Identification de nouveaux membres dans des familles d'interleukines

Nicolas Beaume

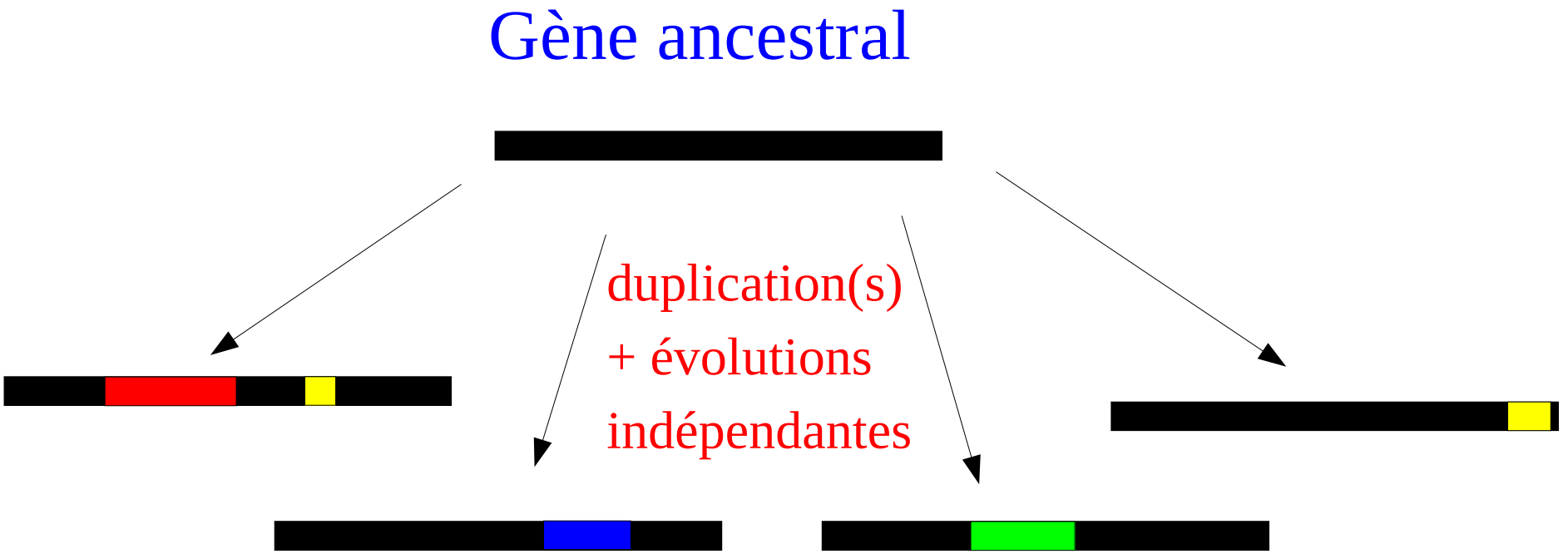
Jérôme Mickolajczak

Gérard Ramstein

Yannick Jacques

1ère partie :
Définition de la problématique

Les familles de gènes

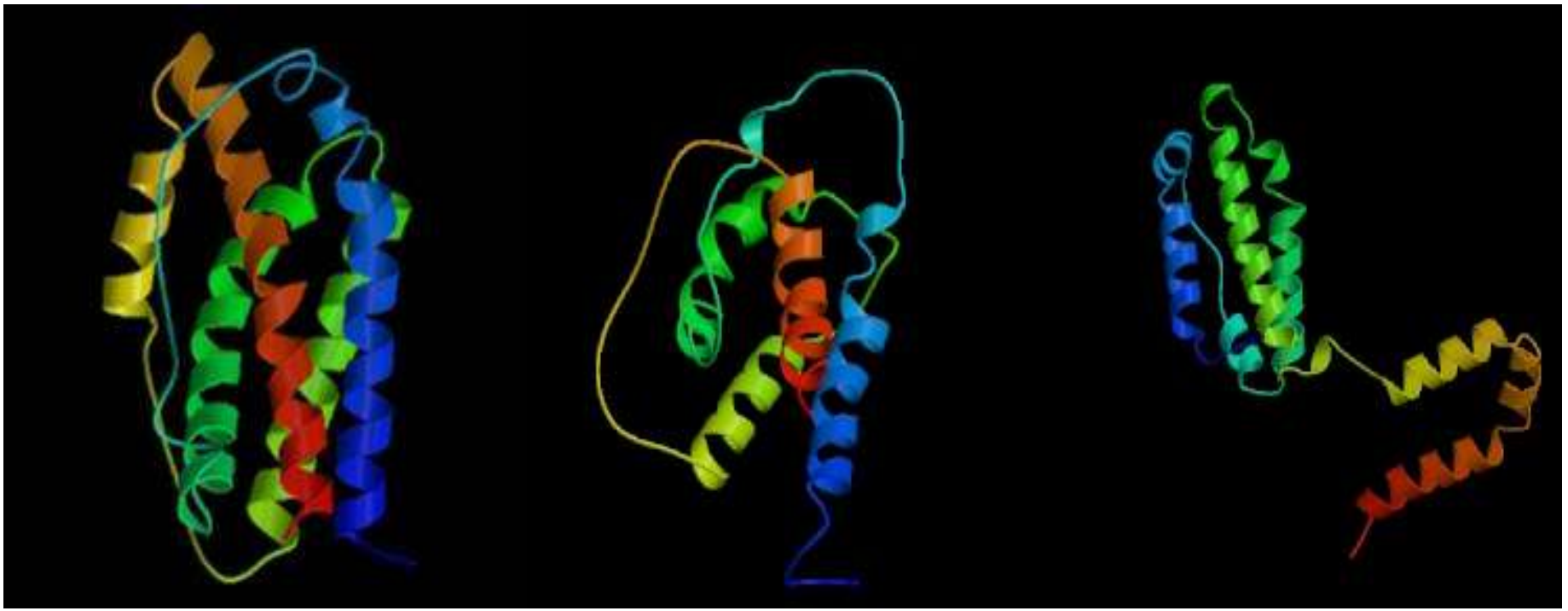


Les cytokines

- 130 « fonctions de cytokines » estimées chez l'Homme.
- Impliquées dans la communication cellulaire, plus particulièrement dans la réaction immunitaire.
- Toutes n'ont pas de séquence connue.

- Classification en fonction du type de récepteurs.
- Gènes souvent en clusters.
- Structure de gènes assez conservée.
- Similarité :
    - Importante dans une sous-famille
    - Faible (< 15%) entre sous-familles

- Structures protéiques relativement conservées autour de 4 hélices α



Enjeux et difficultés

Intérets :

- Applications médicales (psoriasis, inflammation, arthrite, cancer...)
- Étude d'une des plus grandes familles de gènes du génome humain

Problèmes :

- Les cytokines ne se ressemblent pas toutes.
- Pas de caractéristiques universelles identifiées
- Des relations d'homologie qui peuvent être très lointaines

Formulation des objectifs

130 cytokines supposées, 75 cytokines identifiées...

- **Identifier les cytokines manquantes**

- **Mettre au point une méthode générale d'identification de membres d'une famille de gènes**

2ème partie :
Outils et méthodes

Données

Données annotées

45 cytokines de 3 sous-familles

(IL-6, IL-2 & IL-10/INFs)

+

Contre-exemples tirés de la base SCOP

Données à analyser

5 330 000 séquences humaines provenant d'Unigene

Méthodes existantes

Basées sur les séquences :

- BLAST et PSI-BLAST
- HMM / SVM
- Recherche de profils
- Graphes
- ...

Basées sur les structures :

- Superposition de structures
- HMM / SVM
- ...

Hybrides :

- Association score structural – score de séquences
- ...

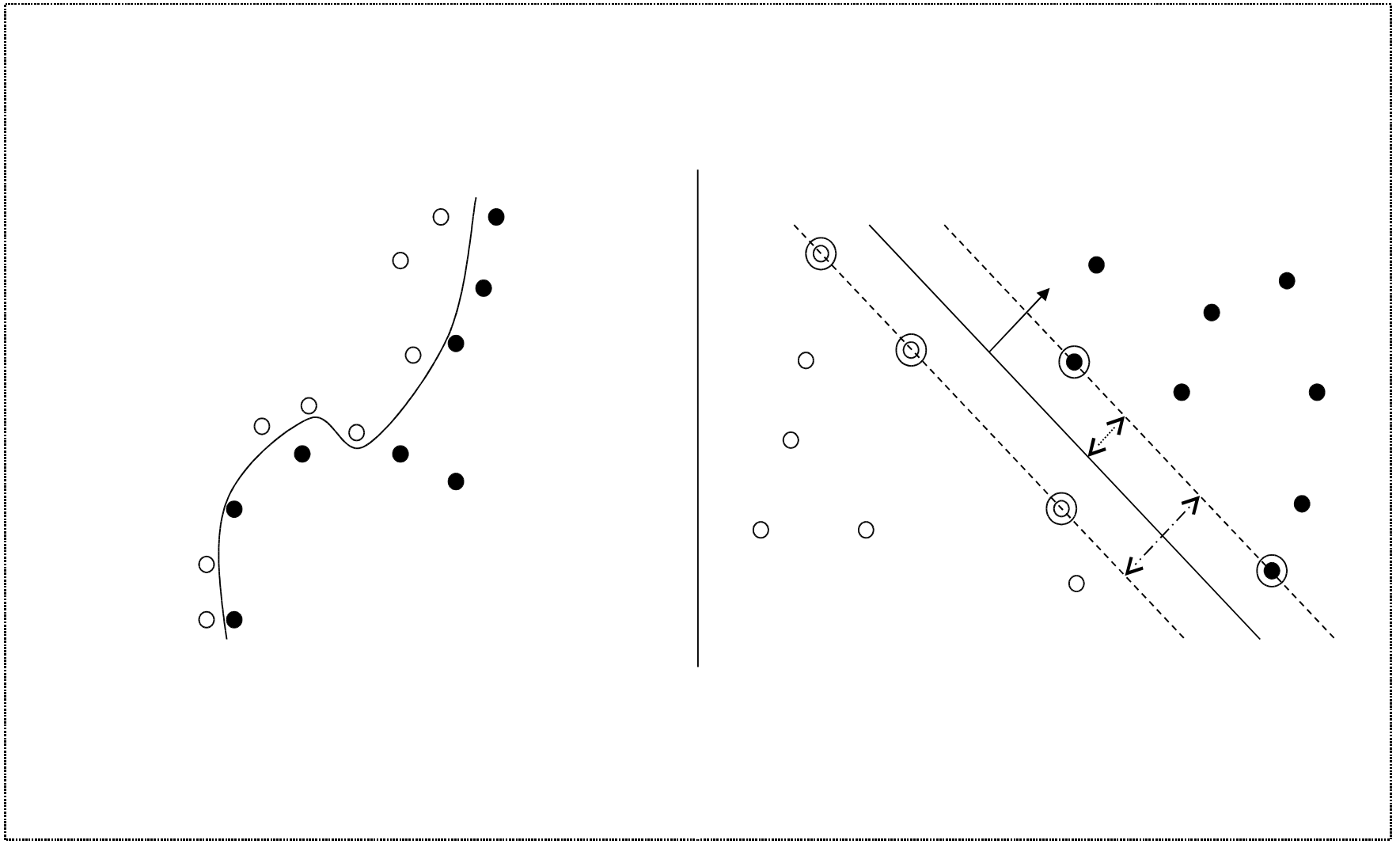
Stratégie choisie

Une méthode d'apprentissage supervisé qui a fait ses preuves : les SVM

Données annotées → Jeu d'apprentissage

Données à analyser → jeu de recherche

Les SVM



Hyperplan sous la forme : $w.x + b = 0$ où w est un vecteur de poids, x le vecteur à classer et b représente le biais

Les SVM manipulent essentiellement des produits scalaires.

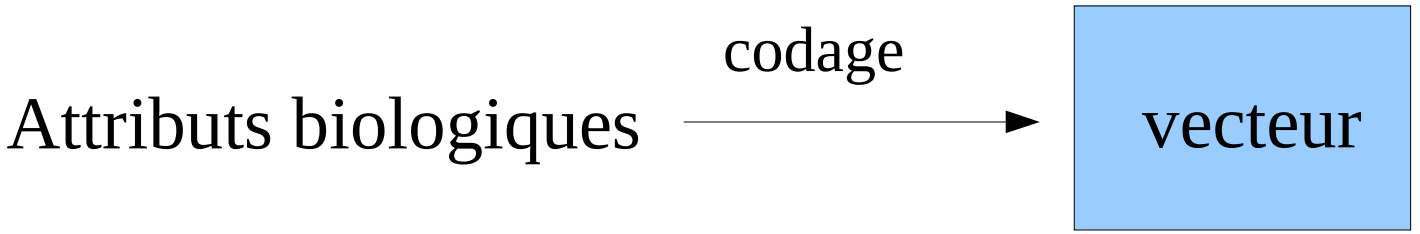
Calculs complexes dans un espace à grande dimension

Utilisation de fonctions noyaux (ou « kernels ») pour les rendre transparents

Fonction noyaux classiques :

- Linéaire : $K(x,y) = x.y$
- Polynomiale : $K(x,y) = (a.x.y+b)^{degré}$
- RBF : $K(x,y) = e^{-a\|x-y\|^2}$
- Sigmoïde : $K(x,y) = \tanh(a.x.y+b)$

Les classifieurs en biologie



Codages possibles :

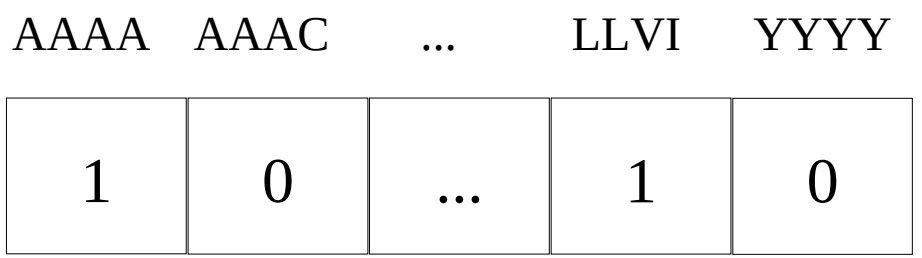
- Composition en sous-séquences
- Présence de motifs
- Scores de similarité
- Structures protéiques
- ...

Sur la composition en sous-séquences

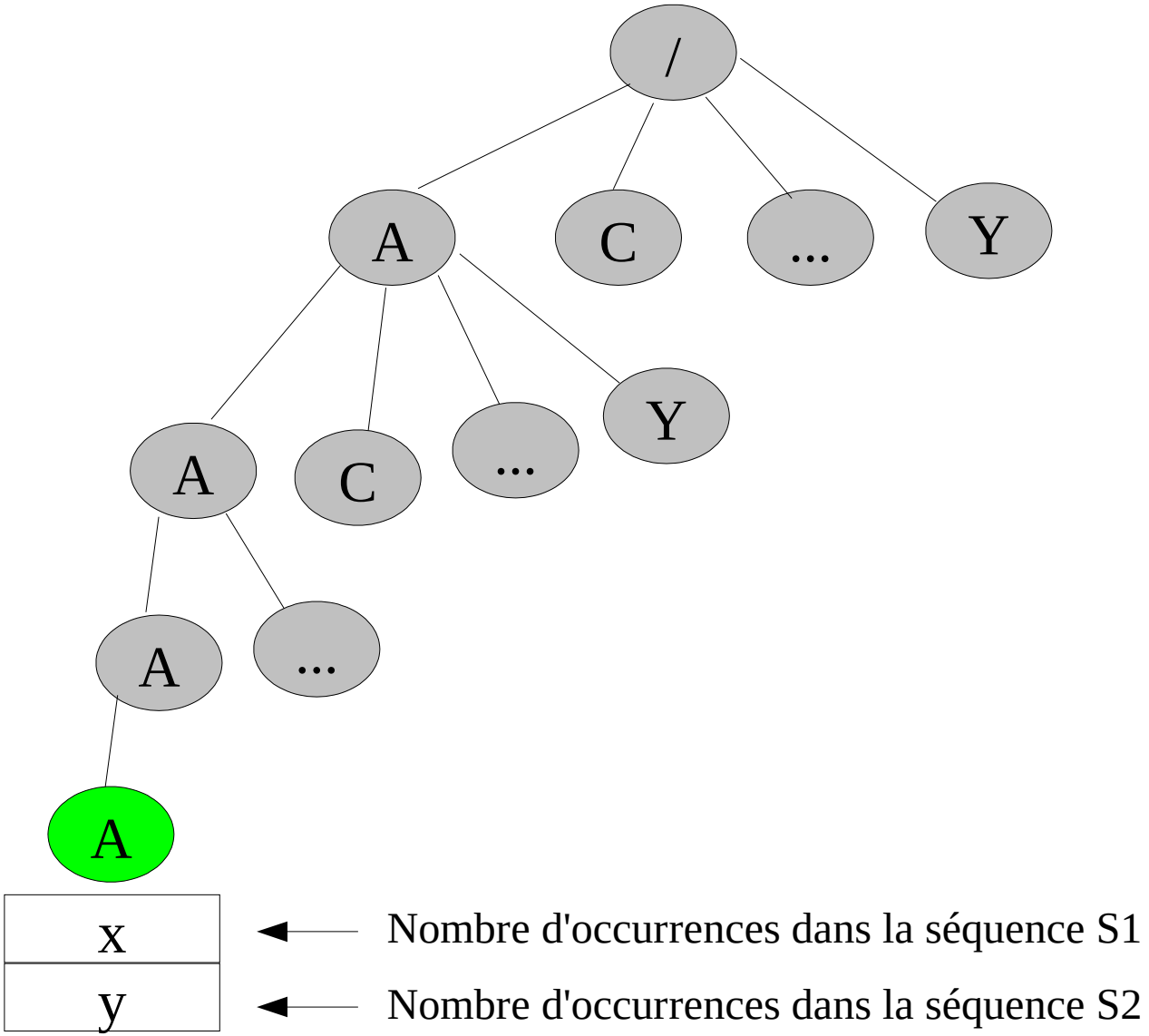
Séquence à vectoriser :

AAAAYGLLLLVITS

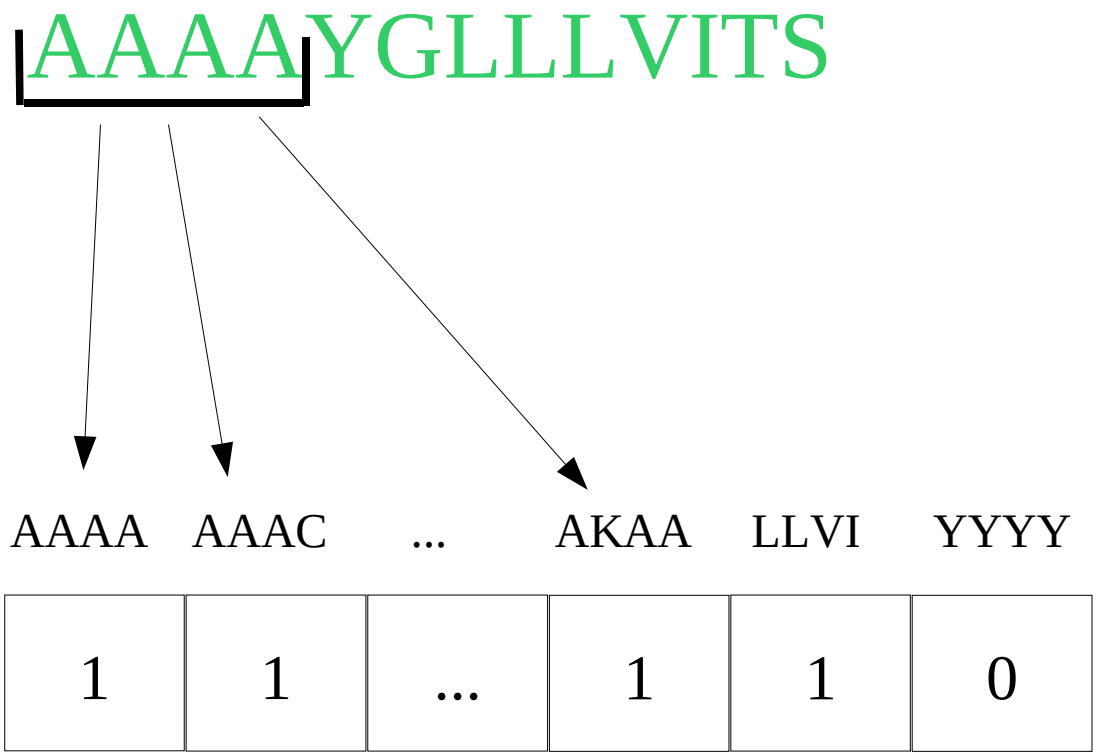
Vecteur :



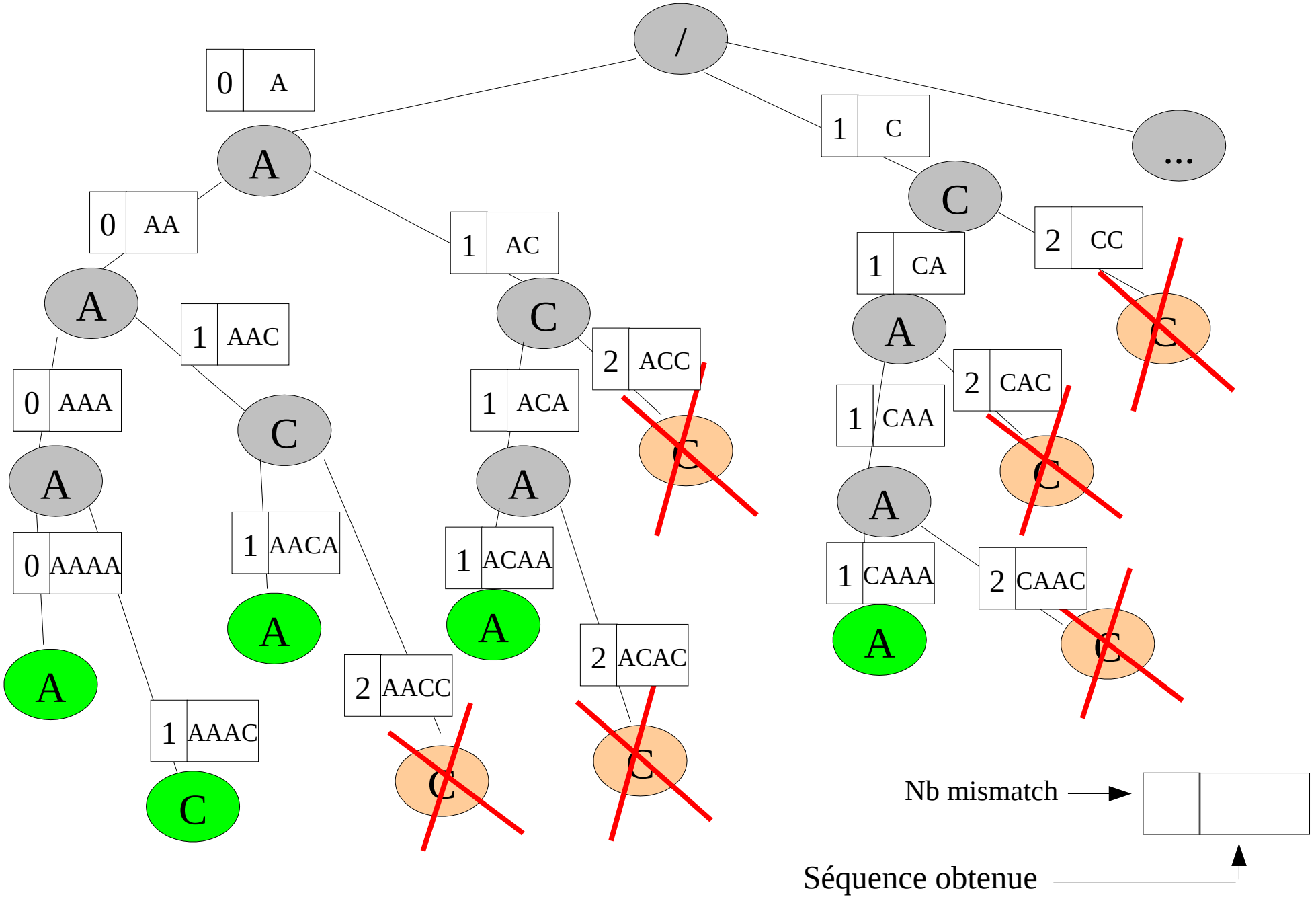
un arbre des suffixes permet de calculer rapidement les produits scalaires



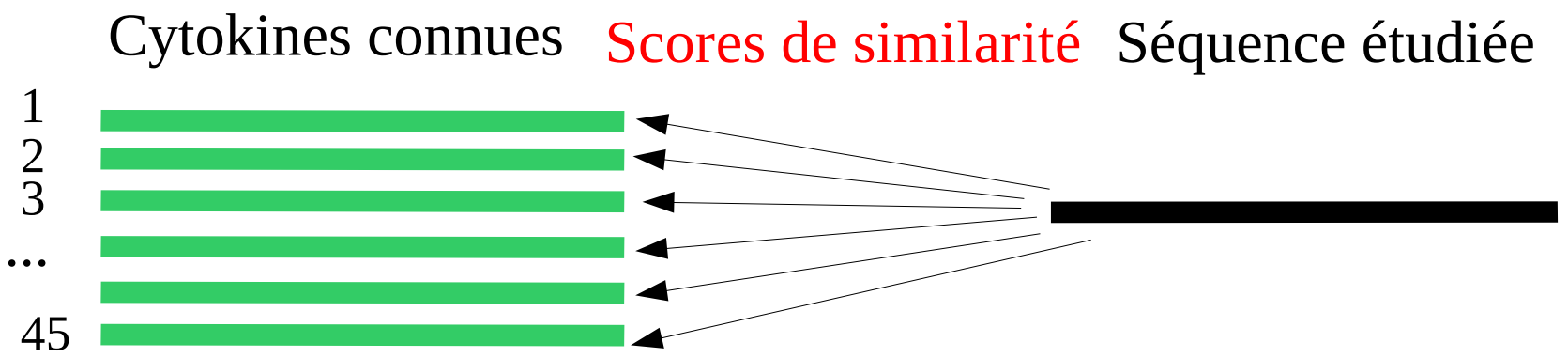
Sur la composition en sous-séquences avec mésappariement



Séquence lue : AAAA, mismatch autorisés : 1



Sur des scores de similarités



Score avec cytokine 1	Score avec cytokine 2	Score avec cytokine 3	...	Score avec cytokine 45
S1	S2	S3	...	S45

2 méthodes possibles pour vectoriser avec les scores de similarité :

- Pairwise : prendre directement le SW score

  ↳ Efficace empiriquement mais n'est pas un kernel théoriquement valide !!

- LA kernel : tenir compte de tout les alignements sous-optimaux possibles

  ↳ kernel théoriquement valide

Sur les motifs

Séquence

Motifs connus

Présence/absence	oui	non	oui	oui	non	non	oui
Vecteur	1	0	1	1	0	0	1

Propriété des motifs

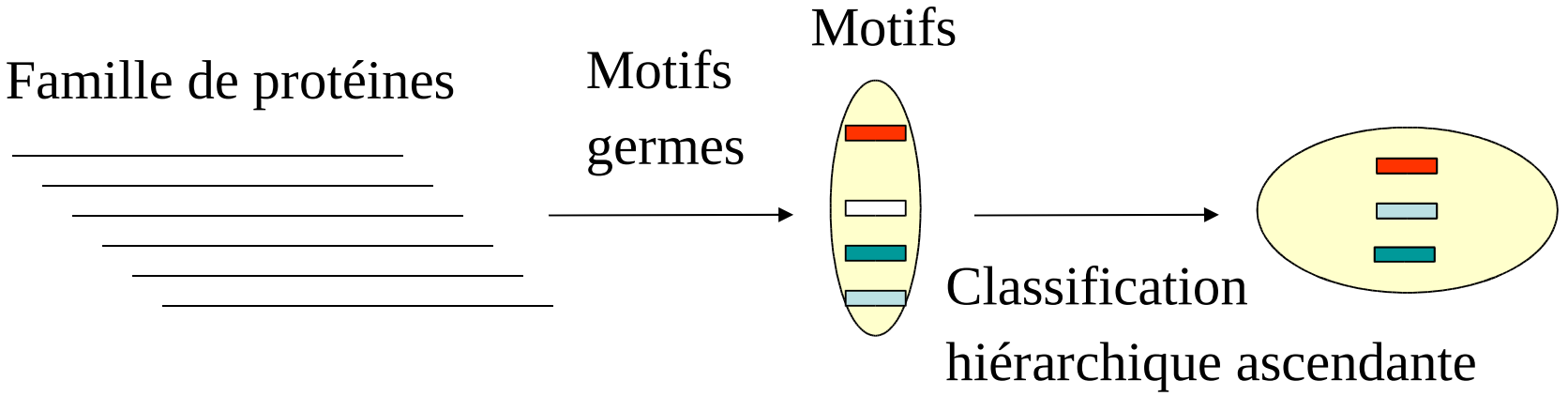
Support : Nombre de séquences vérifiant le motif

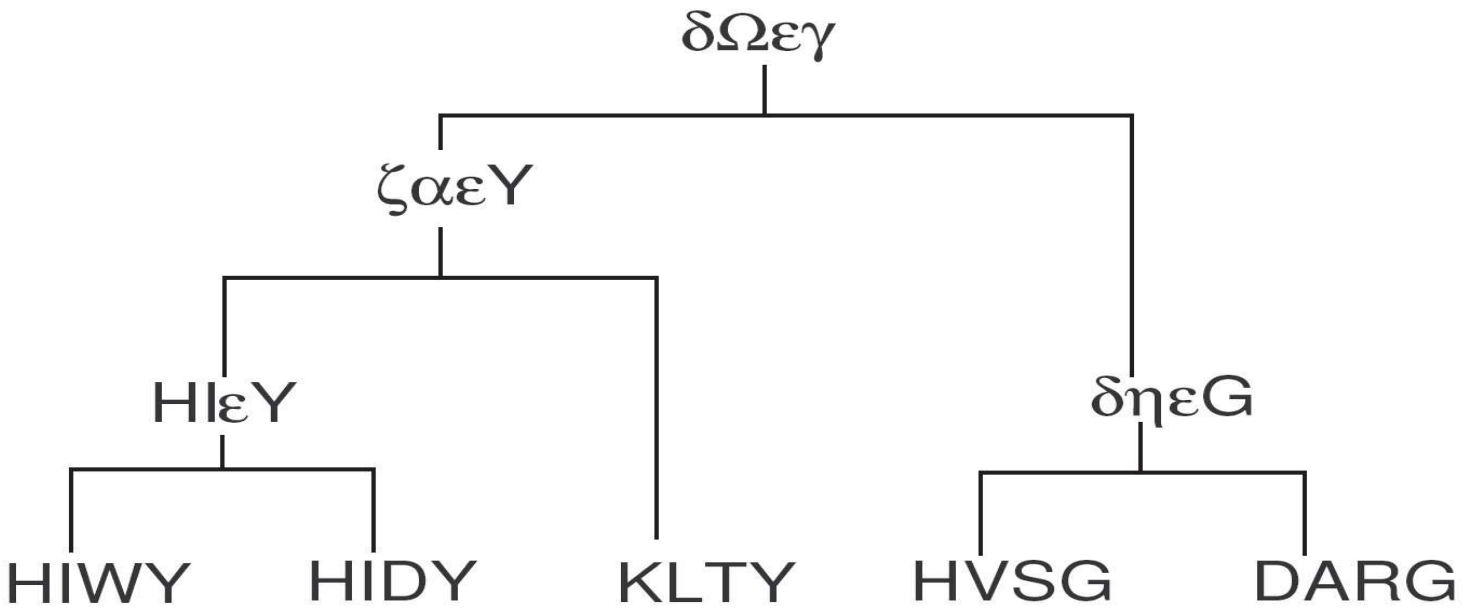
Spécificité : liée à la faible probabilité de trouver le k-motif au hasard dans une séquence.

Fonction de coût : $c(m)=\Pi_{i=1}^{k} f(m_i)$ Spécificité =-log(c(m))

motif	chaîne	support	spécificité
m^1	HIWY	1	14.25
m^2	HIDY	1	12.83
m^3	KLTY	1	11.44
m^4	HVSG	1	11.79
m^5	DARG	1	10.96
$m^6 = m^{1,2}$	HIεY	2	10.64
$m^7 = m^{3,6}$	$\zeta\alpha\varepsilon$Y	3	7.55
$m^8 = m^{4,5}$	$\delta\eta\varepsilon$G	2	5.26
$m^9 = m^{7,8}$	$\delta\Omega\varepsilon\gamma$	5	2.56

Trouver les motifs





Symbole	Classe	Membres
α	Aliphatique	ILV
β	Aromatique	FHWY
γ	Hydrophobe	ACFGHIKLMVWY
δ	Chargé	DEHKR
ε	Polaire	CDEHKNQRSTWY
ζ	Charge positive	HKR
η	Chaîne latérale courte	ACDGNPSTV
θ	Chaîne latérale très courte	ACGST

Evaluation des classifieurs

Validation croisée :

Validation croisée (en 10 lots)	***Vrai positifs***	***faux négatifs***	***Vrai négatifs***	***Faux positifs***	***efficacité***
Spectrum	79,0%	21,0%	84,0%	15,6%	81,9%
Mismatch	86,7%	13,3%	86,7%	13,3%	86,7%
Hmotifs	100,0%	0,0%	100,0%	0,0%	100,0%
Pairwise	100,0%	0,0%	97,8%	2,2%	98,9%
LAkernel	97,8%	2,2%	100,0%	0,0%	98,9%

Nature des données

Jeu d'apprentissage → Séquences protéique de cytokine et contre-exemples

Jeu de recherche → Séquences d'EST traduites en protéines

↳ Insertion dans les séquences = bruitage

Jeu d'apprentissage $\neq$ Jeu de recherche

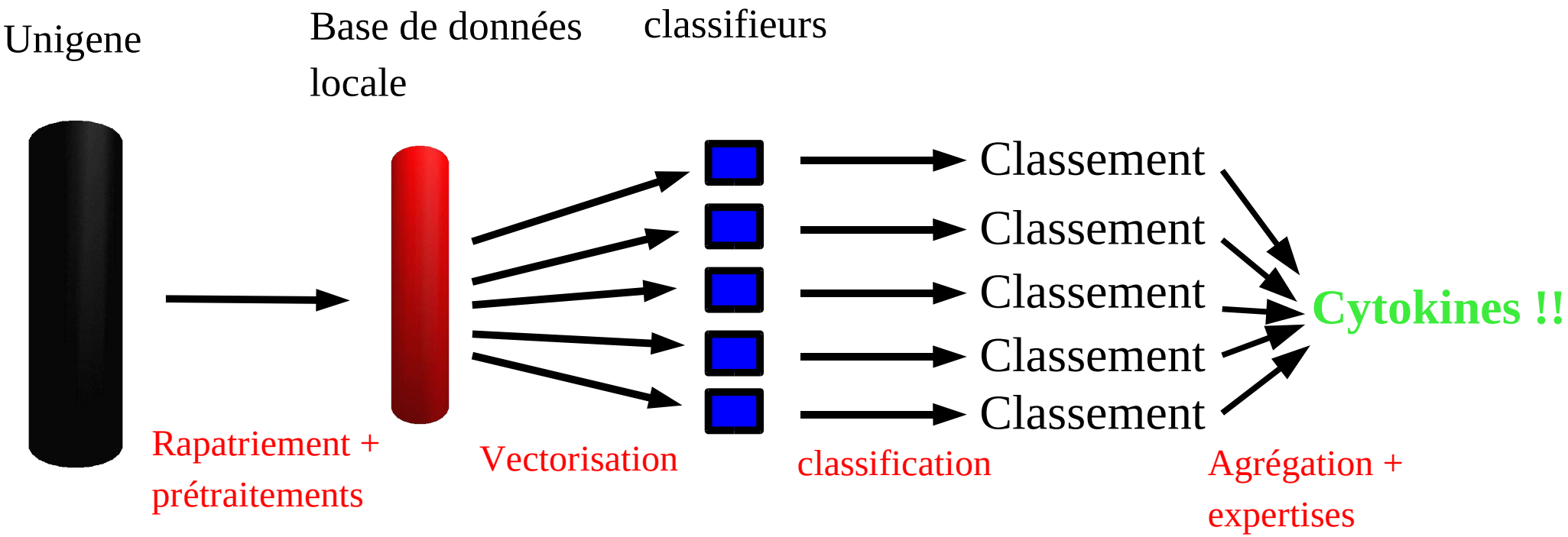
Nécessité de tester les classifieurs sur des données de même type que celles qui seront traitées.

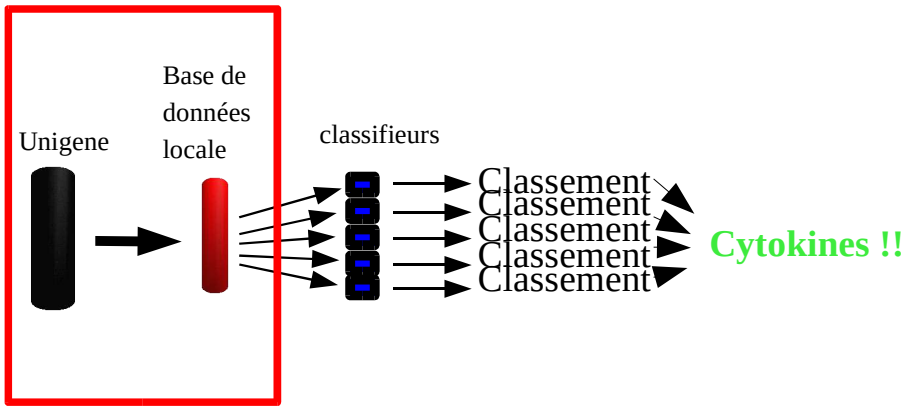
Sur des données Unigene (cytokines + contre-exemples) :

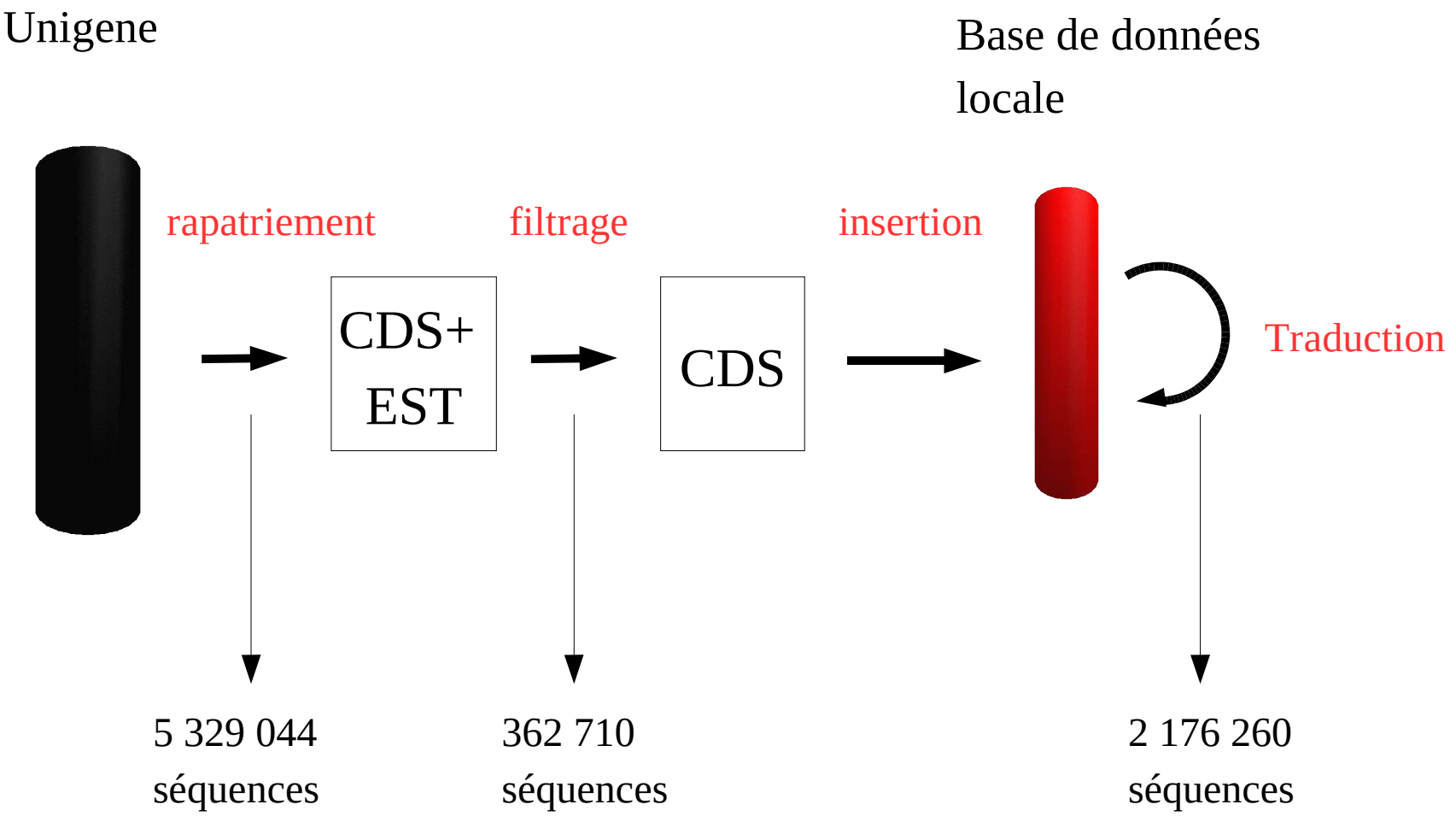
	ROC	***ROC50***	***Médian RFP***
Spectrum	0.98	0.85	0.019
Mismatch	0.98	0.81	0.022
HMotif	0.95	0.87	0.013
Pairwise	0.94	0.69	0.032
LAkernel	0.98	0.85	0.022

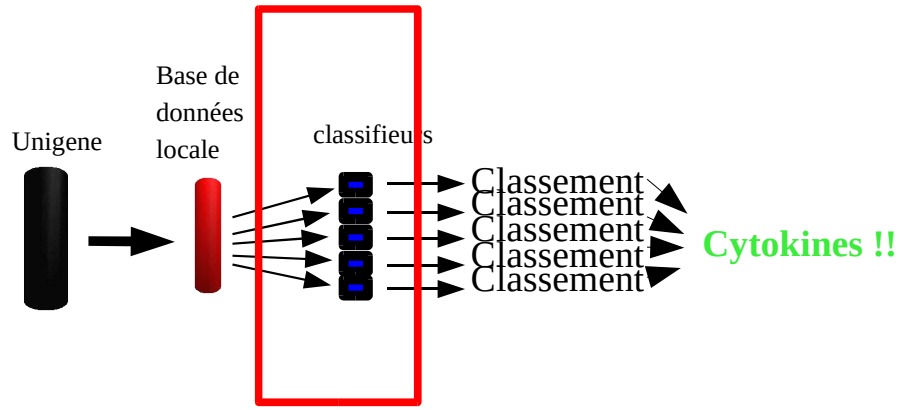
3ème partie :
PRHoD

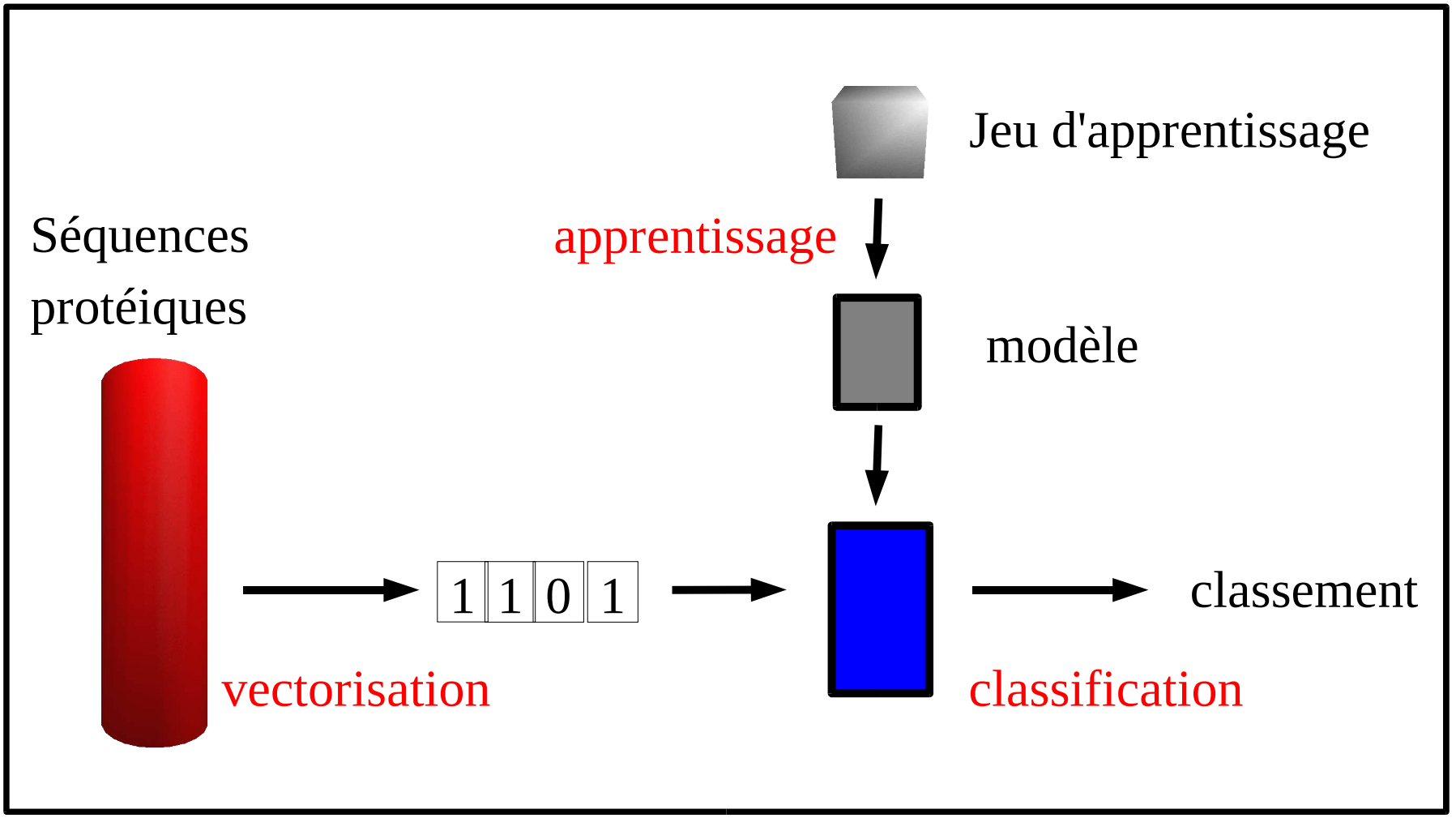
Fonctionnement général

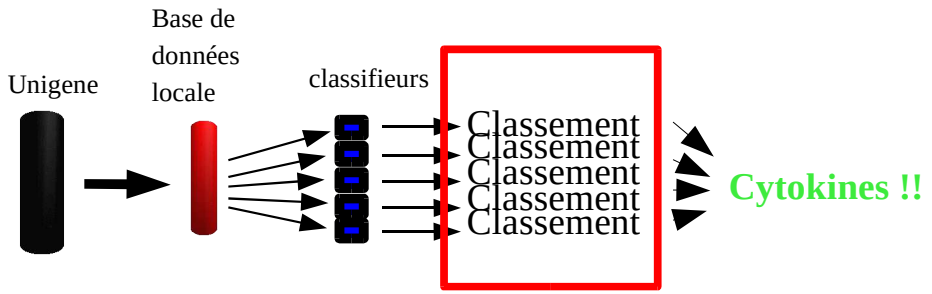



Unigene
Base de données locale
classifieurs
Classement
Classement
Classement
Classement
Classement
Cytokines !!


Unigene
Base de données locale
rapatriement
filtrage
insertion
CDS+ EST
CDS
Traduction
5 329 044 séquences
362 710 séquences
2 176 260 séquences


Unigene
Base de données locale
classifieurs
Classement
Classement
Classement
Classement
Classement
Cytokines !!


Jeu d'apprentissage
Séquences protéiques
apprentissage
modèle
1 1 0 1
vectorisation
classification
classement
X 5

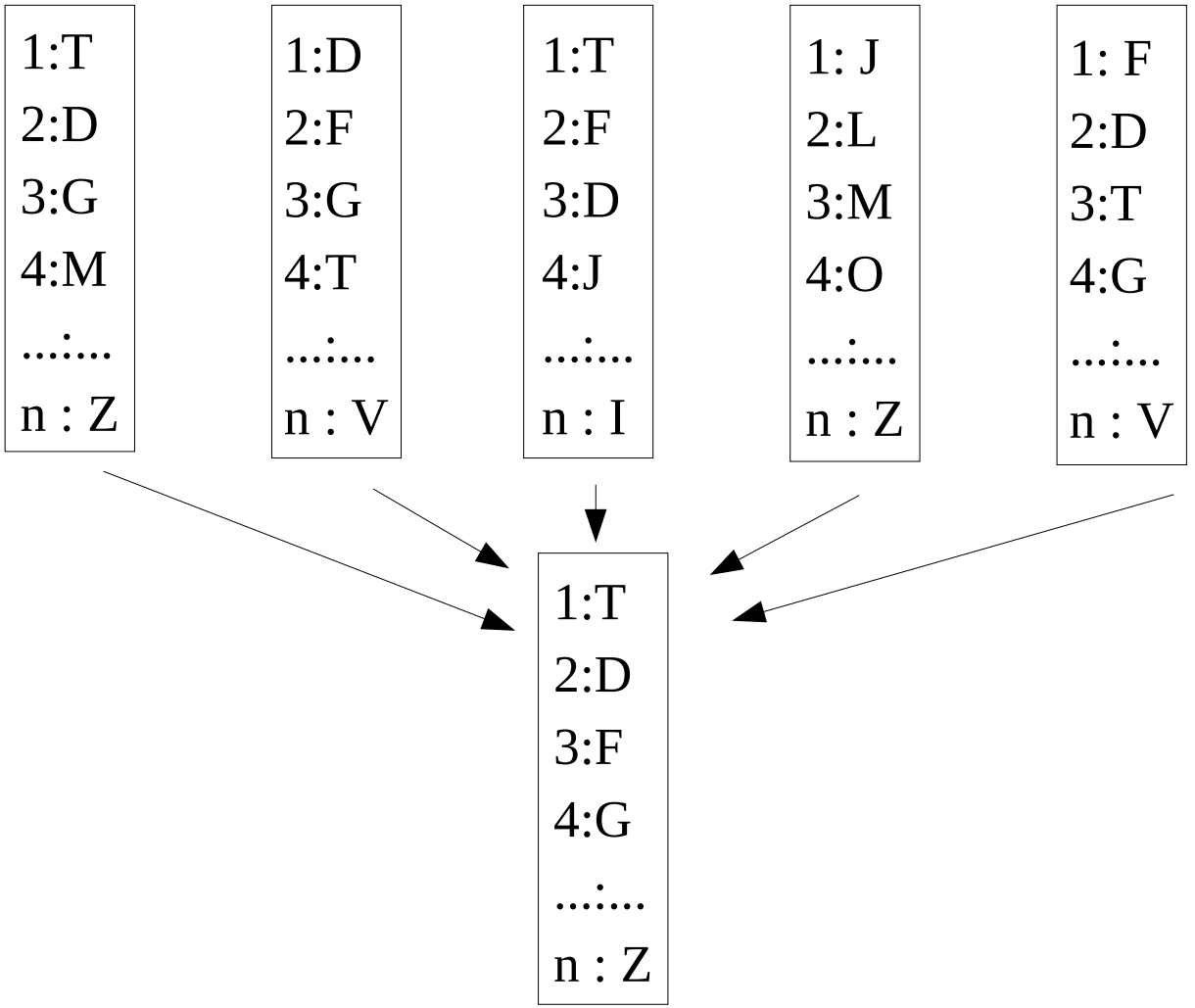


De la séquence qui ressemble le plus au jeu d'apprentissage à celle qui y ressemble le moins.

Classifieur 1	Classifieur 2	Classifieur 3	Classifieur 4	Classifieur 5
1er : seq T	1er : seq T	1er : seq V	1er : seq K	1er : seq Y
2éme : seq V	2éme : seq Z	2éme : seq T	2éme : seq T	2éme : seq F
3éme : seq E	3éme : seq E	3éme : seq Z	3éme : seq V	3éme : seq G
...	...	...	...	...
Nième : seq Y	Nième : seq Y	Nième : seq Y	Nième : seq H	Nième : seq O

4ème partie :
Post-traitements

Agréger les classements

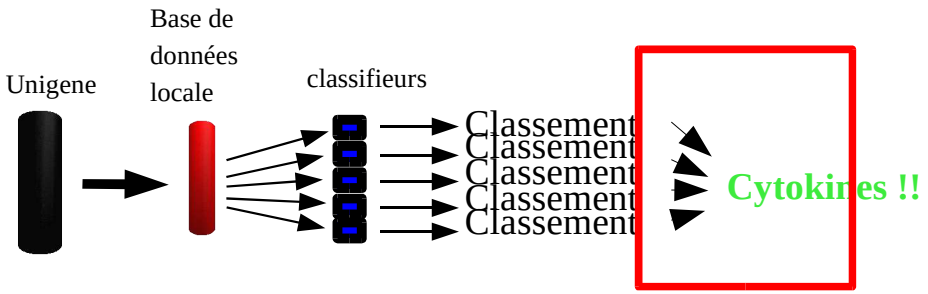


Problème de décision multi-critères !!

Pour y répondre de manière optimale :

- Pondérer les classifieurs selon leurs efficacités relatives
  - ➔ Evaluation sur une base de test
- Tenir compte de la nature des classifieurs
  - ➔ Pondération en fonction des liens entre classifieurs

Il existe des outils mathématiques tels que l'intégrale de Choquet qui permettent de résoudre ces problèmes



Nombreux candidats → Nécessité de filtrer

1) Enlever les cytokines connues

2) Regarder les candidats les plus intéressants

➔Position dans le classement

➔Expertise

Experts

But : Réunir une collection d'indices biologiques pour mettre en évidence les candidats les plus intéressants

Exemples d'experts :

- Taille de la séquence
- BLAST contre le jeu d'apprentissage
- Présence de ponts disulfures
- Ressemblance de la structure secondaire avec une structure de cytokine
- Ressemblance de la structure du gène avec celle d'un gène de cytokine
- Position dans le génome (en cluster avec des cytokines ?)
- Phylogénie du candidat par rapport aux cytokines
- ...

Conclusion

- Recherche de nouveaux membres d'une famille de gènes vaste et diversifiée
- Utilisation des SVM pour classer les séquences d'Unigene
- Mise en place d'un outil gérant 5 classiffieurs différents
- Agrégation des résultats des classifieurs
- Collection d'experts pour compléter l'analyse des candidats