Learning a peptide-protein binding affinity predictor with kernel ridge regression

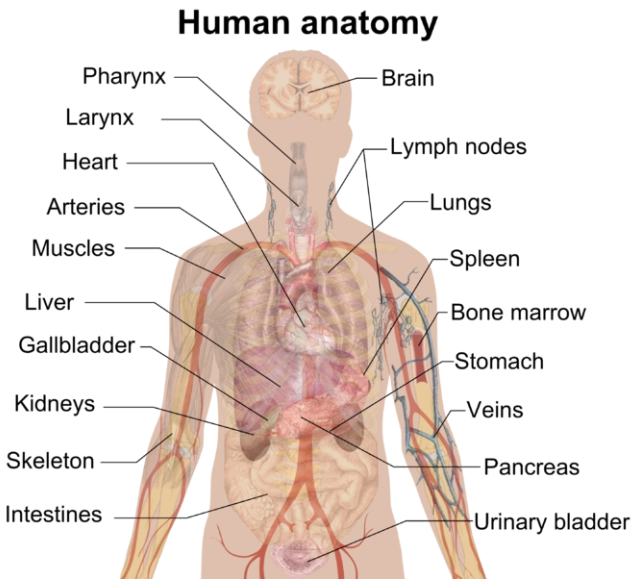
Sébastien Giguère, Mario Marchand, François Laviolette,
Alexandre Drouin, Jacques Corbeil*

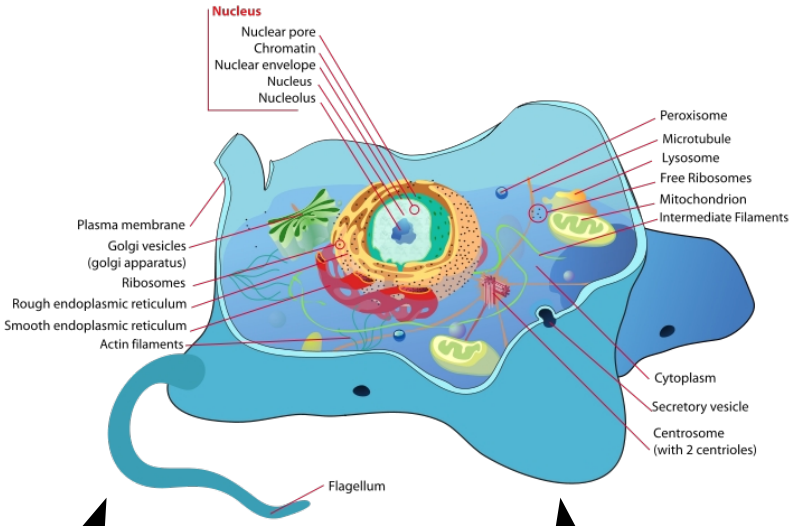
Department of Computer Science and Software Engineering, Université Laval, Québec, Canada
*Department of Molecular Medicine, Université Laval, Québec, Canada

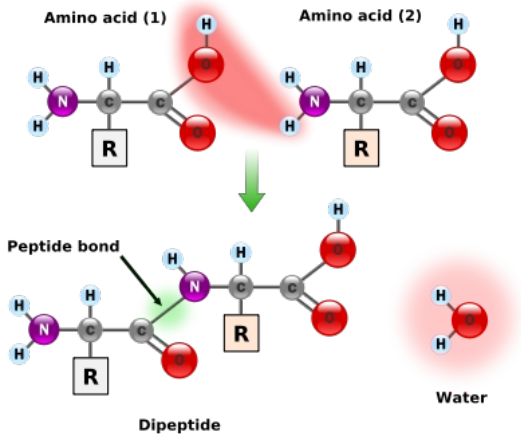
Déroulement

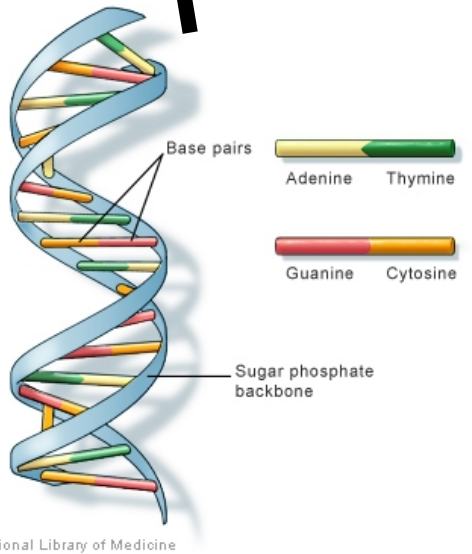
- Notions de bases
  - Biologie
  - Apprentissage automatique
- Notre approche
  - Algorithme d'apprentissage automatique
  - Noyaux adapté à la biologie
- Résultats
  - Interaction peptides-protéines
  - Complexes Majeurs d'Histocompatibilité de type II
  - 2012 Machine Learning Competition in Immunology
- Collaboration en chimie et pharmacie (FQRNT Équipe)

Biologie 101









Traduction

GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGACTCCTCTTC

DNA

(transcription)

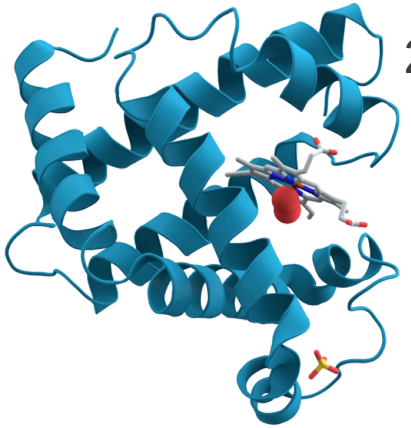
GUGCAUCUGACUCCUGAGGAGAAG

RNA

(translation)

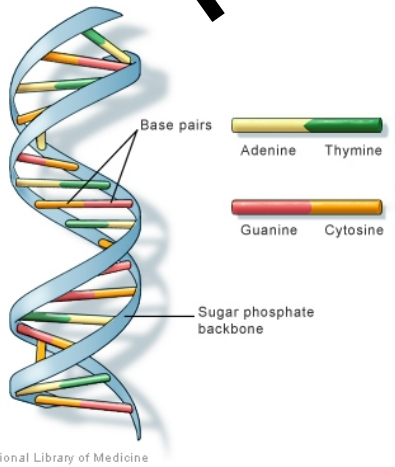
V H L T P E E K

protein



21 Acides Aminées

Fonction bijective

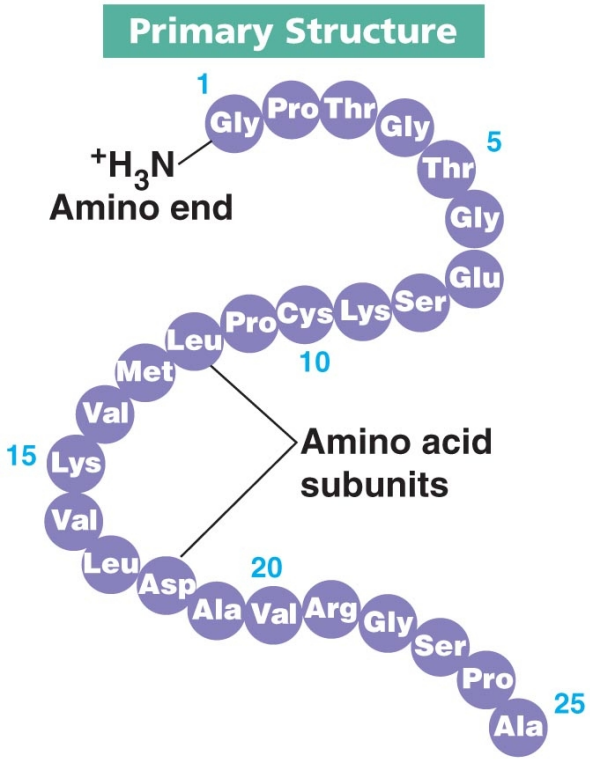


4 Nucléotides

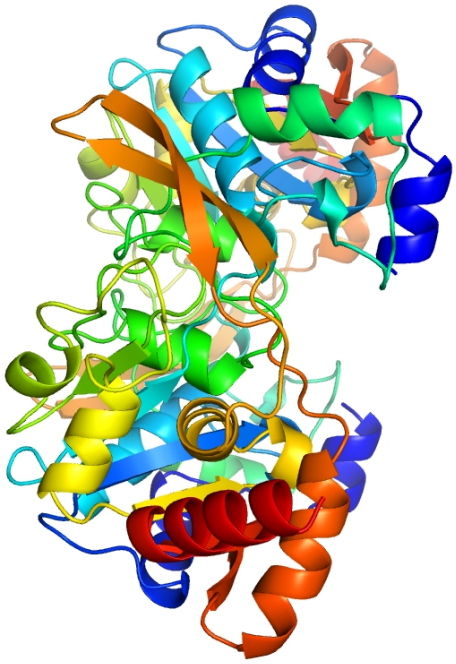
C'est un automate déterministe!

Structure des protéines Video

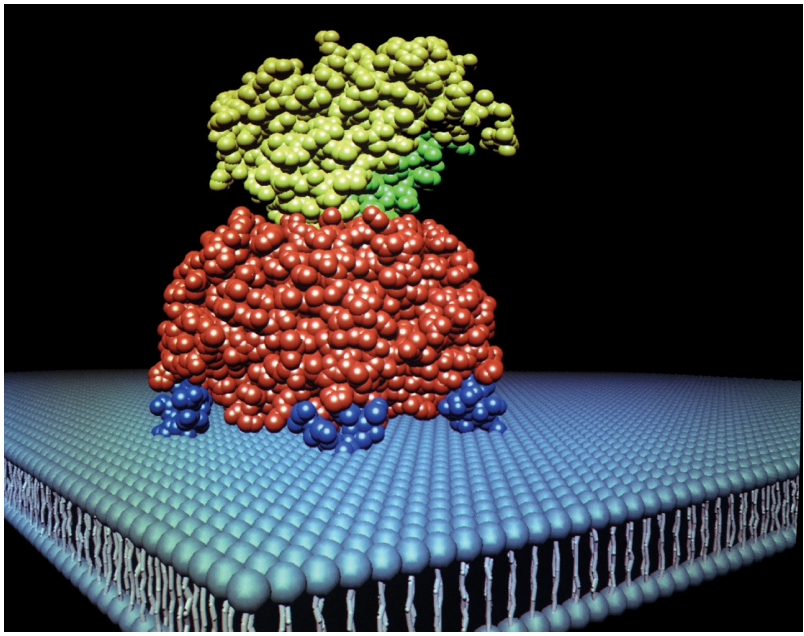
Structure primaire



Structure secondaire

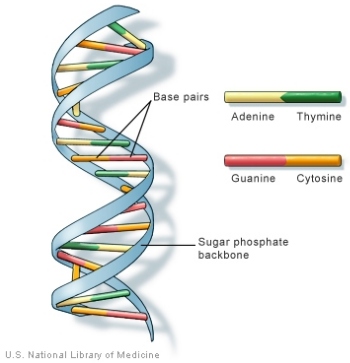


Structure tertiaire



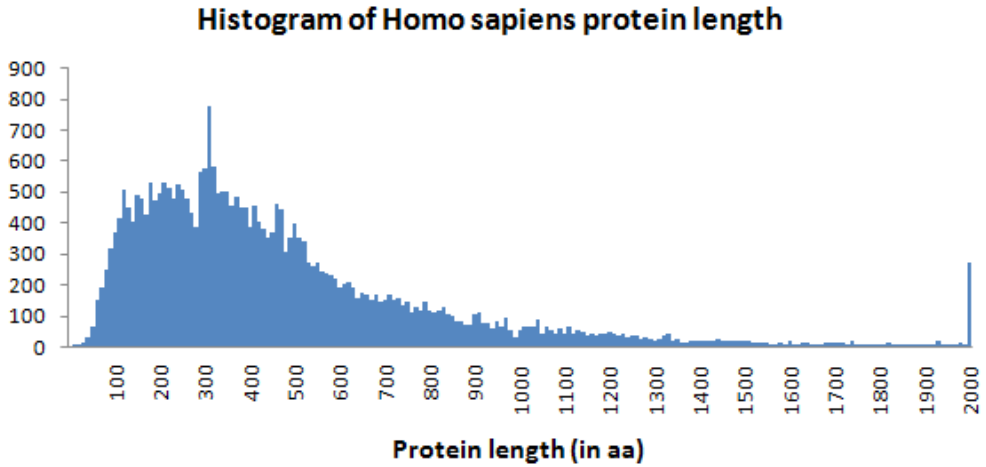
Taille des bio-molécules

Génome :

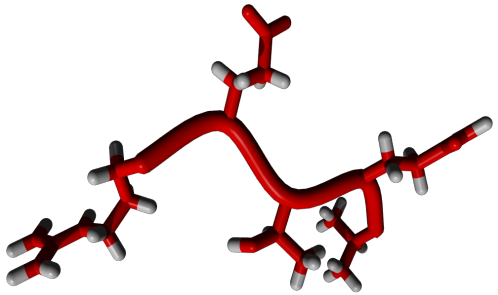


3×10^9 Paires de bases (génome humain)

Protéines :



Peptide :



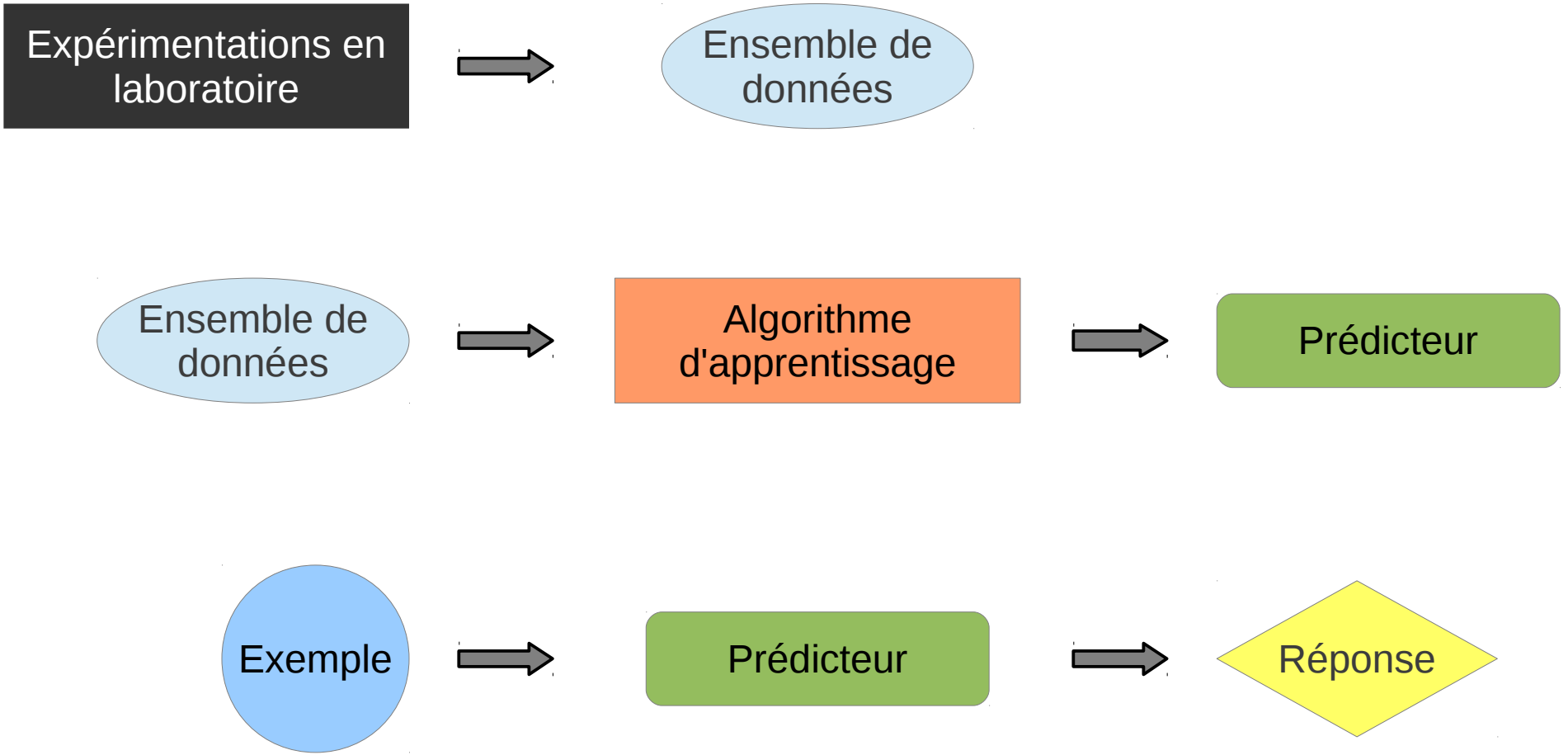
5 à 20 Acides Aminées

Machine Learning 101

"Field of study that gives computers the ability to learn without being explicitly programmed"

Arthur Samuel,1959

Procédure


Expérimentations en laboratoire
Ensemble de données
Ensemble de données
Algorithme d'apprentissage
Prédicteur
Exemple
Prédicteur
Réponse

Nos exemples

- Viennent sous la forme ((***x***,***y***),*e*)

  x : un peptide (structure primaire)

  y : une protéine (structure primaire, secondaire et tertiaire)

  e : valeur réelle de l'affinité de liaison

  (***x***,***y***) : Couple peptide-protéine

Kernel

- La similarité entre deux peptides ***x*** et ***x'*** est donnée par :

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') \longrightarrow \mathbb{R}$$

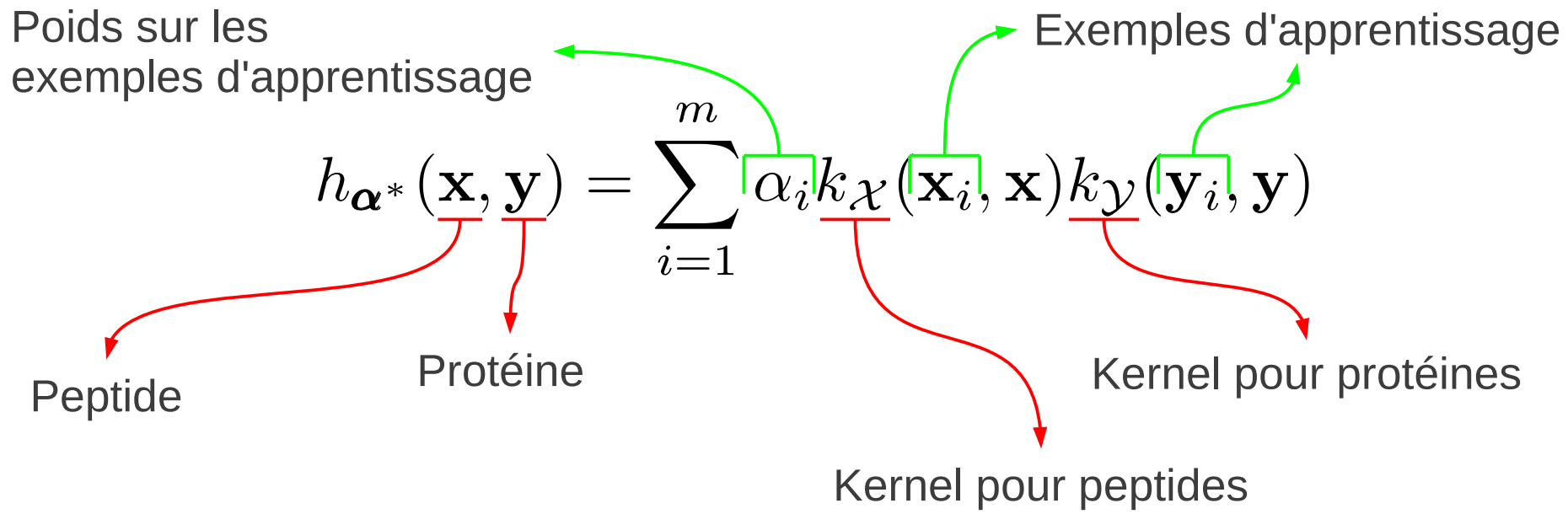
- La similarité entre deux protéines ***y*** et ***y'*** est donnée par :

$$k_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') \longrightarrow \mathbb{R}$$

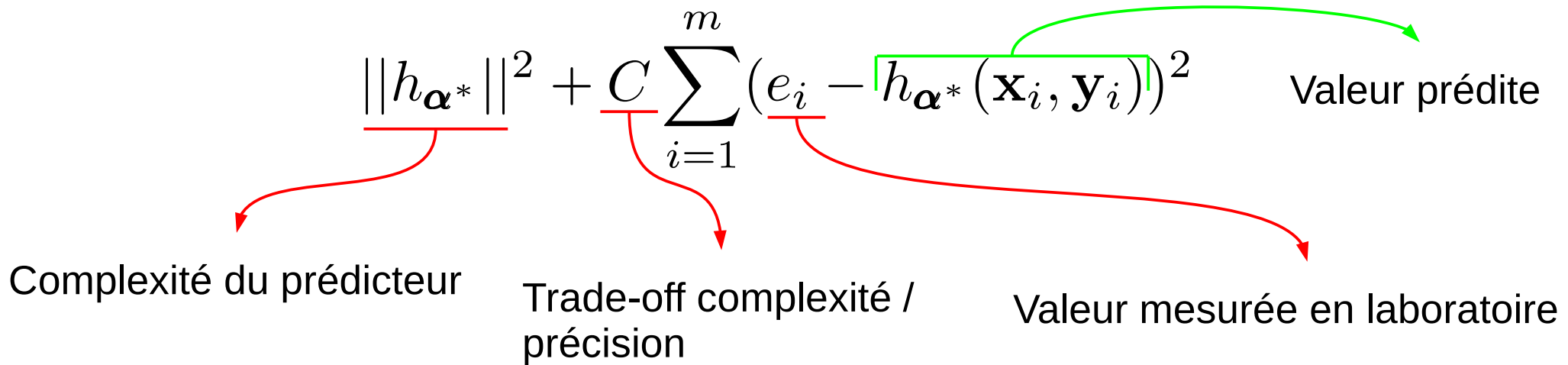
- La similarité entre deux couples (x,y) et (x',y') est donnée par :

$$k((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) = k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}')k_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}')$$

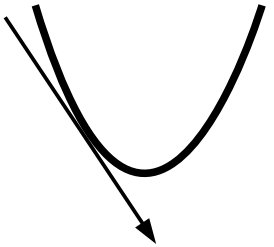
Prédicteur :



L'algorithme d'apprentissage consiste à minimiser :



Algorithme d'apprentissage



- La fonction à minimiser est convexe!
- La solution optimale :
  - Unique
  - Garantie d'exister

- Solution : inverser une matrice de *m x m*

$$\mathcal{O}(m^{2.376})$$

Algorithme de Coppersmith-Winograd

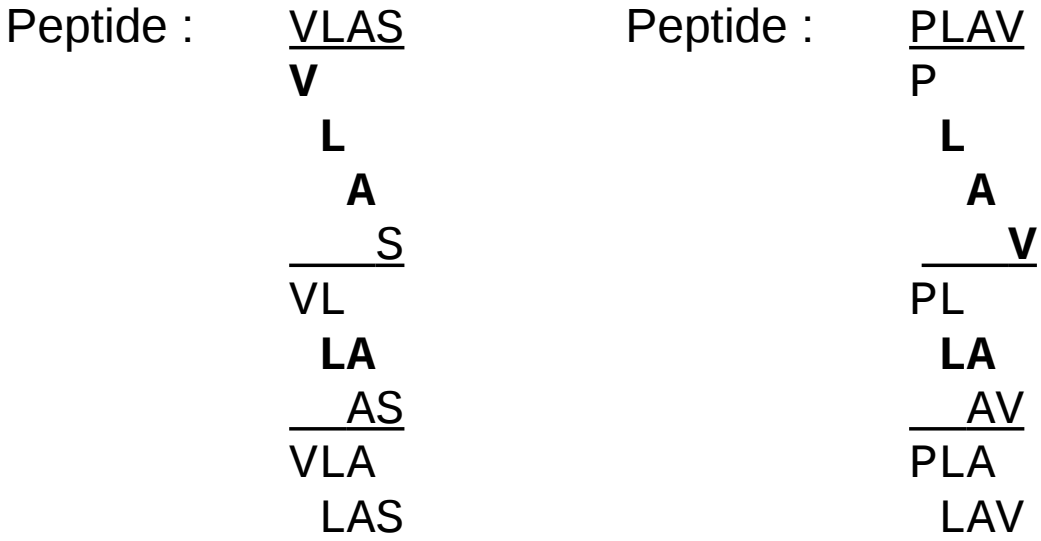
Choix et conception de kernel

- Critique pour l'obtention d'un prédicteur précis
- Doit être adapté à la tâche à apprendre
- Nouvelle application → nouveau kernel

- Nous proposons un kernel spécialisé aux
  - Peptides
  - Petites séquences d'acides aminées

String kernel

Comparer deux chaînes en comptant les sous-chaînes communes!

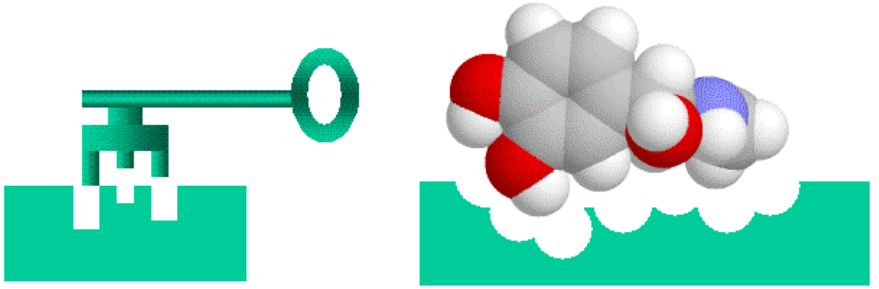


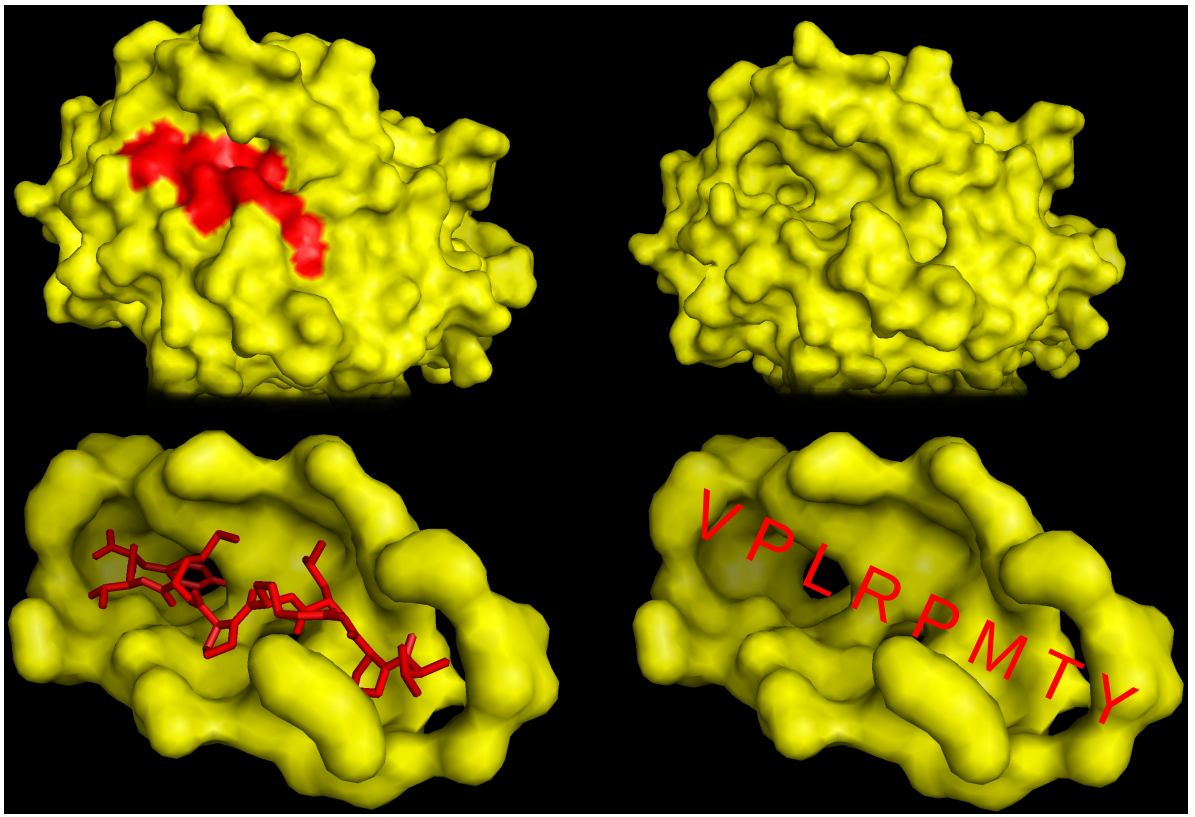
Paramètre L, contrôle la longueur des sous-chaînes. Ici L=3.

Importance de la position des acides aminées

Analogie de la clef et de la serrure

ABCD != DCBA

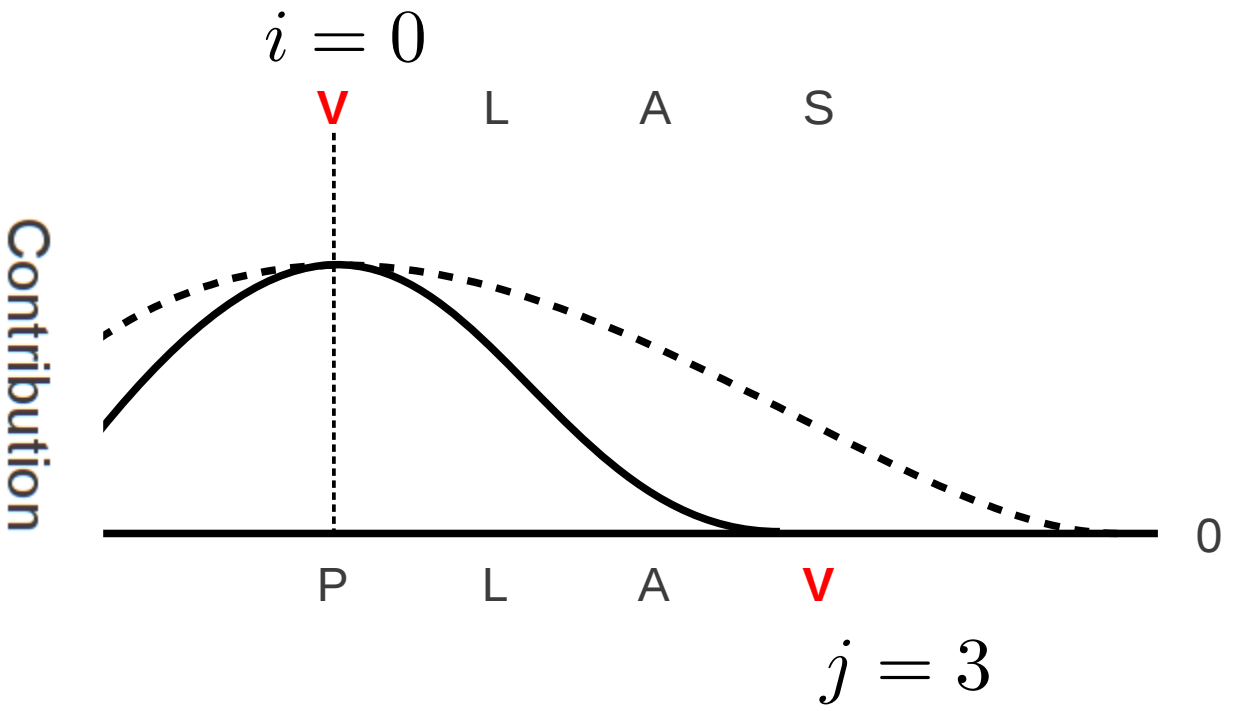




Pénaliser selon la position

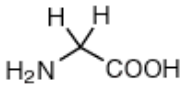
Terme de contribution : $e^{\left(\frac{-(i-j)^2}{2\sigma_p^2}\right)}$

Contrôle la contribution

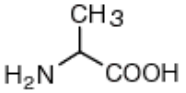


Propriétés physico-chimiques

Small

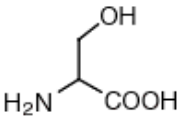


Glycine (Gly, G)
MW: 57.05

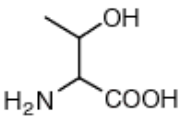


Alanine (Ala, A)
MW: 71.09

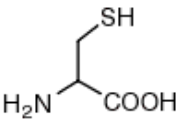
Nucleophilic



Serine (Ser, S)
MW: 87.08, pK_a ~ 16

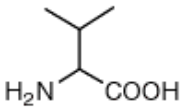


Threonine (Thr, T)
MW: 101.11, pK_a ~ 16

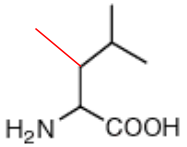


Cysteine (Cys, C)
MW: 103.15, pK_a = 8.35

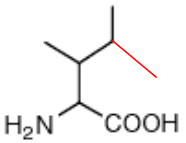
Hydrophobic



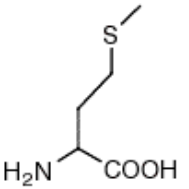
Valine (Val, V)
MW: 99.14



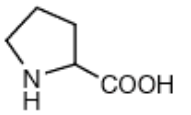
Leucine (Leu, L)
MW: 113.16



Isoleucine (Ile, I)
MW: 113.16

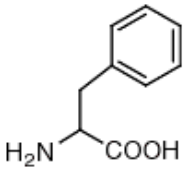


Methionine (Met, M)
MW: 131.19

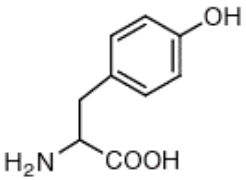


Proline (Pro, P)
MW: 97.12

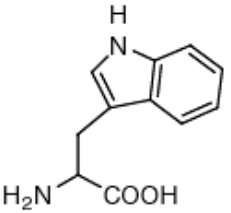
Aromatic



Phenylalanine (Phe, F)
MW: 147.18

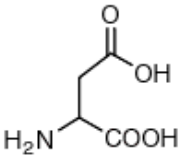


Tyrosine (Tyr, Y)
MW: 163.18

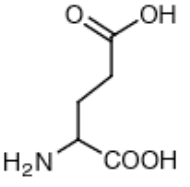


Tryptophan (Trp, W)
MW: 186.21

Acidic

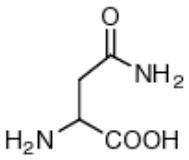


Aspartic Acid (Asp, D)
MW: 115.09, pK_a = 3.9

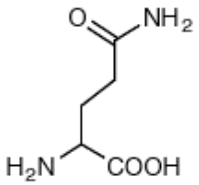


Glutamic Acid (Glu, E)
MW: 129.12, pK_a = 4.07

Amide

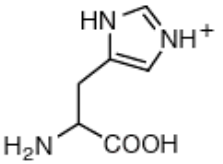


Asparagine (Asn, N)
MW: 114.11

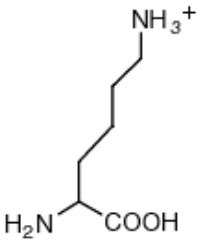


Glutamine (Gln, Q)
MW: 128.14

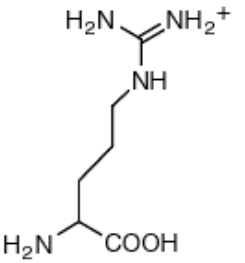
Basic



Histidine (His, H)
MW: 137.14, pK_a = 6.04



Lysine (Lys, K)
MW: 128.17, pK_a = 10.79



Arginine (Arg, R)
MW: 156.19, pK_a = 12.48

Propriétés physico-chimiques

Peptide : VLAS

V
L
A
S
VL
LA
AS
VLA
LAS

Peptide : PLAV

P
L
A
V
PL
LA
AV
PLA
LAV

La Proline et la Leucine sont tous les deux hydro-phobiques.

$\psi : \Sigma \longrightarrow \mathbb{R}^d$ Encode les propriétés des acides aminées.

Terme de contribution: $e\left(\frac{-\|\psi(L)-\psi(P)\|^2}{2\sigma_c^2}\right)$

Contrôle la contribution

Generic String (GS) kernel

$$GS(\mathbf{x}, \mathbf{x}', L, \sigma_p, \sigma_c) \overset{\text{def}}{=} \sum_{l=1}^{L} \sum_{i=0}^{|x|-l} \sum_{j=0}^{|x'|-l} e^{\left(\frac{-(i-j)^2}{2\sigma_p^2}\right)} e^{\left(\frac{-\|\boldsymbol{\psi}^l(x_{i+1},..,x_{i+l}) - \boldsymbol{\psi}^l(x'_{j+1},..,x_{j+l})\|^2}{2\sigma_c^2}\right)}$$

Deux peptides

Paramètres

Sous-chaînes de taille 1 à L

Peptide x

Peptide x'

Contribution selon la position relative des sous-chaînes

Contribution selon les propriétés physico-chimiques

Complexité algorithmique

Complexité : $\mathcal{O}(L^2 \cdot \boldsymbol{\psi} \cdot x \cdot x') \approx \mathcal{O}(n^5)$

$n \geq L, \boldsymbol{\psi}, x, x'$

Programmation dynamique : $\mathcal{O}(L \cdot x \cdot x') \approx \mathcal{O}(n^3)$

Approximation : $\mathcal{O}(L \cdot \max(x, x')) \approx \mathcal{O}(n^2)$

Généralisation

Fixed parameters	Free parameters	Kernel name
$L = 1, \sigma_p \to 0, \sigma_c \to 0$		Hamming distance
$L \to \infty, \sigma_p \to 0, \sigma_c \to 0$		Dirac delta
$\sigma_p \to \infty, \sigma_c \to 0$	L	Blended Spectrum [12]
$\sigma_p \to \infty$	L, σ_c	Blended Spectrum RBF [22]
$\sigma_c \to 0$	L, σ_p	Oligo [13]
$L \to \infty, \sigma_p \to 0$	σ_c	Radial Basis Function (RBF)
$\sigma_p \to 0, \sigma_c \to 0$	L	Weighted degree ($\star$) [14]
$\sigma_p \to 0$	L, σ_c	Weighted degree RBF ($\star$) [22]
	L, σ_p, σ_c	Generic String (GS)

($\star$) Substituting $\boldsymbol{\psi}^l$ by $\boldsymbol{\psi}^l \sqrt{-\ln \beta_l}$ where the β_l's are the weighted degrees defined in [14].

Bonus :

Preuve que le GS kernel est symétrique positif semi-définie!

Métriques et validation

- On utilise toujours un ensemble de validation caché!
- Root Mean Squared Error (RMSE) :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(e_i - h_{\boldsymbol{\alpha}^*}(\mathbf{x}_i, \mathbf{y}_i))^2}{n}}$$

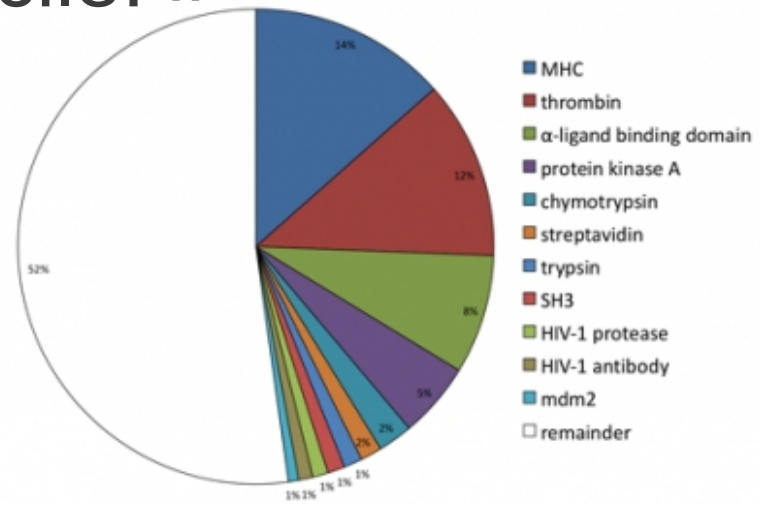
- Pearson product-moment correlation coefficient (PCC) :

$$PCC = \sqrt{1 - \frac{\sum_{i=1}^{n}(e_i - h_{\boldsymbol{\alpha}^*}(\mathbf{x}_i, \mathbf{y}_i))^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2}} \in [0, 1]$$

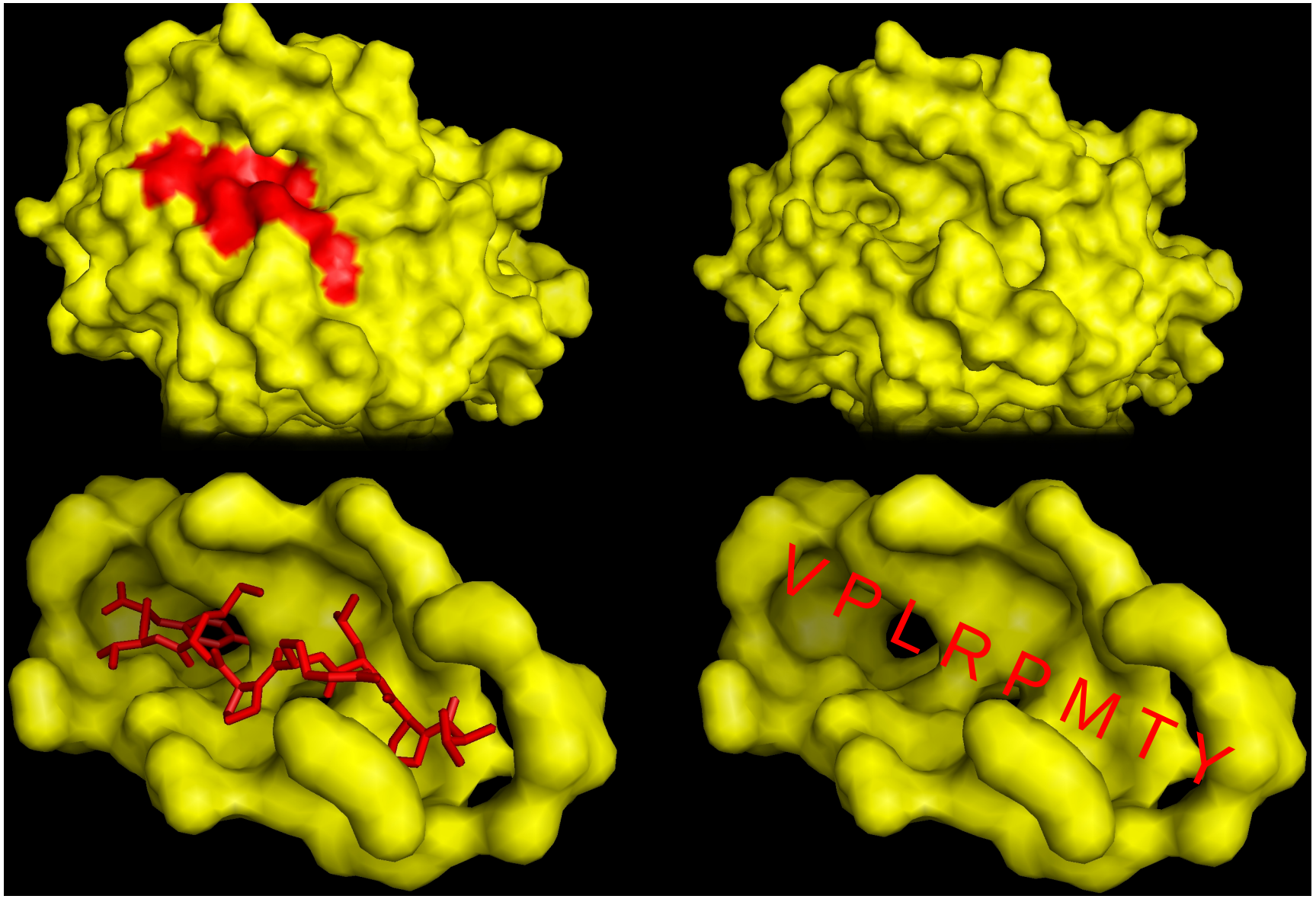
PepX

- 1431 Complexes peptide-protéine cristallisé à haute-définition
- Pleine diversité de tous les protéines connus
- But : Prédire l'énergie de liaison en kcal/mol

- Les bio-chimistes : « C'est difficile! »



Binding pocket kernel (Brice Hoffman et al.)



Applications

- C'est un problème fondamental en biologie
- Réduire les coûts des expérimentations en laboratoire
- Accélérer le développement de composés pharmaceutiques
- Assister les biologistes dans la compréhension des processus cellulaires

Prédicteur pour PepX

Prédicteur est donnée par :

$$h_{\boldsymbol{\alpha}^*}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \alpha_i k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}) k_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y})$$

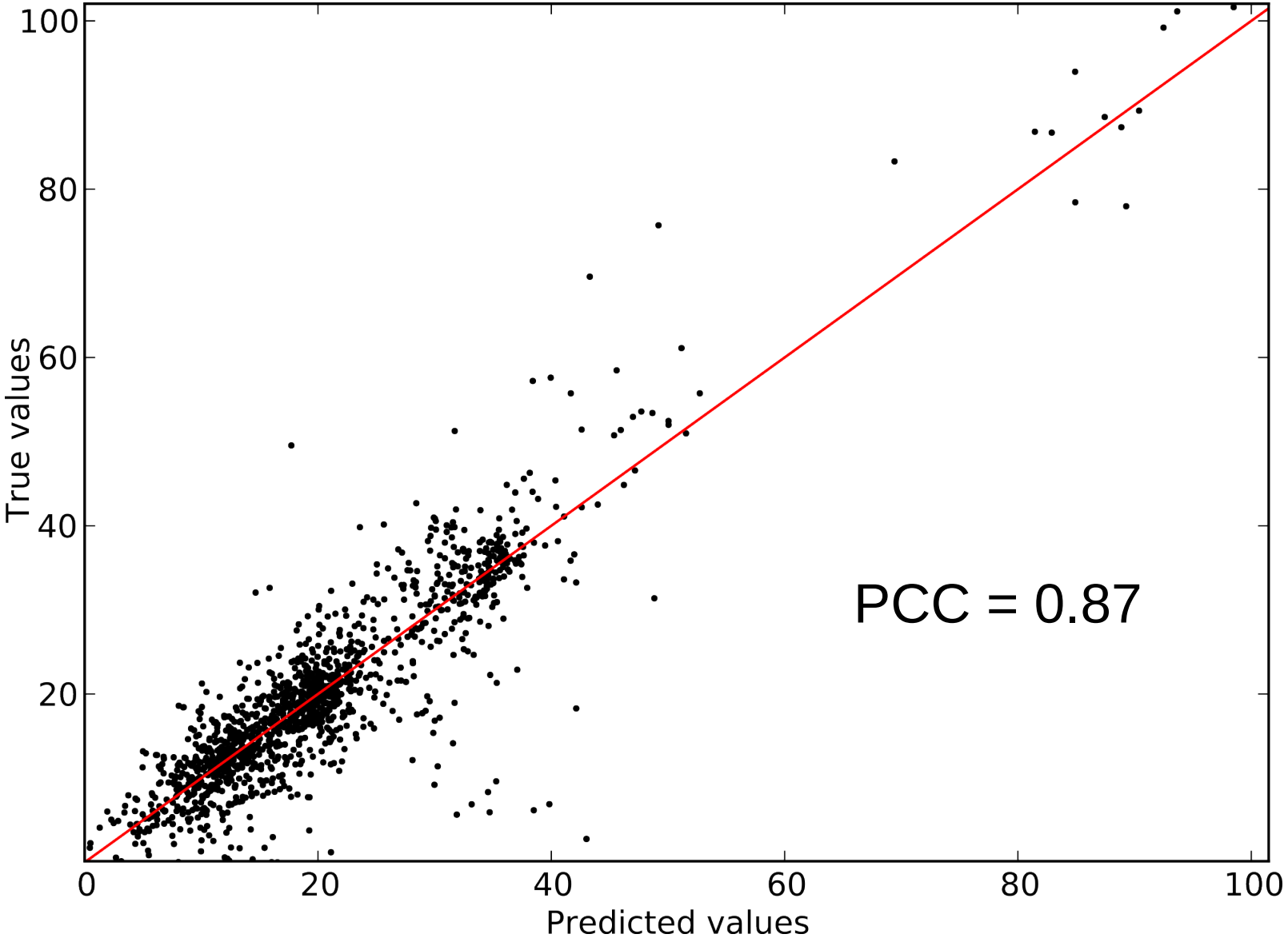
Peptide

Structure cristallographié d'une protéine

GS kernel

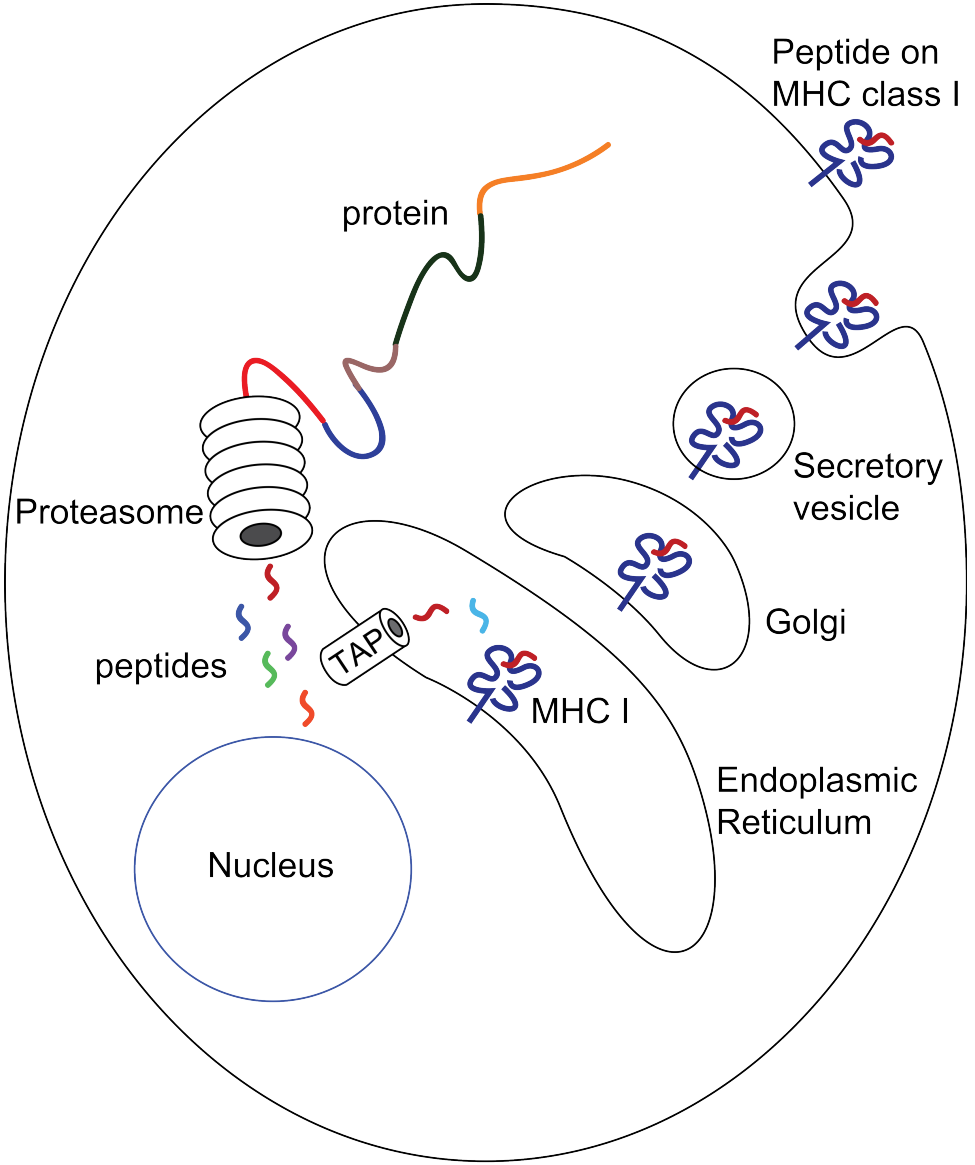
Binding Pocket Kernel, Hoffman et al.

Résultats


100
80
60
40
20
True values
0
20
40
60
80
100
Predicted values
PCC = 0.87

Complexes Majeurs d'Histocompatibilité

- Protéine à la base de votre système immunitaire
- Responsable de reconnaître les antigène
- Détermine la compatibilité pour la transplantation d'organes
- Protéine extrèmement polymorphique pour s'adapter aux pathogènes



Applications

- Accélérer le développement de vaccins
- Mieux comprendre le système immunitaire

- Traitements thérapeutiques contre certaines formes de cancers
- Traitements thérapeutiques pour les maladies auto-immunes
  - Maladie de Crohn
  - Polyarthrite rhumatoïde
  - … 50 autres ...

Prédiction « single-target »

- Pour plusieurs molécules de MHC, nous connaissons déjà un certain nombre de peptide et l'affinité de liaison.
- But : prédire pour ces MHC, l'affinité de liaison de nouveaux peptides

- Prédicteur :

$$h_{\boldsymbol{\alpha}^*}(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x})$$

Peptide

GS kernel

Résultats

Table 3: Best results for each metric are highlighted in bold. The PCC results show that the proposed method (KRR+GS) outperforms the RTA method with a p-value of 0.0308. The RMSE results show that KRR+GS outperforms the RTA method on all 16 allotypes with a p-value of 0.0005.

	PCC		RMSE (kcal/mol)		
MHC β chain	KRR+GS	RTA	KRR+GS	RTA	# of examples
DRB1*0101	**0.632**	0.530	**1.20**	1.43	5648
DRB1*0301	**0.538**	0.425	**1.16**	1.46	837
DRB1*0401	**0.430**	0.340	**1.44**	1.72	1014
DRB1*0404	**0.491**	0.487	**1.25**	1.38	617
DRB1*0405	**0.530**	0.442	**1.09**	1.35	642
DRB1*0701	**0.645**	0.484	**1.24**	1.62	833
DRB1*0802	**0.469**	0.412	**1.19**	1.34	557
DRB1*0901	0.303	**0.369**	**1.55**	1.68	551
DRB1*1101	**0.550**	0.450	**1.17**	1.45	812
DRB1*1302	**0.468**	0.464	**1.51**	1.64	636
DRB1*1501	**0.502**	0.438	**1.41**	1.53	879
DRB3*0101	0.380	**0.425**	**1.03**	1.13	483
DRB4*0101	**0.613**	0.522	**1.10**	1.33	664
DRB5*0101	**0.541**	0.434	**1.20**	1.57	835
H2*IA$_b$	**0.603**	0.556	**1.00**	1.15	526
H2*IA$_d$	0.325	**0.563**	**1.44**	1.53	306
Average:	**0.501**	0.459	**1.25**	1.46	

Prédiction «multi-target »

- Il existe 72 874 molécules de MHC
- Pour la majorité, on ne connaît pas ou peu de peptides
  - Aucune donnée pour s'entraîner!
- Solution : Entraîner un prédicteur universel!

- Prédicteur :

$$h_{\boldsymbol{\alpha}^*}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \alpha_i k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}) k_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y})$$

Peptide Molécule de MHC GS kernel

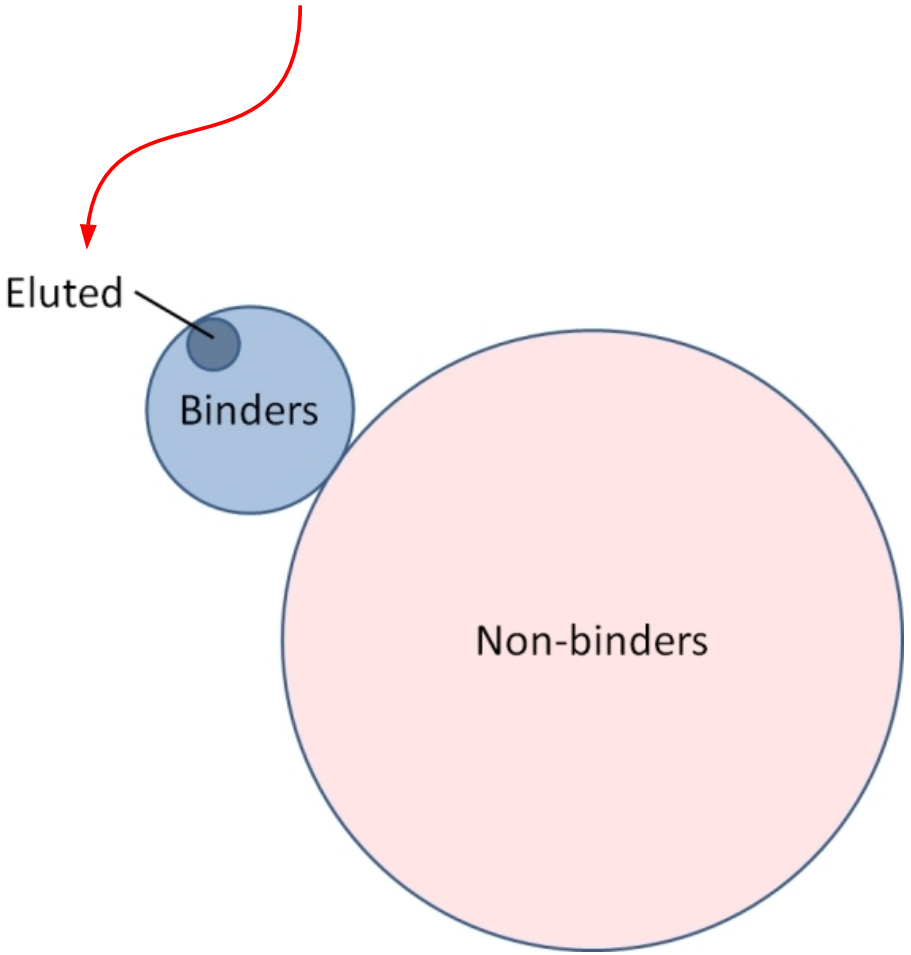
Résultats

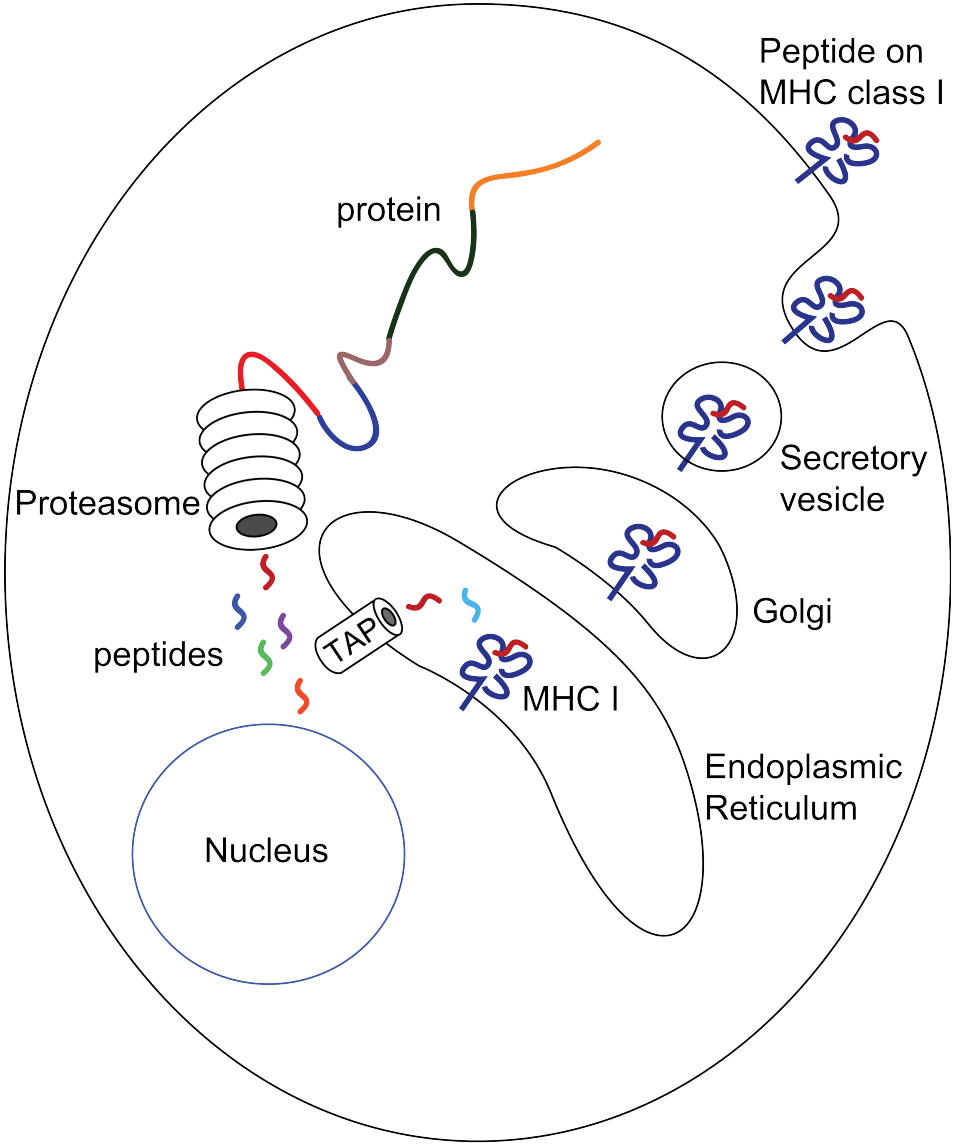
Table 4: Best results for each metric are highlighted in bold. The PCC results show that the proposed method (KRR+GS) outperforms MultiRTA with a p-value of 0.001 and NetMHCIIpan-2.0 with a p-value of 0.0574. The RMSE results indicate that KRR+GS outperforms MultiRTA with a p-value of 0.0466.

	PCC			RMSE (kcal/mol)		
MHC β chain	KRR+GS	MultiRTA	NetMHCIIpan-2.0	KRR+GS	MultiRTA	# of examples
DRB1*0101	**0.662**	0.619	0.627	1.48	**1.33**	5166
DRB1*0301	**0.743**	0.438	0.560	**1.29**	1.36	1020
DRB1*0401	**0.667**	0.534	0.652	**1.36**	1.56	1024
DRB1*0404	0.709	0.623	**0.731**	**1.18**	1.33	663
DRB1*0405	0.606	0.566	**0.626**	**1.25**	1.28	630
DRB1*0701	0.694	0.620	**0.753**	**1.34**	1.51	853
DRB1*0802	**0.728**	0.523	0.700	**1.23**	1.45	420
DRB1*0901	0.471	0.375	**0.474**	**1.53**	2.01	530
DRB1*1101	**0.786**	0.603	0.721	**1.16**	1.46	950
DRB1*1302	**0.416**	0.365	0.337	1.73	**1.68**	498
DRB1*1501	**0.612**	0.513	0.598	**1.46**	1.57	934
DRB3*0101	**0.654**	0.603	0.474	1.52	**1.10**	549
DRB4*0101	**0.540**	0.508	0.515	**1.41**	1.61	446
DRB5*0101	**0.732**	0.543	0.722	**1.28**	1.60	924
Average:	**0.644**	0.531	0.606	**1.37**	1.49	

2012 Machine Learning Competition in Immunology

But : prédire les peptides « Eluted ».





Résultats

	SUM-HUM	SUM-MUS	SUM-ALL
	4.823	1.489	6.162
PREDICTOR			
1D-BENCH	4.757	1.368	6.125
2D	**4.823**	1.339	**6.162**
2E	4.474	1.348	5.900
2F	4.474	**1.489**	6.138
3D	4.474	1.348	5.822
4D	4.444	1.103	5.548
7D	4.782	1.267	6.049
8D	4.816	1.267	6.083
9D	4.816	1.277	6.093
10D	4.754	1.231	5.984
11D	4.762	1.274	6.036
12D	4.814	1.276	6.090
15A	4.751	1.294	6.045
15B	4.789	1.327	6.116
15C	4.707	1.292	5.999
16A	4.745	1.319	6.064
16B	4.742	1.361	6.103
16C	4.706	1.371	6.077
20A	4.452	0.799	5.251
20B	4.396	0.820	5.216
20C	4.358	0.822	5.181
20D	4.698	0.998	5.696
21D	4.784	1.243	6.027
22D	4.798	1.288	6.086
23D	4.710	1.255	5.965

Les données de validation proviennent de nouvelles expérimentations faites expressément pour valider les méthodes.

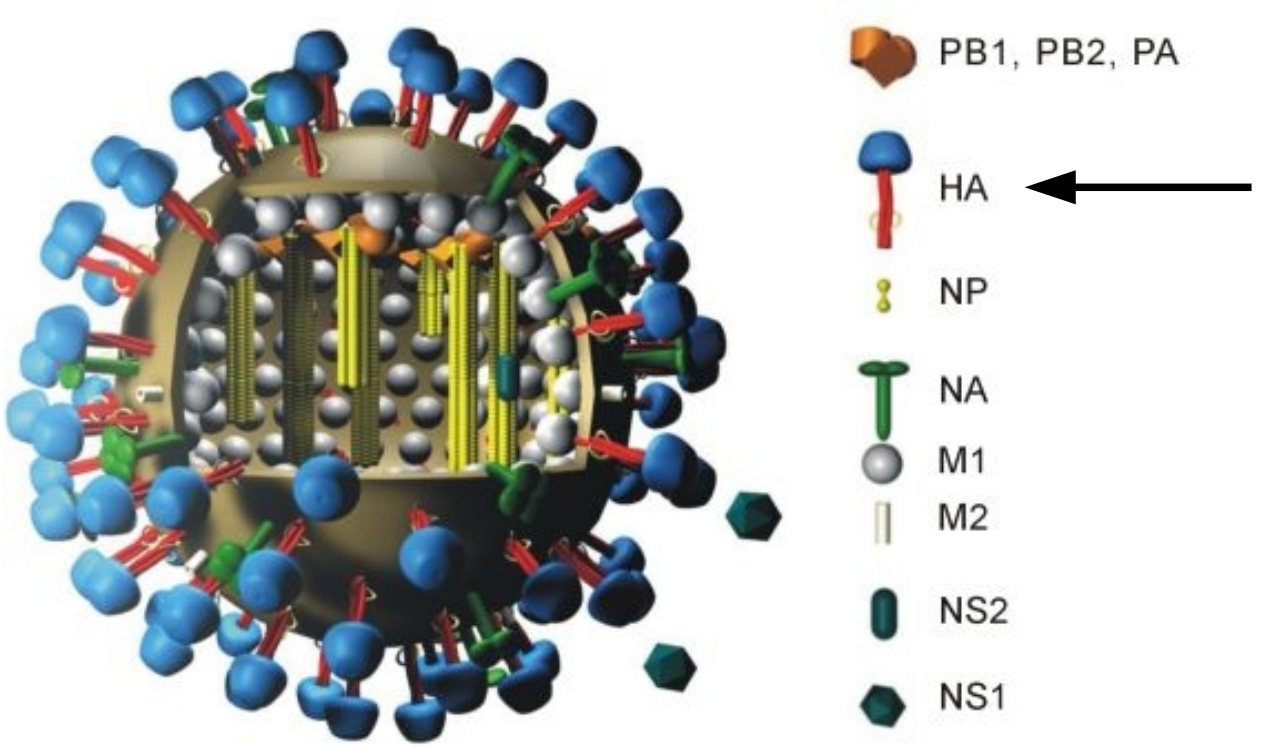
L'ensemble de validation à été mis publique suite à la compétition.

Collaboration

- Deux chercheurs du CHUL : Jacques Corbeil et Éric Biron
- Générer un ensemble de peptides qui se lient à une protéine d'intérêt
  - Techniques des chimie combinatoire
- Apprendre un prédicteur à partir de ces exemples
  - Techniques présentés aujourd'hui
- Prédire le meilleur peptide inhibiteur pour la protéine d'intérêt
  - Dans le cas général, ce problème est NP-Difficile … mais ...!

Notre protéine : l'Hémagglutinine

Responsable de la liaison du virus de l'influenza à la cellule infectée.



Question?