**Développement d’un outil de recherche de mutations**

**de type fsATI**

Julie Gorse Bogoin

Projet Long

M2 BIB

2019-2020

**TABLE DES MATIERES**

[INTRODUCTION 1](#_Toc30095660)

[Contexte 1](#_Toc30095661)

[Description des mutations fsATI 2](#_Toc30095662)

[État de l’art 3](#_Toc30095663)

[Mécanisme d’initiation de la traduction 3](#_Toc30095664)

[Stratégie 3](#_Toc30095665)

[Ambition du projet 4](#_Toc30095666)

[Calcul de la force de Kosak 4](#_Toc30095667)

[Utilisation de PhyloP pour évaluer la conservation au niveau des bases 5](#_Toc30095668)

[MATERIEL ET METHODES 5](#_Toc30095669)

[Récupération des séquences ARNm 5](#_Toc30095670)

[Extraction des ORFs à partir des séquences d’ARNm 6](#_Toc30095671)

[Identification des variants fsATI 6](#_Toc30095672)

[Validation des variants fsATI 7](#_Toc30095673)

[RESULTATS 8](#_Toc30095674)

[fsATI validés 8](#_Toc30095675)

[Scores PhyloP 9](#_Toc30095676)

[Forces de Kosak 9](#_Toc30095677)

[CONCLUSION ET PERSPECTIVES 10](#_Toc30095678)

# INTRODUCTION

## Contexte

Dans leur publication de 2018, l’équipe de Royal *et al.* a démontré qu’un nouveau mécanisme moléculaire pouvait causer des maladies chez l’humain : l’initiation alternative de la traduction causée par des mutations changeant le cadre de lecture (fsATI). Des mutations qui changent le cadre de lecture se trouvant après le premier codon d’initiation peuvent mener à une traduction alternative pour produire une seconde protéine pouvant être délétère. La pertinence de ce mécanisme a été démontrée dans le contexte des canaux ioniques multi-mériques où une mutation fsATI de 2 paires de base du gène TRESK a pu être associée à la migraine.

La migraine est une affection neurologique handicapante qui affecte 15 % de la population. Une attaque migraineuse se traduit par des maux de tête d’une durée de 4 à 72 h accompagnés d’une hypersensibilité́ à de nombreux stimulus, tels que la lumière ou le son, pouvant mener à des nausées, voire des vomissements. La migraine est une maladie complexe liée, entre autres, à l’hyperexcitabilité́ électrique des neurones sensoriels de la face, appelés neurones trigéminaux. Cette excitabilité est sous le contrôle de canaux ioniques. Les canaux « potassiques à deux- domaines pore » ou K2P sont inhibiteurs et servent de frein à l’excitabilité́ neuronale. Lorsque ces protéines dysfonctionnent, une hyperexcitabilité́ des neurones survient. Une version mutée du canal TRESK (TWIK-related spinal cord potassium 1 channel) , un membre de la famille des canaux K2P, présentant une délétion de 2 paires de bases entrainant un décalage du cadre de lecture, dénommée TRESK-MT, a récemment été incriminée dans la migraine. Cette mutation induit la formation d’un canal tronqué qui, en plus d’être non fonctionnel, agit comme dominant négatif du canal TRESK sauvage (TRESK-MT abolit le courant généré́ par TRESK). Les neurones trigéminaux qui expriment TRESK-MT voient leur excitabilité́ augmentée, ce qui explique l’induction de la migraine.

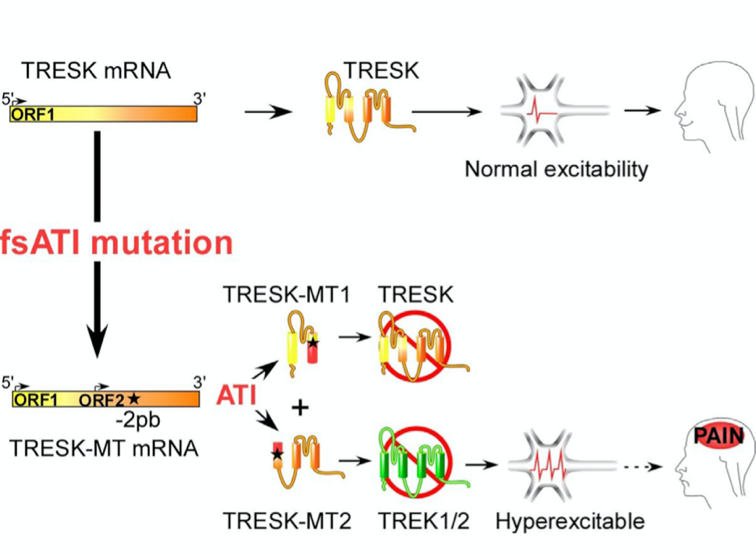


Figure 1 : Représentation du rôle des fsATI dans l’apparition des migraines

Source : Sandoz *et al.*

La mutation TRESK-MT provoque une migraine via l’inhibition de TREK1 et TREK2 induite par TRESK-MT2. Dans la partie supérieure est représentée l’excitabilité́ des neurones trigéminaux en présence de la version sauvage de TRESK. Dans la partie inférieure, les personnes portant la version mutée du canal TRESK présentent une sensibilité liée à la migraine. Ce gène muté, par un mécanisme appelé fsATI est traduit en deux protéines, TRESK-MT1 et TRESK-MT2. TRESK-MT1 inhibe le canal TRESK alors que TRESK-MT2 cible TREK1 et TREK2. L’inhibition de TREK1 et TREK2 conduit à une hyperexcitabilité́ neuronale et à un phénotype migraineux.

Le mécanisme de fsATI pourrait permettre de mieux comprendre le mécanisme par lequel certaines mutations de type frameshift peuvent influencer la fonction du gène. Normalement, une mutation de type frameshift devrait mener à un changement du cadre de lecture puis à une protéine sévèrement tronquée. La production d’une protéine plus légèrement tronquée amène d’autres pistes qui pourraient aider à mieux comprendre l’étiologie de certaines maladies.

## Description des mutations fsATI

En général, l’ARN messager (ARNm) transcrit à partir du gène permet la traduction d’une protéine unique. Dans le cas des fsATI, cet adage n’est pas respecté. Un site alternatif d’initiation de la traduction peut, en effet, être reconnu sur l’ARNm par les ribosomes et induire la synthèse d’une deuxième protéine à partir du même ARNm.

Une mutation de type fsATI sur un ARNm peut être définie comme suit :

* Un site alternatif d’initiation (ATI) qui n’est pas dans le cadre de lecture sauvage.
* Une mutation frameshift (fs) qui permet de retomber dans le cadre de lecture sauvage.
* La fin de séquence est identique à la séquence sauvage.
* Ainsi, seule la section de la protéine entre l’ATI et le frameshift est problématique.

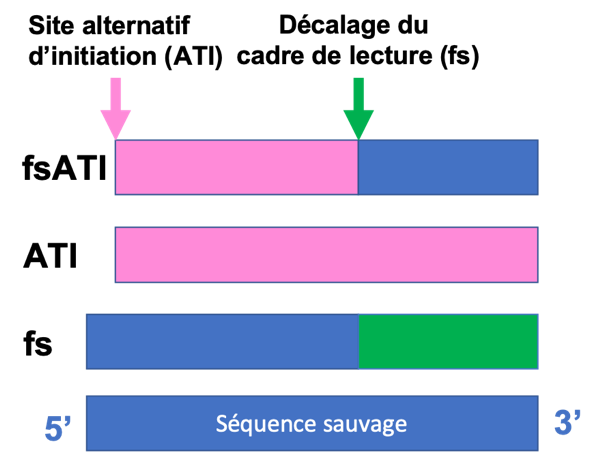


Figure 2 : Représentation des mutations de type fsATI

Sans frameshift l’ATI mènerait à une protéine « aléatoire », sans lien avec la séquence originale. Sans l’ATI le frameshift causerait un changement dans le cadre de lecture qui mènerait probablement à une protéine tronquée à la fin de sa séquence et non fonctionnelle. Les deux ensembles mènent à une protéine tronquée au début de sa séquence, avec un petit bout de séquence “aléatoire”, mais le reste de la séquence est identique à la séquence sauvage.

## État de l’art

Le mécanisme de fsATI vient tout juste d’être publié ce qui explique qu’on ne retrouve que très peu de littérature à ce sujet.

### Mécanisme d’initiation de la traduction

Il est possible d’avoir plusieurs codons d’initiation dans des cadres de lecture différents. A priori, l’initiation de la traduction pourrait se faire à l’un ou l’autre des sites AUG. En réalité, un seul est souvent utilisé exclusivement. Le mécanisme classique d’initiation de la traduction est l’entrée du ribosome en 5’ de l’ARN, puis le « scanning » de la séquence par le ribosome, jusqu’au premier AUG. Mais la traduction peut commencer au nième AUG, en ignorant les n-1 précédents « masqués » dans une structure secondaire (« entrée interne du ribosome » ou « IRES »). La préférence pour un AUG parmi plusieurs en concurrence peut dépendre de la séquence de nucléotides autour. Dans un cas sauvage, un seul AUG est utilisé pour initier la traduction (et pas forcément le premier), même si la séquence en contient plusieurs.

Il existe des mécanismes d’initiation alternative : choix d’un AUG différent, qui peut être dans un autre cadre de lecture, ce qui conduit à une protéine raccourcie (si même cadre de lecture) ou complètement différente (si changement de cadre de lecture).

Un des challenges de ce projet est déterminer si un site d’initiation alternatif de la traduction créé par un fsATI est légitime.

### Stratégie

La publication de Whiffin et al. de février 2019 a pour ambition de présenter l’étude systématique menée à l'échelle du génome, des variants qui créent et perturbent les upstream open reading frames (uORF) humains afin d’explorer leurs rôles dans les maladies.

15708 séquences de génome entier ont été collectées par le projet de base de données d'agrégation de génomes (gnomAD).

* Une forte sélection négative a été montrée pour les 14 897 variants qui créent de nouveaux codons « start » en amont de la séquence de codage canonique (CDS) et les 2 406 variants qui perturbent le site d'arrêt des ORF existants.
* Une forte sélection été démontrée pour les variants créant des uORF qui chevauchent le CDS et pour les mutations faux-sens.
* Les variants perturbant l'uORF sont sous forte sélection lorsqu'ils sont positionnés en amont de gènes délétères connus et de gènes intolérants aux variants de perte de fonction.

Ces résultats mettent en évidence les variants perturbant l'uORF comme une classe fonctionnelle importante qui peut contribuer aux maladies humaines pénétrantes, et démontrent la puissance des données de séquençage de population à grande échelle pour étudier la nocivité de classes spécifiques de variantes non codantes.

Un parallèle peut être fait entre cette étude et le projet réalisé ici peut être fait facilement puisqu’ils cherchent à évaluer tous les deux les conséquences pathologiques des ATI. C’est pourquoi il la stratégie de validation des fsATI tenant compte de la force de Kosak et des scores de PhyloP a été inspiré par ce papier. En revanche, il est important de noter que la différence majeure entre ces deux projets réside dans le recadrage qui a lieu (fs) en aval de l’ATI, dans le cas des fsATI.

## Ambition du projet

L’objectif du projet est de déterminer si les mutations de type fsATI sont un phénomène fréquemment impliqué dans les maladies humaines.

Les différentes étapes du projet sont :

1. Développer un outil de génération des ORF possibles et crédibles à partir de séquences d’ARNm.
2. Mettre en place une méthodologie de priorisation des variants pour sélectionner ceux ayant la plus forte probabilité d’être des variants de type fsATI.
3. Évaluation des mutations grâce à l’étude de la conservation des régions ciblées par des calculs de score de force de Kosak et de PhyloP.

### Calcul de la force de Kosak

La séquence Kozak est une séquence conservée que l'on trouve sur les ARN messagers eucaryotes au niveau du site de démarrage de la traduction, autour du codon de démarrage AUG. Chez les vertébrés, la séquence de Kozak a pour consensus gccRccAUGG où R représente une purine (le codon de démarrage est en gras et les nucléotides en minuscules sont moins conservés que le reste de la séquence). C'est une séquence très conservée lors de l'évolution, sa présence indique donc une bonne conservation et est un argument en faveur de la validation de la mutation fsATI délétère.

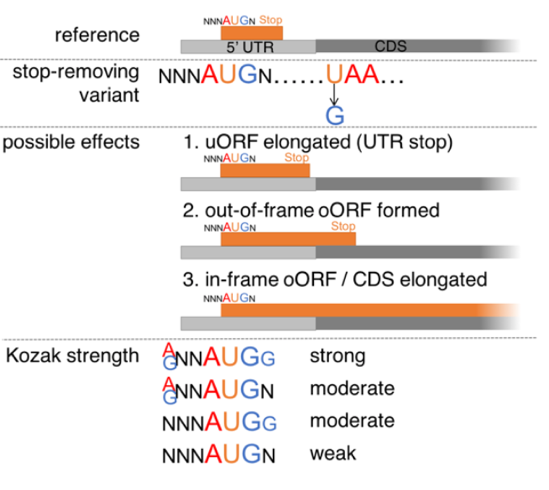


Figure 3 : Schéma des variants uORF et la manière dont la force de la séquence Kozak est déterminée

Source : Whiffin et al., 2019

### Utilisation de PhyloP pour évaluer la conservation au niveau des bases

Les scores phyloP mesurent la conservation évolutive sur les sites d'alignement individuels. Les interprétations des scores sont comparées à l'évolution attendue sous dérive neutre.

* Lorsque le score est positif, cela indique une conservation, qui est une évolution plus lente que prévu, aux sites qui devraient être conservés.
* Lorsque le scores est négatif, cela indique une accélération, qui est une évolution plus rapide que prévu, sur les sites qui devraient évoluer rapidement.

Les scores phyloP sont utiles pour évaluer les signatures de sélection au niveau des nucléotides. Les valeurs absolues des scores phyloP représentent des valeurs de –log p sous une hypothèse nulle d'évolution neutre. Les bases conservées ont été définies comme celles avec un score de PhyloP supérieur ou égal à 2.

# MATERIEL ET METHODES

Tous les scripts ont été codés en Python 3 et en utilisant la bibliothèque les modules BioPython et Pandas.

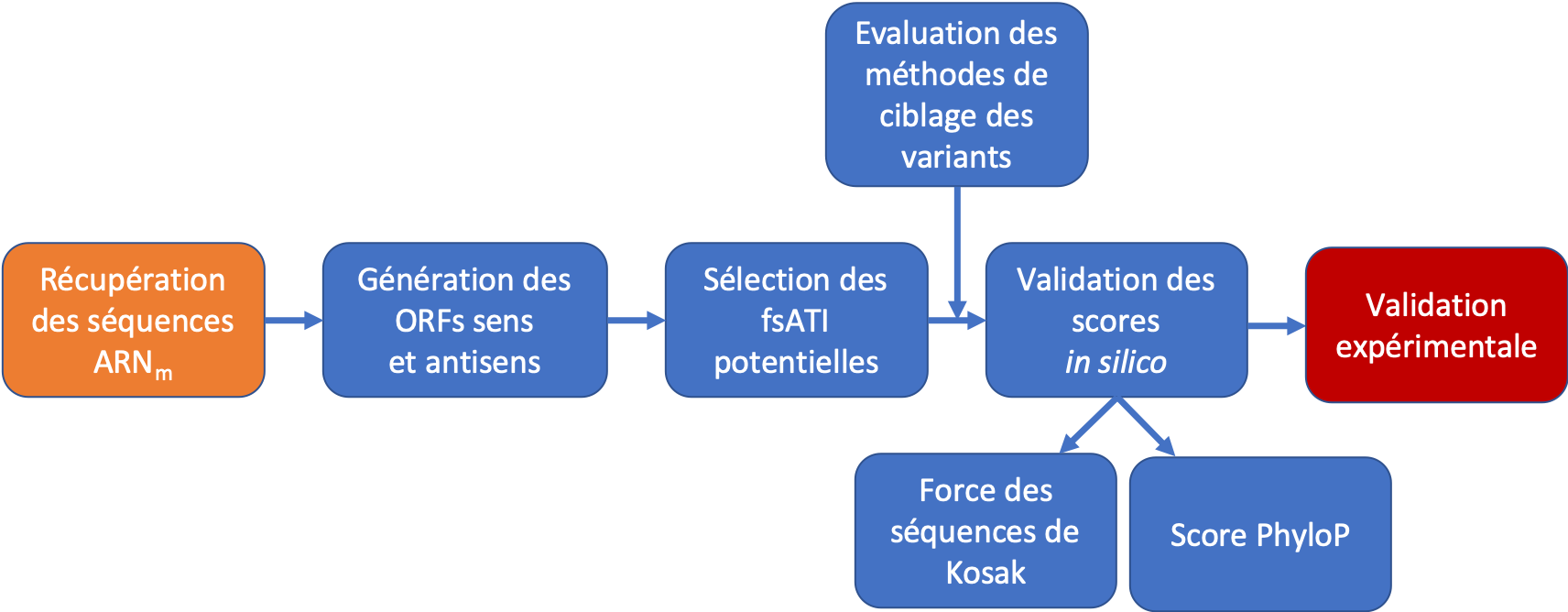


Figure 4 : Étapes du projet.

En orange, le travail effectué en amont, en rouge celui qui reste à faire et en bleu celui qui a été fait.

## Récupération des séquences ARNm

Les séquences d’ARNm des gènes test KCNQ1, KCNQ2, KCNQ3, KCNQ4 et KCNQ5 ont été générées à partir des séquences d’ADN génomique en utilisant le package ANNOVAR.

ANNOVAR (http://annovar.openbioinformatics.org/en/latest/) est un outil logiciel efficace pour utiliser des informations à jour pour annoter fonctionnellement des variants génétiques détectées à partir des génomes humain (hg18, hg19 et hg38) de souris, de ver, de mouche, de levure…).

## Extraction des ORFs à partir des séquences d’ARNm

A partir des séquences d’ARNm récupérées, le script **mRNA\_to\_ORF.py** va permettre de générer et stocker les différentes séquences ORF crédibles dans le fichier **ORF.csv**.

**Entrée :**

Séquence ARNm au format fasta.

**But :**

Extraire des listes d’ORF.

**Fonctionnement :**

Le programme parcourt la séquence selon les trois cadres de lecture possible. Il génère ensuite toutes les séquences possibles entre un codon « start » et un codon « stop ». Un data-frame est créé pour stocker toutes les informations à extraire vers le csv de sortie.

**Quand est ce qu’il marche :**

Le programme fonctionne bien quand la séquence est au format fasta et quand elles sont toutes stockées dans un répertoire appelé : **mRNA\_fasta**.

**Quand est ce qu’il ne marche pas :**

Si le format de la séquence n’est pas fasta.

**Combien de temps d’exécution :**

Quelques secondes pour les 5 gènes KCNQ.

**Sortie :**

Un fichier csv contenant les informations suivantes :

gene, annovar\_line, transcript, mutation, mRNA\_sens, mRNA\_antisens, longueur\_mRNA, nombre\_ORF\_sens, ORF\_sens, nombre\_ORF\_antisens, ORF\_antisens.

**Limites :**

Les fichiers fasta de départ ne doivent pas contenir d’espace dans leurs noms.

**Améliorations possibles :**

Réduire le nombre de boucles pour améliorer la complexité du script. L’utilisation des objets semblent être une bonne piste.

## Identification des variants fsATI

Le script **find\_fsATI.py** a pour objectif d’identifier parmi les ORF, celles qui contiendraient des mutations de type fsATI. Ces séquences sont stockées dans le fichier **fsATI.csv**.

**Entrée :**

Le fichier csv issu du script précédent

**But :**

Identifer les variants de type fsATI parmi une liste d’ORFs.

**Fonctionnement :**

Le programme passe en revue les séquences sélectionnées pour vérifier si elles répondent bien à la définition d’une fsATI (comme définie dans l’introduction). Si c’est le cas, la séquence est stockée dans un data-frame avec toutes les autres informations à extraire. Les scores de force de Kosak et de PhyloP sont calculés pour chaque séquence.

**Quand est ce qu’il marche :**

Le programme fonctionne bien quand le fichier csv de départ est bien complet et que les séquences mutées sont en dessous de la séquence sauvage associée. Il est également necessaire de posséder pour chaque gène les positions génomiques de chaque base.

**Quand est ce qu’il ne marche pas :**

Si l’ordre des entrées dans le fichier csv est arbitraire.

Le programme ne pourrait pas fonctionner si le fichier de score de PhyloP n’était pas stocké dans le répertoire **PhyloP (**https://ccg.epfl.ch/mga/hg19/phylop/phylop.html).

**Combien de temps d’exécution :**

10 minutes pour les 5 fasta KCNQ, le temps de lecture et le parcours du fichier d’entrée **phylop\_vert.sga** de PhyloP à chaque tour de boucle est très chronophage.

**Sortie :**

Un fichier csv contenant les colonnes suivantes :

gene, transcript, mutation, sens, ORF, Pos\_ATI, Pos\_fs, Kosak strenght, PhyloP score

**Limites :**

Interrogation sur la capacité à gérer un très grand nombre de séquences. L’indexation du fichier de score PhyloP parait indispensable.

**Améliorations possibles :**

Réduire le nombre de boucles pour améliorer la complexité du script. L’utilisation des objets semblent être une bonne piste.

## Validation des variants fsATI

Le script **validate\_fsATI.py** a pour objectif d’identifier parmi les séquences fsATI potentielles, celles qui seraient effectivement de type fsATI. Ces séquences sont stockées dans le **fichier fsATI\_validated.csv**.

**Entrée :**

Le fichier csv issu du script précédent

**But :**

Valider les variants qui seraient de type fsATI en fonction des scores de Kosak et de PhyloP.

**Fonctionnement :**

Le programme passe en revue les séquences sélectionnées et sélectionne celles qui ont un score de Kosak de 3 et un score de PhyloP supérieur ou égal à 2. Ces informations sont stockées dans un data-frame.

**Quand est ce qu’il marche :**

Le programme fonctionne correctement.

**Combien de temps d’exécution :**

Quelques secondes pour les 5 gènes KCNQ.

**Sortie :**

Un fichier csv contenant les colonnes suivantes :

gene, transcript, mutation, sens, ORF, Pos\_ATI, Pos\_fs, Kosak strenght, PhyloP score

Les scripts et les fasta de départ peuvent être trouvées et téléchargées à l’adresse suivante :

<https://github.com/ArnaudDroitLab/ProjetLong_201911>

Le fichier phylop\_vert.sga peut être téléchargé à l’adresse suivante :

<https://ccg.epfl.ch/mga/hg19/phylop/phylop_vert.sga.gz>

# RESULTATS

### fsATI validés

Figure 5 : Proportion des fsATI parmi les ORFs

143 séquences d’ARNm ont été générées préalablement par ANNOVAR.

Le script mRNA\_to\_ORF.py a produit 324 ORF sens et 183 ORF antisens soit 507 ORF au total.

Le script find\_fsATI.py a sélectionné 94 séquences susceptibles d’être des fsATI parmi les 507 ORF.

Le script validate\_fsATI.py a validé 6 séquences qui seraient des fsATI parmi les 94 fsATI potentielles.

### Scores PhyloP

Sur les 94 séquences fsATI potentielles, la grande majorité (90%) ont un score de PhyloP supérieur et égal à 2 (2, 3 ou 4). Les autres ont un score de 1. Cela représente 10% des séquences soit 9 au total.

### Forces de Kosak

Figure 6 : Répartition des scores de force de séquence de Kosak parmi les fsATI potentielles

Dans le script, il a été établi qu’un score de –1 correspondait à l’absence de séquence de Kosak à proximité de l’ATI (+ ou – 15 paires de bases). Les deux tiers des fsATI potentielles sont dans ce cas de figure. 10 séquences ont un score de 1, 8 ont un score de 2 et seulement 6 un score de 3.

C’est clairement la force de la séquence de Kosak qui a été discriminante dans la validation des séquences fsATI.

# CONCLUSION ET PERSPECTIVES

Ce projet a permis de réaliser un outil de détermination et d’évaluation des mutations de type fsATI à partir de séquences d’ARNm. Cette méthode a été établie à partir de 5 gènes : KCNQ1, KCNQ2, KCNQ3, KCNQ4 et KCNQ5.

Cette détermination des scores de séquences fsATI *in silico* devra être poursuivie expérimentalement par une évaluation expérimentale. Il faudra également analyser l’ensemble des gènes connus via les bases de données gnomAD, SNPedia et dbSNP pour discuter la pertinence de la stratégie choisie et du code. Il pourrait être intéressant de confronter les fsATI sélectionnés avec les bases de données cliniques pour voir si ces séquences sont référencées comme impliquées dans des pathologies, cela validerait encore un peu plus cet outil. Une étude de la structure des protéines et de leur conservation pourrait apporter également de l’information.

Le code devra être relu et amélioré pour le simplifier et pour diminuer sa complexité. Dans l’état actuel la récupération des scores de PhyloP est trop chronophage et trop demandeuse en ressource. Il sera nécessaire de trouver une façon de s’affranchir du parcours de l’immense fichier de données PhyloP pour chaque ORF. Il faudra également trouver comment récupérer facilement les coordonnées génomiques de chaque base d’ARNm.

Scientifiquement, la stratégie choisie et son implantation sont pertinentes et justifiées. C’est techniquement qu’il reste encore du travail. Ce projet était novateur et il devrait permettre de comprendre de nouveaux mécanismes provoquant des maladies et pourquoi pas mettre en place de nouveaux traitements thérapeutiques.

**BIBLIOGRAPHIE**

Jain, Ankur, Ruijie Liu, Yang K Xiang, et Taekjip Ha. « Single-Molecule Pull-down for Studying Protein Interactions ». *Nature Protocols* 7, no 3 (mars 2012): 445‑52. <https://doi.org/10.1038/nprot.2011.452>.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. « Variation across 141,456 Human Exomes and Genomes Reveals the Spectrum of Loss-of-Function Intolerance across Human Protein-Coding Genes ». Preprint. Genomics, 28 janvier 2019. <https://doi.org/10.1101/531210>.

Kozak, Marilyn. « Initiation of Translation in Prokaryotes and Eukaryotes ». *Gene* 234, no 2 (juillet 1999): 187‑208. <https://doi.org/10.1016/S0378-1119(99)00210-3>.

Ramani, Ritika, Katie Krumholz, Yi-Fei Huang, et Adam Siepel. « PhastWeb: A Web Interface for Evolutionary Conservation Scoring of Multiple Sequence Alignments Using PhastCons and PhyloP », s. d., 3.

Royal, Perrine, Alba Andres-Bilbe, Pablo Ávalos Prado, Clément Verkest, Brigitte Wdziekonski, Sébastien Schaub, Anne Baron, et al. « Migraine-Associated TRESK Mutations Increase Neuronal Excitability through Alternative Translation Initiation and Inhibition of TREK ». *Neuron* 101, no 2 (janvier 2019): 232-245.e6. <https://doi.org/10.1016/j.neuron.2018.11.039>.

Siepel, A. « Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes ». *Genome Research* 15, no 8 (1 août 2005): 1034‑50. <https://doi.org/10.1101/gr.3715005>.

Whiffin, Nicola, Konrad J Karczewski, Xiaolei Zhang, Sonia Chothani, Miriam J Smith, D Gareth Evans, Angharad M Roberts, et al. « Characterising the Loss-of-Function Impact of 5’ Untranslated Region Variants in 15,708 Individuals ». Preprint. Genomics, 7 février 2019. <https://doi.org/10.1101/543504>.