathcal{X}_2} \cup \ldots \cup h_{\mathcal{X}_n}$$

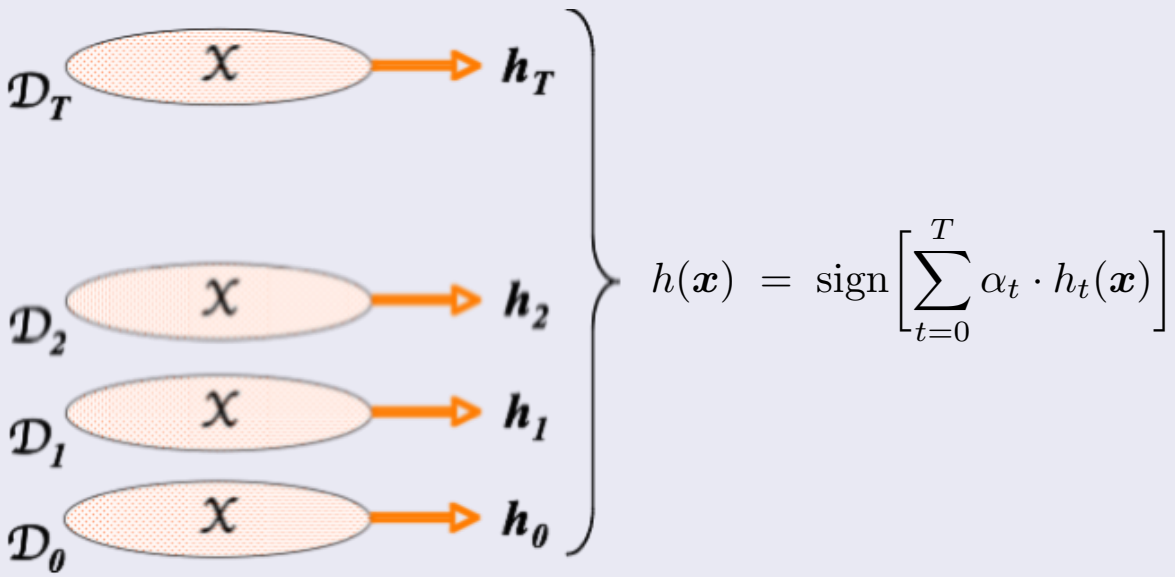
Exemple : Arbres de décisions



Apprentissage supervisé

Méthodes : Fonctions de décision par vote

Exemple : Boosting



Apprentissage supervisé

Étude de cas : Pima Indians Diabetes Data Set

Problème

Prédire la classe `diabète` ou `pas diabète` à partir de :

- mesures sur 8 descripteurs numérique
- population 768 femmes

Problème de classification

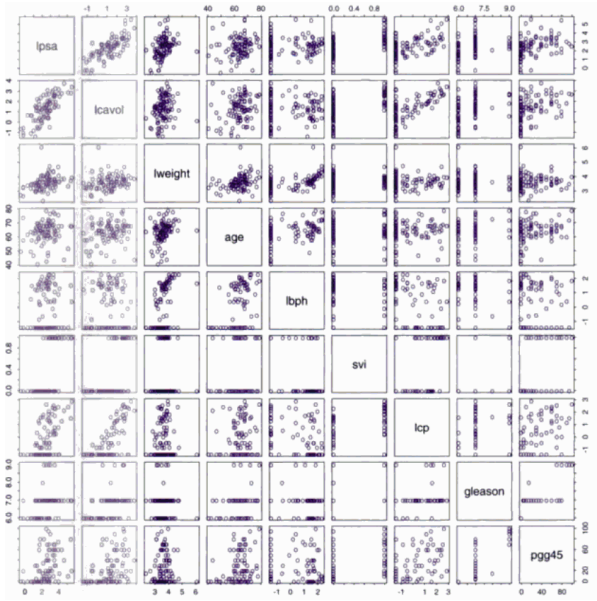
Apprentissage supervisé

Étude de cas : cancer de la prostate

Problème

Prédire le log de PSA (*prostate specific antigen* à partir de :

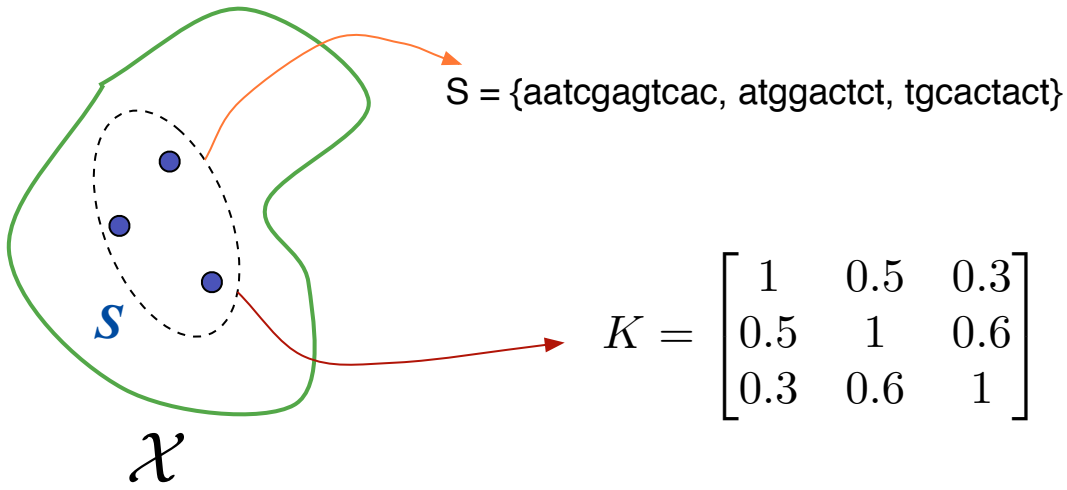
- mesures sur 8 descripteurs
- population 97 hommes



Problème de régression

Si en plus, estimation de la probabilité : Problème de calibration

Apprentissage supervisé
Méthodes à noyaux



Les méthodes à noyau remplacent les informations sur les objets comparés par la matrice de leurs distances : la matrice de Gram. C'est un changement de perspective radical.

Apprentissage supervisé
Méthodes à noyaux

Fonction noyau 3_spectre

Soient deux chaînes construites sur l'alphabet {A, C, G, T} :

$$s = \texttt{GAGTTCTAAT}$$
$$t = \texttt{GGATCACTAA}$$

Les sous-chaînes de longueur 3 présentes dans ces deux chaînes sont :

GAG, AGT, GTT, TTC, TCT, CTA, TAA, AAT
GGA, GAT, ATC, TAC, CAC, ACT, CTA, TAA

Les sous-chaînes en commun sont : CTA et TAA, ce qui donne un produit interne $\kappa(s,t) = 2$.

Des méthodes de calcul récursives permettent de calculer efficacement cette fonction noyau.

Apprentissage supervisé
Méthodes à noyaux

Fonction noyau 3_spectre prenant en compte les trous

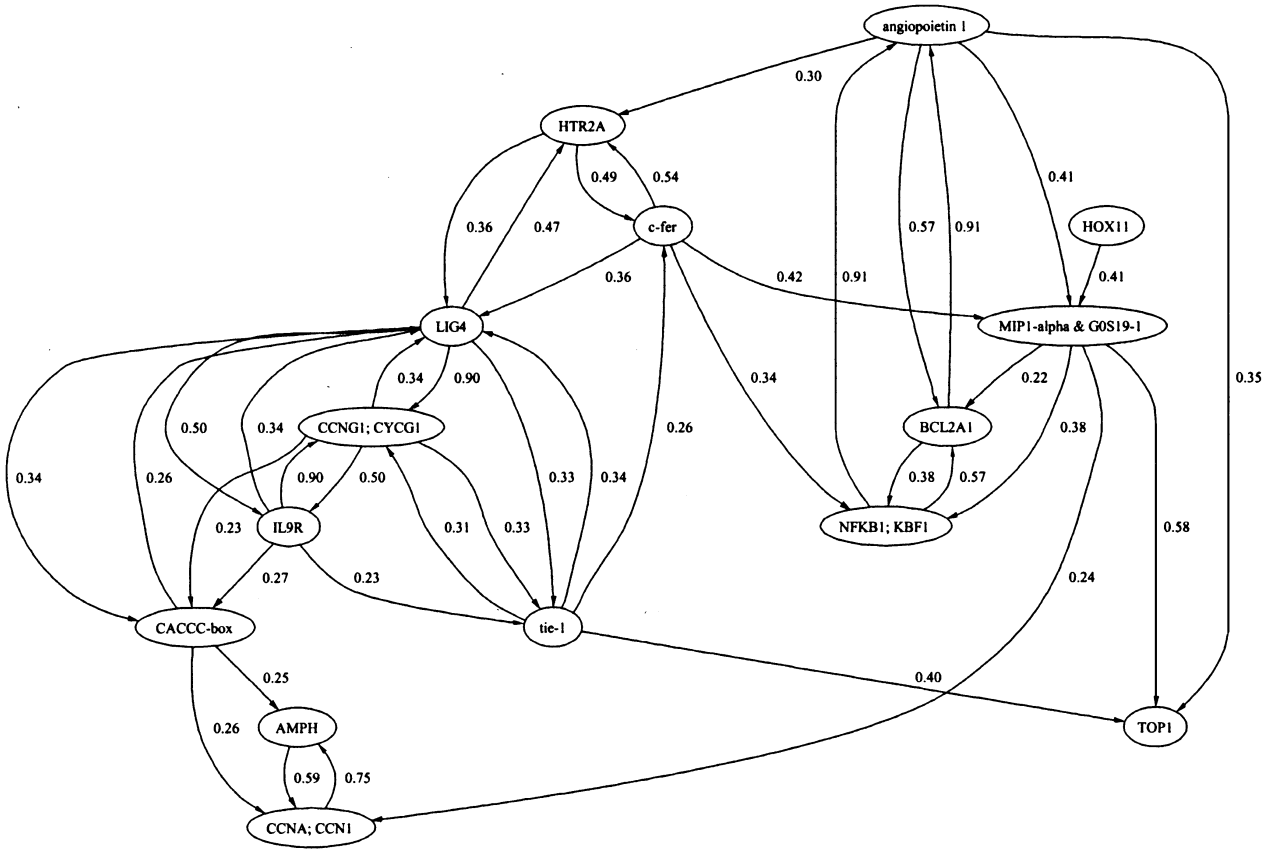
Par exemple, la séquence ACG est présente dans les séquences ACGT, ACTTGA et AGGCATGA, mais la première occurrence est plus significative car elle apparaît comme une sous-séquence consécutive de la séquence considérée, alors que la dernière est la moins significative car impliquant une grande dispersion de la sous-séquence ACG dans AGGCATGA.

On introduit un facteur d'amoindrissement $\lambda \in [0, 1]$.
Si la sous-séquence u de longueur $|u|$ est trouvée dans la séquence s avec une longueur $l(u)$, alors on pondère la sous-séquence u avec le poids $\lambda^{l(u)-|u|}$.

Avec $\lambda = 0.9$, on aurait le poids de la sous-séquence $u =$ ACG valant 1 dans ACGT, $0.9^{5-3} = 0.81$ dans ACTTGA et $0.9^{7-3} = 0.9^4 \approx 0.66$.

Apprentissage pour les réseaux d'interaction

Modèles graphiques probabilistes



An example of influences in a small network containing 15 genes. Each **arrow** represents the influence of a gene on another gene. The **number** next to the arrow is the magnitude of the influence. Only those influences that are above 0.2 are shown. (Tiré de [Shmulevitch, Dougherty and Zhang, 2002] [SDZ02])

Plan

Conclusion

Leçons : ce que j'ai appris

- Impossible sans **expert**
- Projets à **longue durée**
  - Multi-disciplinaires
  - Recueil des données (choix de descripteurs)
  - Choix de la tâche
- Notion d'**hypothèse nulle**
  - Sélection des attributs
  - Exemples positifs seuls (Motifs fréquents et surprenants)
- Importance de la **redescription**
  - Être imaginatif
- **Étiquettes** elles-mêmes peu fiables

Conclusion
Bilan

Champ immense pour l'AA

- Enormément de données
- Des "utilisateurs" motivés
- Utilisation de toutes les techniques d'apprentissage
- Générateurs de nouveaux problèmes (grande dimension, incertitude, ...)
- ...

Beaucoup de résultats inaccessibles autrement

Conclusion
Domaines d'application à venir

Domaines d'avenir

- Examen du biotope du tract gastro- intestinal
- Examen du biotope des surfaces (*peau, poumons, tracts, ...*)
- Dynamique des réseaux (*de régulation, ...*)
- ...

Conclusion
Ce que nous n'avons pas traité

Ce que nous n'avons pas traité

- Le traitement des arbres (arbres minimaux couvrants)
- Le traitement de texte bio-médical
- Le traitement d'images
- ...

Conclusion
Petite bibliographie

P. Baldi and S. Brunak.

Bioinformatics. The machine learning approach.
MIT Press, 1998.

N. Cristianini and M. Hahn.

Introduction to computational genomics. A case studies approach.
Cambridge University Press, 2007.

M. Eisen, P. Spellman, P. Brohn, and D. Botstein.

Cluster analysis and display of genome-wide expression patterns.
Proceedings of the National Academy of Sciences of the United States of America, 95(25):14863–14868, 1998.

I. Guyon and A. Elisseeff.

An introduction to variable and feature selection.
Journal of Machine Learning Research, 3:1157–1182, 2003.

N. Jones and P. Pevzner.

An introduction to bioinformatics algorithms.
MIT Press, 2004.

W. Majoros.

Methods for computational gene prediction.
Cambridge University Press, 2007.

I. Shmulevitch, E. Dougherty, and W. Zhang.

From boolean to probabilistic boolean networks as models of genetic regulatory networks.
Proceedings of the IEEE, 90(11):1778–1792, 2002.