CLASSIFICATION DES SÉQUENCES GÉNOMIQUES VIRALES PAR UNE APPROCHE D'APPRENTISSAGE AUTOMATIQUE

Mohamed Amine Remita

Université du Québec à Montréal

Fonds de recherche Nature et technologies Québec

CRSNG NSERC

BIF7002 – Hiver2017

M.A. Remita, A. Halioui, A.A.M. Diouara, B. Daigle, G. Kiani et A. B. Diallo. *A machine learning approach for viral genome classification*. BMC Bioinformatics (sous presse)

Plan

UQÀM

- Motivation et problématique
- Apprentissage automatique
- Méthode
  - Ensembles de données, attributs et algorithmes
- Résultats
- CASTOR – plateforme web
- Conclusion et perspectives

Introduction

Motivation - problématique

- Étant donné une séquence génomique (partielle ou complète) d'un virus nouvellement séquencée ou extraite à partir d'une base de données :

- peut-on identifier le type du virus ?
  - En utilisant ses caractéristiques génomiques
    - Sans refaire un alignement de séquences ni une phylogénie

Motivation - problématique

- Classifications - typages
  - **Classification taxonomique**
  - Classification géographique
  - Classification par pouvoir pathogène

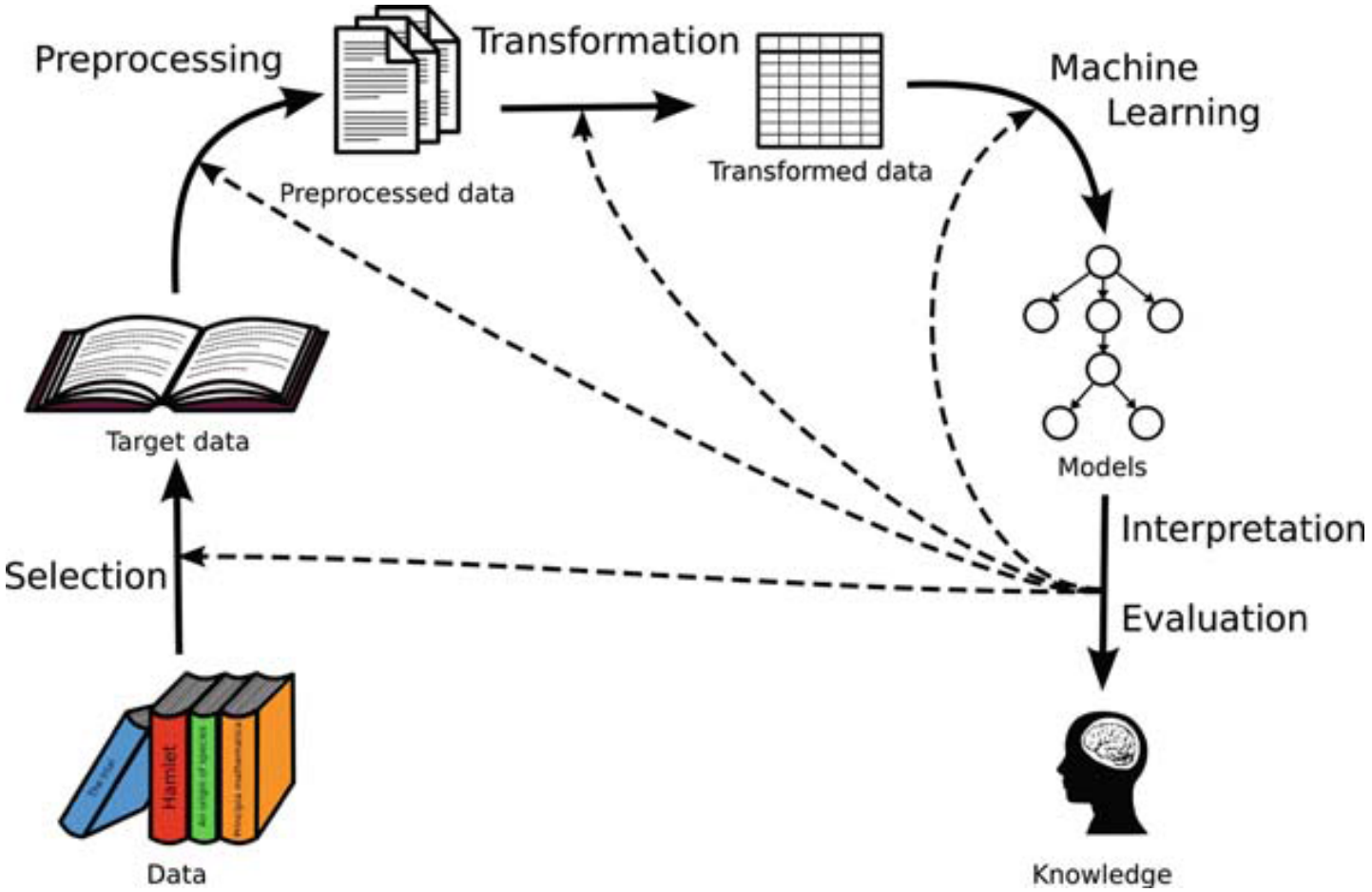
Motivation - travaux associés

- Méthodes basées sur l'alignement de séquences
  - NCBI Pairwaise Sequence Comparison (PASC) (Bao *et al.* 2015)
  - Diversity partitioning by hierchical clustering (DEmARC) (Lauber et Gorbalenya 2012)
  - BLAST (Altschul et al. 1997)
- Méthodes basées sur une phylogénie
  - REGA (Alcantara et al., 2009; de Oliveira et al., 2005)
  - Pplcer (Matsen et al. 2010)
- Méthodes indépendantes de l'alignement de séquences
  - COMET (Struck et al. 2014)
  - Natural vector based on the distributions of nucleotides (Deng *et al.* 2011)

Motivation

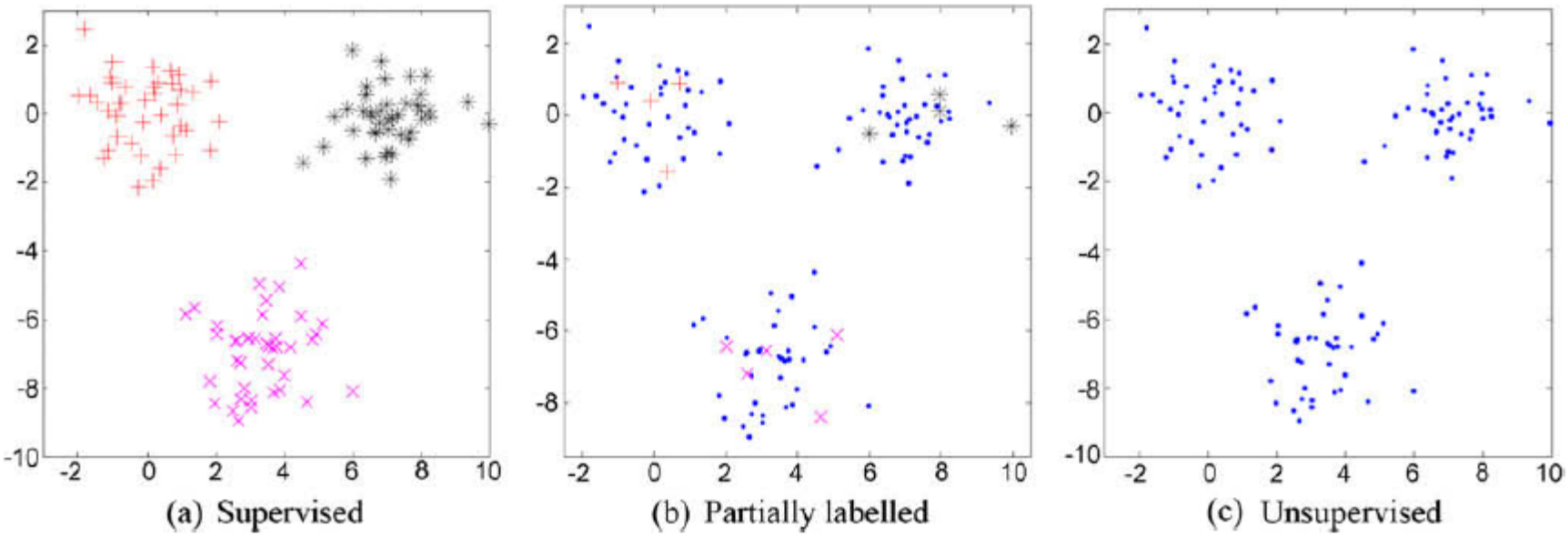
- Application
  - Efficace et rapide
  - Automatique
  - Réutilisable et reproductible
  - Accessibilité

Apprentissage automatique



The general chain of work of a common data mining task (Inaki Inza et al. 2010)

Apprentissage automatique



Jain A.K (2010)

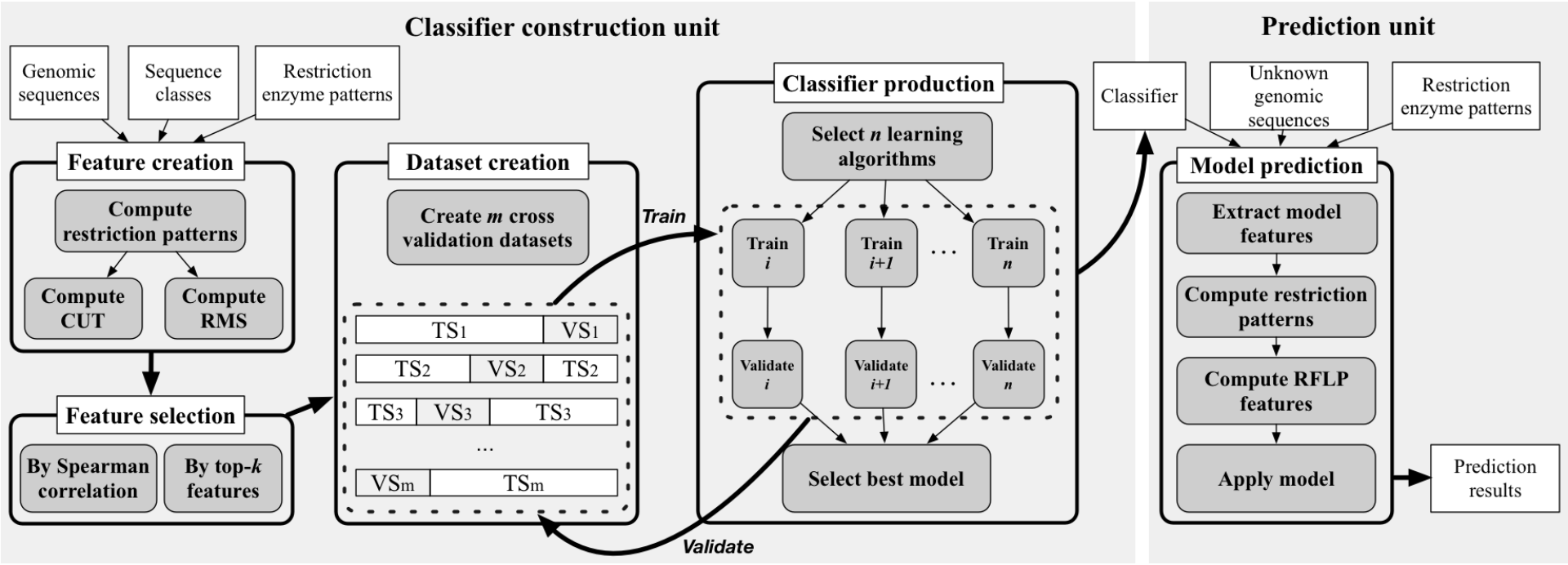
- Apprentissage supervisé (Classification et régression)
  - Arbre de décision, SVM, KNN, bayésiens, réseaux de neurone etc.
- Apprentissage non supervisé (Clustering)
  - Partitionnements K-means
  - Partitionnements hiérarchiques, basés sur la densité, des graphes etc.

Méthode

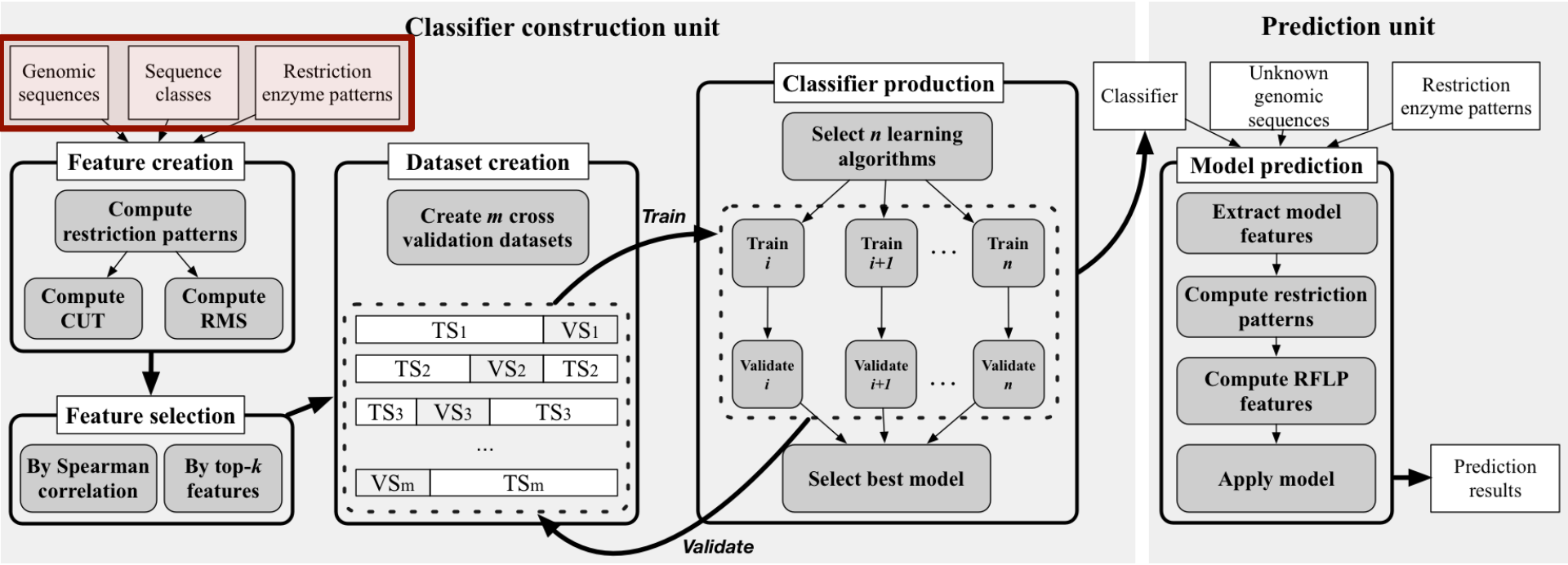
Les étapes de la classification

1. Construire des **jeux de données** représentatifs
2. Déterminer un ensemble **d'attributs** (features) pertinents et non redondants
3. Identifier des **algorithmes d'apprentissage** adéquats et performants

Approche de classification



1. Jeu de données



1. Jeu de données

- Virus du papillome humain (VPH)
  - ADN double brin circulaire
  - ~ 8000 nt
- Virus de l'hépatite B (VHB)
  - ADN circulaire
  - Partiellement double brin
  - ~ 3200 nt
- Virus de l'immunodéficience humaine type 1 (VIH-1)
  - ARN simple brin en double exemplaire
  - ~ 9700 nt

Smith B et *al.* 2011

1. Jeu de données

- Classification inter-genres des VPHs

Classe (genre)	nombre
Alphapapillomavirus	457
Bétapapillomavirus	48
Gammapapillomavirus	27
Mupapillomavirus	02
Nupapillomavirus	01

- Classification inter-espèces des VPHs

Classe (espèce)	nombre	Classe (espèce)	nombre	Classe (espèce)	nombre	Classe (espèce)	nombre
AlphaPV 1	03	AlphaPV 5	04	AlphaPV 9	249	AlphaPV 14	07
AlphaPV 2	11	AlphaPV 6	34	AlphaPV 10	72	-	-
AlphaPV 3	11	AlphaPV 7	49	AlphaPV 11	02	-	-
AlphaPV 4	09	AlphaPV 8	04	AlphaPV 13	02	-	-

1. Jeu de données

- Classification inter-génotype des VHBs

Classe (Génotype)	nombre
A	378
B	540
C	1085
D	825
E	228
F	129
G	29
H	21

1. Jeu de données

- Classification des sous-types M du VIH-1

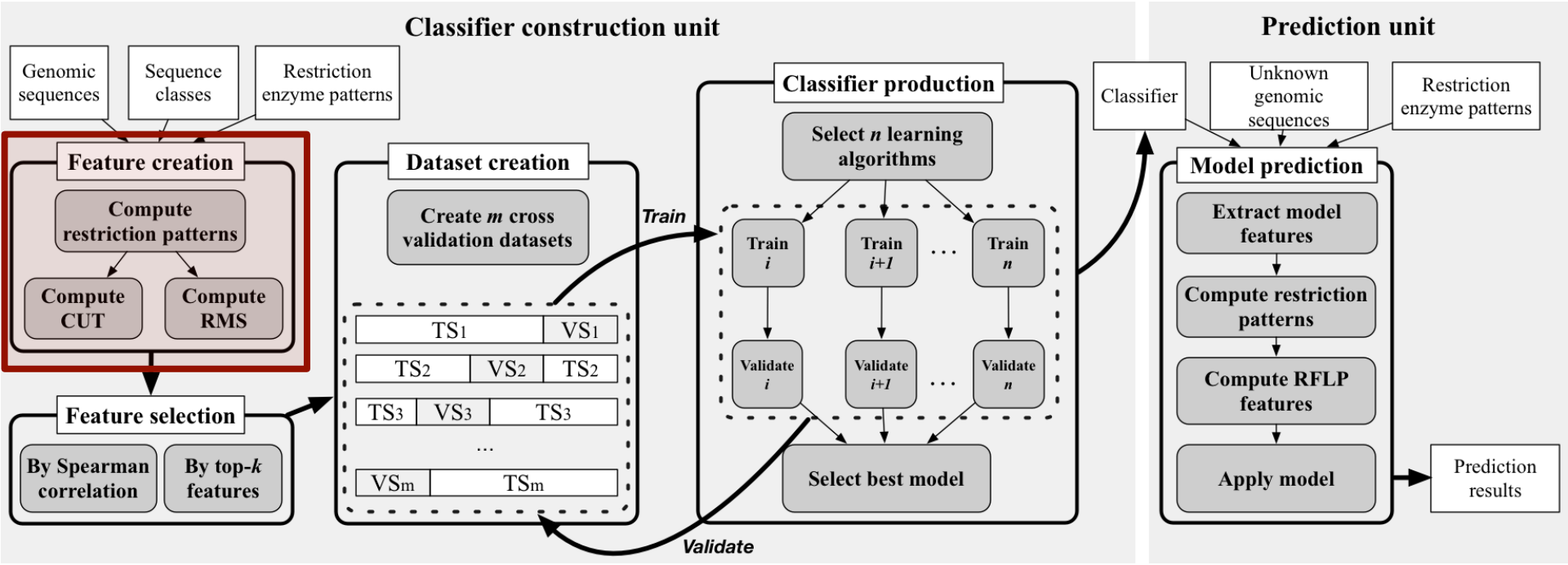
	Classe (sous-type)	nombre
PURS	A	314
	B	2113
	C	1065
	D	76
	F	55
	G	53
	H	4
	J	3
	K	2

	Classe (sous-type)	nombre
CRFs	01_AE	712
	02_AG	93
	07_BC	44
	08_BC	36
	35_AD	22
	42_BF	17
	22_01A1	16
	11_cpx	14
	14_BG	13

1. Jeu de données

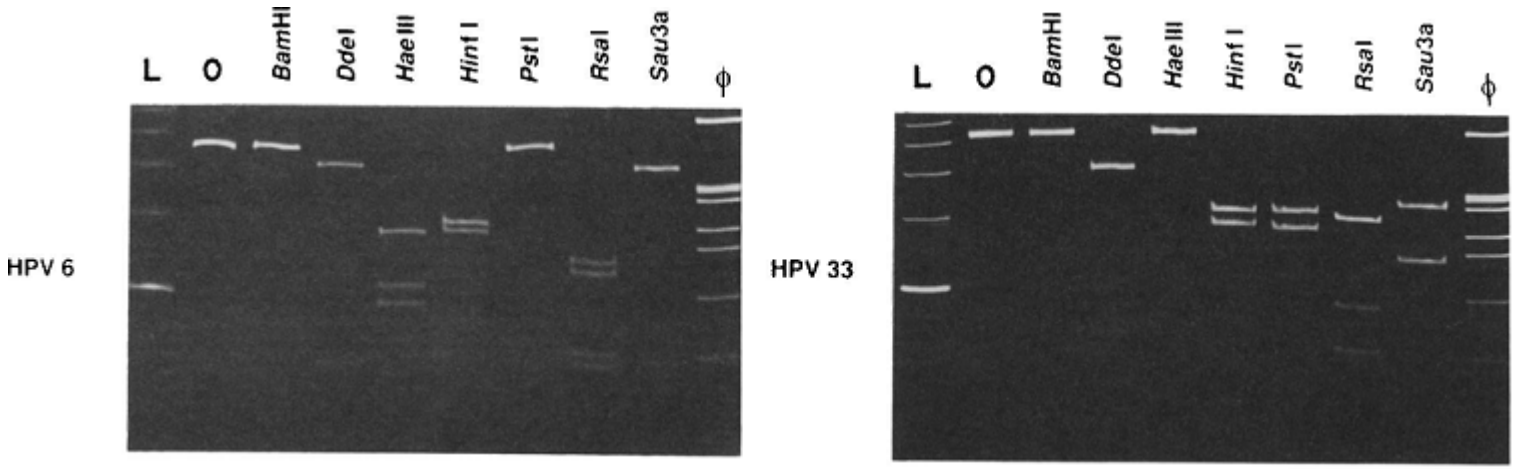
- Source des données
  - *NCBI Taxonomy*
  - *NCBI Nucleotide*
  - *NCBI RefSeq*
  - *Papillomavirus Episteme (PaVE)*
  - *Los Alamos HIV databases*

2. Attributs


Classifier construction unit
Genomic sequences
Sequence classes
Restriction enzyme patterns
Feature creation
Compute restriction patterns
Compute CUT
Compute RMS
Feature selection
By Spearman correlation
By top-k features
Dataset creation
Create m cross validation datasets
TS1
VS1
TS2
VS2
TS2
TS3
VS3
TS3
...
VSm
TSm
Train
Classifier production
Select n learning algorithms
Train i
Train i+1
Train n
Validate i
Validate i+1
Validate n
Select best model
Validate
Prediction unit
Classifier
Unknown genomic sequences
Restriction enzyme patterns
Model prediction
Extract model features
Compute restriction patterns
Compute RFLP features
Apply model
Prediction results

2. Attributs - RFLP

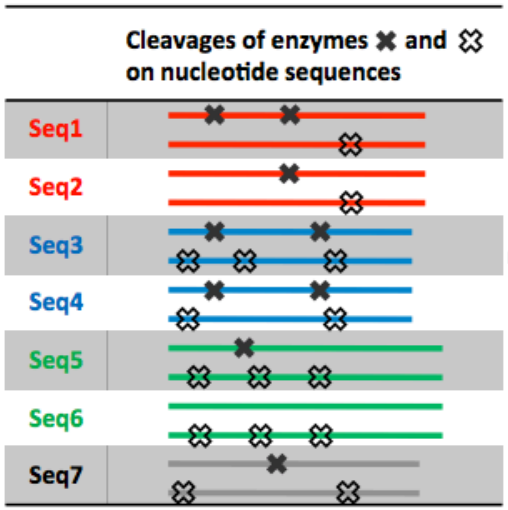
- Polymorphisme de longueur des fragments de restriction (RFLP)
- Technique de biologie moléculaire
- Coupure de l'ADN par des enzymes de restrictions
- Empreinte génétique



Motifs RFLP d'un segment de la séquence du L1 pour l'identification des VPH génitaux (Bernard, et al., 1994).

2. Attributs - RFLP

- Technique bioinformatique
- Restriction Enzyme dataBASE (REBASE)
  - **172** prototypes d'enzymes de type II



Extraction

	Fragments of enzyme ✖	Fragments of enzyme ⛒
Seq1	660, 780, 1560	2280, 720
Seq2	1440, 1560	2280, 720
Seq3	660, 1754, 386	327, 493, 1678, 302
Seq4	660, 1754, 386	327, 2171, 302
Seq5	820, 2480	537, 680, 679, 1304
Seq6	3300	537, 680, 679, 1304
Seq7	1340, 1640	273, 2107, 600

Transformation

	Enzyme ✖	Enzyme ⛒	CLASS
Seq1	1076.7	1690.7	RED
Seq2	1501.2	1690.7	RED
Seq3	1104.7	902.3	BLUE
Seq4	1104.7	1279.5	BLUE
Seq5	1847	853.3	GREEN
Seq6	3300	853.3	GREEN
Seq7	1497.5	1274.6	?

2. Attributs - métriques

- Attributs numériques
- Pour chaque couple virus – enzyme, on calcule
  - CUT : nombre de coupures (attributs entiers)
  - RMS : la moyenne quadratique des longueurs des fragments (attributs réels)

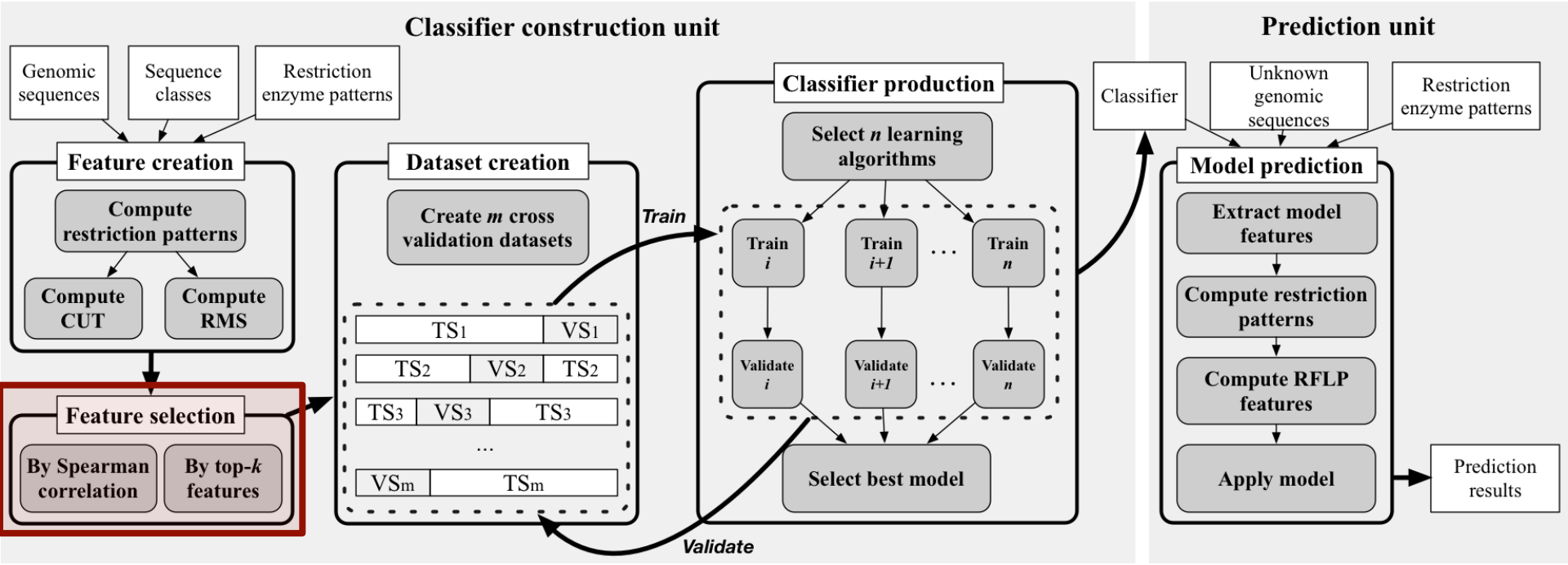
$$\overline{x} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

2. Attributs

- Exemple de données d'apprentissage

ID	CUT_AccI	RMS_AccI	CUT_AcII	RMS_AcII	CUT_AcyI	RMS_AcyI	CUT_AfIII	RMS_AfIII	Class
X74483	5	1977,92	1	7844,00	3	2650,41	0	7844,00	**PSV_PV**
X77858	4	2645,70	3	3269,98	2	5423,81	2	3998,11	**PSV_PV**
X94164	5	2141,91	0	7988,00	7	1504,92	1	7988,00	**PSV_PV**
X94165	6	1866,19	4	2267,66	1	7700,00	1	7700,00	**PSV_PV**
Y15173	5	1748,69	2	4818,36	1	7537,00	0	7537,00	**PSV_PV**
Y15174	4	2182,74	1	7549,00	1	7549,00	1	7549,00	**PSV_PV**
Y15175	8	1227,69	2	3947,36	8	1456,80	2	5264,39	**PSV_PV**
NC_000852	254	1830,14	178	2718,91	177	2824,55	51	9170,54	**NGV_PV**
NC_000866	59	3783,18	64	3840,31	32	6939,67	27	7755,87	**NGV_PV**
NC_000867	1	10079,00	7	1949,85	7	1524,67	2	6809,43	**NGV_PV**
NC_000871	19	2230,87	16	2759,61	3	10656,71	4	10107,46	**NGV_PV**
NC_000872	22	2239,47	20	2822,73	6	8522,07	6	7608,48	**NGV_PV**
NC_000896	14	3942,09	15	3928,77	9	6618,34	12	4500,37	**NGV_PV**
NC_000898	101	2269,26	53	4339,03	124	2612,96	15	15151,01	**NGV_PV**

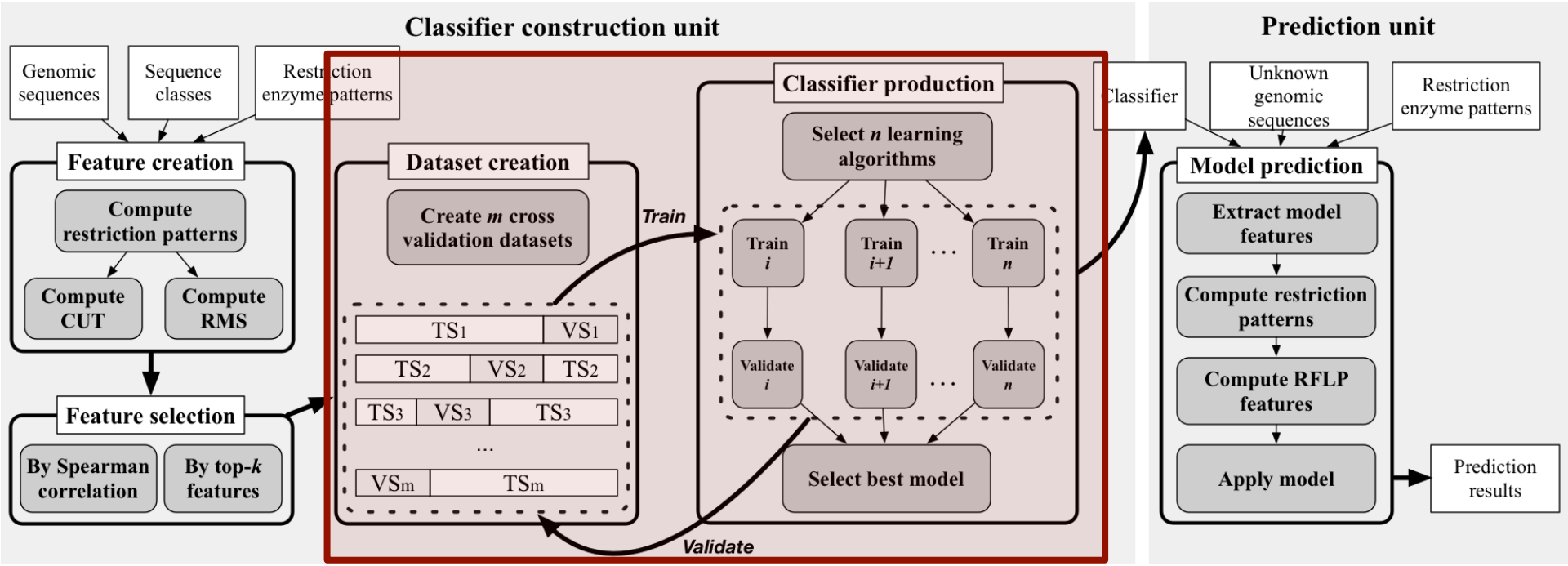
2. Attributs - Sélection



2. Attributs - Sélection

- *Pertinence : gain d'information*
  - *Information mutuelle entre un attribut et la classe*
  - *Information gain evaluator (avec Ranker search method)*
  - *Top-k*
- *Redondance : corrélation*
  - *Deux attributs corrélés sont redondants*
  - *Corrélation de Spearman (rho)*

2. Attributs - Sélection



3. Algorithmes d'apprentissage

Algorithms	Weka modules
Arbres de décision	weka.classifiers.trees.J48
Random Forest	weka.classifiers.trees.RandomForest
Machines à vecteurs de support (SVM)	weka.classifiers.functions.LibSVM
KNN	weka.classifiers.lazy.IBk
Bagging	weka.classifiers.meta.Bagging
AdaBoost	weka.classifiers.meta.AdaBoostM1
Naive Bayes	weka.classifiers.bayes.NaiveBayes

3. Algorithmes d'apprentissage

- La classification et l'évaluation sont effectuées avec la plateforme Weka (Waikato Environment for Knowledge Analysis)

- Les entrainements des modèles sont réalisés avec une validation croisée de 10 itérations

4. Évaluation – Cohésion des classes

- *Compacité (cohésion interne) : les objets appartenant à un cluster sont les plus similaires*
- *Séparabilité (isolation externe) : les objets appartenant aux autres clusters sont les plus distincts*

- *Indice de* Silhouette (Rousseeuw 1987) : indice $\in [-1, 1]$
- *Indice de cohésion* (Daigle et al. 2015) : indice $\in [0, 1]$

4. Évaluation – modèles d'apprentissage

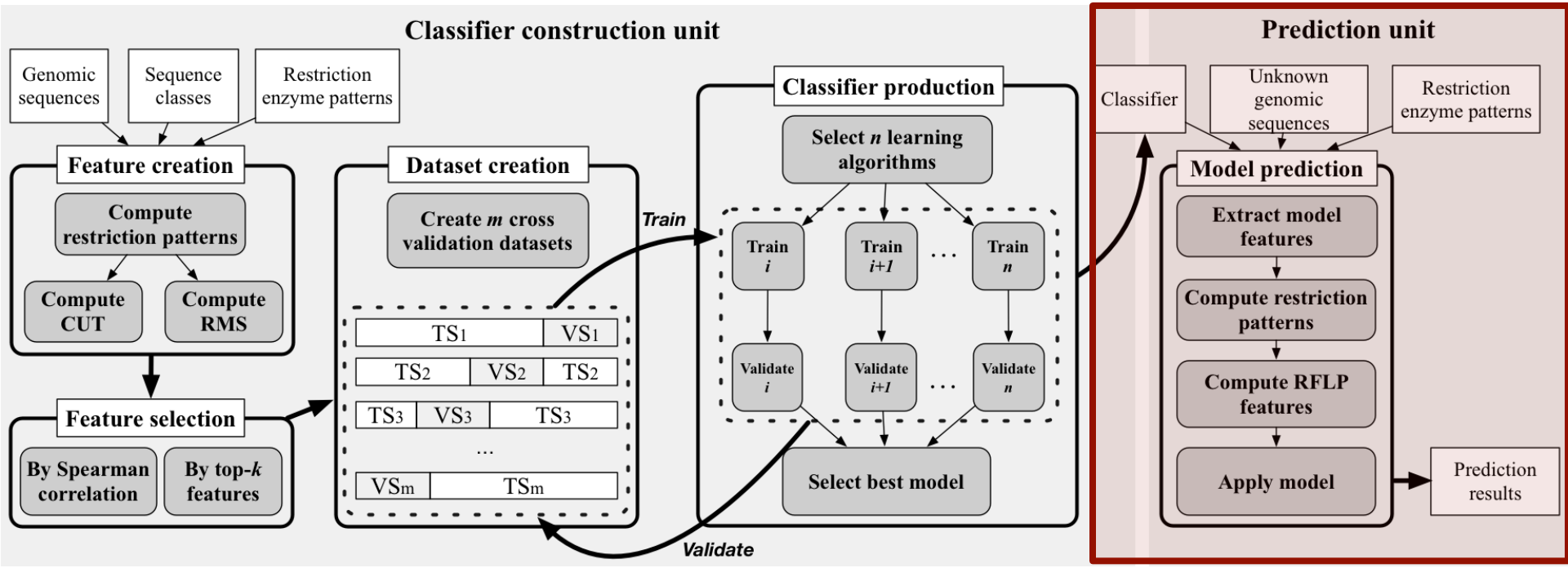
		CLASSE RÉELLE	
CLASSE PRÉDITE		**Condition positive**	**Condition négative**
	Condition positive	Vrais positifs (TP)	Faux positifs (FP)
	Condition négative	Faux négatifs (FN)	Vrais négatifs (TN)

Mesure	Formule
Taux de vrais positifs (rappel, sensibilité)	$TPR = R = \frac{TP}{TP + FN}$
Taux de faux positifs (FPR, 1 - spécificité)	$FPR = \frac{FP}{FP + TN}$
Précision	$P = \frac{TP}{TP + FP}$
F-mesure	$f - measure = \frac{2P \times R}{P + R}$

5. Simulations

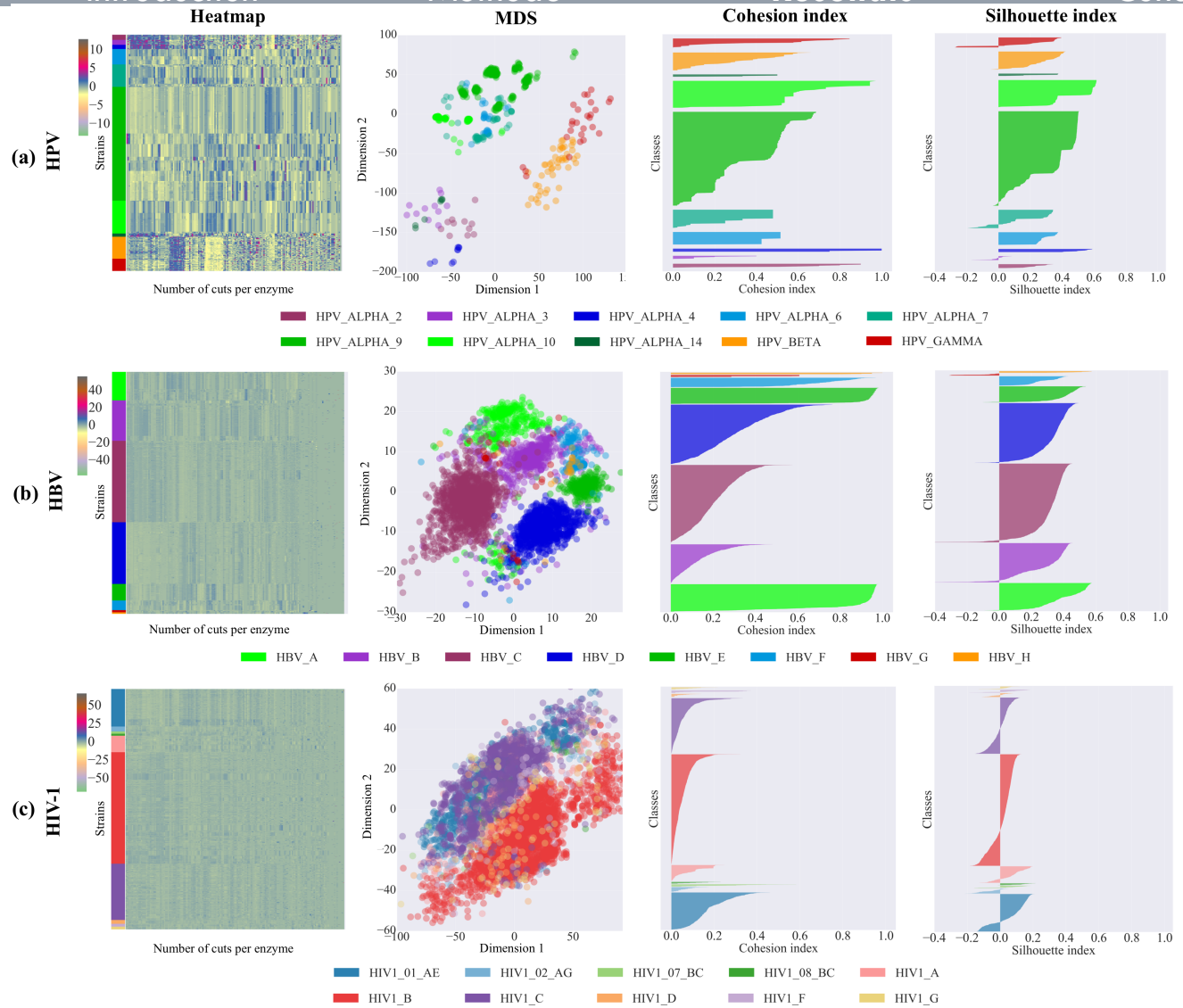
- *Cinq jeux de données principaux de génomes viraux*
- *Pour chaque jeu de données :*
  - *Générer 10 échantillons (sans remise)*
  - *Pour chaque échantillon :*
    - *Construire des modèles avec validation croisée en combinant :*
      - *2 métriques d'attributs (CUT et RMS)*
      - *2 méthodes de sélection d'attributs (topAttributes et correlation)*
      - *7 algorithmes d'apprentissage (J48, SVM, ADA, etc.)*
    - *Au total 280 modèles*

Prédiction

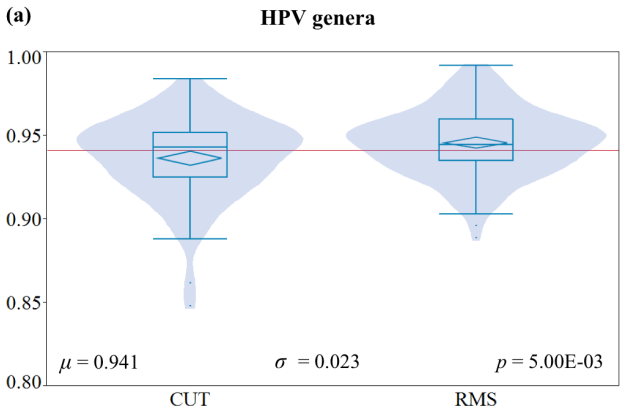

Classifier construction unit
Genomic sequences
Sequence classes
Restriction enzyme patterns
Feature creation
Compute restriction patterns
Compute CUT
Compute RMS
Feature selection
By Spearman correlation
By top-k features
Dataset creation
Create m cross validation datasets
TS1
VS1
TS2
VS2
TS2
TS3
VS3
TS3
...
VSm
TSm
Train
Classifier production
Select n learning algorithms
Train i
Train i+1
Train n
Validate i
Validate i+1
Validate n
Select best model
Validate
Prediction unit
Classifier
Unknown genomic sequences
Restriction enzyme patterns
Model prediction
Extract model features
Compute restriction patterns
Compute RFLP features
Apply model
Prediction results

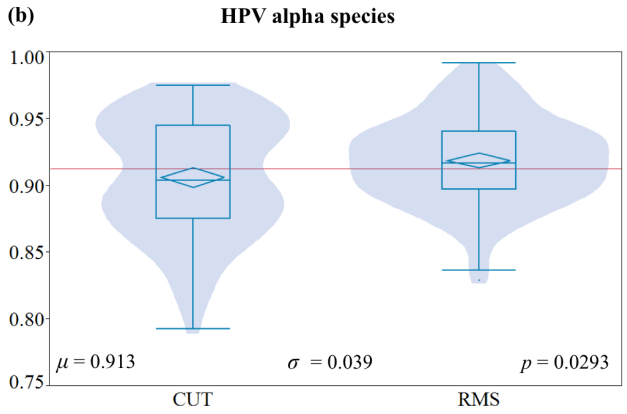
Résultats

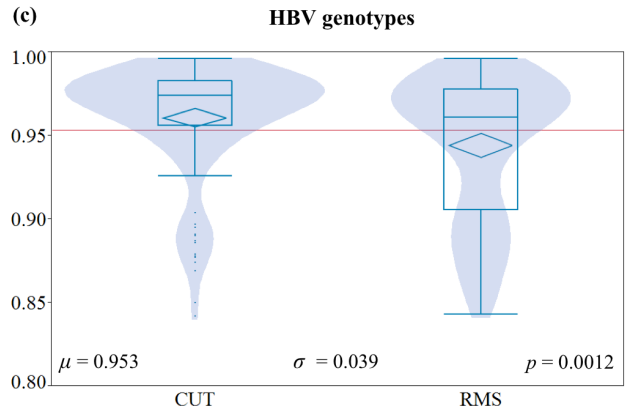
Cohésion des classes

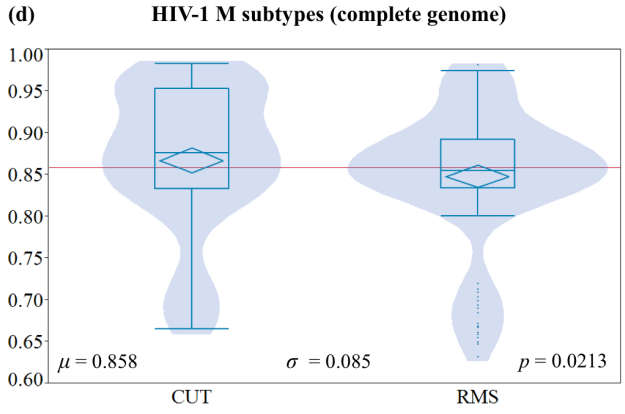


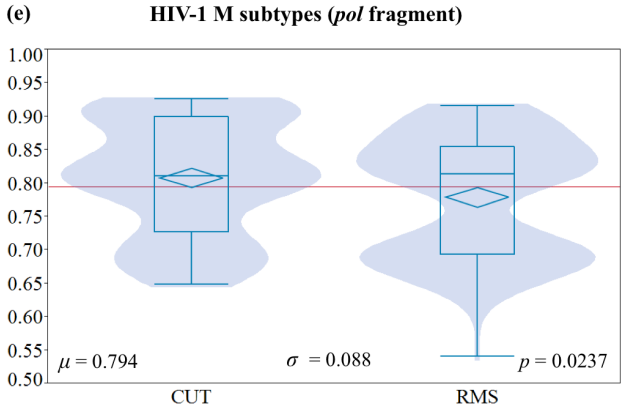
Simulation - comparaison des métriques



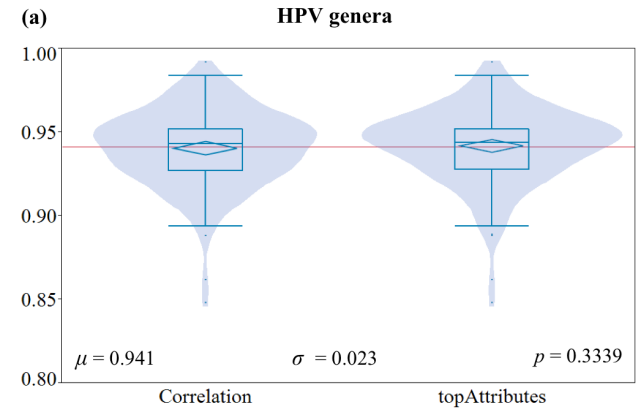


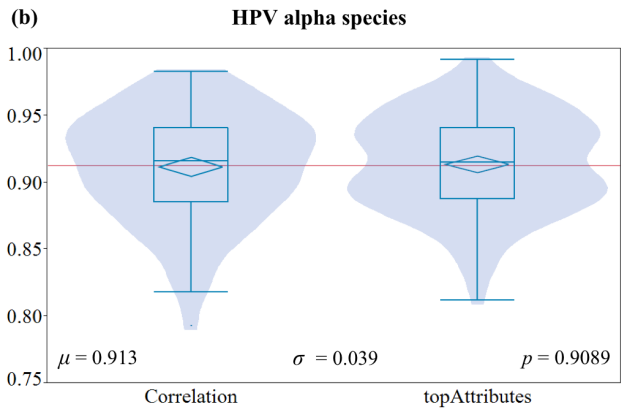


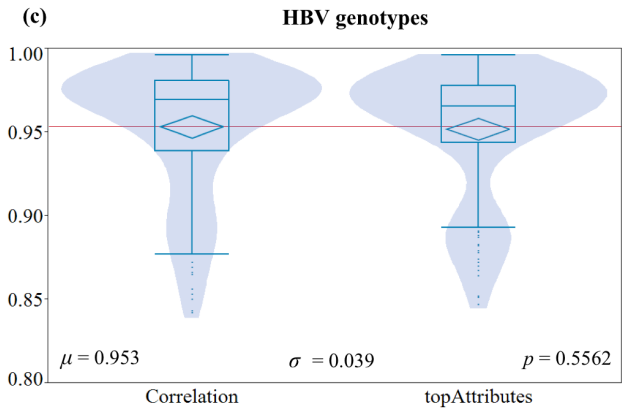


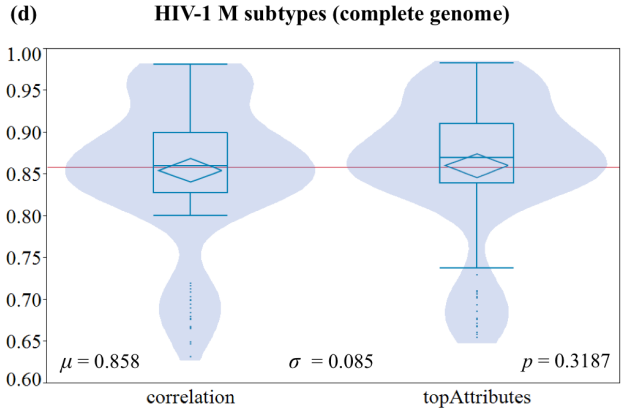


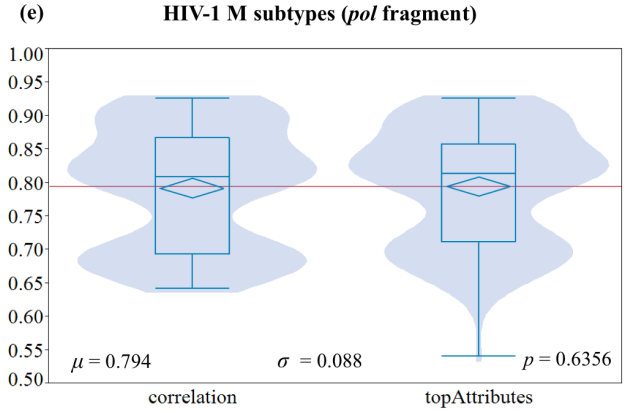
Simulation - comparaison des méthodes de sélection d'attributs



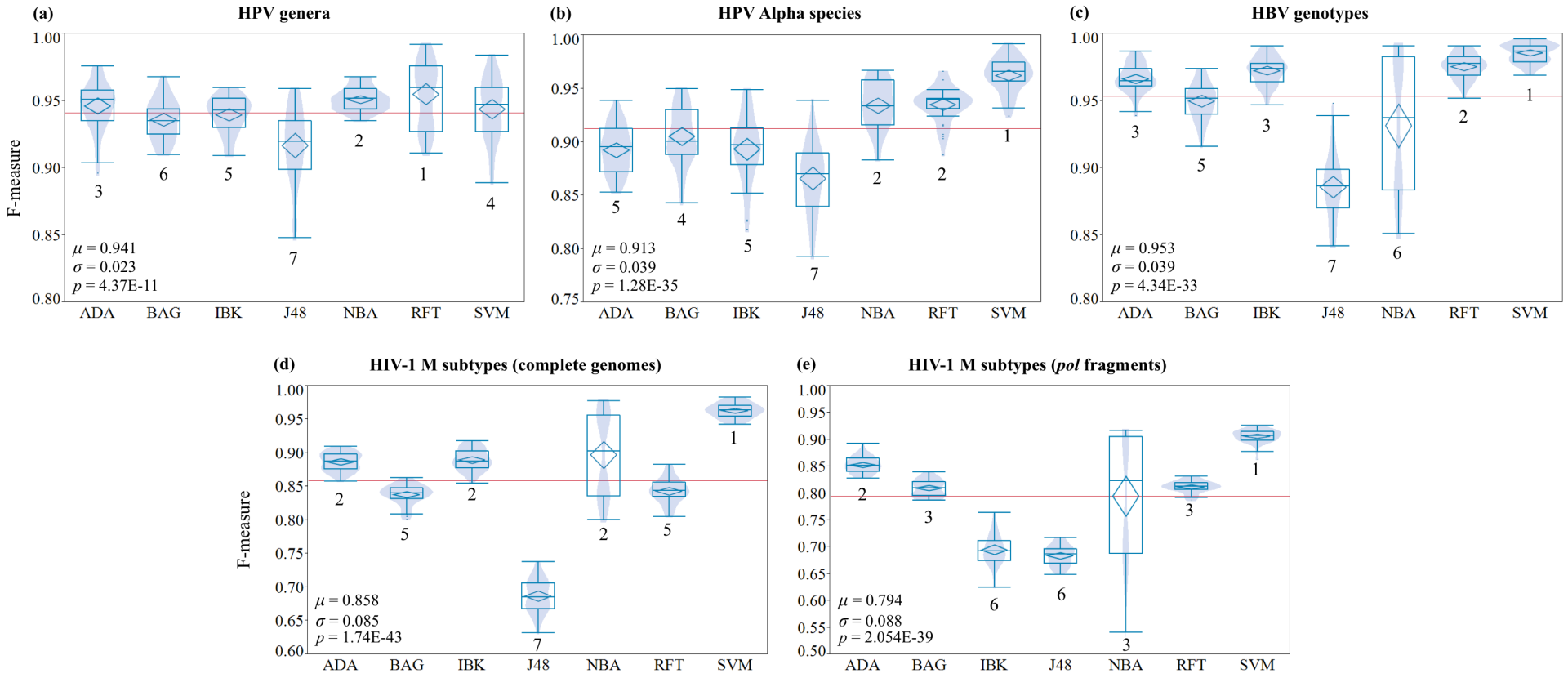




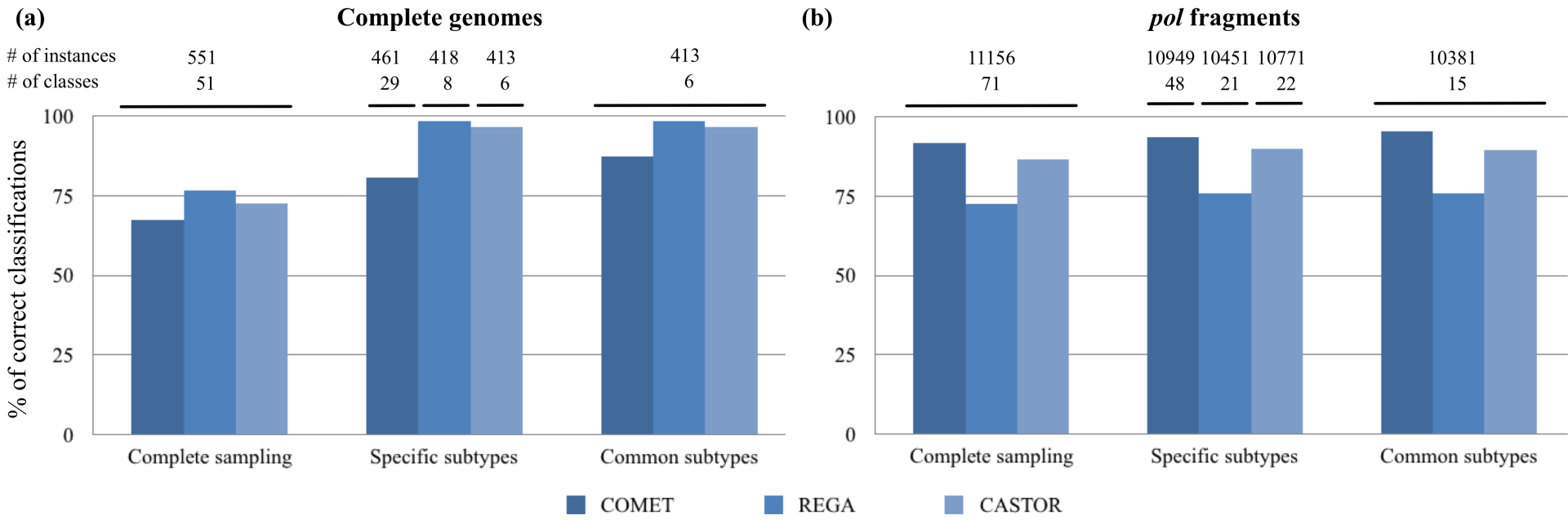




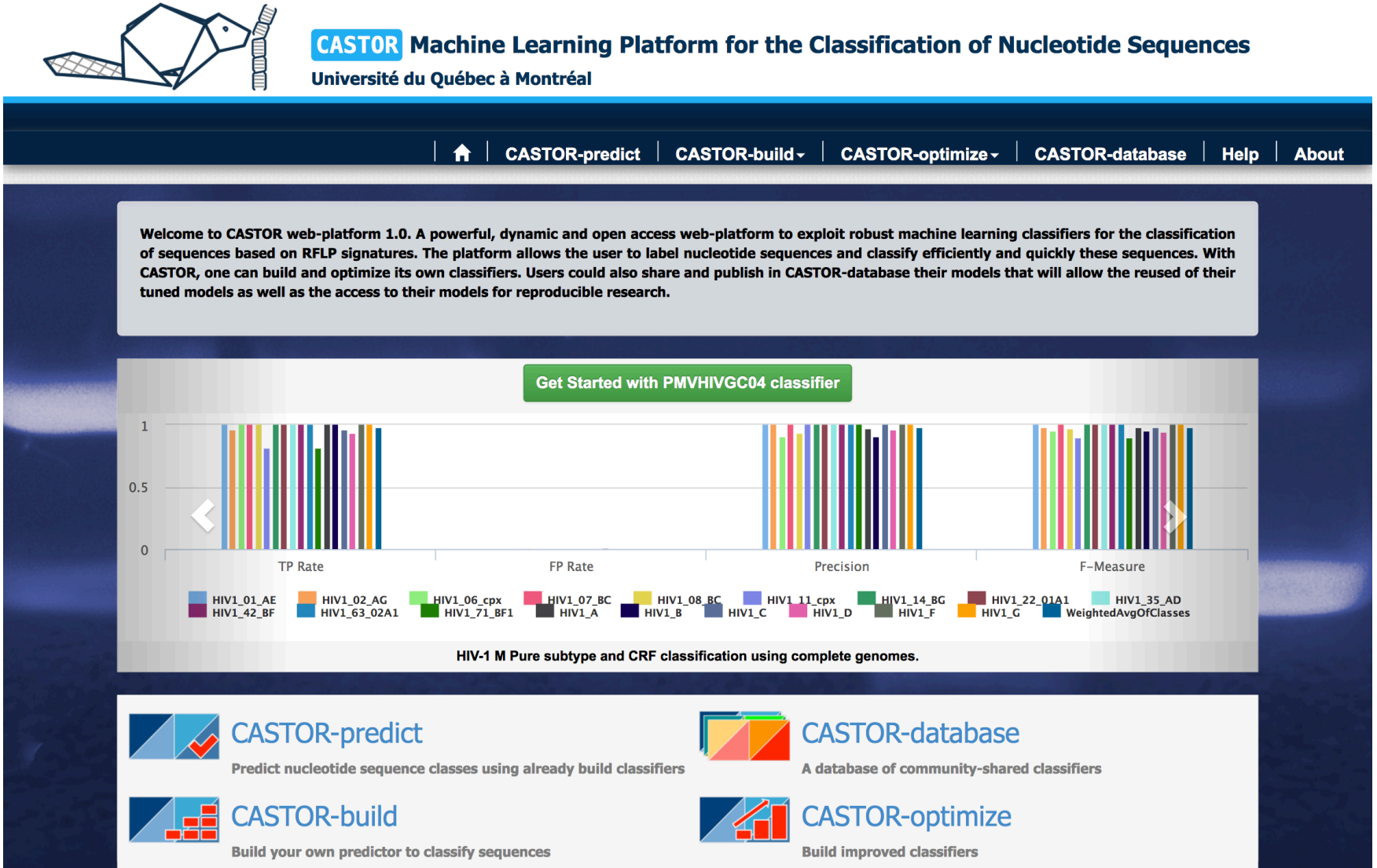
Simulation - comparaison des algorithmes d'apprentissage



Comparaison CASTOR avec REGA et COMET dans la classification de VIH



Plateforme CASTOR



Perspectives

- Typage d'autres virus et organismes
- Identifier l'ensemble d'enzymes qui a un pouvoir discriminant pour un type de classification
- Autres types de classification
  - Géographique
  - Pouvoir pathogène

Références

UQÀM

- Alcantara, L. C. J., Cassol, S., Libin, P., Deforche, K., Pybus, O. G., Van Ranst, M., Galvo-Castro, B., Vandamme, A. M., and de Oliveira, T. (2009). A standardized framework for accurate, high-throughput genotyping of recombinant and nonrecombinant viral sequences. Nucleic Acids Research, 37(SUPPL. 2), 634–642.
- Altschul, S.F., Madden, T.L., Sch □a↵er, a.a., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research 25(17), 3389–402 (1997)
- Bao, Y., Chetvernin, V., & Tatusova, T. 2014. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. *Archives of virology*, *159*(12), 3293-3304.
- Bernard, H. et al., 2010 . Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology,* 401(1), pp. 70-9.
- Deng, M., Yu, C., Liang, Q., He, R. L., & Yau, S. S. T. 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PloS one*, *6*(3), e17293.
- Daigle, B., Makarenkov, V., Diallo, A.B.: E↵ect of hundreds sequenced genomes on the classification of human papillomaviruses. In: Data Science, Learning by Latent Structures, and Knowledge Discovery, pp. 309–318. Springer, ??? (2015)
- de Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E. J., Wensing, A. M. J., van de Vijver, D. A., Boucher, C. A., Camacho, R., and Vandamme, A. M. (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics, 21(19), 3797–3800.
- de Villiers, E. et al., 2004. Classification of papillomaviruses. *Virology,* Volume 324.
- de Villiers, E. M. 2013. Cross-roads in the classification of papillomaviruses.*Virology*, *445*(1), 2-10.

Références

UQÀM

- Hall, M. et al., 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations,* 11(1).
- Lauber, C., & Gorbalenya, A. E. 2012. Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses.*Journal of virology*, *86*(7), 3890-3904.
- Matsen, F.a., Kodner, R.B., Armbrust, E.V.: pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11(1), 538 (2010)
- Roberts, R., Vincze, T., Posfai, J. & Macelis, D., 2010 . REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.,* 38(Database issue), pp. D234-6.
- Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Comp. and Applied Math. 20(C), 53–65 (1987)
- Struck, D., Lawyer, G., Ternes, A.-M., Schmit, J.-C., Berco↵, D.P.: Comet: adaptive context-based modeling for ultrafast hiv-1 subtype identification. Nucleic Acids Research 42(18), 144 (2014)
- Santiago, E., Camacho, L., Junquera, M. & Vázquez, F., 2006 . Full HPV typing by a single restriction enzyme. *J Clin Virol,* 37(1), pp. 38-46.
- Zheng, Z. & Baker, C., 2006. Papillomavirus genome structure, expression, and post-transcriptional regulation. *Front Biosci,* Volume 11, pp. 2286-302.

Merci

LABORATOIRE BIOINFORMATIQUE UQÀM

UQÀM

CRSNG NSERC

Fonds de recherche Nature et technologies Québec