Projet - Etude des facteurs influençant l'asymétrie structurale de la protéase du VIH-2

Introduction

La protéase du VIH-2 est une enzyme du virus qui joue un rôle dans le cycle viral permettant la formation de nouveaux virus grâce au clivage protéolytique. Elle peut être à l'état sauvage ou muté, induisant des asymétries structural. L'objectif du TP est d'identifier quels sont les facteurs qui influencent l'asymétrie structurale de la protéase du VIH-2 parmi les 96 facteurs proposés. La protéase du VIH-2 peut avoir 3 conformations fermée, semi-ouverte et ouverte ayant pour fonction la fixation de ligand qui induit des changement de structure protéique ou l'hydrolyse des substrats. La problématique est : Quels facteurs influencent l'asymétrie structural da la protéine ? Un modèle de prédiction doit être crée en utilisant les données fournis du fichier matrice_descripteurs.dat composé de 201 modèles de structure 3D de la protéase du VIH-2 (101 modèles) et de mutants (100 modèles), ainsi que 96 descripteurs pour répondre à cette problématique. Ce qui permettra de chercher des inhibiteurs de la protéase du VIH-2 qui empêcheront de catalyser le clivages des polypeptides et donc empêcheront la production de virions infectieux.

Matériel et méthode

Le jeu de données contient 201 modèles de la structure 3D de la protéase du VIH-2, dont 101 proviennent d'une simulation dynamique de la structure cristallographique de le protéase du VIH-2 sauvage. Une simulation à partir d'un modèle de la structure de protéase ayant la mutation I82F a permis d'obtenir 101 modèles. La variable à expliquer est asymPos.nbr correspondant au nombre de position asymétrique qui est de type quantitative. Les 96 variables descriptives sont également de type quantitative, qui sont le nombre de poches extraites de la structure, 51 de ces descripteurs permettent de caractériser les propriétés physico-chimiques et géométriques de la poche centrale et de la structure étudier. Ainsi que 24 descripteurs avec la distances calculés entre deux carbones alpha de la structure étudiée. Ainsi que des descripteurs de la fréquence des 20 acides aminés dans la structure étudier.

Une variable quantitative à prédire et plusieurs variables quantitatives descriptives donc nous réalisons une régression linéaire multiple. Nous réalisons les étapes suivante.

Etape1 : Préparation du jeu de données : Ouvrir le jeu de données matrice_descripteurs.dat et décrire les descriptifs. Supprimer les descripteurs contenants uniquement des NA, les descripteurs avec une variance inférieur à 0.1 et les descripteurs avec une forte corrélations soit de 0.9. Représenter les descripteurs restant à l'aide d'un boxplot puis centrer et réduire la matrice contenant le jeu de donnée. Enfin analyser les corrélations entre les variables à l'aide d'un corrplot

utilisant la méthode circle. Les fonctions nécessaire sont read.table(), str(), boxplot(), which(), na.omit(), apply(), cor(), corrplot(), findCorrelation() et scale().

Etape2 : Création des jeux de données d'apprentissage et test : Créer à partir du jeu de données un échantillon d'apprentissage aléatoirement contenant 2/3 des modèles et un échantillon test contenant le 1/3 des modèles restant. Pour cela, créer des vecteurs avec les numéros des modèles des échantillons d'apprentissage et test à l'aide de la méthode de bootstrap. Créer des matrices qui possèdent les descripteurs pour les modèles des échantillons d'apprentissage et test. Les fonctions utilisés sont sample() et as.data.frame().

Etape3 : Validation des échantillons d'apprentissage et de test : Analyser la distribution des échantillons d'apprentissage et test et vérifier que les hypothèses sont validées. Utiliser hist() et for().

Etape4 : Apprentissage du modèle complet : Estimer les paramètre du modèle linéaire permettant de prédire asymPos.nbr en fonction des descripteurs restants. Identifier les descripteurs significative qui ont donc un paramètre β_1 significativement différent de 0, permettant de trouver une relation linéaire entre asymPos.nbr et la descripteur significative au risque α. Test de la nullité du paramètres β_0. Utiliser lm() et summary()

Etape5 : Etude des performances du modèle sur l'échantillon d'apprentissage : Créer les valeurs prédite de l'échantillon d'apprentissage nécessaire pour déterminer les coefficient de détermination (R^2 et R^2ajusted). Créer un vecteur avec les valeurs observer de asymPos.nbr et un vecteur avec les valeurs prédite de asymPos.nbr. Réaliser un plot des valeurs prédite en fonction des valeurs observer dans cet échantillon d'apprentissage. Déterminer le coefficient de corrélation entre les valeurs observer et prédite sur cet échantillon s'il est proche de 1 les valeurs sont correctement prédite. Déterminer les coefficient de détermination (R^2 et R^2ajusted). Enfin, calculer RMSEP qui est un indicateur de la moyenne des erreurs du modèle. Utiliser predict(), plot(), cor(), summary().

Etape 6 : Etude des performances du modèle sur l'échantillon test : Prédire les valeurs de asymPos.nbr sur l'échantillon test puis calculer le coefficient de détermination prédit (R^2pred). Calculer les coefficient de corrélation sur l'échantillon test entre les valeurs observées et prédites. Calculer le RMSEP. Utiliser les mêmes fonction que l'étape précédente.

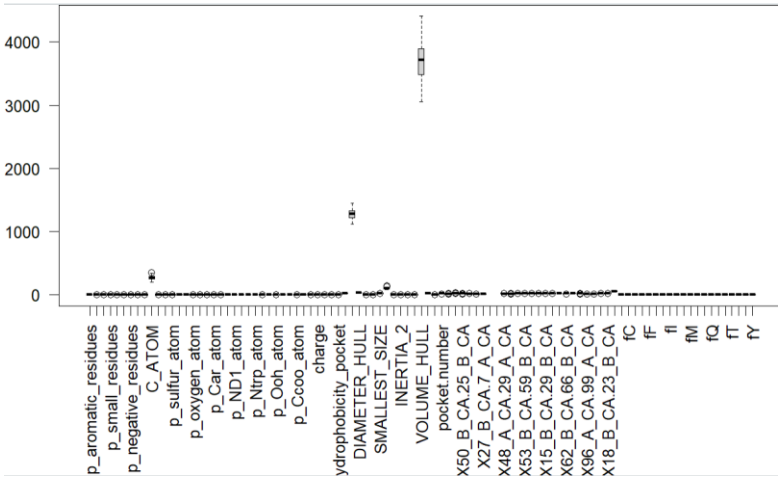
Etape 7 : Etude résidus du modèle : Analyser la distribution des résidus et vérifier la normalités de celui-ci. Calculer l'espérance et vérifier si elle est proche de 0. Analyser si une indépendance et homoscédasticité des résidus est présente. Utiliser hist(), mean(), plot() et summary().

Etape 8 : Sélection des variables les plus significatives et analyse d'un nouveau modèle : Sélectionner les descriptifs significatif et calculer le nouveau modèle.

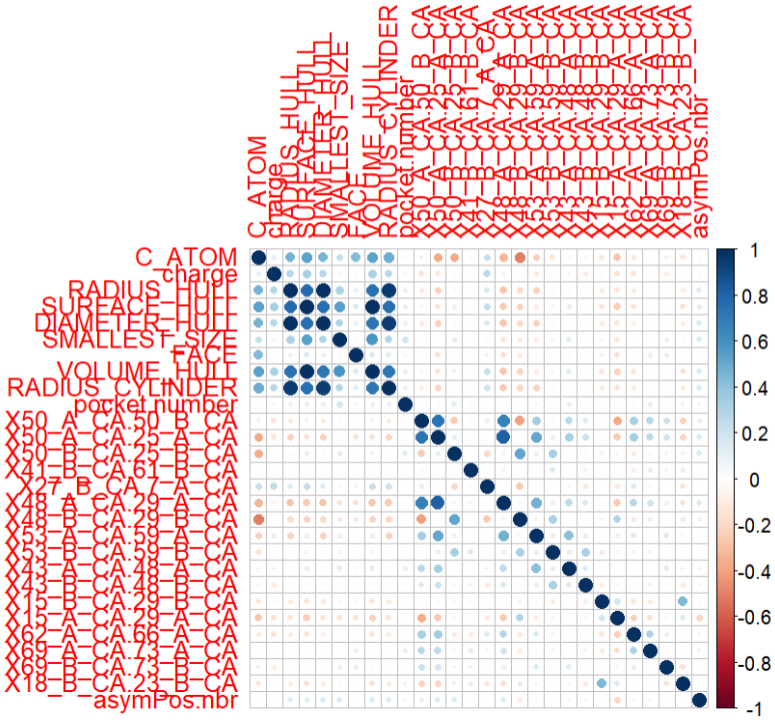
Déterminé les coefficient de détermination et calculer les performances du modèle. Etudier les résidus et calculer la moyenne de celui-ci, analyser l'homoscédasticité et l'indépendance des résidus. Utiliser data.frame(), lm(), summary(), predict(), plot(), cor(), hist() et mean().

Résultats :

Etape1 :



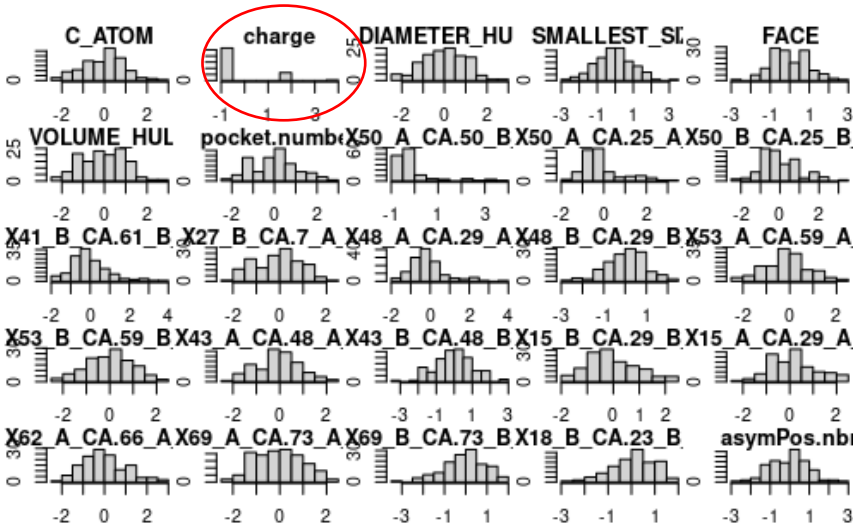
Le boxplot réaliser sur l'ensemble des descriptifs montre que la majorité ont une moyenne proche de 0 avec une faible variance. Excepté pour les variables C_ATOM, SURFACE_HULL et VOLUME_HULL.



L'analyse du corrplot montre que des variables sont fortement corrélées. Par exemple RADIUS_CYLINDER est fortement corrélé à RADIUS_HULL , SURFACE_HULL et DIAMETER_HULL. Le descripteur VOLUME_HULL est également fortement corrélé à ces même descripteurs.

Etape 3 :

Pour l'échantillon d'apprentissage :



Nous observons les distributions des descripteurs de l'échantillon d'apprentissage. Ils suivent une lois normale excepté le descripteur charge entouré en rouge. Nous supprimons donc ce descripteur qui ne suit pas une lois normale. Le même résultat est obtenue pour l'échantillon test.

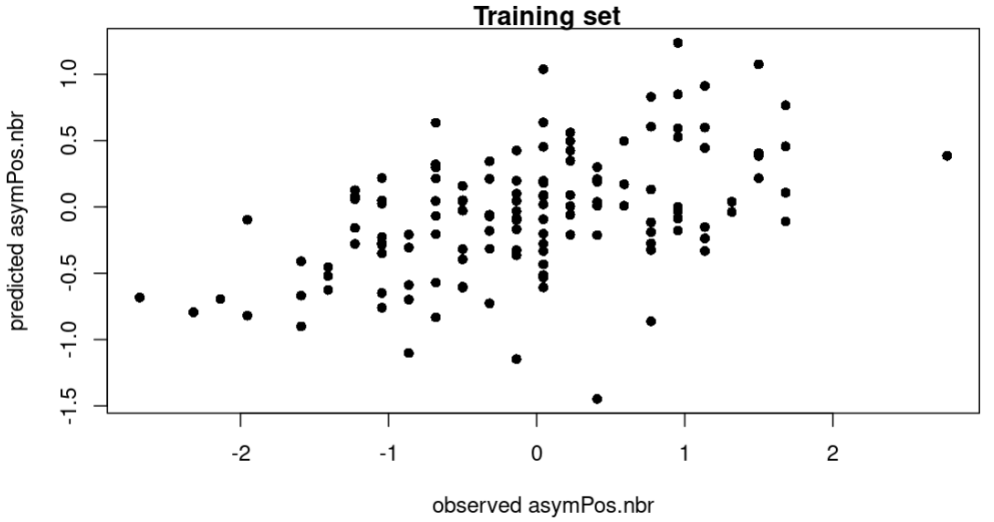
Etape 4 :

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.023014   0.081058  -0.284  0.77701
C_ATOM           0.130705   0.173344   0.754  0.45246
charge          -0.140121   0.096483  -1.452  0.14930
DIAMETER_HULL    0.232469   0.127402   1.825  0.07079 .
SMALLEST_SIZE    0.096302   0.115841   0.831  0.40760
FACE            -0.004705   0.097993  -0.048  0.96179
VOLUME_HULL     -0.058527   0.180003  -0.325  0.74570
pocket.number    0.015965   0.086960   0.184  0.85467
X50_A_CA.50_B_CA -0.080472   0.189834  -0.424  0.67247
X50_A_CA.25_A_CA  0.233158   0.217443   1.072  0.28597
X50_B_CA.25_B_CA  0.070551   0.132906   0.531  0.59661
X41_B_CA.61_B_CA  0.008031   0.086342   0.093  0.92606
X27_B_CA.7_A_CA  -0.053764   0.097702  -0.550  0.58325
X48_A_CA.29_A_CA  0.219001   0.170354   1.286  0.20132
X48_B_CA.29_B_CA  0.017259   0.135279   0.128  0.89871
X53_A_CA.59_A_CA  0.024594   0.118652   0.207  0.83618
X53_B_CA.59_B_CA  0.070626   0.100911   0.700  0.48549
X43_A_CA.48_A_CA -0.158499   0.101869  -1.556  0.12263
X43_B_CA.48_B_CA -0.048964   0.091890  -0.533  0.59522
X15_B_CA.29_B_CA -0.056333   0.103696  -0.543  0.58807
X15_A_CA.29_A_CA -0.030709   0.107360  -0.286  0.77539
X62_A_CA.66_A_CA -0.282166   0.102755  -2.746  0.00706 **
X69_A_CA.73_A_CA  0.061976   0.094059   0.659  0.51135
X69_B_CA.73_B_CA -0.111132   0.088595  -1.254  0.21238
X18_B_CA.23_B_CA  0.077065   0.109643   0.703  0.48363

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.919 on 109 degrees of freedom
Multiple R-squared:  0.2394,    Adjusted R-squared:  0.07188
F-statistic: 1.429 on 24 and 109 DF,  p-value: 0.1104
```

Nous observons β_0 n'est pas significativement différent de 0 entouré en orange. β_1 en bleu est significativement différent de de 0 au risque α=0.01. Les descripteurs DIAMETER_HULL et X62_A_CA.66_A_CA sont significativement différent de 0. R^2 entouré en vert vaux 0.2394 et R^2ajuste entouré en violet vaux 0.07188. Les coefficients de détermination étant inférieur à 1 le modèle n'est pas performant et

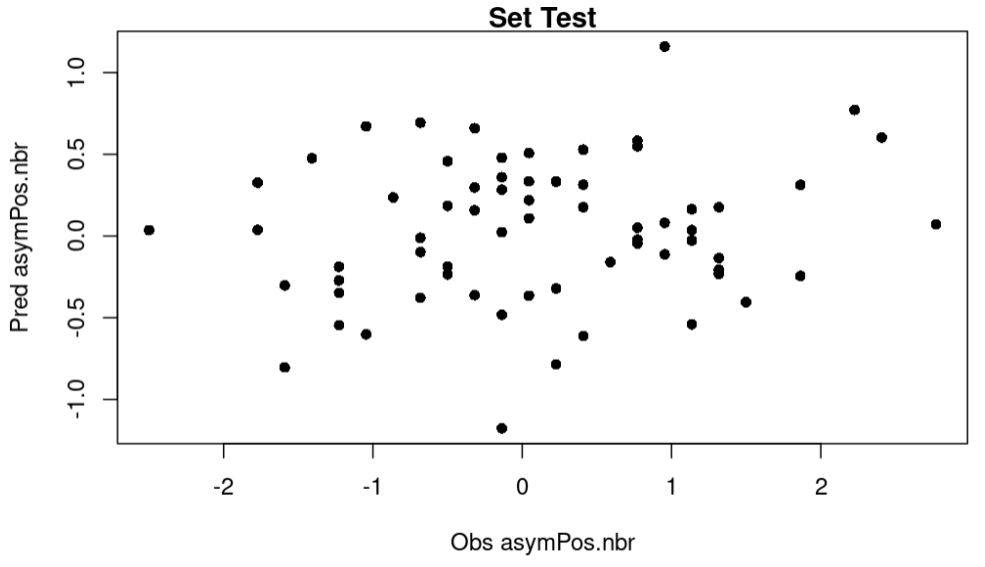
explique seulement 23% de la variabilité des données. Une équation du modèle est asymPos.nbr= -0.282166 [X62_A_CA.66_A_CA].

Etape 5 :



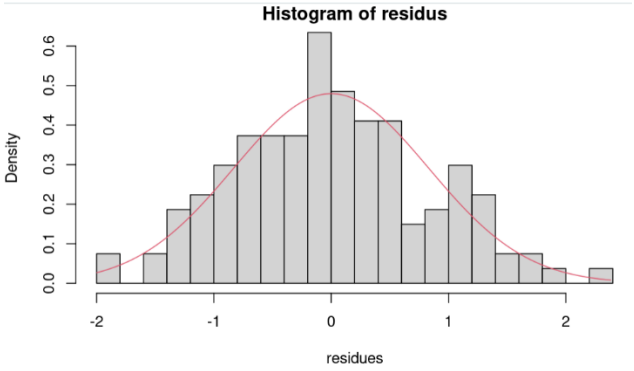
Il y a une dépendance entre les valeurs prédite et observé pour l'échantillon d'apprentissage, il y a un lien positif. La variance des asymPos.nbr prédit n'est pas constante, le coefficient de corrélation vaut 0.49. De plus $R^2 = 0.24$ et le R^2ajusté = 0.07 et RMSEP=0.83 donc le modèle fait beaucoup d'erreur, ce n'est pas un bon modèle. R^2 et R^2ajusté étant très éloigné de 1, le modèle n'est pas performant, il n'a pas bien appris les données.

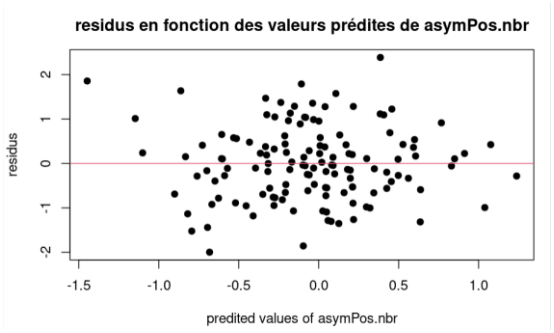
Etape 6 :



Il n'y a pas de dépendance entre les valeurs prédite et observé d'après le plot. Le coefficient de corrélation entre les valeurs prédites et observées vaux 0.16 , étant éloigné de 1, le modèle n'a donc pas un bon pouvoir productif. Nous obtenons R^2pred < 0.5, donc le modèle n'a pas modèle a un bon pouvoir prédictif. RMSEP=1.1, le modèle fait peu d'erreur. Le modèle n'est pas reproductible.

Etape7 :



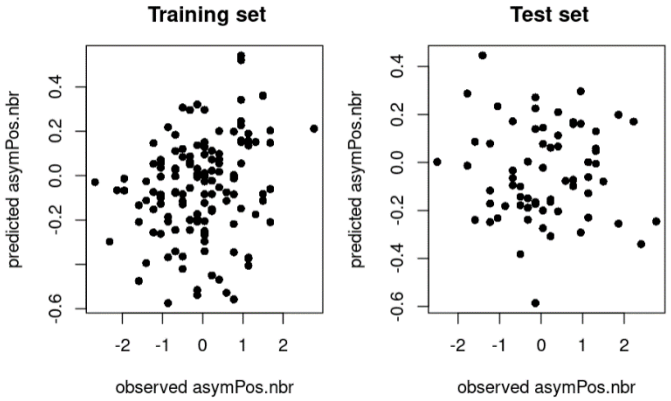


L'histogramme permet de montrer que les résidus suivent une lois normale. La moye nne des résidus vaux -9.149372e-18 étant proche 0 alors les résidus suivent une lois normale centrer en 0. D'après le plot, il ne semble pas exister de dépendance entre l es résidus et le nombre de structure asymétrique prédite, l'homoscédasticité est valid é. La variance des résidus est constante. Les résidus suivent une loi normale de moy enne 0 et de variance sigma².On peut conclure que le modèle est valide.

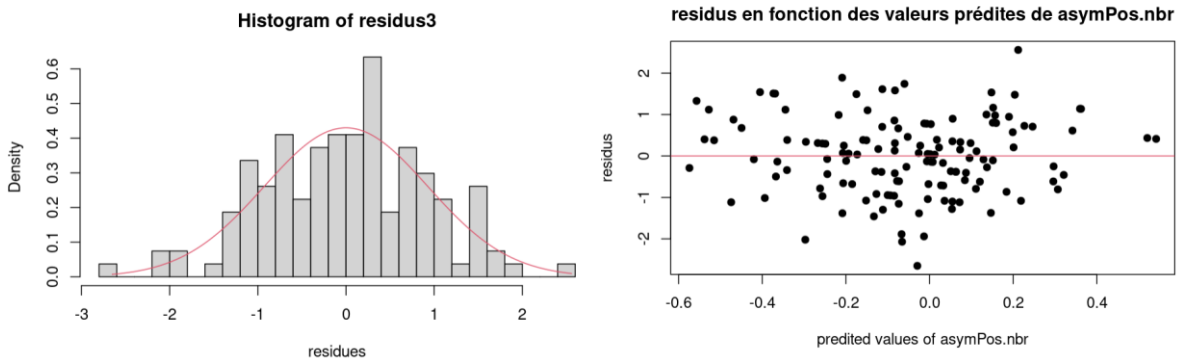
Etape8 :

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.6518 -0.6478  0.0402  0.6738  2.5600

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.04947    0.08078  -0.612   0.5413
X62_A_CA.66_A_CA -0.14092    0.07653  -1.841   0.0678 .
DIAMETER_HUL      0.14727    0.07917   1.860   0.0651 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9348 on 131 degrees of freedom
Multiple R-squared:  0.05407,   Adjusted R-squared:  0.03963
F-statistic: 3.744 on 2 and 131 DF,  p-value: 0.02623
```

Le coefficient de corrélation entre les valeurs prédites et observées pour l'échantillon d'apprentissage vaux 0.23, et pour celui de l'échantillon test -0.03, étant éloigné de 1, le modèle n'a donc pas un bon pouvoir productif. R^2 (en vert) < 1 et R^2ajuste (en violet) <1. Les coefficients de détermination étant inférieur à 1 le modèle n'est pas performant et explique seulement 5% de la variabilité des données. Nous avons β_0 (en orange) qui n'est pas significativement différent de 0 et β_1 = 0.14727 (en bleu) . Donc une équation de ce modèle est asymPos.nbr=0.14727 [DIAMETER_HULL].



Il n'y a pas de dépendance entre les valeurs observé et prédite dans les échantillon d'apprentissage et test.



L'histogramme montre que les résidus suivent une loi normale. La moyenne des résidus est proche de 0, donc les résidus suivent une lois normale centrer en 0. D'après le plot, il ne semble pas exister de dépendance entre les résidus et asymPos.nbr. De plus il semble que la variance des résidus semble constante, les résidus suivent une loi normale de moyenne 0 et de variance $sigma^2$. On peut conclure que le modèle est valide.

Conclusion : Un bon modèle est performant, reproductible, faible complexité, compréhensible et interprétable. Nous avons donc un bon modèle de prédiction. Les facteurs qui influencent sur l'asymétrie structurale de la protéase du VIH-2 sont les descripteurs DIAMETER_HULL et X62_A_CA.66_A_CA donc le diamètre ainsi que la taille du brin supérieur du cantilever des chaînes A et B.