Université Paris Cité

Projet Long - Les affimers

Mohamed Oussarren

Master 2 Biologie-Informatique

15 janvier 2023

Université Paris Cité

Table des matières

1 Introduction

1.1 Contexte

Les virus sont responsables d'innombrables maladies dans le monde. Dans ce que la gestion de la COVID-19 a démontré, ils pouvaient avoir des conséquences dramatiques que ce soit sur le plan sanitaire ou bien même économique. Parmi ces virus se trouvent les Flavivirus qui peuvent provoquer la Dengue, le virus Zika et bien d'autres maladies virales. Ces maladies ont certes un faible taux de mortalité mais actuellement aucun vaccin ou traitement antiviral spécifique n'est efficace, cela laissant les individus atteints bénéficier seulement de soins de soutiens pour en guérir. Le système immunitaire est donc livré à lui-même.

Le système immunitaire est un ensemble extrêmement sophistiqué de mécanismes de reconnaissance et de défense de l'organisme. Sa reconnaissance se base sur des molécules appelées antigènes. Notre organisme utilise ces antigènes qui permettent la liaison avec des anticorps. Les anticorps jouent un rôle important car ils permettent de détecter et neutraliser les agents pathogènes ainsi les empechant de pénétrer ou d'endommager nos cellules. La figure 1 s'interesse au mecanisme de reconnaissance de la COVID-19 par biais d'anticorps spécifiques et illustre bien le rôle d'un anticorp. Cette reconnaissance s'explique notamment par sa structure.

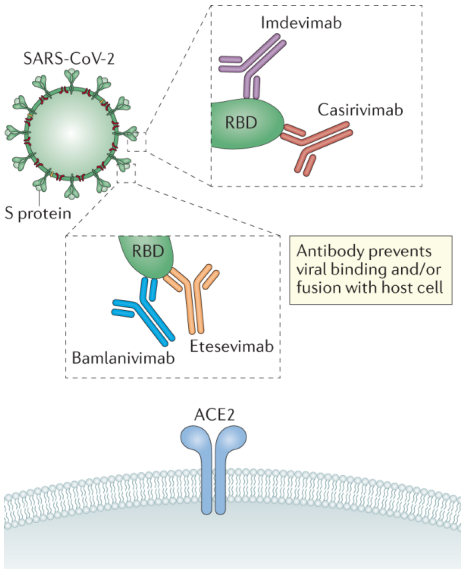

Figure 1 : Inhibition du SRAS-COV-2 par les anticorps monoclonaux [1].

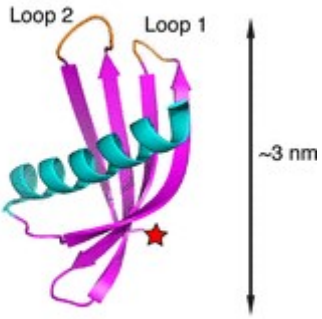

Figure 2 : Structure 3D d'un Affimer [2].

Les anticorps sont constitués de deux chaînes lourdes et de deux chaînes légères qui sont disposées en forme de Y et constituent la partie constante de l'anticorps comme représenté sur la figure ci-dessus. Un paratope, partie variable de l'anticorps, qui est une zone de reconnaissance spécifique d'une partie d'un antigène appelée épitope, est situé à l'extrémité de chaque branche. Nous distinguons ensuite les anticorps monoclonaux des anticorps polyclonaux, qui constituent ensemble une famille d'anticorps tous liés à un épitope différent d'un même antigène. Ces derniers proviennent du même globule blanc qui a été cloné pour produire des anticorps identiques qui ne reconnaissent qu'un seul épitope d'un antigène.

Récemment, un candidat pour les anticorps monoclonaux sous la forme de peptides connus sous le nom d'*Affimers* [7] a été identifié. Ces molécules ont été créées pour avoir une spécificité plus élevée que les anticorps avec une meilleur stabilité et des prix moins chers. Il s'agit d'une petite protéine constituée d'une hélice

alpha et de feuillet bêta comme représenté sur la figure 2. Ils possèdent deux boucles qui peuvent être très spécifiques et qui sont positionnées de sorte qu'elles n'ont aucun effet sur la stabilité de la protéine. Cela s'explique par leur taille plus petite la rendant très accessible et peut être produit en grand nombre. En plus de cela, il n'est pas nécessaire d'utiliser d'animaux[6] pour produire des affimers car ils peuvent être produits dans des cellules bactériennes tout en restant fonctionnels même dans des cellules de mammifères. En outre, la procédure est plus rapide et moins coûteuse, ce qui les rend prometteurs pour une utilisation généralisée.

1.2 Objectif

Pour ce travail, nous avions pour objectif de développer un programme automatisé permettant de modéliser des Affimers dont leur but est de reconnaitre des protéines de Flavivirus spécifiquement identifiés. Nous nous sommes interessé aux protéines de type non structural (NS) [8] qui sont responsables de la réplication virale et nous allons cherché à modifier notre affimer de sorte qu'il puisse reconnaitre ces protéines. Pour ce faire, nous utiliserons des outils informatique et bio-informatique pour identifier les sites d'interactions des anticorps spécifiques à une protéine NS. Ces sites d'interactions seront remplacés dans les boucles d'un affimer sélectionné. Nous proposerons donc un programme automatisé permettant la création d'un affimer modifié. Celui-ci est disponible via le lien suivant : https://github.com/Moohmoo/Affimer.

2 Matériels et méthodes

2.1 Jeux de données

Afin de concevoir notre affimer, il est nécessaire d'avoir un ensemble de jeux de données constituées d'ensemble d'anticorps complexé à un antigène. Cela va permettre d'identifier les chaînes des sites en contacts. Les fichiers utilisés seront donc les fichiers de la Protein Data Bank (PDB) qui ont été annotés et faissant partie de la base de données connue sous le nom de l'*AbDb* [9]. Ils sont choisis en utilisant le *Summary of Antibody Crystal Structures* (SACS) [10] comme point de départ, et passent ensuite par une annotation gardant que les chaînes pertinentes pour l'interaction. Dans l'*AbDb*, nous avons choisi d'utiliser le jeu de données Martin qui est non redondant et ne contient que des anticorps complets. Martin étant la méthode d'annotation la plus fiable. Nous avons dénombré 1205 anticoprs compléxés à un ou des antigène.s.

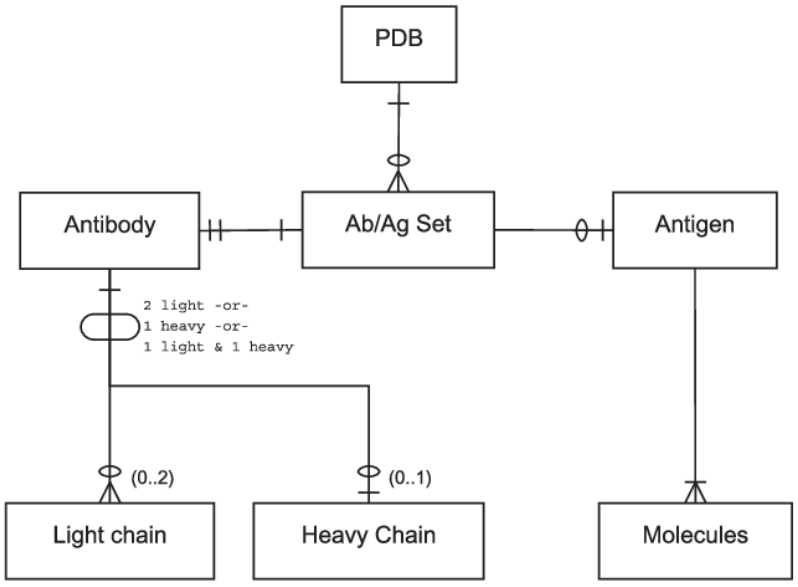

Figure 3 : Diagramme des relations des entités de la base de donnée [3].

2.2 Proximité des résidus

Afin de déterminer les sites de fixation, nous avons cherché à déterminer les atomes de résidus d'une chaîne d'anticorps interragissant avec au moins un autre atome d'une chaîne de l'antigène. Nous avons donc cherché à estimer les distances euclidienne d'atomes appartenant à des chaines distincts d'anticorps et antigènes.

$$d = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2}$$

En appliquant la formule ci-dessus avec b correspond aux coordonnées des atomes associés aux anticorps et a aux coordonnées des atomes d'antigènes, nous avons gardé seulement les résidus dont la distance d'atomes dépasse 5.0 Angström. Tous les residus inclus seront sauvegardés dans un fichier pdb spécifiquement annoté.

2.3 Patch Search

Nous avons utilisé un outil de comparaison structurelle pour les protéines appelé PatchSearch [11, 12]. Il s'agit de la pièce maitresse de ce projet car c'est celui-ci qui nous a permis d'identifier les sites de liaison ayant une structure comparable à celle de NS1. Patch Search cherche à comparer un peptide donné avec la surface de protéines d'une base de données pour déterminer quelles protéines ont le plus de similarité structurelle. Nous avons récupéré cet outil de manière à ce qu'il compare la surface d'une protéine donnée, ici NS1, avec une base de données de sites de liaisons que nous avons générée. Patch Search renvoit un fichier contenant un score pour chaque comparaison fait avec les sites de liaisons. Nous prendrons ceux dont le score est meilleur. Par cette sélection, nous remplacerons les boucles des affimers par les sites identifiés.

2.4 Modélisation de nos affimers et alignement 3D

Les modèles structuraux en format PDB ont été obtenus par les trois outils suivants :

1. Swiss Model [13] est un outil en ligne permettant la mod´elisation d'une protéine par homologie. Son principe est d'organiser le squelette de la séequence identiquement à celui d'une autre séquence ayant une identité supérieure ou égale à 50

2. Modeller est aussi un outil de modélisation par homologie mais ayant une procédure différente à celle de Swiss model. Il mime le principe des RMN des protéines par laquelle un ensemble de critères géométriques sont utilisés pour créer une fonction de densité de probabilité pour l'emplacement de chaque atome dans la protéine [4]. A l'inverse de Swiss model, il est nécessaire de préciser un support homologue.

La visualisation des structures 3D se font à l'aide de Pymol. Il va permettre d'expliquer et de mettre en évidence la structure des modèles.

2.5 Optimisation du codes

Compte tenu de la taille de nos jeux de données, il est impératif d'optimiser le code pour permettre une amélioration significative des performances et donc d'accélérer le processus de création d'affimer. La majorité de nos codes ont été écrit en C, un langage compilé. Nous proposons plusieurs niveaux de parallélisations : une parallélisation au niveau des threads et une parallélisation au niveau vectorielles.

— Parallélisation des threads : OpenMP [14] est un module permettant la parallélisation par un principe de partage de mémoire. Un espace de mémoire dans le CPU est mise en commun pour limiter la redondance des calculs des données. À partir de cela, OpenMP va définir des régions parallèles c'est-à-dire des portions de codes destinées à être exécutées en parallèle.

— Parallélisation vectorielle : Dans le cas de notre étude, le concept de vectorisation consiste à appliquer une seule et même opération sur deux ensembles de valeures distinctes. Dans notre cas, des calculs de distances seront effectués afin d'évaluer la proximité des résidus. La vectorisation des calculs va donc permettre une augmentation considérable en terme de temps de calcul compte tenu de la taille de nos jeux de données.

Une optimisation a notamment été opéré sur des fonctions. Par exemple, des fonctions permettant la conversion de chaines de caractères en valeurs numériques. Seul le code associé à Patch Search n'a pas été optimisé en raison de sa complexité et dû au manque de temps. Il constitue à lui seul la majorité du temps de calcul du programme.

3 Résultats

3.1 Banques de sites de liaison

À l'aide des 1205 fichiers pdb de la base de données *AbDb* nous avons pus construire la banque de sites de liaisons. Un exemple de structure de complexe anticorps-antigène est illustré par la figure 4. Nous avons récupéré les atomes proches pour chaque antigène et chaîne d'anticorps. Nous obtenons plus de 2 fichiers différents. Notre programme a obtenu 2553 sites de liaison. La majorité d'entre eux sont constitués d'une centaine atomes. La chaîne d'antigène qui interagit avec la chaîne d'anticorps est spécifiée dans le nom de fichier suivant un format spécifique.

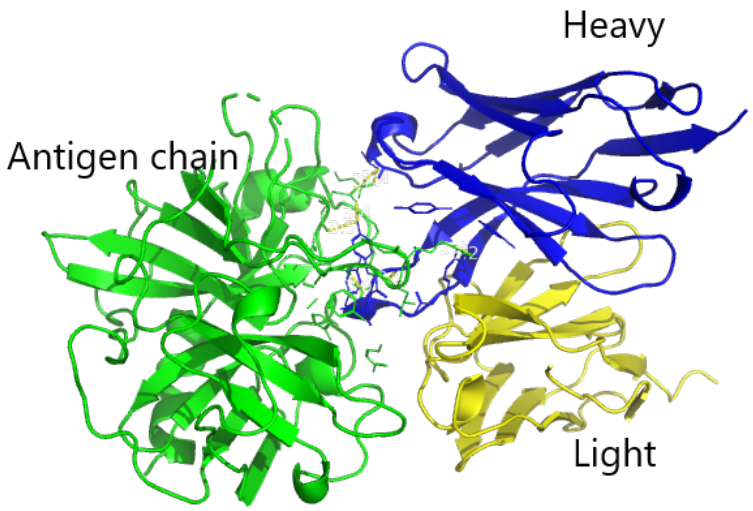

Figure 4 : Structure du complexe 6R8X_1 issue de la base de donnée AbDb

3.2 Recherche de similarité structurale avec Patch Search

Le lancement de Patch Search nécessite de préciser la protéine à comparé avec les sites de liaisons. Pour cela, nous avons comparé à une protéine NS1 obtenue à partir de la PDB. Nous avions besoin d'une protéine sans anticorps associés, avec une bonne résolution. Notre décision s'est donc basée sur la protéine NS1 d'une variation de Zika qui a été identifiée au Brésil et son fichier pdb est 5GS6. Ici, nous n'avons comparé que la partie antigène. L'utilité d'annoter de manières spécifiques les fichiers des sites de liaisons permet de faciliter la

considération de l'atome à choisir par Path Search. PatchSearch a envoyé un fichier contenant plusieurs scores après l'avoir comparé à l'ensemble de la banque. Le score permet de discrimer selon si des résidus d'un site de liaison s'alignent de manière similaire dans la NS1. Par la suite, nous avons regardé la distribution des scores illustrée dans la figure 5.

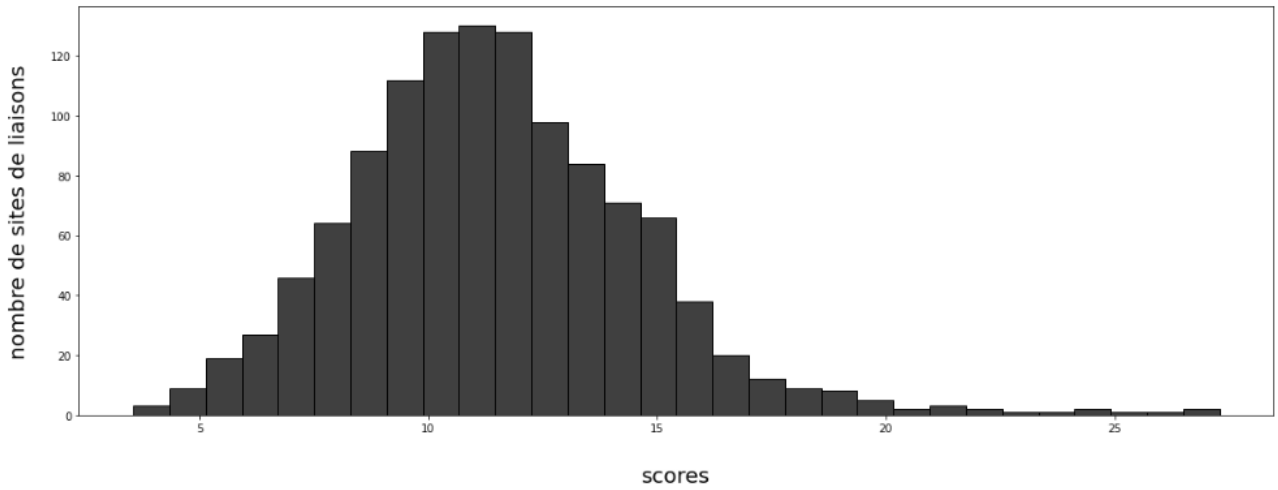

Figure 5 : Distribution des scores obtenus par Patch Search

Nous avons dénombré un total de 1923 scores soit 630 sites de liaisons de moins qu'au départ. Le choix des meilleurs sites de liaisons s'est fait de manière empirique. Nous avons ensuite récupéré cette fois-ci la partie anticorps. Ainsi, pour chacun d'entre eux, notre script a parcouru les atomes de la partie anticorps et a récupéré des ensembles de résidus subséquents qui comprenaient au moins un atome dans le site de liaison. Ils remplaceront les boucles de l'affimer sélectionné et dont leur longueur sera choisie spécifiquement en fonction de la longueur de la boucle pour qu'une boucle puisse interagir. Ici, nous avons choisi l'affimer 7NY8. Nous avons détecté 3 régions d'interactions grâce à Patch Search à savoir aux positions 2-7, 41-44 et 75-80. Ces résidus ont été subsitués par les résidus suivants 'WHIN', 'MYAP', 'ACGLLR'.

3.3 Modélisation par homologie

L'objectif de cette partie consiste dans un premier temps en l'obtention des modèles structuraux par les différents outils présentés dans la partie *Matériels et Méthodes* et dans un deuxième temps à l'évaluation de ces modèles. Les différentes figures suivantes permettent de mettre en évidence la comparaison structurelle de nos affimers.

3.3.1 Modélisation par homologie : Swiss Model

En utilisant Swiss Model, nous avons pu obtenir une structure 3D représentant notre protéine.

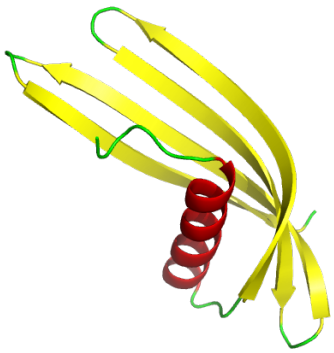

Figure 6 : Structure 3D de l'affimer modifié obtenue avec SwissModel

Le modèle obtenu est conforme à nos attentes. Nous retrouvons bien les feuillets bêta, les héciles ainsi que les boucles. L'avantage de SwissModel est qu'il propose un score, le QMean Z-score [5], permettant de d'écrire l'aspect géométriques des structures. La valeur obtenue pour notre modèle est supérieure à 0 affirmant qu'il s'agit d'un bon modèle. D'autres paramètres permettent d'affirmer la robustesse de ce modèle.

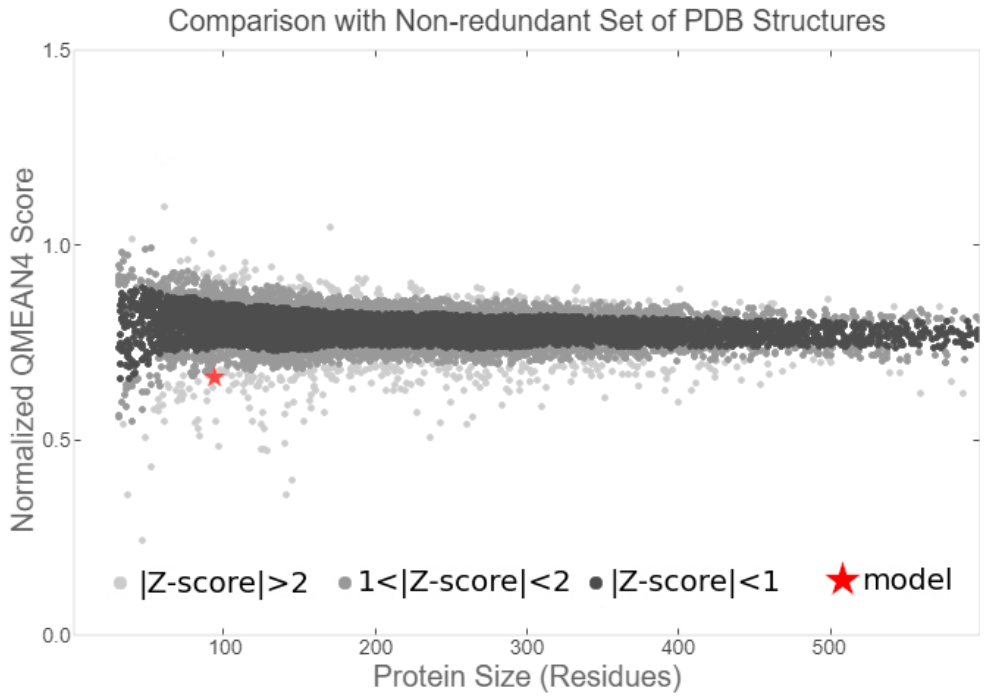

Figure 7 : Plot représentant la valeur des Z-score des protéines

La figure 6 permet d'affirmé ce qui a été dit précédemment. Ici notre modèle est représenté par l'étoile rouge et se situe parmi les points de couleurs gris clair. Plus le Z-score est élevé et plus le modèle est significatif c'est-à-dire est le plus vraisemblable. Ici le Z-Score du modèle est compris entre 1 et 2 ce qui affirme à nouveau qu'il s'agit d'un bon modèle.

3.3.2 Modélisation par homologie : Modeller

L'application de Modeller nécessite un alignement global par paire à l'aide de l'algorithme de Needleman-Wunsch. L'utilisation de cet algorithme nécessite l'entrée de la séquence cible et de la séquence support qui ont été alignées préalablement. La modélisation a permis d'obtenir un ensemble de structures. Notre ensemble devrait être composé de 50 modèles. Ce choix peut varier en fonction du nombre d'itérations choisi. La sélection du modèle s'est faite en fonction d'un score : le score DOPE (Discrete optimized protein energy). Ce score est utilisé pour discriminer les modèles. Le score le plus faible correspond au meilleur modèle parmi tous. Malheuresement, par manque de temps nous n'avons pas pus obtenir un modèle à l'aide de Modeller en raison d'un manque de temps.

4 Conclusion

Dans l'ensemble, ce projet a été une réussite grâce à nos méthodes et à nos expériences malgré tous de même un manque de temps considérable. Les scripts permettent la création des bases de données des sites de liaisons utilisés pour PatchSearch et peuvent être facilement utilisable pour d'autres cas. Pour l'utilisation de ce programme il est nécessaire d'avoir suffisament d'informations structurales sur nos protéines utilisées. Le résultat n'est pas encore tout à fait au point, mais après seulement quelques essais, il va dans la bonne direction. L'étape suivante consisterait à effectuer des simulations de dynamique moléculaire pour voir clairement les interactions,

ce qui confirmerait que la robustesse de nos affimers. Nous aurions pu aussi évaluer par un docking notre affimer et la NS1 notamment à l'aide d'HDOCK et notamment ajouter des paramètres quantifiant la stabilité électrostatique.

Références

[1] Taylor, P.C., Adams, A.C., Hufford, M.M. et al. Neutralizing monoclonal antibodies for treatment of COVID-19. Nat Rev Immunol 21, 382–393 (2021). https ://doi.org/10.1038/s41577-021-00542-x

[2] Carrington G, Tomlinson D, Peckham M. Exploiting nanobodies and Affimers for superresolution imaging in light microscopy. Mol Biol Cell. 2019 Oct 15 ;30(22) :2737-2740. doi : 10.1091/mbc.E18-11-0694. PMID : 31609674 ; PMCID : PMC6789155.

[3] Ferdous, Saba and Andrew C. R. Martin. "AbDb : antibody structure database—a database of PDB-derived antibody structures." Database : The Journal of Biological Databases and Curation 2018 (2018) : n. pag.

[4] Tunyasuvunakool, K., Adler, J., Wu, Z. et al. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596 (2021). https ://doi.org/10.1038/s41586-021-03828-1

[5] Benkert P, Tosatto SC, Schomburg D. QMEAN : A comprehensive scoring function for model quality assessment. Proteins. 2008 Apr ;71(1) :261-77. https ://pubmed.ncbi.nlm.nih.gov/17932912/. PMID : 17932912.

[6] Birch, J. R., Racher, A. J. (2006). Antibody production. In Advanced Drug Delivery Reviews (Vol. 58, Issues 5–6, pp. 671–685). Elsevier BV. https ://doi.org/10.1016/j.addr.2005.12.006

[7] Klont, F., Hadderingh, M., Bischoff, R. (2018). Affimers as an Alternative to Antibodies in an Affinity LC–MS Assay for Quantification of the Soluble Receptor of Advanced Glycation End-Products (sRAGE) in Human Serum. In Journal of Proteome Research (Vol. 17, Issue 8, pp. 2892–2899). American Chemical Society (ACS). https ://doi.org/10.1021/acs.jproteome.8b00414

[8] Caraballo, G. I., Rosales, R., Viettri, M., Ding, S., Greenberg, H. B., Ludert, J. E. (2021). The dengue virus non-structural protein 1 (NS1) interacts with the putative epigenetic regulator DIDO1 to promote flavivirus replication. Cold Spring Harbor Laboratory. https ://doi.org/10.1101/2021.09.01.458517

[9] Ferdous S, Martin ACR. AbDb : antibody structure database-a database of PDB-derived antibody structures. Database (Oxford). 2018 Jan 1 ;2018 :bay040. doi : 10.1093/database/bay040. PMID : 29718130 ; PMCID : PMC5925428.

[10] Allcorn, L. C., Martin, A. C. R. (2002). SACS–Self-maintaining database of antibody crystal structure information. In Bioinformatics (Vol. 18, Issue 1, pp. 175–181). Oxford University Press (OUP). https ://doi.org/10.1093/bioinformatics/18.1.175

[11] I. Rasolohery, G. Moroy, F. Guyon PatchSearch : a fast computational method for off-target detection, Journal of chemical information and modeling, 2017

[12] J. Rey, I Rasolohery, P. Tufféry, F. Guyon, G. Moroy PatchSearch : a web server for off-target protein identification, Nucleic acids research, 2019

[13] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, Torsten Schwede, SWISS-MODEL : homology modelling of protein structures and complexes, Nucleic Acids Research, Volume 46, Issue W1, 2 July 2018, Pages W296–W303, https ://doi.org/10.1093/nar/gky427

[14] Instructions about OpenMP : https ://learn.microsoft.com/fr-fr/cpp/parallel/openmp/reference/openmp-directives ?view=msvc-170