Mini Projet Entrepôt de Données

Analyse des Variants Génétiques pour les Maladies Héréditaires

Préparé par :

SAIDI BOUCHRA
REZIG WASSILA
MEDDOUR MERIEM
HAMADOU Lydia
OUAIL Lydia

Professeur : Mme L. BERKANI

Université des Sciences et Technologies

Année Universitaire 2024-2025

Table des matières

Table des figures

1 Introduction

L'analyse des variants génétiques dans les maladies héréditaires est essentielle pour identifier les mutations responsables de pathologies transmises génétiquement. Ces études permettent de mieux comprendre les mécanismes moléculaires, d'améliorer les diagnostics précoces et de perfectionner les traitements et la prévention.

Les maladies héréditaires, souvent graves ou fatales, sont liées à des mutations dans des gènes spécifiques. Les progrès du séquençage de l'ADN ont permis d'analyser ces variations et de déterminer leur rôle dans l'apparition de ces pathologies.

Ce projet se concentre sur l'étude de ces variants en utilisant des outils bioinformatiques pour identifier, analyser et interpréter les mutations significatives, dans le but d'améliorer le diagnostic et de proposer de nouvelles approches thérapeutiques.

1.1 Thème

Ce projet porte sur l'analyse des variants génétiques dans les maladies héréditaires. Ces maladies, transmises par les gènes, représentent un défi majeur en médecine en raison de leur diagnostic et traitement complexes.

Les progrès du séquençage de nouvelle génération (NGS) ont permis des découvertes clés dans l'identification des mutations responsables de ces pathologies, qu'elles soient rares ou courantes, avec des conséquences parfois graves pour la santé.

L'objectif du projet est d'analyser les bases de données génétiques pour comprendre les liens entre mutations génétiques et maladies héréditaires, afin de faciliter le diagnostic et d'ouvrir la voie à des solutions thérapeutiques innovantes.

1.2 Problématique

Les maladies héréditaires représentent une problématique majeure de santé publique. Selon l'OMS, 1 personne sur 20 est atteinte d'une maladie génétique rare, et 80% d'entre elles sont d'origine génétique [3]. Cependant, la majorité des mutations responsables restent inconnues, et environ 30% des maladies rares sont liées à des gènes non identifiés [2].

Le diagnostic des maladies héréditaires est compliqué par la diversité des mutations, certaines bénignes, d'autres graves. De plus, l'augmentation des données génétiques générées par les technologies de séquençage rend difficile la distinction entre mutations pathogènes et variants bénins.

Les technologies de séquençage de nouvelle génération (NGS) ont facilité la collecte de données massives. Cependant, leur analyse reste un défi, avec 70% des professionnels de santé éprouvant des difficultés à interpréter les tests génétiques [1], soulignant la nécessité d'outils bioinformatiques plus performants.

2 État de l'art

2.1 Concepts de base

L'analyse des variants génétiques se concentre sur l'identification et l'étude des mutations génétiques présentes dans l'ADN, responsables de nombreuses maladies héréditaires. Ces maladies résultent généralement de mutations dans des gènes spécifiques, qui peuvent être transmises d'une génération à l'autre.

Concept de l'analyse génétique L'analyse génétique repose principalement sur le séquençage de l'ADN pour identifier différentes variations génétiques, telles que :

- Mutations ponctuelles
- Délétions
- Duplications

Une fois ces variations identifiées, l'impact fonctionnel de ces mutations sur les protéines codées par les gènes doit être analysé. Ces analyses permettent de comprendre les effets des mutations sur la fonction des gènes et des protéines.

Technologies de séquençage Le séquençage de nouvelle génération (NGS) a radicalement transformé le domaine de l'analyse génétique. Cette technologie permet une analyse plus rapide, plus précise et à grande échelle des génomes, facilitant ainsi la découverte de variants génétiques associés à diverses maladies.

Rôle des outils bioinformatiques Les outils bioinformatiques sont essentiels dans ce processus. Ils permettent :

- L'annotation des mutations génétiques
- L'analyse des relations entre les gènes et les maladies
- La prédiction de l'impact clinique des mutations

Ces outils aident également à gérer les grandes quantités de données générées par les technologies de séquençage.

Défis de l'analyse génétique L'un des principaux défis reste l'identification des mutations pathogènes parmi les nombreuses variations génétiques trouvées. En effet, les données générées sont vastes, ce qui rend difficile la distinction entre les mutations bénignes et celles responsables de pathologies. La gestion de ces grandes quantités de données reste donc un défi majeur dans ce domaine.

2.2 Architecture

Notre projet suit une architecture en plusieurs étapes bien définies :

- **Téléchargement des données** : Récupération du fichier `clinvar.vcf`, contenant des informations détaillées sur les variants génétiques et leurs associations cliniques.
- **Conversion VCF vers CSV** : Transformation du fichier VCF en format CSV pour faciliter la manipulation et le traitement des données à l'aide d'outils comme `Pandas`.
- **ETL (Extract-Transform-Load)** :
  - **Extraction** : Chargement des données depuis les fichiers CSV.

  - **Transformation** : Nettoyage initial, sélection des colonnes pertinentes, gestion des valeurs manquantes.
  - **Chargement** : Organisation des données prêtes à être analysées.
- **Nettoyage avancé** : Application d'un modèle `Random Forest` pour l'évaluation de l'importance des variables et la suppression des attributs ou enregistrements non pertinents.
- **Analyse multidimensionnelle** : Construction de cubes OLAP permettant d'explorer les données selon plusieurs dimensions telles que les gènes, les types de mutations et les maladies associées.
- **Visualisation** : Développement d'une interface simple permettant de visualiser les résultats de l'analyse.

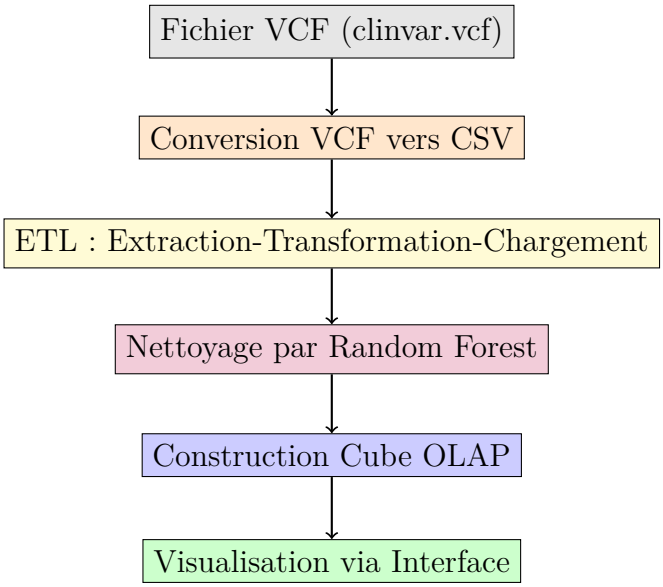

FIGURE 1 – Architecture complète du projet

2.3 Pourquoi avons-nous choisi cette architecture ?

Nous avons adopté cette architecture pour plusieurs raisons :
- **Manipulation facilitée** : Le format CSV permet une manipulation rapide et efficace des données comparé au format VCF brut.
- **Modularité** : En séparant l'ETL, le nettoyage, l'analyse et la visualisation, chaque étape du processus est indépendante et optimisable.
- **Efficacité de nettoyage** : L'utilisation d'un modèle Random Forest permet d'identifier automatiquement les variables les plus pertinentes, garantissant une meilleure qualité des données pour l'analyse.
- **Analyse multidimensionnelle performante** : Les cubes OLAP facilitent l'exploration des mutations génétiques selon plusieurs axes d'analyse (par exemple, par gène, par maladie, par type de variant).
- **Accessibilité des résultats** : Le développement d'une interface de visualisation rend l'interprétation des résultats accessible même aux utilisateurs non spécialistes en Big Data.

3 Étude de cas

3.1 Présentation du cas choisi

L'étude de cas porte sur l'analyse d'un fichier VCF (Variant Call Format) provenant de la base de données ClinVar. Ce fichier contient des informations sur les variants génétiques associés à des maladies héréditaires. L'objectif de cette étude est de proposer une solution pour mieux comprendre l'impact de ces mutations sur la santé et de pouvoir identifier des mutations génétiques liées à des maladies rares.

3.2 Description du problème

Dans le domaine de la génétique, il existe une grande quantité de données relatives aux mutations génétiques, mais ces données sont souvent difficiles à analyser en raison de leur volume et de leur complexité. Les outils actuels ne permettent pas toujours de traiter efficacement ces données, notamment pour identifier les variants responsables de maladies rares. Ainsi, l'enjeu est de trouver une solution pour effectuer cette analyse de manière plus efficace.

3.3 Notre solution proposée

Pour répondre à ce besoin, nous avons développé une solution qui consiste à :
- **Télécharger les données génétiques** : Nous avons récupéré le fichier `clinvar.vcf` qui contient des informations sur les variants génétiques.
- **Convertir le fichier VCF en CSV** : Cette transformation permet une manipulation plus facile des données à l'aide d'outils comme Pandas.
- **Appliquer un processus ETL** (Extraction, Transformation, Chargement) pour nettoyer et préparer les données.
- **Utiliser des modèles d'apprentissage automatique**, comme le modèle Random Forest, pour sélectionner les variables importantes et nettoyer les données.

— **Construire un cube OLAP** pour une analyse multidimensionnelle des relations entre les gènes, les mutations et les maladies associées.
— **Visualiser les résultats** à l'aide de graphiques générés avec Seaborn, ce qui permet une interprétation plus simple des données.

3.4 Technologies et outils utilisés

Les technologies clés utilisées pour cette solution comprennent :
— **Pandas** : pour la manipulation des données.
— **Random Forest** : pour le nettoyage avancé et la sélection des variables importantes.
— **OLAP cubes** : pour l'analyse multidimensionnelle des données.
— **Seaborn** : pour la visualisation des résultats.

3.5 Analyse des résultats

Les premiers résultats montrent que cette solution permet d'identifier des clusters de variants génétiques associés à des pathologies rares. Par exemple, certains variants sont fortement corrélés avec des maladies comme la fibrose cystique et la dystrophie musculaire de Duchenne. L'analyse multidimensionnelle à l'aide des cubes OLAP a permis de faire ressortir ces relations de manière claire et compréhensible.

3.6 Limites et perspectives

Bien que cette solution soit prometteuse, elle présente certaines limites. Par exemple, le fichier VCF utilisé peut contenir des données manquantes ou incorrectes, ce qui peut affecter la qualité de l'analyse. De plus, certaines mutations génétiques peuvent être mal caractérisées.

Pour les perspectives futures, il serait intéressant de tester cette approche sur d'autres bases de données génétiques et d'introduire des algorithmes plus avancés, comme les réseaux neuronaux, pour améliorer la prédiction des relations entre les mutations génétiques et les maladies. Une validation des résultats par des études cliniques ou des analyses fonctionnelles serait également nécessaire.

3.7 Conclusion

Dans cette étude de cas, nous avons proposé une solution originale pour analyser les variants génétiques issus de fichiers VCF. Notre approche repose sur une série d'étapes méthodologiques et l'utilisation d'outils modernes pour traiter et analyser les données, offrant ainsi un potentiel pour améliorer la compréhension des maladies génétiques rares.

4 Implémentation

4.1 Configuration de la machine

Machine 1

- **Système d'exploitation :** Windows 10
- **Mémoire RAM :** 6 Go
- **Processeur :** Intel Core i5

Machine 2

- **Système d'exploitation :** (à compléter)
- **Mémoire RAM :** (à compléter)
- **Processeur :** (à compléter)

4.2 Environnement de développement

- **Langage :** Python 3.12.7
- **Environnement :** Jupyter Notebook

4.3 Bibliothèques principales utilisées

- `pandas` : pour la manipulation des données
- `scikit-learn` : pour l'application du modèle Random Forest
- `seaborn` : pour la visualisation graphique
- `matplotlib` : pour les graphiques complémentaires

4.4 Architecture logicielle

1. **Téléchargement du fichier** `clinvar.vcf`.
2. **Conversion** du fichier VCF en format `CSV` pour faciliter la manipulation.
3. **ETL (Extraction, Transformation, Chargement)** :
   - Extraction des colonnes pertinentes
   - Nettoyage des valeurs manquantes et aberrantes
4. **Nettoyage et prétraitement** avec un modèle Random Forest pour sélectionner les variables importantes.
5. **Construction d'un cube OLAP** pour l'analyse multidimensionnelle.
6. **Visualisation** des résultats à travers une interface graphique simple (graphiques Seaborn, tableaux).

4.5 Détails de l'implémentation

- **Conversion VCF → CSV :** Utilisation d'un script Python pour parser les données brutes.
- **Nettoyage des données :**
  - Suppression des colonnes inutiles
  - Traitement des valeurs nulles
  - Transformation des types de données
- **Sélection des attributs :** Utilisation d'un classifieur Random Forest pour estimer l'importance des variables.
- **Cube OLAP :** Structuration des données pour permettre des analyses par dimensions (gènes, maladies, types de mutation...).
- **Interface de visualisation :**
  - Graphiques de répartition (`seaborn.histplot`, `seaborn.countplot`)
  - Diagrammes de corrélation (`seaborn.heatmap`)
  - Tableaux dynamiques

5 Conclusion

L'analyse des variants génétiques pour les maladies héréditaires constitue un domaine important dans la bioinformatique, permettant de mieux comprendre les mutations génétiques responsables de certaines pathologies. Le projet a permis de démontrer l'efficacité d'une approche basée sur des bases de données génétiques volumineuses, permettant d'extraire, d'analyser et de stocker ces données dans un format structuré. Bien que le projet ait utilisé des outils des techniques pour le traitement des données, des solutions plus avancées peuvent être envisagées pour une analyse plus rapide et plus complexe à l'avenir.

Références

[1] Fondation des Maladies Génétiques. *Difficultés d'interprétation des tests génétiques en milieu clinique*. Accessed : 2025-04-17. 2022.

[2] Global Rare Diseases Research. *Etude des mutations génétiques dans les maladies rares*. Accessed : 2025-04-17. 2020.

[3] Organisation mondiale de la santé. *Rapport sur les maladies rares*. Accessed : 2025-04-17. 2021.