**Projet long :**

**Impact de l’épissage alternatif sur la structure protéique**

PRETET Maël

Encadrant : M. Alexandre DE BREVERN

Université de Paris

M2 BI 2020-2021

[Introduction 3](#_Toc67073391)

[Matériel et méthodes 3](#_Toc67073392)

[Base de données 3](#_Toc67073393)

[Clustal Omega [1] 3](#_Toc67073394)

[PSIPRED [2] 4](#_Toc67073395)

[DISOPRED [3] 5](#_Toc67073396)

[Résultats 5](#_Toc67073397)

[Clustal Omega 5](#_Toc67073398)

[Psipred 7](#_Toc67073399)

[Disopred 8](#_Toc67073400)

[Discussion 9](#_Toc67073401)

[Conclusion 9](#_Toc67073402)

[Bibliographie 9](#_Toc67073403)

# Introduction

Depuis maintenant plusieurs années, le nombre de gènes identifiés chez l’humain codant pour une protéine est estimé à environ 20 000. Cependant, il est connu également que le nombre de type de protéines retrouvées dans ce même organisme dépasse largement cette donnée : celui-ci est en effet estimé à plusieurs centaines de milliers. Ce phénomène est expliqué principalement par l’épissage alternatif : il a été montré que les gènes des organismes eucaryotes étaient organisés en succession d’exon et d’intron. Lors de la transcription, ce sont uniquement les premiers qui vont être conservés lors de l’épissage : les introns seront alors supprimés et l’ARN messager ainsi obtenu pourra être transcrit. Cependant, il a été observé que lors de cet épissage, le même ARN messager pouvait ne pas être obtenu deux fois à partir du même gène initial. En effet, une réorganisation des exons est possible lors du processus de création des ARNm, ce qui même donc à la traduction d’une protéine à partir d’une séquence différente, appelée isoforme. Ce phénomène a été appelé épissage alternatif, et a été observé en 1977 [5] pour la première fois. Celui-ci permet d’augmenter de façon importante la diversité du protéome, notamment chez l’humain, ce qui explique la différence considérable entre gènes identifiés et protéines connues notée précédemment. Cependant, cela a ouvert la voie à un nombre de recombinaison possible tout aussi important, et si aujourd’hui il est plus aisé de connaître l’ensemble d’un génome humain, il devient bien plus complexe d’analyser l’ensemble de son protéome, et les estimations du nombre de protéines totales n’ont ainsi fait qu’augmenter au cours des années. Il est reconnu qu’actuellement, une partie du protéome humain est encore à découvrir, autant par les protéines qui le composent que par la diversité fonctionnelle que l’épissage alternatif implique.

Si ce phénomène intéresse aujourd’hui autant, c’est notamment dû à son facteur explicatif de différentes maladies : en effet, de nombreux sites subissant l’épissage alternatif sont retrouvés comme étant impliqués, par exemple dans la mucoviscidose, dont certaines des mutations retrouvées dans la maladie affectent directement les phénomènes d’épissage.

Certaines études étudiant l’impact de l’épissage alternatif sur la diversité fonctionnelle chez un organisme proposent aussi ce phénomène comme mécanisme lié à la pression évolutive [6], permettant à des organismes dont le temps de vie est considéré comme long par rapport à d’autres (organisme eucaryotes multicellulaires, par rapport à des procaryotes par exemple) d’obtenir une plus grande diversité génétique potentielle à partir d’un gène qui n’évoluera que peu, pour ainsi réduire une potentielle pression sélective.

Avec l’épissage alternatif, un autre phénomène lié à la structure des protéines a pu être corrélé : le désordre. En effet, si certaines régions des protéines sont rigides et ne modifient pas leurs conformations une fois repliées, d’autres sont au contraire très libres, et possèdent une flexibilité importante et essentielle à leur rôle protéique. Ainsi, plus de 30% du protéome humain au moins est estimé comme étant intrinsèquement désordonné, une caractéristique qui complique évidemment la résolution expérimentale des structures de ces protéines. Mais, comme énoncé précédemment, ce phénomène a été étudié en lien avec l’épissage alternatif. En effet, un article [6] a pu mettre en évidence une augmentation significative de région intrinsèquement désordonnée dans des régions protéiques subissant un épissage alternatif, comparée à l’ensemble de la protéine. La flexibilité de ces régions pourrait donc être corrélée avec leur propension à subir ce mécanisme. Il serait donc intéressant d’observer si les isoformes issus de l’épissage alternatif possèdent un désordre intrinsèque différent par rapport à la protéine initiale (nb : la protéine issue de la traduction de tous les exons ordonnés, et donc la plus longue), et si la structure secondaire de ces isoformes pourraient être impactée par l’épissage alternatif.

# Matériel et méthodes

## Base de données

Les analyses effectuées ont été menées sur un jeu de données obtenu via Uniprot[[1]](#footnote-1). Celui-ci a été sélectionné selon différents critères. Tout d’abord, les protéines retenues sont uniquement celles dont certains isoformes sont connus, et présentes dans le protéome humain. Une deuxième sélection a ensuite été effectuée, pour retenir des protéines dont une partie au moins de la séquence a été résolue, et présente sur la Protein Data Bank.

Enfin, pour des raisons techniques liées à l’utilisation en local de différents algorithmes comme psipred ou disopred, une autre sélection a été effectuée sur ce jeu de données : entre 10 et 15 protéines ont été retenues de manière aléatoire (ainsi que leurs isoformes, pour comptabiliser une centaine de séquences dans le jeu de données final). Les analyses effectuées ci-dessous ont donc pour objectif avant tout de présenter un pipeline de différents scripts pouvant être appliqués sur des bases de données plus importantes : en effet l’échantillonnage fait de manière aléatoire ne doit pas être oublié comme pouvant biaiser les résultats obtenus sur des données plus importantes.

## Clustal Omega [1]

Clustal Omega est un algorithme d’alignement de séquences, utilisés lorsqu’il est nécessaire d’aligner plus de deux séquences nucléiques ou protéiques. Dans ce cas d’application, étant donné que ces sont des protéines résultant de l’épissage alternatif qui sont étudiées, il semble approprié d’utiliser ce type d’alignement. L’algorithme de Clustal Omega est utilisé en local, et avec différents scripts pour obtenir les résultats d’alignement.

En premier lieu, un script python de traitement de fichiers permet de découper une base de données. En effet, les bases de données disponibles au format fasta sur uniprot contiennent l’ensemble des protéines avec leurs isoformes concaténés en un fichier. Pour permettre un meilleur traitement via Clustal, la base de données est séparée en différents fichiers, contenant chacun une protéine avec l’ensemble de ses isoformes. Un script bash permet ensuite d’appliquer l’algorithme Clustal à l’ensemble de ces fichiers.

Enfin, pour vérifier le bon fonctionnement de l’algorithme, un script R permet de vérifier le nombre de gaps moyens par séquence inséré par séquence via l’algorithme, ainsi que leur longueur (l’analyse est effectuée uniquement sur les séquences présentant des gaps). Il est possible de modifier le répertoire de travail dans le script, ou de le passer en argument. Étant donné que les protéines étudiées sont des isoformes, il serait logique d’observer un faible nombre de gaps par séquence, avec une longueur moyenne élevée.

## PSIPRED [2]

PSIPRED est un algorithme de prédiction de structures secondaires (principalement les trois types de structures secondaires majoritaires, à savoir les boucles, les hélices α et les brin β), se basant sur le calcul de matrice PSSM à l’aide de Psi-BLAST. On estime que les résultats obtenus via PSIPRED sont fiables à environ 80%. L’algorithme va ici être utilisé pour déterminer si l’épissage alternatif semble modifier le type de structure secondaire prédit sur les isoformes, et dans quelle proportion. Dans ce projet, PSIPRED a été utilisé en local, le script bash utilisé est présent dans le même répertoire que les résultats (psipred/results\_psipred/ results\_reduce/run\_all\_psipred.sh). Celui permet d’appliquer l’algorithme à la base de données indiquées, et d’obtenir les résultats dans le répertoire local. Comme PSIPRED a été utilisé en local, les chemins vers la base de données blast et les programmes NCBI doivent être modifiés si c’est ce programme qui est réutilisé.

Pour analyser les résultats obtenus, un autre script R ("analyse\_psipred\_disopred.R") est utilisé (le répertoire de travail doit également être modifié). Celui-ci utilise les résultats de Clustal et de Psipred pour être fonctionnel. En effet, à partir des alignements Clustal, il va être possible de modifier les séquences psipred pour insérer les gaps calculés par Clustal. Ce script est donc basé principalement sur de la manipulation de chaînes de caractères avec le package stringr de R[[2]](#footnote-2). Il effectue ensuite deux analyses pour comparer les proportions de structures secondaires par isoforme par rapport à la protéine de référence, et pour calculer le nombre de position possédant la même prédiction de structure secondaire en omettant les gaps dans les séquences.

## DISOPRED [3]

DISOPRED est un algorithme de machine-learning permettant la détection des résidus ordonnés et désordonnés dans une séquence protéique. La version 3.1 est utilisée ici. À partir d’une séquence au format fasta, il va ainsi calculer et associer un score à chaque résidu : un score compris entre 0 et 0.49 est associé à un résidu ordonné, et sera représenté avec un point (.) dans le fichier de sortie, tandis qu’un résidu désordonné aura un score compris entre 0.50 et 1, et sera représenté par un astérisque (\*). Tout comme PSIPRED, DISOPRED utilise les programmes NCBI et une base de données BLAST : la présence et le chemin vers ces outils sont donc nécessaires.

Un script bash présent dans le répertoire "DISOPRED" est présent pour appliquer l’algorithme à un jeu de données choisi, et stocker les résultats utiles dans un répertoire de résultats fixé.

Le script R permettant d’analyser les résultats de DISOPRED est le même que pour les résultats de PSIPRED, et se base donc sur les mêmes principes qu’évoqués précédemment : les séquences de caractères sont manipulées pour que les gaps de CLUSTAL soient insérés ; les taux de match et les similarités de proportion sont ensuite calculées et stockées dans le même data frame.

# Résultats

## Clustal Omega

Lors de l’analyse des alignements de CLUSTAL, il est constatable que ceux-ci sont totalement cohérents avec les attendus. En effet, dans un premier temps, il est observable sur le boxplot que le nombre médian de gaps par séquence est de 2. Le résumé statistique fourni par R donne plus précisément les informations suivantes :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1er Quartile | Médiane | Moyenne | 3ème Quartile | Max |
| 1 | 1 | 2 | 2.766 | 4 | 10 |

Sur le jeu de données, il y a donc globalement très peu de gaps qui ont été ouverts par séquences, et 75% des séquences au moins n’ont eu que 4 ouvertures de gap lors de l’alignement.

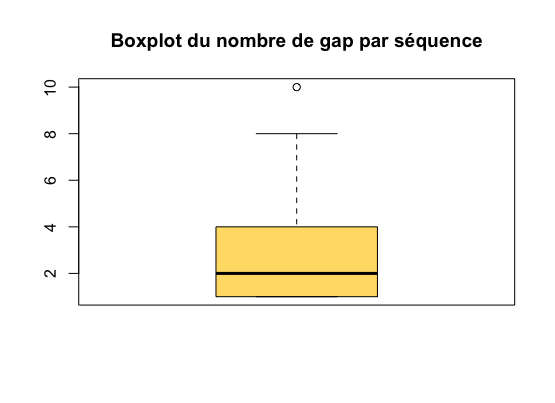


Figure 1 : Résumé du nombre de gaps ouverts par séquence sur le jeu de données étudié

La longueur de gap confirme la cohérence de l’alignement pour des isoformes. Le résumé statistique à partir duquel le boxplot a été construit est présent ci-dessous :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1er Quartile | Médiane | Moyenne | 3ème Quartile | Max |
| 1 | 6 | 16 | 54.64 | 45 | 966 |

Les longueurs de gap observées semblent être assez élevées, avec un nombre de valeurs extrêmes non négligeables ( > 200) correspondant avec un phénomène d’épissage alternatif, et donc la disparition d’une portion continue de la protéine initiale dans l’isoforme.

Ces résultats peuvent donc être utilisés lors des comparaisons avec PSIPRED et DISOPRED, en considérant que les isoformes correspondent bien à des parties continues des protéines non épissées sur lesquelles elles ont été correctement alignées.

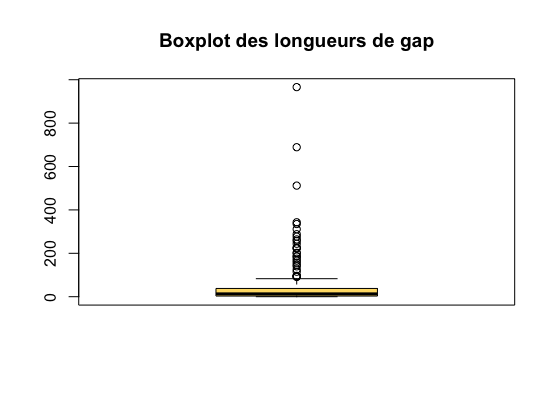


Figure 2 : Résumé de la longueur des gaps ouverts lors de l'alignement par CLUSTAL sur l'ensemble du jeu de données étudié

## Psipred

Lors de cette analyse, a été considéré comme match : la même structure secondaire prédite à la même position dans les deux séquences. Un mismatch est comptabilisé lorsqu’à une même position dans les séquences, la même structure secondaire n’est pas retrouvée, et qu’il n’y a pas de gap à cette position. La proportion de match correspond donc à la proportion de structures secondaires prédites identiques entre l’isoforme et la protéine de référence aux positions qui n’ont pas été affectées par l’épissage alternatif.

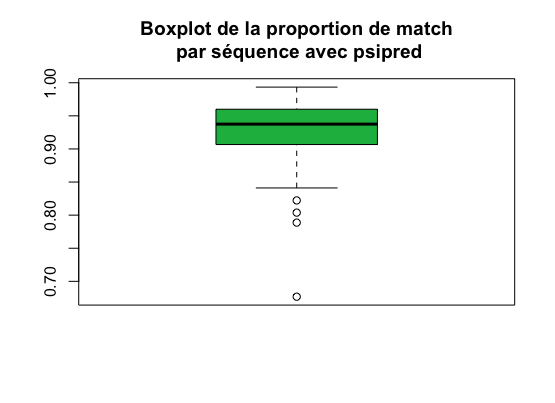


Figure 3 : Résumé des proportions de matchs psipred sur le jeu de données étudié

Sur la figure 3, la presque totalité des séquences isoformes du jeu de données ont un taux de match supérieur à 80%. Seule 2 séquences ont un taux de match inférieur, qui reste cependant supérieur à 60%. Globalement, il semble donc que sur ce jeu de données, le phénomène d’épissage alternatif n’impacte pas fondamentalement la prédiction et l’organisation de la structure secondaire protéique des isoformes.

La seconde analyse consiste à évaluer si les proportions de structures secondaires entre les protéines de référence et leurs isoformes sont significativement différentes. Pour cela, le nombre d’acides aminés impliqués dans les différents types de structures secondaires a été calculé pour chaque séquence, et un test du chi2 d’homogénéité a été ensuite appliqué au risque α = 5%, avec pour hypothèses :

* H0 : les proportions observées sont homogènes
* H1 : les proportions observées ne sont pas homogènes et suivent des lois de probabilité différentes.

Lorsque la p-value calculée est inférieure à 0.05, H0 est rejetée et le test est significatif au risque α %.

Le test est effectué sur l’ensemble des protéines isoformes : est calculée ensuite la proportion de tests significatifs dans le jeu de données étudié. Ainsi, environ 30.4% des tests se révèlent significatifs, ce qui signifie que 30% des isoformes ont des proportions de structures secondaires qui diffèrent de la protéine initiale. Ainsi, même si le phénomène d’épissage ne semble pas modifier de façon significative la structure secondaire des acides aminés, la proportion des types de structure secondaire semble être impactée de façon plus importante, ce qui peut être mis en relation avec une modification fonctionnelle importante.

On retrouve également dans ces protéines ayant une proportion significativement différente l’isoforme Q6P4Q7-2, issu d’une protéine ayant un rôle de transporteur lié à un ion métallique : cet isoforme est également celui qui possède un taux de match de 60% environ lors de la première analyse.

## Disopred

Les matchs et mismatchs de DISOPRED ont été considérés de la même manière que pour PSIPRED : deux résidus prédits de la même manière sont considérés comme un match, et lorsque la prédiction est différente et qu’il y a absence de gap, c’est un mismatch qui est comptabilisé.

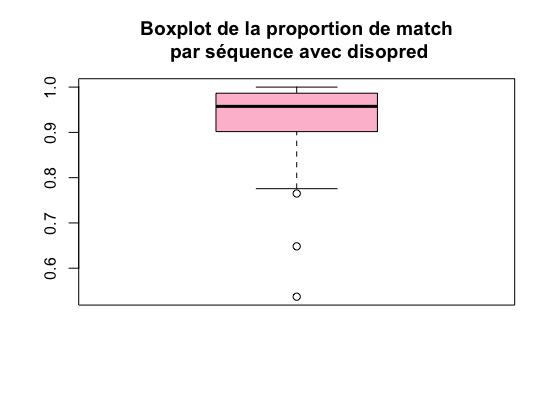


Figure 4 : Résumé des proportions de matchs disopred sur le jeu de données étudié

Comme précédemment, la majorité des séquences a un taux de match très élevé, supérieur à 80%. Il semble donc que sur ce jeu de données, l’épissage alternatif n’ait pas d’impact significatif non plus sur la prédiction de résidus désordonnés.

La seconde analyse est également effectuée de la même façon que précédemment, en effectuant un test du chi2 d’homogénéité sur les proportions de résidus ordonnés et désordonnés au risque α = 5%.

Pour 40% des séquences, il y a une différence significative du désordre protéique par rapport aux protéines de référence. Même si l’épissage alternatif ne semble pas avoir d’impact fondamental sur la prédiction du désordre des résidus, il est notable une nouvelle fois que la conséquence de ce phénomène sur les proportions de résidus désordonnés n’est pas négligeable.

On vérifie également la proportion de résidus désordonnés dans les protéines références et leurs isoformes.

|  |  |
| --- | --- |
|  |  |

*Tableau 1 : Comparatif des proportions de résidus désordonnés dans les protéines de référence et leurs isoformes*

Dans les représentations en boxplot du tableau 1, il ne semble pas y avoir de différence importante dans les proportions de résidus désordonnés retrouvés dans les isoformes par rapport aux protéines initialement connues. On peut noter une augmentation de points extrêmes dans le groupe des isoformes cependant.

# Discussion et conclusion

Certaines études effectuées précédemment proposent un lien de corrélation entre région intrinsèquement désordonnée et épissage alternatif. En effet, entre des régions protéiques ne subissant pas d’épissage et leurs opposés, une augmentation significative du désordre était observée dans la deuxième catégorie. Ici, les analyses se concentrent sur les protéines connues pour subir un épissage alternatif. Les observations préliminaires sur un jeu de données échantillonné ont pu mener aux conclusions que sur les protéines étudiées, l’épissage alternatif n’impactait pas de manière significative le type de structure secondaire ainsi que le désordre protéique sur les résidus retrouvés dans les isoformes. Cependant, un nombre non négligeable d’isoformes possède des proportions des facteurs cités précédemment significativement différents par rapport à la protéine de référence : cette observation semble être correctement corrélée à une augmentation de diversité fonctionnelle observée lors de nombreuses études sur l’épissage alternatif. Cependant, il pourrait être intéressant de se questionner sur les implications vis-à-vis des isoformes de la faible différence observée sur les prédictions de résidus. Un biais du jeu de données est évidemment une piste à ne pas négliger. Cependant, la très forte proportion de matchs supérieurs à 80% semblent indiquer une tendance qui pourrait être approfondie : les résidus sont-ils prévus également dans les mêmes conformations sur des protéines issues d’autres organismes ? Est-ce qu’une catégorisation des protéines selon des propriétés fonctionnelles définies avant analyse pourrait modifier ces observations ? Il pourrait être également intéressant de se concentrer sur les protéines chez qui on ne retrouve pas ces tendances : les isoformes avec des taux de matchs plus faibles ont-ils par exemple été retrouvés dans des fonctions bien plus éloignées de la protéine initiale que d’autres isoformes ? Ou, sur des jeux de données plus importants, peut-on retrouver des points communs entre les protéines retrouvées comme cas extrêmes ? Les limites des connaissances liées à l’épissage alternatif entrent ainsi en jeu : de nombreuses protéines résultant de l’épissage alternatif, bien que connues, ont des fonctions qui n’ont pas encore été découvertes, ou ne sont qu’hypothétiques et n’ont pas été observées. Enfin, des approches de prédiction de structure 3D pourraient être utilisées pour observer si les conformations prédites peuvent différer de manière bien plus importante. Cependant, celles-ci doivent aussi être utilisées avec précaution : si certaines portions des protéines ont été résolues, et permettent des résolutions par threading par exemple, les isoformes peuvent ne pas correspondre aux régions déjà connues, et la pluralité des approches nécessaires pour résoudre les structures peuvent induire un biais dans la comparaison de structures.

# Bibliographie

[1] Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Research. 2019 Jul;47(W1):W636-W641. DOI: 10.1093/nar/gkz268.

[2] Jones DT. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195-202.

[3] Jones DT, Cozzetto D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015 Mar 15;31(6):857-63. doi: 10.1093/bioinformatics/btu744. Epub 2014 Nov 12.

[4] Fabian Birzele, Gergely Csaba, Ralf Zimmer, Alternative splicing and protein structure evolution, Nucleic Acids Research, Volume 36, Issue 2, 1 February 2008, Pages 550–558, <https://doi.org/10.1093/nar/gkm1054>

[5] Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell. 1977 Sep;12(1):1-8. doi: 10.1016/0092-8674(77)90180-5. PMID: 902310.

[6] Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms Pedro R. Romero, Saima Zaidi, Ya Yin Fang, Vladimir N. Uversky, PredragRadivojac, Christopher J. Oldfield, Marc S. Cortese, Megan Sickmeier, Tanguy LeGall, Zoran Obradovic, A. Keith Dunker Proceedings of the National Academy of Sciences May 2006, 103 (22) 8390-8395; DOI:10.1073/pnas.0507916103

[7] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure. 2003 Nov;11(11):1453-9. doi: 10.1016/j.str.2003.10.002. PMID: 14604535.

[8] The complexity of alternative splicing of *hagoromo* mRNAs is increased in an explosively speciated lineage in East African cichlids

Yohey Terai, Naoko Morikawa, Koichi Kawakami, Norihiro Okada

Proceedings of the National Academy of Sciences Oct 2003, 100 (22) 12798-12803; DOI:10.1073/pnas.2132833100

[9] Colak R, Kim T, Michaut M, Sun M, Irimia M, Bellay J, et al. (2013) Distinct Types of Disorder in the Human Proteome: Functional Implications for Alternative Splicing. PLoS Comput Biol 9(4): e1003030. https://doi.org/10.1371/journal.pcbi.1003030

[10] Sara Light, Arne Elofsson, The impact of splicing on protein domain architecture, Current Opinion in Structural Biology, Volume 23, Issue 3, 2013, Pages 451-458, ISSN 0959-440X, <https://doi.org/10.1016/j.sbi.2013.02.013>.

[11] Garcia-Blanco, M., Baraniak, A. & Lasda, E. Alternative splicing in disease and therapy. *Nat Biotechnol***22,**535–546 (2004). https://doi.org/10.1038/nbt964

[12] Pauline L. Lee, Terri Gelbart, Carol West, Carol Halloran, Ernest Beutler, The Human Nramp2 Gene: Characterization of the Gene Structure, Alternative Splicing, Promoter Region and Polymorphisms, Blood Cells, Molecules, and Diseases, Volume 24, Issue 2, 1998, Pages 199-215, ISSN 1079-9796, https://doi.org/10.1006/bcmd.1998.0186.

[13] Vladimir N. Uversky, Chapter One - Protein intrinsic disorder and structure-function continuum, Editor(s): Vladimir N. Uversky, Progress in Molecular Biology and Translational Science, Academic Press, Volume 166, 2019, Pages 1-17, ISSN 1877-1173, ISBN 9780128168516, <https://doi.org/10.1016/bs.pmbts.2019.05.003>.

Lien vers les codes : https://github.com/Lugulabre/Projet\_long

1. https://www.uniprot.org [↑](#footnote-ref-1)
2. https://cran.r-project.org/web/packages/stringr/index.html [↑](#footnote-ref-2)