UE SV16Y010 Biostatistiques 5: Estimation, modèles statistiques

Projet 1

Leslie REGAD

Consignes générales

Ce travail est à faire en binôme.

Ce travail doit être rendu **au plus tard le dimanche 05/03/2023 à 23h59 sous moodle**, aucun délai supplémentaire ne sera accordé. Vous devez déposer votre fichier dans le dossier "Projet1".

Ce travail doit être fait avec le logiciel R. Le fichier que vous devez rendre doit être au format PDF (Aucun autre format ne sera toléré). Les consignes pour le fichier à rendre sont les suivantes :

- votre fichier doit se nommer **NOM1__NOM2__projet1.pdf**, avec "NOM1" et "NOM2" correspondant aux noms des deux étudiants du binôme,
- vous devez indiquer les deux noms sur la première page de votre document,
- vous devez numéroter les pages de votre document.
- le fichier PDF doit contenir :
  - les commandes R,
  - les explications des commandes R,
  - la justification des différentes analyses,
  - les sorties R,
  - les graphiques et les tables,
  - la description, l'analyse et l'interprétation des résultats obtenus,
  - les conclusions aux différentes analyses.

Bon courage,

Consignes pour les études statistiques

Pour les différentes parties du projet, la rédaction des tests statistiques doit se faire de la même manière qu'en TD. N'oubliez pas :

- de définir la ou les variables aléatoires,
- de définir les paramètres dans la population et dans les échantillons,
- d'énoncer le test que vous allez réaliser,
- de vérifier les conditions de validité des tests réalisés,
- d'énoncer la règle de décision des tests réalisés,
- de conclure à la question biologique.

Pour les différents tests, vous devez prendre un risque de première espèce de 5%.

Contexte de l'étude et but du projet

Il existe deux types de virus VIH: le virus de type 1 (VIH-1) et celui de type 2 (VIH-2). Aujourd'hui, plus de trois millions de personnes sont infectées par le VIH-2 dans le monde. La distribution du VIH-2 est limitée à l'Afrique de l'Ouest et aux pays ayant des liens historiques avec cette région (France, Portugal, Brésil...). L'arsenal thérapeutique antirétroviral utilisé contre le VIH-2 correspond à celui utilisé contre le VIH-1. Cependant, le VIH-2 est naturellement résistant à certains antirétroviraux dont des inhibiteurs de protéase (IPs). De plus, certaines mutations de résistance ont été observées après traitement antirétroviral correspondant aux IPs chez la protéase du VIH-2 (PR2). Ainsi, avec un arsenal thérapeutique beaucoup plus limité que pour le VIH-1, il est nécessaire de développer de nouvelles molécules thérapeutiques contre le VIH-2. Une approche possible repose sur la recherche de nouvelles molécules inhibitrices spécifiques de PR2. Cette protéine est une cible thérapeutique importante car elle intervient dans la maturation des virus.

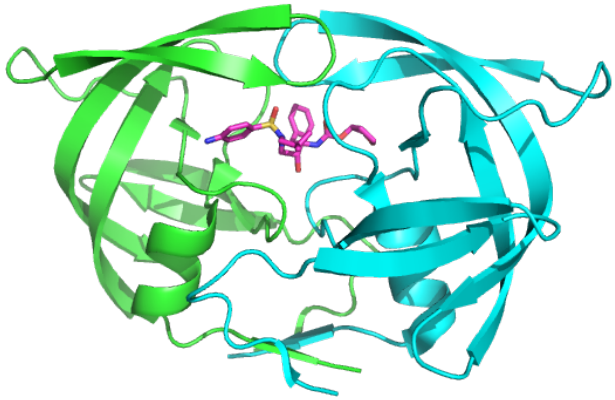

Figure 1: Structure de la PR2 complexée à l'amprenavir

La protéase (PR1 et PR2) est un homodimère de 99 résidus par chaîne (Figure 1). La fixation des substrats et inhibiteurs se fait au niveau d'une poche centrale localisée à l'interface des deux monomères. La fixation d'un ligand dans cette poche entraîne de fort changements structuraux de la protéine : la PR passe d'un état ouvert, permettant la fixation des ligands, à un état fermé qui permet l'hydrolyse des substrats(Menéndez-Arias and Alvarez 2014). L'analyse des structures cristallographiques de PR1 et de celles obtenues par des approches de simulations de dynamique moléculaire (approche bioinformatique qui permet de modéliser l'évolution d'une structure protéique au cours du temps) montre que la PR1 peut adopter 3 conformations (Figure 2) : conformation fermée, conformation semi-ouverte, et conformation ouverte (Mulichak et al. 1993, Tie et al. (2004), Chen et al. (2014), Kar and Knecht (2012), Triki et al. (2018)).

Ces trois états se caractérisent par des conformations particulières des régions flaps qui correspondent à un feuillet β antiparallèle, composé de deux brins, localisé dans la partie supérieure de la protéine. Lorsqu'un ligand vient se fixer sur la PR (substrat ou inhibiteur) cette région vient se refermer sur le site liaison pour favoriser les interactions entre la PR2 et le ligand aboutissant à la forme fermée de la PR. Dans la littérature, il existe plusieurs métriques permettant de différencier la forme de la PR2 (Hornak et al. 2006, Chen et al. (2014), Toth and Borics (2006), Chen et al. (2014), Sadiq and Fabritiis (2010)). La plus utilisée correspond à la distance entre les carbones α des résidus 50 des deux chaînes, notée d_{50}, et mesure l'écartement entre les deux flaps dans un dimère, (Figure 3). Dans leur étude Hornak and al. (2006) ont caractérisé la conformation semi-ouverte de PR1 comme présentant une distance d50 de 4.3 Å et la conformation fermée comme présentant une distance de 5.8 Å (Hornak et al. 2006).

Ces études de fermeture des flaps ont été principalement faites sur des structures de PR1. L'objectif du projet est d'étudier les mécanismes de fermeture des flaps de la PR2. Pour cela, des simulations de dynamique

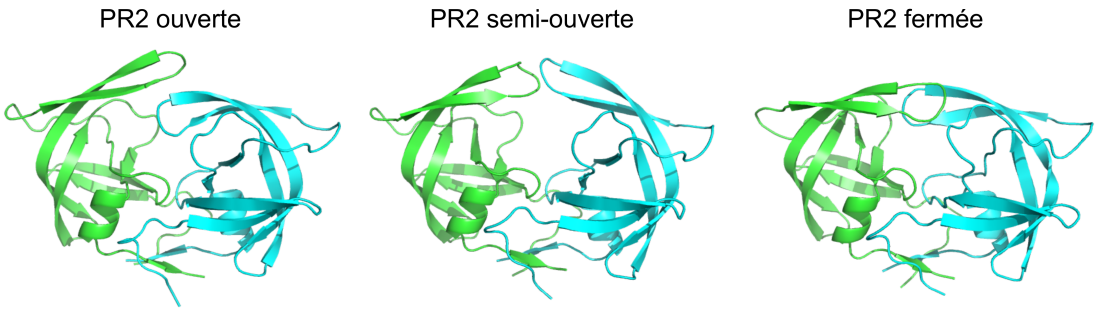

Figure 2: Présentation des trois conformations de la PR2

moléculaire de la PR2 dans différentes conditions ont été lancées. Le but du projet sera de comparer les structures obtenues au cours des simulations en focalisant particulièrement sur la distance $d50$ pour comprendre l'impact de la conformation initiale sur cette distance.

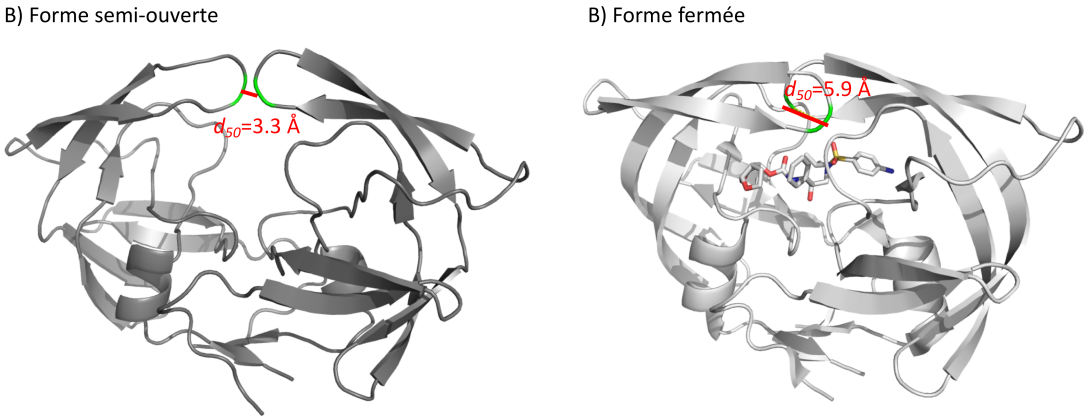

Figure 3: Représentation de la distance d_{50} dans les formes semi-ouverte et fermée de la PR2. Les résidus 50 de chaque chaîne sont colorés en vert et la distance est représentée en rouge.

Présentation des données

La dynamique moléculaire est une technique de simulation numérique permettant de modéliser l'évolution d'un système de particules au cours du temps. Dans ce projet, vous disposez d'un jeu de 2002 structures de la PR2 sauvage extraites de deux trajectoires de simulation de dynamique moléculaire :

- une lancée à partir de la forme semi-ouverte de la PR2 (code PDB 1HSI : PR2 non complexée à un ligand)
- une lancée sur la PR2 complexée au darunavir (code PDB : 3EBZ) où le ligand a été enlevé.

Ces deux simulations ont été réalisées en solvant explicite en utilisant le logiciel Gromacs sur une durée de 1 μs. Les structures sont nommées `trjconv_[3ebz/1hsi]_X.pdb` où `X` correspond au temps d'extraction de la structure : les structures ont été extraites toutes les nanosecondes de chacune des trajectoires.

Calcul des distances $d50$

Pour chaque structure la distance $d50$ a été calculée et stockée dans le fichier `data/d50_3ebz_1hsi.csv`. Ce fichier contient les valeurs de $d50$ pour les 2002 structures extraites des deux simulations :

- structures extaites de la simulation lancée à partir de la structure PDB 1HSI se nomme : `trjconv_1hsi_X.pdb` où X correspond au temps d'extraction en nanoseconde. Par exemple la structure `trjconv_1hsi_498.pdb` a été extraite à la 498 ns de la simulation de 1000 ns.
- structures extaites de la simulation lancée à partir de la structure PDB 3EBZ se nomme : `trjconv_3ebz_X.pdb` où X correspond au temps d'extraction en nanoseconde. Par exemple la structure `trjconv_1hsi_498.pdb` a été extraite à la 498 ns de la simulation de 1000 ns.

Le fichier `data/d50_3ebz_1hsi.csv` contient deux colonnes :

- colonne 1 : nom de la structure. Ce nom vous permettra de récupérer le temps d'extraction et la protéine sur laquelle a été lancée la simulation,
- colonne 2 : la distance $d50$ (en Å).

Calcul des distances $d_{SL-C\alpha}$

Pour caractériser les positions des carbones α par rapport au site de liaison dans une structure, 198 distances, notées $d_{SL-C\alpha}$, ont aussi été calculées (Figure 4). Ces distances correspondent à la distance euclidienne entre les carbones α et le barycentre du superligand (SL). Le SL est un ligand virtuel constitué de ligands co-cristallisés extraits de 28 structures superposées de PR1 et PR2 qui a été utilisé pour estimer la poche commune de liaison aux inhibiteurs des structures de PR (Triki et al. 2018, Laville, Petitjean, and Regad (2021)) de façon indépendante de la taille de l'inhibiteur et en prenant en compte la diversité des ligands et la déformation locale induite par la liaison de l'inhibiteur.

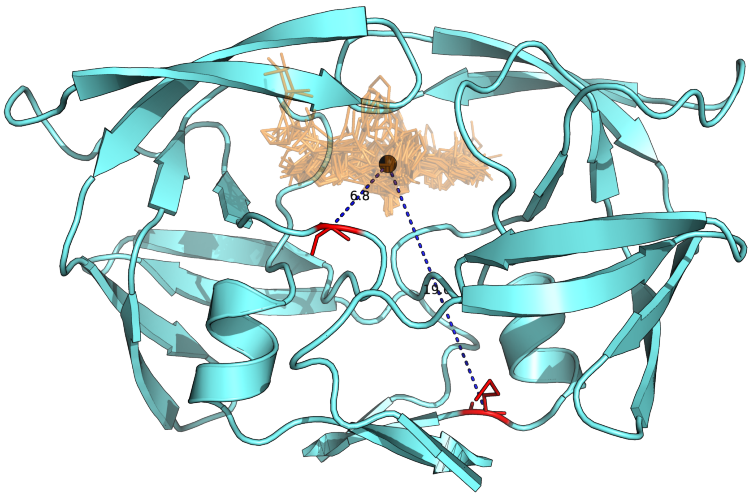

Figure 4: Représentation de la distance $d_{SL-C\alpha}$ des résidus ALA28 ($d_{SL-C\alpha} = 6.8$ Ängstrum) et MET95 ($d_{SL-C\alpha} = 19$ Ängstrum) dans la structure trjconv_3ebz_6.pdb. La structure de PR2 est représentée en mode cartoon et colorée en cyan. Le superligand est représenté en mode sticks et coloré en orange. Le barycentre du superligand est représenté par la sphère noire. Les deux distances $d_{SL-C\alpha}$ illustrées sont représentées par des lignes (tirets) bleues.

Pour calculer les distances $d_{SL-C\alpha}$, le protocole suivant a été utilisé :

1. Superposition optimale des 2002 structures extraites des simulations de dynamique moléculaire à la structure 1HSI (PDB code).

2. Calcul du barycentre du superligand, qui a été défini comme le point de référence pour localiser les atomes dans chaque structure.
3. Calcul de la distance euclidienne entre chaque carbone α et le barycentre du SL.

Pour un carbone $C\alpha$, la distance $d_{SL-C\alpha}$ renseigne sur son emplacement par rapport au site de liaison : plus la distance $d_{SL-C\alpha}$ est grande, plus l'atome $C\alpha$ est éloigné du site de liaison.

Comme la PR2 contient 198 atomes $C\alpha$, 198 distances $d_{SL-C\alpha}$ ont été calculées par structure. Au total, 198×2002 distances $d_{SL-C\alpha}$ ont été calculées à partir des 2002 structures de PR2 extraites des deux simulations. Ces distances $d_{SL-C\alpha}$ ont été stockées dans les fichiers suivants :

- `data/dist_SLCalpha_superlig_1HSI.csv` pour les structures extraites de la simulation lancée à partir de la structure PDB 1HSI,
- `data/dist_SLCalpha_superlig_3EBZ.csv` pour les structures extraites de la simulation lancée à partir de la structure PDB 3EBZ.

Ces fichiers contiennent trois colonnes :

- colonne 1 : nom de la structure,
- colonne 2 : nom du carbone α, note X_A_CA où X correspond au numéro du résidu dans le fichier PDB. Par exemple, l'atome 1_A_CA correspond au carbone $C\alpha$ du résidu 1 de la chaîne A.
- colonne 3 : la distance entre le carbone $C\alpha$ et le barycentre du superligand ($d_{SL-C\alpha}$).

Questions du projet

Partie 1 : Impact de la structure de départ des simulations sur la distance $d50$ et la forme de la PR2

Dans cette première partie, il vous est demandé d'étudier et de comparer la distance entre les deux simulations.

1. Représenter la distribution de la distance $d50$ pour les structures extraites des deux simulations. Commenter les résultats obtenus.
2. Comparaison de la distance $d50$ moyenne dans les deux simulations. Quel est l'impact de la structure de départ sur la distance $d50$ de la PR2 ?
3. Etudier et comparer l'évolution de la distance $d50$ au cours du temps pour les deux simulations.

Partie 2 : Impact de la conformation de départ sur la structure de la PR2

Dans cette partie vous allez comparer les distances $d_{SL-C\alpha}$ des structures extraites des simulations lancées sur 1HSI et 3EBZ (codes PDB).

4. La première étape va consister à supprimer les distances $d_{SL-C\alpha}$ qui ont une variance nulle dans les 2002 structures. Pourquoi supprime-t-on ces distances ? Combien de distances $d_{SL-C\alpha}$ avez-vous supprimées ?
5. Réaliser une analyse en composante principale des 2002 structures à partir de distances $d_{SL-C\alpha}$ avec une variance non nulle. Analyser les résultats obtenus.

L'analyse en composante principale a montré qu'il existait une forte corrélation entre certaines les distances $d_{SL-C\alpha}$. L'étape suivante de l'analyse va consister à supprimer les distances corrélées.

6. Calculer et représenter la matrice de corrélation des distances $d_{SL-C\alpha}$ calculées avec les 2002 structures. Pour la représentation, vous pouvez utiliser la fonction `corrplot()` de la librairie `corrplot()`. Commenter les résultats obtenus.

7. Supprimer les distances $d_{SL-C\alpha}$ ayant un coefficient de corrélation supérieur à 0.85 en utilisant la la fonction `findCorrelation(cutoff = 0.85)` du package `caret`. Pourquoi supprimer ces distances $d_{SL-C\alpha}$? Combien de distances $d_{SL-C\alpha}$ avez-vous supprimées ?

8. A l'aide de PyMOL, réaliser une figure qui localise sur la structure les carbones $C\alpha$ impliqués dans les distances $d_{SL-C\alpha}$ sélectionnées. Commenter la localisation de ces résidus sur la structure de la PR2.

9. A partir des distances $d_{SL-C\alpha}$ sélectionnées, calculer une nouvelle ACP des 2002 structures. Commenter les résultats obtenus.

10. A partir des distances $d_{SL-C\alpha}$ sélectionnées, calculer une classification hiérarchique des 2002 structures. Commenter la classification obtenue.

11. A partir de la classification, extraire 6 groupes en utilisant la fonction `cutree()`. Commenter les différents groupes obtenus. Etudier et commenter la distribution des deux types de structures (structures extraites de la simulation lancée avec 1HSI ou avec 3EBZ) dans les différents groupes.

Remerciements

Merci à Yann Vander Meersche pour la relecture de l'énoncé.

Bibliographie

Chen, J., Z. Liang, W. Wang, C. Yi, S. Zhang, and Q. Zhang. 2014. "Revealing Origin of Decrease in Potency of Darunavir and Amprenavir Against Hiv-2 Relative to Hiv-1 Protease by Molecular Dynamics Simulations." *Sci. Rep.* 4: 6872. doi:10.1038/SREP06872.

Hornak, V., A. Okur, R.C. Rizo, and C. Simmerling. 2006. "HIV-1 Protease Flaps Spontaneously Open and Reclose in Molecular Dynamics Simulations." *Proc. Natl. Acad. Sci. USA* 103: 915–20. doi:10.1073/pnas.0508452103.

Kar, P., and V. J. Knecht. 2012. "Origin of Decrease in Potency of Darunavir and Two Related Antiviral Inhibitors Against Hiv-2 Compared to Hiv-1 Protease." *Phys. Chem. B.* 116: 2605–14. doi:10.1021/JP211768N.

Laville, P., M. Petitjean, and L. Regad. 2021. "Structural Impacts of Drug-Resistance Mutations Appearing in Hiv-2 Protease." *Molecules* 26: 611. doi:doi: 10.3390/molecules26030611.

Menéndez-Arias, L., and M. Alvarez. 2014. "Antiretroviral Therapy and Drug Resistance in Human Immunodeficiency Virus Type 2 Infection." *Antiviral Res.* 102: 70–86. doi:10.1016/j.antiviral.2013.12.001.

Mulichak, A. M., J.O. Hui, A.G. Tomasselli, R.L. Heinrikson, K.A. Curry, C.S. Tomich, S. Thaisrivongs, T.K. Sawyer, and K.D. Watenpaugh. 1993. "The Crystallographic Structure of the Protease from Human Immunodeficiency Virus Type 2 with Two Synthetic Peptidic Transition State Analog Inhibitors." *J. Biol. Chem.* 268: 13103–9.

Sadiq, S. K., and G. de Fabritiis. 2010. "Explicit Solvent Dynamics and Energetics of Hiv-1 Protease Flap Opening and Closing." *Proteins Struct. Funct. Bioinforma.* 78: 2873–85.

Tie, Y., P.I. Boross, Y.F. Wang, L. Gaddis, A.K. Hussain, S. Leshchenko, A.K. Ghosh, J.M. Louis, R.W. Harrison, and I.T. Weber. 2004. "High Resolution Crystal Structures of Hiv-1 Protease with a Potent Non-Peptide Inhibitor (Uic-94017) Active Against Multi-Drug-Resistant Clinical Strains." *J Mol Biol.* 23: 341–52. doi:10.1016/j.jmb.2004.02.052.

Toth, G., and A. Borics. 2006. "Closing of the Flaps of Hiv-1 Protease Induced by Substrate Binding: A Model of a Flap Closing Mechanism in Retroviral Aspartic Proteases." *Biochemistry* 45: 6606–14.

doi:doi.org/10.1021/bi060188k.

Triki, D., T. Billot, D. Flatters, B. Visseaux, D. Descamps, A.C. Camproux, and L. Regad. 2018. "Exploration of the Effect of Sequence Variations Located Inside the Binding Pocket of Hiv-1 and Hiv-2 Proteases." *Sci. Rep.* 8: 5789. doi:DOI:10.1038/s41598-018-24124-5.