LISTE DES ACRONYMES ET ABRÉVIATIONS

ACP	Analyse en Composantes Principales
AFDM	Analyse Factorielle de Données Mixtes
AFND	Allele Frequency Net Database
ACM	Analyse en Composantes Multipes
BMDW	Bone Marrow Donors Worldwide
CBICA	Center for Biomedical Image Computing and Analytics
CDC	Codage Disjonctif Complet
CHU	Centre Hospitalier Universitaire
CMH	Complexe Majeur d'Histocompatibilité
CNIL	Commission Nationale de l'Informatique et des Libertés
DIVAT	Données Informatisées et Validées en Transplantation
DL	Deep Learning
DME	Dossier Médical Electronique
DNN	Deep Neural Networks
EM	Estimation-Maximisation
FAMD	Factor Analysis Of Mixed Data
FATE	Federated AI Technology Enabler
FLS	Federated Learning System
FL	Federated Learning
GSB	Gold, Silver, Bronze
HDFS	Système des Fichiers Distribués Hadoop
HIPAA	Health Insurance Portability and Accountability
HI	Human Immunology journal
HLA	Human Leucocyte Antigens
IaaS	Infrastructure en tant que service
IHWS	International HLA and Immunogenetics Workshops
IMC	Indice de Masse Corporelle
IoT	Internet des Objets

IPD	Base de données Immuno Polymorphism
KTD-innov	Kidney Transplantation Diagnostics - Innovation
MO	Moelle Osseuse
ML	Machine Learning
MSE	Multi Source Estimation
OSE	One Source Estimation
PaaS	Plate-forme en tant que service
POC	Proof Of Concept
POI	Patient Of Interest
RDD	Resilient Distributed Datasets
POR	Population Of Reference
RGPD	Règlement Européen sur la Protection des Données
SaaS	Logiciel en tant que service
SD	Standard Deviation
SGD	Descente de Gradient Stochastique
SPR	Short Population Report
SVM	Support Vector Machine
WMDA	World Marrow Donor Association

SOMMAIRE

TABLE DES FIGURES

LISTE DES TABLEAUX

CHAPITRE 1

INTRODUCTION

Contents

1.1 Motivations

Dû à la croissance énorme de la quantité des données produites dans le monde, nous nous trouvons devant une évolution très rapide d'analyses de données massives dans pratiquement tous les domaines biomédicaux et d'ingénierie. Des plus grandes quantités de données permettent généralement aux analyses d'obtenir des résultats plus précis et atteindre une puissance statistique plus grande. Elle se place à la base de l'amélioration de connaissances dans de nombreux secteurs tels que la médecine, le secteur bancaire, le transport, etc.

La médecine de précision est un domaine médical émergent qui a pour objectif de cibler des diagnostics et des traitements médicaux à des petits groupes de personnes ou même des individus. Ainsi, la prévention et les traitements des maladies peuvent tenir compte de la variabilité de facteurs tels que génétiques, métaboliques ou environnementaux. Cela contraste avec l'approche médicale actuelle, qui appliquent des stratégies de prévention et des traitements développés pour des représentants « moyens » de (très) larges groupes de personnes, sans pouvoir prendre en compte les différences entre les individus. [122].

Dans le domaine de la médecine de précision, la disponibilité de larges quantités de données, associée à l'accélération de la technologie et à une précision accrue des tests, permet aux chercheurs de découvrir, entre autres, de nouveaux bio-marqueurs et de mieux comprendre les maladies au niveau moléculaire et génétique. Cela permettra à terme de fournir des traitements ciblés à des populations de patients stratifiées pour des bénéfices médicaux optimaux. [95]

Dans le cadre de l'évolution de la médecine de précision, les collaborations nationales et internationales se sont notamment développées afin d'améliorer les analyses biomédicales. Les architectures classiques centralisées régissant les analyses actuelles ne sont pas utiles dans ce contexte car les données nécessaires ne peuvent très souvent pas être partagées pour des raisons de protection de données, des problèmes de sécurité et/ou des problèmes de performances. Afin de déployer ces analyses biomédicales dans des contextes de distribution large, voir géo-distribués, de nouveaux algorithmes distribués doivent être développés. Les calculs doivent être, en particulier, réalisés par rapport à des bases de données médicales distribuées à grande échelle, qui sont placées sur des sites différents.

Dans le contexte des systèmes et des algorithmes médicaux, l'analyse et le partage de données parmi différents sites sont souvent restreints pour des raisons réglementaires, de gouvernance de données, ainsi que de contraintes techniques et scientifiques. Les politiques de protection des données comme le RGPD [42], au niveau européen, et le HIPAA [55] aux Etats-Unis, imposent des restrictions strictes sur le partage des données. Par conséquent, les analyses sont souvent réalisables seulement sur les sites des propriétaires des données. De plus, des quantités massives de données sont difficiles à partager ou à transférer à cause des coûts de l'utilisation des ressources de calcul, de stockage et de réseau.

1.2 Contributions

Le point de départ de cette thèse multidisciplinaire ont été deux observations de l'équipe du CR2TI (Center for Research in Transplantation and Translational Immunology) dont j'ai fait partie : le manque d'algorithmes distribués, notamment fondés sur des calculs statistiques, et le manque de contrôle sur des données, en particulier, en utilisant des bases de données médicales. Dans ce contexte, nous nous intéressons principalement à des nouveaux algorithmes distribués qui réconcilient protection de données, sécurisation et performance. Pour chaque nouvelle analyse, nous évaluons la correction des résultats, la performance et des propriétés relatives aux analyses classiques centralisées.

Cette thèse est structurée autour de trois projets :

1. Le projet de recherche Kidney Transplantation Application (KITAPP) [53], qui s'inscrit dans une coopération nationale, KTD (Kidney Transplantation Diagnostics). Il s'agit d'un projet de médecine de précision pour suivre l'état des transplantés rénaux au cours du temps (pour une présentation plus détaillée, voir la section 3.1). KITAPP permet la contextualisation de problèmes rénaux de patients individuels par rapport à des cohortes

entières de patients sous traitement dans les hôpitaux. En particulier, elle fournit une contextualisation référentielle où l'on compare un patient à des groupes, définis avec l'aide de cliniciens qui, sont sujets, par exemple, à un rejet aigu de greffe, un rejet humoral ou un rejet cellulaire.

Actuellement, le calcul des centiles [59] et l'analyse factorielle des données mixtes (FAMD) [96], qui font partie des algorithmes utilisés dans ce projet pour la contextualisation d'un patient, sont réalisés de manière centralisée sur un seul site.

Dans le cadre de ce projet, nous avons développé deux nouvelles contributions pour une contextualisation distribuée et sécurisée :

— Un nouveau modèle pour un calcul des centiles distribué et sécurisé pour une évaluation collaborative des données des patients. Ce travail a été présenté dans la conférence «AICCSA'20» [106].

— Une nouvelle méthode distribuée et sécurisée pour l'analyse factorielle des données mixtes. Cette méthode permet de réduire les dimensions des données d'origine en un sous-espace inférieur afin de faciliter la phase de la contextualisation. Ce travail a été présenté dans la conférence «AINA'21» [108].

2. La base de données Allele Frequency Net Database (AFND) [2] fournit à la communauté scientifique un référentiel librement disponible pour le stockage des fréquences des gènes immunitaires dans différentes populations à travers le monde. Les utilisateurs intègrent les résultats de leur travail dans une base de données commune et y cherchent les informations déjà disponibles. Aujourd'hui, l'AFND impose la centralisation de la base de données qui impose le transfert de données individuelles ou de données agrégées (fréquence allélique à la section 3.2.1 ou fréquences des haplotypes à la section 3.2.3), cela peut engendrer une possibilité de dégradation des données. La dégradation est imposée en respectant la confidentialité due aux patients (les données individuelles ne peuvent pas être centralisées) donc, ce sont des données agrégées sous la forme de fréquences. De plus, la centralisation n'est pas seulement un risque de perdre toutes les données en cas d'incident, mais aussi une perte de contrôle sur les usages.

   Comme solution à ce problème, nous proposons un nouveau modèle distribué pour ces analyses (calculs de la fréquence allélique, la fréquence des haplotypes et les données génotypiques individuelles). Ces algorithmes évitent de partager les données de chaque site entre eux, afin de garantir le contrôle d'usage des données et, bien évidemment, la confidentialité des données. Nous considérons également des problèmes de contrôle d'utilisation des données dans ce contexte. Ce travail a été accepté pour publication dans

le journal «Exploration of Immunology».

3. Le développement de nouvelles techniques pour l'estimation de la fréquence des haplotypes, sans avoir accès aux génotypes des parents, est très utile pour les cliniciens. L'utilisation de l'algorithme Expectation-Maximization (EM) est parmi les méthodes qui ont montré leurs succès pour analyser la vraisemblance des génotypes à partir des haplotypes *HLA* sur une base de données centralisée.

   Par la suite, nous proposons une nouvelle version distribuée pour l'estimation des fréquences des haplotypes en se basant sur l'algorithme EM distribué. Cette contribution a été soumise comme papier court dans le journal "International Journal of Immunogenetic".

1.3 Publications

Papiers soumis dans des journaux

Sayadi S., Vince N., Südholt M.& Gourraud P. A. (2022). Distributed *HLA* Haplotype Frequency Estimation Using Expectation Maximization. In *International Journal of Immunogenetics* (**International Journal of Immunogenetics**).

Papiers publiés dans des journaux

Sayadi S., Douillard V., Vince N., Südholt M.& Gourraud P. A. (2022). Distributing *HLA* database in histocompatibility : a shift in *HLA* data governance. In *Exploration of Immunology Journal* (**Exploration of Immunology Journal**) [107].

Papiers publiés dans des conférences

Sayadi S., Geffard E., Südholt M., Vince N. & Gourraud, P. A. (2021). Secure distribution of Factor Analysis of Mixed Data (FAMD) and its application to personalized medicine of transplanted patients. In *35-th International Conference on Advanced Information Networking and Applications", vol 225, Springer, Cham, Toronto, Canada* (**AINA-2021**) [108].

Sayadi S., Geffard E., Südholt M., Vince N. & Gourraud P. A. (2020). Distributed Contextualization of Biomedical Data : a case study in precision medicine. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications* (**AICCSA-2020**) [106].

1.4 Aperçu de la thèse

Cette section présente le contenu des chapitres de ce mémoire.

Le chapitre 2 présente l'état de l'art de trois différentes domaines. Il présente les travaux actuels liés aux questions de recherche présentées dans la section 1.2. La section 2.2 présente une revue de la littérature existante concernant les technologies les plus utilisées pour les analyses distribuées à grande échelle et plus précisément pour les analyses biomédicales. Dans 2.3, nous présenterons les mécanismes de confidentialité existants les plus utilisés selon la littérature pour sécuriser les analyses biomédicales distribuées.

Le chapitre 3 introduit quelques outils et méthodes d'analyses en bio-informatique qui sont importants pour nos travaux. La section 3.1 présente l'application qui permet le suivi de transplantés rénaux KITAPP qui offre des techniques de contextualisations personnalisées. Cette application est un nouveau service pour l'amélioration des soins effectués par les cliniciens.

Les quatre chapitres suivants sont consacrés à nos contributions.

Le chapitre 4 présente une première version de contextualisation distribuée et sécurisée sur des données biomédicales. Cette contribution permet d'aider les cliniciens à évaluer l'état des données et à avoir un autre point de vue pour prendre des décisions. Notre nouvelle approche a été appliquée dans un cadre d'une application dans le domaine de la transplantation rénale KITAPP.

Le chapitre 5 présente une nouvelle méthode distribuée et sécurisée pour la réduction des dimensions de données quantitatives et qualitatives. Cette méthode permet de transformer les données complexes (mixtes), provenant de plusieurs sites, en un sous-espace de dimension inférieure, tout en préservant les caractéristiques importantes des données originales.

Le chapitre 6 décrit une nouvelle version alternative distribuée et sécurisée de la ressource AFND. Notre nouveau modèle distribué fournit les même calculs que l'AFND centralisée concernant les parties des analyses effectuées sur les données *HLA*.

Le chapitre 7 présente une nouvelle méthode distribuée pour l'estimation des fréquences des haplotypes *HLA* sans accéder aux génotypes des parents en utilisant l'algorithme EM distribué.

Enfin, le dernier chapitre conclu ce manuscrit. Nous y présenterons une synthèse de nos contributions et nos perspectives.

CHAPITRE 2

ÉTAT DE L'ART

Contents

Ce chapitre résume la littérature scientifique existante liée aux analyses biomédicales géo-distribuées et sécurisées. Nous présenterons dans un premier temps une revue de littérature sur des technologies qui permettent des analyses géo-distribuées, en précisant pour chaque technologie les travaux existants liés aux analyses biomédicales distribuées pour améliorer la médecine de précision. Ensuite, nous définirons quelques mécanismes de sécurité et de confidentialité les plus adaptés aux calculs distribués.

2.1 Introduction

Pour réaliser notre état de l'art, nous avons choisi d'étudier des technologies qui permettent d'exécuter des analyses distribuées à grande échelle. Dans ce contexte, nous avons identifié les quatre technologies les plus importantes selon leurs utilisations dans la bibliographie :

— Cloud computing,

— Blockchain,
— Les technologies de calcul Big data telles que Map Reduce, Hadoop, Spark et Apache Storm
— Apprentissage et analyse fédérés.

Pour chaque technologie, nous allons présenter ces concepts importants, quelques travaux existants relatifs aux analyses distribuées à grande échelle et des travaux liés à des analyses biomédicales distribuées pour la médecine de précision. Nous étudierons à la fin quelques techniques de sécurité qui permettent de renforcer la privatisation des données sensibles pour les analyses distribuées.

2.2 Des technologies pour des analyses distribuées à grande échelle

Dans cette section, nous allons décrire les technologies citées ci-dessus en présentant les travaux les plus importants relatifs à notre problématique.

2.2.1 Cloud computing

Traditionnellement défini comme la fourniture des services informatiques (des serveurs, du stockage, des bases de données, de gestion réseau, des logiciels, des outils d'analyse, d'intelligence artificielle) via Internet (le cloud) dans le but d'offrir une innovation plus rapide, des ressources flexibles et des économies d'échelle [18]. Grâce à sa propriété de mutualisation des ressources, le cloud computing a permis de démocratiser l'informatique dans les entreprises. Il a contribué à la centralisation de l'infrastructure pour un grand nombre d'utilisateurs pour plusieurs applications en même temps, ce qui facilite les collaborations entre les équipes. Il a aussi permis d'adapter la demande des ressources informatiques nécessaires selon l'évolution du besoin du métier, ce qui permet de payer les ressources selon le besoin [41].

Les types du Cloud

Pour les types de Cloud, on peut trouver aujourd'hui :

— Cloud public : C'est une infrastructure informatique dans laquelle les ressources d'un fournisseur de services sont mises à la disposition de tous via internet comme "DropBox" [27].

— Cloud privé : Contrairement au cloud public, ce type concerne un petit ensemble de groupes qui font partie à un cercle fermé. L'hébergement de l'infrastructure informatique dans ce type de cloud n'est pas partagé. Le cloud privé est connu par un niveau de sécurité et de contrôle le plus élevé car un réseau privé est utilisé.
— Cloud hybride : Ce cloud utilise des clouds privés et publics, en fonction de leur objectif. On peut héberger les applications les plus importantes sur nos propres serveurs, pour les garder plus sécurisées et héberger les applications secondaires ailleurs.
— Cloud communautaire : Le cloud communautaire est partagé entre des organisations ayant un objectif commun ou qui s'intègrent dans une communauté spécifique (communauté professionnelle, communauté géographique, etc.).

Les services Cloud

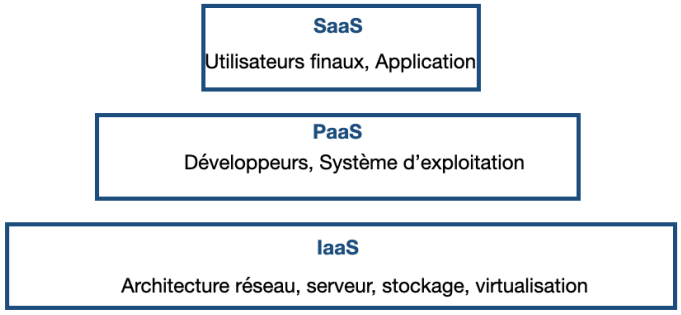


FIGURE 2.1 – Les différents services de Cloud computing.

Pour les principaux services de Cloud, comme illustré dans la figure 2.1, on peut citer :

— Infrastructure en tant que service (IaaS) : IaaS est la catégorie basique des services informatiques en cloud qui permet de louer à la carte une infrastructure informatique (serveurs ou machines virtuelles) auprès d'un fournisseur de services informatique en cloud. Ce service ne fournit aucune infrastructure logicielle. L'utilisateur doit fournir lui-même le système d'exploitation et les services distribués de base.
— Plateforme en tant que service (PaaS) : La plate-forme en tant que service (PaaS) est destinée aux développeurs ou aux entreprises de développement. En plus des services IaaS, il ajoute un système d'exécution et des services distribués de base. Il fait référence à la fourniture d'un environnement à la demande pour le développement, le test, la fourniture et la gestion d'applications logicielles. Il est conçu pour créer rapidement

des applications Web ou mobiles, sans se soucier de la configuration ou de la gestion de l'infrastructure sous-jacente de serveurs, de stockage, de réseau et de bases de données nécessaires au développement.

— Logiciel en tant que service (SaaS) : Le logiciel en tant que service (SaaS) est une méthode permettant de fournir des applications logicielles sur Internet à la demande et sur abonnement. SaaS permet de faire héberger et de gérer l'application logicielle et l'infrastructure sous-jacente, ainsi que de faire gérer toute maintenance (mises à niveau logicielles et correctifs de sécurité).

Cloud computing pour le domaine biomédical

Dans le cadre des projets medicaux collaboratifs, on se trouve devant la nécessité d'une réelle distribution des analyses biomédicales. Dans ce contexte, on peut citer le projet I-CAN [63] dans lequel 34 hôpitaux français apportent leurs données à des serveurs centralisés pour un premier stockage. Le traitement des données, qui comprend les données cliniques, d'imageries et génétiques, est alors effectué sur un serveur différent mais également de manière centralisée, ce qui implique une transmission de données à grande échelle au moment du calcul, qui nécessite la disponibilité d'un réseau à haut débit.

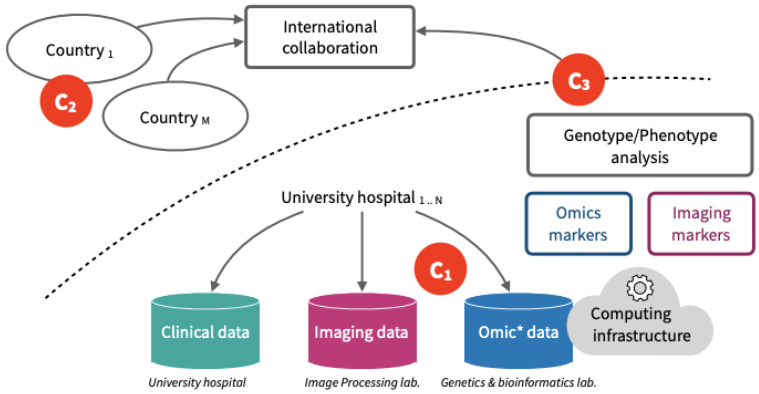


FIGURE 2.2 – Architecture informatique distribuée basée sur le Cloud computing © 2019 IEEE [13].

Dans le cadre de ce projet, Boujdad et al. [13] ont montré l'importance de l'utilisation des clouds publics sécurisés et des clouds communautaires pour résoudre les problèmes de reproductibilité, de confidentialité et d'évolutivité dans le contexte des études multi-centres. Leur architecture présentée dans la figure 2.2 a montré que les données génétiques séquencées au laboratoire de génétique et bioinformatique, peuvent être apportées à une cohorte internationale

dans le cadre d'une analyse collaborative internationale, en utilisant le cloud computing. C1, C2 et C3 sont les 3 propriétés de reproductibilité, de confidentialité et d'évolutivité que cette architecture a garanti aux données.

2.2.2 La technologie Blockchain

La technologie Blockchain, bien que datant des années 1980, a été décrite et implémentée pour la premiere fois par Nakamoto [92] "bitcoin a peer-to-peer electronic cash system". Comme le montre la figure 2.3, il s'agit d'une technologie de stockage de grands registres décentralisés, utile comme technique de sécurité pour l'authentification, l'autorisation et la vérification des données générées. Avec la technologie blockchain, le concept de consensus est devenu un mécanisme garantissant la confiance dans la communication entre deux entités sans l'intervention d'un intermédiaire. Aujourd'hui, la blockchain est omniprésente dans la crypto-monnaie, les contrats intelligents, la gestion des identités numériques, les applications de contrôle d'accès, l'assurance automatisée entre homologues, dans les banques et dans de nombreuses autres applications [105].

FIGURE 2.3 – Le concept général de la technologie blockchain.

Les enjeux de la blockchain pour le domaine biomédical

Nous présentons dans cette partie les propriétés de la blockchain qui peuvent justifier l'utilisation de cette technologie comme une solution prometteuse pour les analyses biomédicales distribuées.

— **Décentralisation** : La nature des soins de santé, dans laquelle il y a des intervenants qui sont répartis, nécessite des systèmes de gestion décentralisés. La technologie blockchain peut devenir ce réseau central de gestion de données de santé décentralisé à partir duquel tous les participants peuvent avoir un accès contrôlé aux mêmes données de santé, sans jouer le rôle d'autorité centrale sur les données de santé mondiales.
— **Transparence** : Grâce à sa nature ouverte et transparente, la blockchain crée une atmosphère de confiance autour des applications de soins de santé distribuées. Cela facilite l'acceptation de telles applications par les acteurs de la santé.
— **Immuabilité, sécurité des données et confidentialité** : Cette propriété de blockchain permet d'améliorer considérablement la sécurité des données de santé qui y sont stockées, car les données, une fois enregistrées dans la blockchain, ne peuvent pas être corrompues, modifiées ou récupérées. Toutes les données de santé sur la blockchain sont cryptées, horodatées et ajoutées dans un ordre chronologique. De plus, les données de santé sont sauvegardées dans la blockchain à l'aide de clés cryptographiques qui aident à protéger l'identité et les données privées des patients.
— **Les contrats intelligents** : Les contrats intelligents sont simplement des programmes stockés sur une blockchain qui s'exécutent lorsque des conditions prédéterminées sont remplies. Ils sont généralement utilisés pour automatiser l'exécution d'un accord afin que tous les participants puissent être immédiatement certains du résultat, sans intervention d'un intermédiaire ni perte de temps. Grâce à cette caractéristique, les patients peuvent préserver la propriété de leurs données et maîtriser leurs utilisations. Les patients ont besoin de l'assurance que leurs données de santé ne sont pas utilisées à mauvais escient par d'autres parties prenantes et doivent disposer d'un moyen de détecter les cas d'abus. La blockchain aide à répondre à ces exigences via des protocoles cryptographiques puissants et des contrats intelligents bien définis.
— **Disponibilité/ Robustesse** : Puisque les enregistrements de la blockchain sont répliqués dans plusieurs nœuds, la disponibilité des données de santé stockées dans la blockchain est garantie car le système est robuste et résiste aux pertes de données, à la corruption des données et a certaines attaques de sécurité portant sur la disponibilité des données.
— **Intégrité des données** : Même sans accéder au texte en clair, des enregistrements sont stockés dans la blockchain. L'intégrité et la validité de ces enregistrements peuvent être vérifiées. Cette fonctionnalité est très utile dans les domaines de la santé dans lesquels la vérification des enregistrements est une exigence.

Blockchain pour le domaine biomédical

Inspirés par la littérature [3], nous présentons maintenant l'utilisation de la technologie blockchain dans le domaine biomédical :

Gestion des dossiers médicaux éléctroniques (DME) : Les propriétés de la blockchain comme la décentralisation, l'immuabilité, la provenance des données, la fiabilité, la robustesse, les contrats intelligents, la sécurité et la confidentialité la rendent très appropriée pour le stockage et la gestion des dossiers médicaux électroniques des patients. Conformément au règlement général européen sur la protection des données (RGPD), qui interdit le traitement de données personnelles et sensibles des patients, la blockchain est proposée comme une technologie viable permettant de créer une plateforme pour faciliter le partage de données des patients entre différents intervenants du secteur de la santé tout en responsabilisant les patients pour maîtriser le partage, le traitement et l'utilisation de leurs données. On peut trouver plusieurs applications de DME basées sur une Blockchain : MedBlock [37] , BlockHIE [65], FHIRChain [130], MeDShare [125] etc.

Réclamation d'assurance maladie : Le traitement des réclamations d'assurance dans le secteur de la santé peut bénéficier de la transparence de la blockchain, de sa décentralisation, de son immuabilité et de son auditabilité des enregistrements qui y sont stockés. Un certain nombre de papiers identifient le traitement des demandes d'assurance comme un domaine très prometteur pour l'application de la blockchain des soins de santé.

Chaîne d'approvisionnement pharmaceutique : Une autre utilisation de la blockchain concerne la gestion de la chaîne d'approvisionnement en santé, en particulier dans le domaines de l'industrie pharmaceutique où la livraison de médicaments contrefaits ou non conformes aux normes peut avoir des conséquences désastreuses pour les patients. L'idée générale est d'enregistrer toutes les transactions relatives à la prescription de médicaments sur le réseau blockchain auquel toutes les parties prenantes (fabricants, distributeurs, médecins, patients et pharmaciens) sont connectés. De cette manière, toute modification malveillante de la prescription par l'une des parties peut être détectée.

Recherche biomédicale et éducation : La recherche et l'éducation biomédicale constituent aussi des cas d'utilisation pour la blockchain. Benchoufi et al. [82] expliquent la mise en œuvre de la traçabilité du consensus dans les essais cliniques utilisant le protocole de blockchain.

Nugent et al. [116] présentent leurs recherches dans lesquelles ils démontrent comment des contrats intelligents sur la plateforme Ethercha blockchain peuvent être utilisés pour améliorer la transparence des données dans les essais cliniques. La plateforme Ethereum [91] est également utilisée pour implémenter une autre solution basée sur une blockchain, qui est proposée pour authentifier des documents extraits de bases de données biomédicales.

Analyse des données de santé : Blockchain offre également une occasion unique de tirer parti de la puissance d'autres technologies émergentes telles que le deep learning et les techniques d'apprentissage par transfert pour réaliser une analyse prédictive des données de santé et faire progresser la recherche dans le domaine de la médecine de précision. Juneja et Marefat [68] ont mené une recherche expérimentale dans laquelle la blockchain est proposée en tant que gestionnaire de contrôle d'accès, pour stocker en toute sécurité, et accéder aux données requises par le classificateur de l'arythmie en temps réel à partir d'un stockage de données externe.

Surveillance des patients à distance : La surveillance des patients à distance implique la collecte de données biomédicales par le biais de capteurs de zone corporelle (ou appareils IoT) et d'appareils mobiles afin de pouvoir surveiller à distance l'état du patient en dehors des environnements de soins de santé traditionnels tels que l'hôpital. La Blockchain a été proposée comme moyen de stockage, de partage et de récupération des données biomédicales collectées à distance.

Toutes les propriétés et les travaux cités ci-dessus ont montré que la blockchain pourrait être une solution prometteuse pour notre problématique. Cependant, certains avantages surpassent nos besoins et peuvent alors constituer des inconvénients. La garantie de la disponibilité des données par exemple est très importante, mais nous n'avons pas besoin de stocker toutes les transactions et nous n'avons pas besoin non plus que toutes les parties possèdent les mêmes transactions stockées. En plus, tout cela peut dégrader les performances et augmenter la consommation en énergie de nos résultats.

2.2.3 Big Data

Le big data ou mégadonnées désignent l'ensemble des données numériques produites par l'utilisation des nouvelles technologies à des fins personnelles ou professionnelles. Cela recoupe les données en (courriels, documents, bases de données, historiques de processeurs métiers...) aussi bien que les données issues de capteurs, des contenus publiés sur le web (images, vidéos, sons, textes), des transactions de commerce électronique, des échanges sur les réseaux

sociaux, des données transmises par les objets connectés (étiquettes électroniques, compteurs intelligents, smartphones...), des données géolocalisées, etc.

Les technologies Big Data et leurs utilisations dans le domaine des analyses biomédicales distribuées

MapReduce : Un modèle pour le traitement parallèle de grands volumes de données en deux phases, Map et Reduce. En phase de projection, la fonction mappeur est appliquées à des données d'entrée des listes de n-uplets et renvoie une liste intermédiaire. En phase de réduction, le réducteur fusionne la liste des n-uplets retournés, créant un ensemble de paires [26, 17]. MapReduce est un moyen pour traiter de gros volumes de données en parallèle sur des supercalculateurs parallèles ou des ordinateurs distribués.

Dans le domaine biomédical, MapReduce est couramment utilisé dans de nombreux outils pour résoudre des problèmes biomédicaux, par exemple la cartographie du séquençage de nouvelle génération, l'identification des SNP et l'analyse du séquençage [117].

Hadoop : C'est le framework le plus répandu pour le traitement de grands ensembles de données utilisant des clusters dans le cadre d'un modèle de programmation simple.

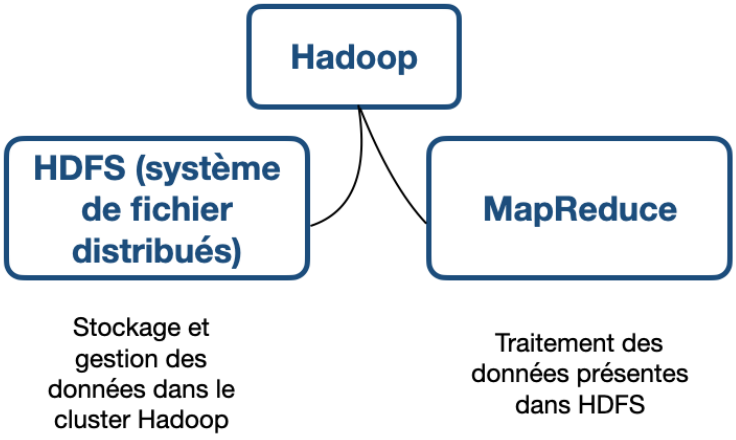


FIGURE 2.4 – L'architecture de Hadoop.

Comme montré dans la figure 2.4, l'architecture de haut niveau est composée de deux composants : le système de fichiers distribués Hadoop (HDFS) (pour le stockage) et MapReduce (pour le traitement) [26]. Hadoop permet de traiter de gros volumes de données en les partition-

nant parmi divers nœuds, chacun résolvant ensuite un problème spécifique qui contribue à la résolution générale du problème. Hadoop est souvent utilisé par la communauté bio-informatique. Par exemple, Driscoll et al. [94] proposent une catégorisation des approches existantes basées sur Hadoop. Dwork et al. [27] présentent une méthode qui utilise HADOOP pour une analyse efficace des données de soins de santé.

SPARK : Un framework d'analyse de données volumineuses dans le cadre d'un modèle de programmation efficace en mémoire. Il implémente des RDD (Resilient Distributed Datasets) permettant des opérations rapides sur de grands volumes de données. Un RDD peut stocker en mémoire cache et mettre plusieurs opérations parallèles à réutiliser dans les nœuds [128]. Spark est basée sur des nœuds Master et Workers, qui comportent trois composants principaux : Driver, Worker et Executor. Contrairement à Hadoop, Spark prend en charge le flux de données et le traitement de données structurées en temps réel.

Dans le domaine biomédical, Spark a été largement utilisé dans les problèmes de séquençage, de découverte de médicaments et d'analyse de séquence de nouvelle génération [51]. Kouanou et al. [118] ont proposé deux architectures pour la classification d'images en utilisant les deux frameworks Hadoop et Spark.

Apache Storm : Un système distribué sous licence libre, utilisé pour traiter de gros volumes de données en temps réel. L'architecture Storm comprend des nœuds master et workers similaires aux technologies précédentes.

Yadav et al. [126] proposent un système de surveillance des soins médicaux utilisant Storm dans le cloud. Ce système garantit un temps de réponse court, offrant une solution adaptée à la télémédecine. Contrairement à Hadoop, Storm s'annonce comme le cadre idéal pour l'analyse de gros volumes de données en temps réel.

La figure 2.5 montre les similarités et les particularités des technologie Big Data présentées précédemment. Toutes ces technologies ont montré leur efficacité pour gérer des problématiques autour des analyses distribuées, mais la distribution des analyses dans les technologies Big Data est effectuée sur différents clusters et pas sur des sites géo-distribués. Cela ne répond pas efficacement à nos besoins. La combinaison de plusieurs technologies Big data dans un seul modèle est une idée pour résoudre notre problématique. Cependant, cette idée peut être compliquée et gourmande en ressources.

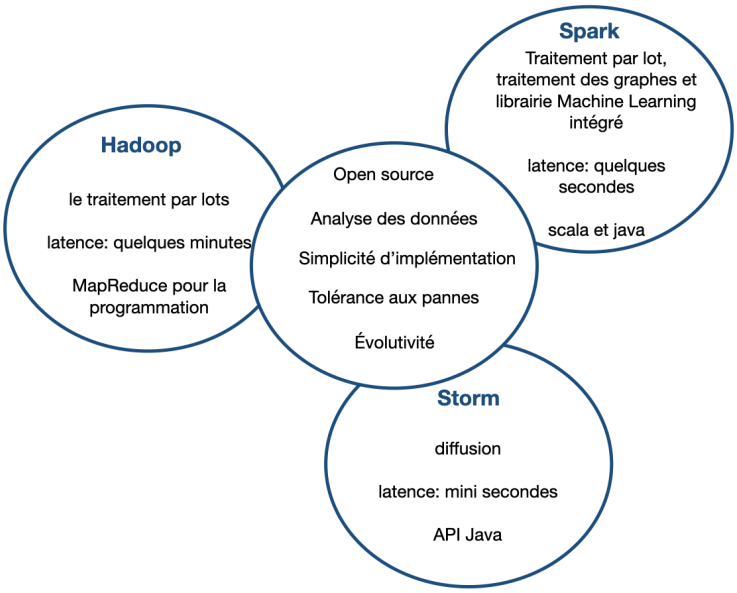


FIGURE 2.5 – Les points communs et les différences entre les technologies Big Data.

2.2.4 L'apprentissage fédéré

L'apprentissage ou apprentissage automatique (machine learning en anglais) est une application de l'intelligence artificielle qui permet aux systèmes d'apprendre et de s'améliorer à partir de l'expérience sans être explicitement programmés. L'apprentissage automatique se concentre sur le développement de programmes informatiques capables d'accéder aux données et de les utiliser pour apprendre par eux-mêmes.

L'apprentissage fédéré (federated learning en anglais) (FL) vise à former un algorithme d'apprentissage automatique sur plusieurs ensembles de données locaux contenus dans des nœuds locaux sans échanger d'échantillons de données. Le principe général consiste à former des modèles locaux sur des échantillons de données locales et à échanger des paramètres entre ces modèles locaux à une certaine fréquence pour générer un modèle global. Chaque cycle de ce processus consiste à transmettre l'état actuel du modèle global aux nœuds participants, à calculer les modèles locaux sur les nœuds locaux pour produire un ensemble de mises à jour potentielles du modèle sur chaque nœud, puis à agréger ces mises à jour locales en une seule mise à jour globale et l'appliquer au modèle global.

L'analyse fédérée (federated analysis en anglais) est une analyse décentralisée des données brutes stockées sur les appareils des utilisateurs. Il est utilisé pour les calculs de base sur le comportement des utilisateurs, qui ne nécessitent pas d'apprentissage automatique avec le

même principe de l'apprentissage fédérée.

Apprentissage horizontal et vertical

En fonction de la répartition des données, les systèmes d'apprentissage fédéré (FLS) peuvent généralement être classés en FLS horizontaux, verticaux et hybrides (voir figure 2.6).

Dans le FL horizontal, les jeux de données de différentes organisations ont le même espace d'entités mais peu d'intersection sur l'espace d'échantillon. Le système utilise un serveur pour agréger les informations des appareils et pour adopter une confidentialité différentielle et une agrégation sécurisée afin améliorer les propriétés de confidentialité. La reconnaissance des mots de réveil, comme "Hey Siri" et "OK Google", est une application typique de la partition horizontale car chaque utilisateur prononce la même phrase avec une voix différente.

Dans le FL vertical, les jeux de données de différentes organisations ont le même espace d'échantillonnage mais diffèrent dans l'espace d'entités. Le FLS vertical, adopte généralement des techniques d'alignement d'entités, pour collecter les échantillons chevauchants des organisations. Ensuite, les données qui se chevauchent sont utilisées pour former le modèle d'apprentissage automatique à l'aide de méthodes de chiffrement.

Les FLS existants se concentrent principalement sur un seul type de partition (horizontal ou vertical). La partition de données en utilisant les deux types est considérée hybride.

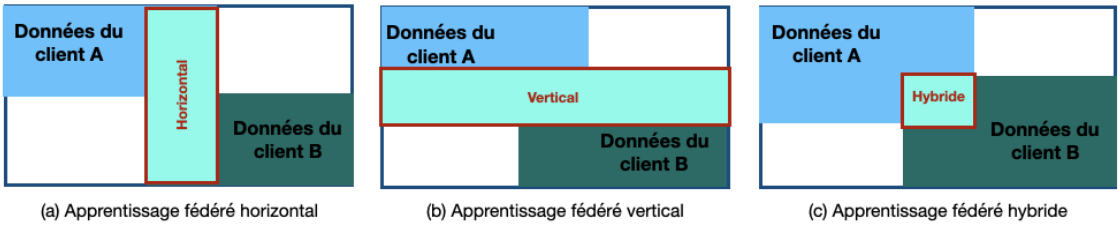


FIGURE 2.6 – Les types des partitions de l'apprentissage fédéré.

Federated Learning pour les analyses des données distribuées

Le FL est aujourd'hui une technologie algorithmique utilisée dans beaucoup de domaines d'applications :

L'évolution des technologies de communication : Google est parmi les premiers qui ont proposé un FLS évolutif qui permet à plus de dizaines de millions d'appareils Android d'ap-

prendre un réseau neuronal profond basé sur TensorFlow afin d'améliorer l'efficacité de leurs services [12, 86]. Dans la conception du modèle, on trouve un serveur qui agrège les mises à jour du modèle avec la moyenne fédérée. Ces moyennes sont calculées par les périphériques localement en cycles synchrones. La confidentialité différentielle et l'agrégation sécurisée sont utilisées dans ce travail pour améliorer les garanties de confidentialité. Corbacho aussi a implémenté PhotoLabeller [83], qui constitue un autre exemple de FLS. Il utilise des téléphones Android pour calculer des modèles localement et utilise la moyenne fédérée sur le serveur pour agréger le modèle. Enfin, le modèle formé est partagé entre tous les clients pour l'étiquetage des photos.

Secteur bancaire : La société WeBankFinTech a implémenté la plate-forme Federated AI Technology Enabler (FATE) [38], qui prend en charge plusieurs types de partitionnement de données et d'algorithmes. Les protocoles de calcul sécurisés sont basés sur un cryptage homomorphe et un calcul multi-parties. De nombreux algorithmes d'apprentissage automatique ont été pris en considération, y compris la régression logistique, les arbres de décision de renforcement du gradient, etc.

Gestion des ressources : Wang et al. [120] ont défini un algorithme d'apprentissage fédéré sur des systèmes à ressources limitées pour aborder le problème de l'utilisation efficace des ressources limitées de calcul et de communication. À l'aide de la moyenne fédérée, ils ont implémenté de nombreux algorithmes d'apprentissage automatique, y compris la régression linéaire, SVM et CNN.

Domaine biomédical : En 2018, Intel a entamé une collaboration avec le Center for Biomedical Image Computing and Analytics (CBICA) de l'Université de Pennsylvanie pour montrer la première application de validation de principe de l'apprentissage fédéré à l'imagerie médicale réelle. Cette application est utilisée concrètement pour la segmentation des tumeurs cérébrales [112, 61]. Cette étude initiale a démontré que l'apprentissage fédéré pouvait entraîner un modèle d'apprentissage en profondeur (U-Net) à 99 pourcent de la précision du même modèle formé avec la méthode traditionnelle de partage de données. Les chercheurs de NVIDIA du King's College London ont utilisé les GPU V100 Tensor Core de Nvidia pour la formation et l'inférence de modèles [33]. L'objectif est le déploiement d'une plateforme d'apprentissage fédérée et sécurisée, qui permettrait une « médecine de précision basée sur les données à grande échelle ». Brisimi et al. [15] ont développé des modèles de prédiction des hospitalisations pour

événements cardiaques à l'aide d'un algorithme distribué fédéré. Ils ont développé un cadre d'optimisation décentralisé général permettant à plusieurs détenteurs de données de collaborer et de converger vers un modèle prédictif commun, sans échanger explicitement les données brutes.

2.3 Des mécanismes de confidentialité pour les calculs distribués

La sécurité de l'information est un aspect très important de nos jours, surtout avec l'évolution des domaines du traitement et de l'analyse des données. La confidentialité, l'intégrité et la disponibilité qui composent la triade CIA sont indispensables à la protection de l'information dans les structures professionnelles. La confidentialité est la technique de sécurité la plus utilisée pour la protection des données sensibles. Elle permet de protéger les informations confidentielles pour qu'elles ne puissent être accédées que par les personnes qui y sont autorisées. La confidentialité est aussi un problème important pour les analyses distribuées et il y a eu de nombreuses attaques contre les modèles de distribution. De plus, il existe aujourd'hui de nombreux mécanismes de confidentialité, tels que la confidentialité différentielle [29] qui offre différentes garanties de confidentialité.

Nous présentons ici quelques approches populaires qui sont adoptées pour la protection des données au sein des systèmes distribués : l'agrégation et la fragmentation des données, les méthodes de cryptographie, la confidentialité différentielle et la synthétisation des données.

2.3.1 L'agrégation et la fragmentation des données

L'agrégation des données (ou moyennage des données) est un principe général qui implique l'union des base de données pour un envoi agrégé d'un message de beaucoup petites données pour des raisons d'efficacité. En calcul distribué, l'agrégation des données [85, 84] est un cadre largement utilisé pour éviter la communication de données brutes. Plus précisément, un modèle global est formé en agrégeant les paramètres du modèle des parties locales. Un algorithme typique pour l'apprentissage fédéré est une moyenne fédérée basée sur la descente de gradient stochastique (SGD), qui agrège les modèles calculés localement, puis met à jour le modèle global à chaque tour.

Contrairement à l'agrégation des données, la fragmentation des données est un processus de division de la base de données entière en plusieurs sous-tables afin que les données puissent être

stockées dans différents systèmes. Les petits morceaux de sous-tables sont appelés fragments. Ces fragments sont appelés unités de données logiques et sont stockés sur différents sites. Du fait qu'une seule information soit divisé et stocké sur différents sites, un attaquant ne peut pas y accéder en accédant à un seul site. La fragmentation des données est un mécanisme très important pour maintenir la sécurité et la confidentialité du système de base de données.

2.3.2 Les méthodes de cryptographie

Les méthodes cryptographiques telles que **le cryptage homomorphique** [52, 97, 81] et **le calcul multi-parties sécurisé** [11, 45] qui sont largement utilisées dans les calculs distribués pour préserver la confidentialité. Fondamentalement, les parties doivent crypter leurs messages avant de les envoyer, opérer sur les messages cryptés et décrypter la sortie cryptée pour obtenir le résultat. En appliquant les méthodes ci-dessus, la confidentialité des utilisateurs des analyses distribuées peut généralement être bien protégée. Par exemple, le calcul multi-parties sécurisé garantit que toutes les parties ne peuvent rien apprendre, sauf le résultat d'un calcul. Cependant, ces systèmes ne sont généralement pas efficaces et impliquent une surcharge de calcul et de communication importante.

2.3.3 La confidentialité différentielle

La confidentialité différentielle [29] est un mécanisme de sécurité qui permet aux chercheurs et aux analystes de données de bénéficier d'une facilité pour obtenir les informations utiles des bases de données, contenant les informations personnelles des individus, sans divulguer leurs identifications personnelles. Ceci peut être réalisé en introduisant une distraction minimale dans les informations fournies par la base de données. La distraction introduite est suffisamment importante pour être capable de protéger la vie privée, et en même temps suffisamment limitée pour que les informations fournies aux analystes soient toujours utiles. La confidentialité différentielle rend les données anonymes, en injectant soigneusement du bruit dans l'ensemble de données. Elle permet aux experts en données d'exécuter toutes les analyses statistiques possibles sans identifier aucune information personnelle. On peut trouver plusieurs systèmes basés sur la confidentialité différentielle dans l'état de l'art. Bonawitz et al. par exemple [12] adoptent une confidentialité différentielle pour la protection de la confidentialité des données, où les parties ne peuvent pas savoir si leurs données individuelles participent ou non à une analyse d'apprentissage. En ajoutant du bruit aléatoire aux données ou aux paramètres du modèle, la confidentialité différentielle supporte des calculs statistiques pour les données individuelles et

une protection contre l'attaque par inférence sur le modèle. En raison des bruits dans le processus d'apprentissage, ces systèmes ont tendance à produire des modèles moins précis.

2.3.4 La synthétisation des données

Les données synthétiques sont des informations construites artificiellement plutôt qu'extraites du monde réel. Les données synthétiques sont créées de manière algorithmique et sont utilisées comme substitut pour tester des ensembles de données de production ou de données opérationnelles, pour valider des modèles mathématiques et, de plus en plus, pour former des modèles d'apprentissage automatique.

Les avantages de l'utilisation de données synthétiques incluent la réduction des contraintes liées à l'utilisation de données sensibles ou réglementées, l'adaptation des besoins en données à certaines conditions qui ne peuvent pas être obtenues avec des données authentiques [30]. Les inconvénients incluent des incohérences lors de la tentative de reproduction de la complexité trouvée dans l'ensemble de données d'origine et l'incapacité de remplacer les données authentiques, car des données authentiques précises sont toujours nécessaires pour produire des exemples synthétiques utiles de l'information [31].

Les services financiers et la santé sont deux secteurs qui bénéficient des techniques de données synthétiques. Les techniques peuvent être utilisées pour fabriquer des données avec des attributs similaires à des données sensibles ou réglementées. Cela permet aux professionnels des données d'utiliser et de partager les données plus librement. Par exemple, les données synthétiques permettent aux professionnels des données de santé d'autoriser l'utilisation publique des données au niveau des enregistrements tout en préservant la confidentialité des patients.

Parmi les méthodes de synthétisation de données biomédicales, on peut notamment citer les avatars [50]. L'avatarisation est une nouvelle méthode pour générer les données synthétiques de granularité individuelle tout en respectant la vie privé des individus. Cette méthode utilise un modèle local pour générer de nouvelles données synthétiques aléatoires, appelées « avatar ». Cette méthode permet de garder la même structure (30 variables par exemple), la même granularité (30 individus par exemple), les mêmes agrégats (moyenne, somme ou médiane des variables par exemple) et les mêmes liens (corrélation) par rapport aux données d'origine. L'une des particularités de la méthode est qu'elle ne dispose que d'un lien temporaire entre chaque enregistrement original et son avatar pour éviter la réidentification des données. Comme les avatars conservent la même structure et la même granularité que les données d'origine, ils permettent d'obtenir des résultats similaires en utilisant les mêmes analyses statistiques et algorithmes de machine learning sur les avatars que sur les données d'origine.

2.4 Conclusion

Les progrès de la médecine de précision et des systèmes d'information ont entraîné une augmentation massive du volume des données et de la complexité des calculs, posant de problèmes aux modèles traditionnels pour la recherche médicale. Cependant, d'autres technologies présentes permettent de nouvelles méthodes de partage et d'analyse des données en toute sécurité.

Dans ce chapitre, nous avons présenté quelques travaux avec les technologies présentées comme le Cloud computing, la blockchain, les technologies Big Data et Federated Learning. Nous avons détaillé les caractéristiques de chaque modèle et leurs capacités pour contribuer à la privatisation des analyses biomédicales distribuées. Les travaux présentés dans ce chapitre valident la capacité de chaque technologie pour évoluer aux travaux autour de la distribution des analyses biomédicales, mais cela ne répondent pas exactement ou répondent beaucoup plus aux exigences de notre problématique.

Beaucoup de travaux affirment que ces technologies peuvent être utilisées à l'aide des mécanismes de confidentialités pour des analyses de données médicales distribuées de manière sécurisée et vérifiable. Pour cela, nous avons aussi présenté quelques techniques de sécurité et de confidentialité utilisées dans le cadre des calculs distribués.

CHAPITRE 3

QUELQUES OUTILS ET MÉTHODES D'ANALYSES DANS LE DOMAINE BIOMÉDICAL

Contents

Ce chapitre présente quelques outils et méthodes biomédicaux importants pour le développement de nos différentes contributions.

3.1 L'application de suivi des transplantés rénaux KITAPP

L'insuffisance rénale chronique touche environ 10% de la population mondiale et peut évoluer progressivement vers une insuffisance rénale terminale nécessitant un traitement substitutif (dialyse ou transplantation). La transplantation rénale est le meilleur traitement pour l'insuffisance rénale terminale [54]. Le taux de survie à un an de la greffe de rein est maintenant de 90% et la demi-vie du greffon est d'environ 10 ans. Le suivi des patients est essentiel pour surveiller l'état du greffon et la santé du patient. Les cliniciens disposent de plusieurs moyens pour surveiller la fonction rénale, en particulière des valeurs biologiques comme la créatinine. Le taux

de créatinine est mesuré pour évaluer le taux de filtration rénale.

Des outils modernes, tels que KITAPP, donnent de nouvelles informations aux cliniciens dans le but d'effectuer les meilleurs soins possibles.

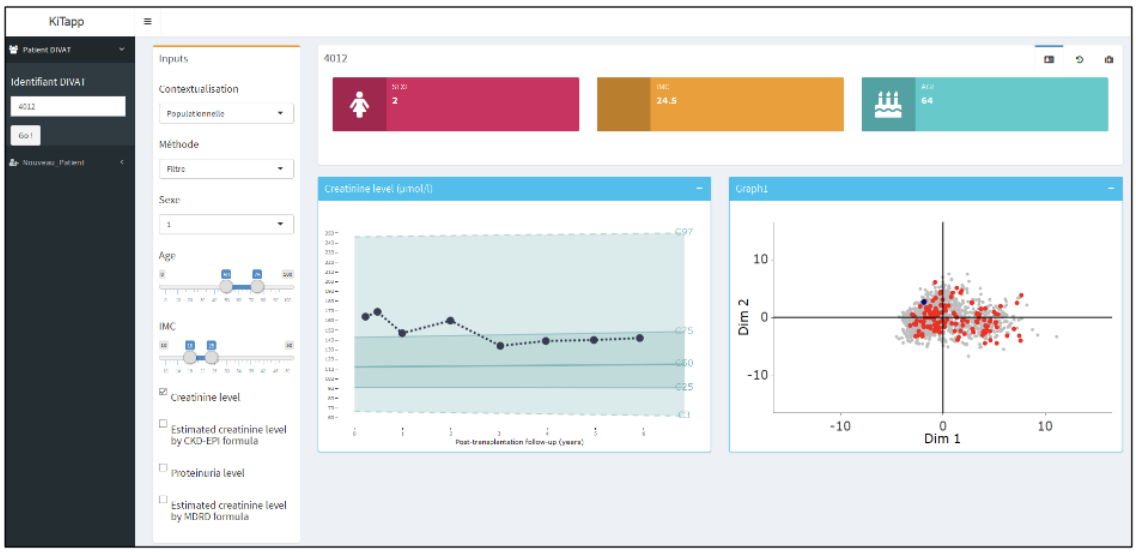


FIGURE 3.1 – L'interface d'un prototype de l'application KITAPP.

Le prototype de l'application KITAPP a été développé par Geffard et al. [43] en utilisant le language R. Les données utilisées dans cette application sont des données d'environ 1500 transplantations rénales, comprenant des éléments cliniques et immunologiques, qui ont été collectées depuis 2008 dans le cadre de la cohorte DIVAT [25]. KITAPP permet d'exploiter un algorithme de contextualisation personnalisé pour comparer les trajectoires de données d'un patient donné (Patient Of Interest POI) à une sous-population aux caractéristiques similaires (Patient Of Reference POR) sélectionnée par des filtres ou des mesures de distance (voir figure 3.1). Les informations relatives à une greffe sont sélectionnées parmi des cas similaires au moment de la greffe.

Avec l'aide des cliniciens et la connaissance du corpus de recherche existant, un ensemble de variables a été bien défini pour sélectionner la sous-population de référence. Celle-si a été sélectionnée soit en filtrant par rapport à une ou plusieurs de ces variables, soit en appliquant une stratégie statistique. Une stratégie statistique effectue une sélection sur la base d'une technique de réduction de dimension statistique qui réduit les données d'un grand espace dimensionnel à des données dans un espace dimensionnel plus petit. Dans notre cas, une population de référence a été déterminée sur la base de différentes méthodes statistiques. Parmi ces méthodes, l'approche du plus proche voisin qui permet de sélectionner les individus les plus similaires à

un POI et l'approche de clustering pour sélectionner les individus dans le même cluster qu'un POI.

La visualisation des informations contextualisées se fait en comparant une donnée biologique POI (le créatinine par exemple) et à son évolution dans le temps selon des visites cliniques à une POR qui est représentée par leurs valeurs médianes et centiles.

3.2 Le système des antigènes des leucocytes humains (*HLA*)

Le système *HLA* (Human Leukocyte Antigen en anglais), le complexe majeur d'histocompatibilité (CMH) chez l'humain, joue un rôle primordial dans la réponse immunitaire. L'histocompatibilité représente le taux de compatibilité entre deux tissus ou organes, qui permet à une greffe de ne pas être rejetée. La description détaillée du système *HLA* a engendré des découvertes majeures en recherche fondamentale et en clinique. Le CMH est un système de présentation d'antigènes (toute molécule étrangère ou non à un organisme capable de déclencher une réponse immunitaire et engendrer la formation d'anticorps). Il permet au système immunitaire de distinguer le soi (ensemble des antigènes de l'organisme ne déclenchant pas de réponse immune en condition normale) du non-soi (ensemble des antigènes étrangers ou pathogènes pouvant déclencher une réponse immune) [44]. D'un point de vue clinique, la conséquence première de la découverte du système *HLA* est son application pour la prise en charge des patients ayant besoin d'une transplantation d'organe ou d'une greffe de moelle osseuse (MO).

3.2.1 Allèle *HLA*

Un allèle est une version alternative d'un même gène (voir la figure 3.2) qui est distinguée par des variations de sa séquence nucléotidique [90]. Les allèles *HLA* sont définis par séquençage de l'ADN et associés à un nom pour les identifier. Pour chaque gène *HLA*, chaque individu possède au maximum 2 allèles par locus, un de son père et un de sa mère. Ces allèles sont exprimés de façon co-dominante, c'est à dire qu'ils sont co-exprimés en même temps dans les mêmes cellules. Si le même allèle a été transmis par les 2 parents, l'individu est dit homozygote ; s'il a reçu 2 allèles différents, il est dit hétérozygote. Le système *HLA* comprend 28320 allèles décrits en octobre 2020 avec 6291 allèles $HLA-A$, 7562 $HLA-B$, 6223 $HLA-C$, 2838 $HLA-DRB1$, 1930 $HLA-DQB1$ [32]. Ces nombres montrent bien le polymorphisme important du *HLA* [100]. Les allèles *HLA* se caractérisent par un spectre de fréquence large, de $0,276$ pour l'allèle le plus fréquent du $HLA-A$ ($HLA-A*02:01$) à $0,00001$ pour les

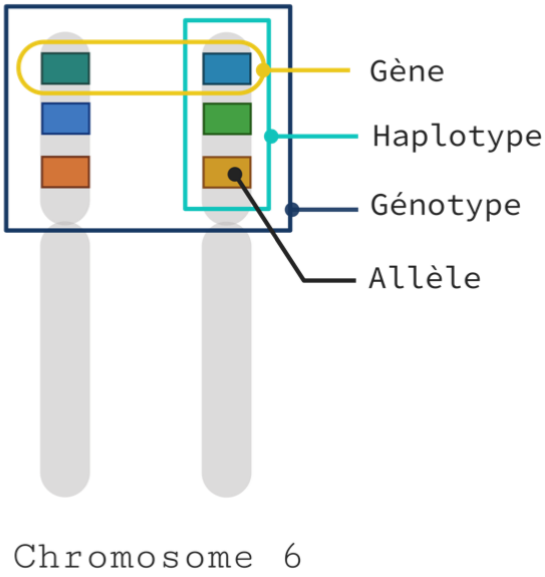


FIGURE 3.2 – Représentation schématique de quelques notions de génétique [43].

allèles les plus rares dans les données d'individus d'ancestralité européenne ($N = 1242890$) [47]. Pour le gène $HLA-A$, $0,16\%$ des allèles (10 allèles) les plus fréquents représentent 90% de la fréquence cumulée de tous les allèles $HLA-A$ des européens.

3.2.2 Génotype *HLA*

La transmission des gènes *HLA* suit les lois de la génétique mendélienne, basée sur la théorie de l'hérédité biologique. Les différents gènes *HLA* sont co-dominants. Le polymorphisme *HLA* à l'échelle individuelle est caractérisé par la présence de 2 allèles pour chacun des locus qui forment un génotype *HLA* (expression à la surface cellulaire des 2 molécules *HLA* pour chaque locus de l'allèle paternel et de l'allèle maternel). Un génotype correspond à l'ensemble des allèles d'un locus (ou gène) d'un individu. Il existe des millions de combinaisons possibles des allèles pour chacun des gènes *HLA*. En théorie, on pourrait avoir plus de 82 millions de milliards de milliards de milliards de génotypes. Mais en réalité, certains génotypes sont beaucoup plus rares que d'autres : cela est dû au déséquilibre de liaisons entre certains allèles et à la pression de sélection face à certains allèles ou haplotypes.

Néanmoins, le *HLA* reste extrêmement divers, ce qui explique la grande diversité des génotypes *HLA* et le défi pour la compatibilité lors d'une transplantation d'organe ou une greffe.

3.2.3 Haplotype *HLA*

Un haplotype est un ensemble d'allèles d'un même chromosome. Les gènes, étant très proches les uns des autres sur le chromosome, peu d'événements de recombinaison ont donc lieu et ces gènes sont transmis en «bloc» à la descendance, que l'on appelle haplotype. Ceci conduit à la notion de déséquilibre de liaison entre 2 allèles proches sur le même chromosome. Il s'agit d'une association préférentielle entre deux ou plusieurs allèles, c'est à dire une association rencontrée avec une fréquence plus élevée que la probabilité d'association si ces deux allèles étaient complètement indépendants l'un de l'autre [69]. La pression de sélection entraîne donc une combinaison préférentielle d'allèles *HLA*, qui ont permis une survie face à certains pathogènes, et donc des haplotypes plus fréquents que d'autres. Connaître la répartition des allèles en haplotypes *HLA* est important dans divers cas comme les études d'association de la maladie, la génétique des populations, ou la greffe.

Pour reconstruire les haplotypes, il existe des méthodes algorithmiques telles que l'estimation-maximisation EM et des méthodes combinatoires. Les méthodes combinatoires reposent sur le principe de tester toutes les combinaisons d'allèles possibles et de les discriminer via un critère de phylogénie. Un individu sera génotypé ainsi que ses deux parents et on va ensuite regarder de quel parent est hérité chacun des allèles de chaque gène. Les allèles étant transmis en haplotype, on va ainsi déduire les 2 haplotypes.

3.3 La base de données AFND

Parmi les bases de données les plus importantes en bio-informatique, la base de données Allele Frequency Net (AFND) [47] est une ressource disponible gratuitement pour le stockage des données de fréquence sur les polymorphismes de plusieurs gènes liés à l'immunité, y compris le *HLA*, les récepteurs tueurs de type immunoglobuline (KIR), la séquence liée aux polypeptides (MIC) du CMH de classe I et plusieurs polymorphismes de gènes de cytokines (ex., IL4, TGFB1, TNF).

L'AFND a été créée en 2003 avec quelques sections et fréquences d'allèles *HLA*. Elle s'est enrichie, au fil des années, de nouveaux outils intégrés dans une nouvelle version majeure en 2015 [98, 46]. En 2020, une nouvelle mise à jour a été effectuée sur les jeux de données disponibles permettant la soumission et le partage des données en utilisant un critère de classification de la qualité des données (GSB : or, argent, bronze) [47].

La ressource AFND a été consultée par plus de 100 000 utilisateurs différents de 186 pays

au cours des dernières années. Cette référence contient des informations provenant de plus de 10 millions d'individus sains et de plus de 1600 populations, qui permettent l'analyse des régions les plus polymorphes du génome humain. Ces données démographiques proviennent de 141 pays du monde entier avec une couverture de population variée. Avec ces données, les utilisateurs peuvent effectuer des analyses sur les fréquences alléliques, les gènes, les génotypes ou les haplotypes pour *HLA*, KIR, MIC et les cytokines.

Les données du site de l'AFND sont issues (1) de publications à comité de lecture, (2) de données démographiques issues d'ateliers internationaux *HLA* et immunogénétique (IHWS), (3) de soumissions de laboratoires du monde entier et (4) de rapports de publication (SPR) en collaboration avec la revue «Human Immunology».

3.4 L'inférence des haplotypes à l'aide de la méthode statistique EM pour le calcul de vraisemblance

Le développement de nouvelles techniques pour connaître la fréquence des haplotypes et expliquer les génotypes sans avoir accès aux génotypes des parents est très utile pour les cliniciens. Les méthodes d'inférence d'haplotypes, basées sur des calculs statistiques, ont montré leurs efficacités. L'une des méthodes se base sur un calcul de vraisemblance grâce à l'algorithme EM. Cela revient à exprimer la vraisemblance des génotypes (i.e. probabilité d'avoir un génotype) en fonction des fréquences de tous les haplotypes qui peuvent expliquer les génotypes observés.

L'algorithme d'estimation-maximisation (EM) est une méthode d'estimation paramétrique s'inscrivant dans le cadre général du maximum de vraisemblance. Cet algorithme itératif a été mis au point par Dempster et al. et permet d'obtenir les estimations du maximum de vraisemblance dans un modèle comprenant des données manquantes [22]. En considérant un échantillon $G = (g_1, ..., g_n)$ suivant une loi $f(g_i, \theta)$ de paramètre θ, on cherche à déterminer le paramètre θ maximisant la vraisemblance (probabilité des paramètres suivant l'échantillon).

La formule de la vraisemblance :

$$L(G;\theta) = \sum_{i=1}^{n} \log f(g_i, \theta) \tag{3.1}$$

L'algorithme EM est composé de 2 étapes :

— L'étape E, va estimer l'espérance, conditionnellement aux données observées, de la log-

vraisemblance sachant les paramètres déterminés à l'itération précédente.

Étape E : évaluation de l'espérance

$$Q\left(\theta \mid \theta_{old}\right) = E_{old}[\log L(X, Z \mid \theta) \mid X = x, Z = z] \tag{3.2}$$

— L'étape M va maximiser le calcul de l'étape précédente en mettant à jour les paramètres pour avoir une nouvelle estimation de celle-ci. Après, une étude de convergence des valeurs estimées et de stabilité des paramètres entre l'itération k et k-1 et si ça nécessite une nouvelle itération.

Étape M : Maximisation

$$\theta_{new} = \operatorname{argmax}_{\theta} Q\left(\theta \mid \theta_{old}\right) \tag{3.3}$$

Dans le cadre d'une application de l'algorithme EM dans le *HLA*, EM est utilisé pour le calcul des fréquences des haplotypes pour les données *HLA*. Excoffier et al. [36] ont utilisé en entrée de l'algorithme EM une liste de génotypes *HLA* ($G = (g_1, ..., g_n)$) et un échantillon de génotypes par rapport à l'algorithme présenté ci-dessus). Cet algorithme permet d'estimer les fréquences des haplotypes d'une population (le paramètre θ dans la formule présentée ci-dessus).

3.5 Conclusion

Dans ce chapitre, nous avons présenté quelques concepts et techniques biomédicaux qui sont utiles pour la compréhension et la réalisation de nouvelles versions des analyses distribuées dans les contributions qui seront développées dans les chapitres suivants.

CHAPITRE 4

CONTEXTUALISATION DISTRIBUÉE DES DONNÉES BIOMÉDICALES

Contents

Les travaux présentés dans ce chapitre ont donné lieu à une publication dans la *17th International Conference on Computer Systems and Applications (AICCSA)* [106].

4.1 Introduction

La médecine de précision repose sur de nouvelles méthodes et outils biomédicaux adaptés au traitement spécifique des patients. La contextualisation est alors utile pour évaluer les données médicales de patients individuels par rapport à des groupes plus larges de patients présentant des caractéristiques similaires. En particulier, elle peut servir à comparer des marqueurs de patients à des échelles représentant différents niveaux de gravité. Les cliniciens peuvent ainsi évaluer l'état d'une patiente et apporter une perspective sur son état passé, présent et futur afin d'aider à décider d'un suivi et d'un traitement[49].

À titre d'exemple récent, nous considérons la pandémie de COVID-19 qui a entraîné la perte de centaines de milliers de vies et d'immenses conséquences sociales et économiques. La dy-

namique de propagation des maladies infectieuses est souvent due aux comportements sociaux et exerce une pression énorme sur les organisations sociales et les infrastructures sanitaires existantes. La contextualisation, la modélisation et la compréhension des données statistiques et des comportements sociaux (panique, distanciation sociale...) aident à la prise de décision d'un meilleur traitement afin d'avoir une meilleure réponse à l'épidémie au niveau international [64].

Nous avons envisagé la contextualisation dans le projet de recherche KITAPP [53] présenté dans la section 3.1. Ce nouvel outil vise à relever les défis de la gestion dynamique d'une maladie chronique complexe, notamment l'interaction cliniciens-patients afin d'étendre la médecine numérique et d'aider à la décision thérapeutique personnalisée. Ce projet se concentre sur des entités étiologiquement hétérogènes qui peuvent conduire à des symptômes et des traitements différents. Il est donc actuellement compliqué, même pour un clinicien expérimenté, de prédire l'évolution clinique d'un patient et d'anticiper précisément la sécurité et l'efficacité des traitements. KITAPP s'appuie sur la contextualisation des résultats de la fonction rénale (créatinémie, protéinurie...) d'un patient d'intérêt (POI) par rapport à une population de patients de référence (POR). Les POR sont basés sur l'identification de sous-groupes de patients présentant des caractéristiques similaires, telles que l'âge, le sexe et l'indice de masse corporelle (IMC).

L'algorithme de contextualisation utilisé dans KITAPP se compose de deux étapes. Lors de la première étape, nous sélectionnons une sous-population POR. Lors de la seconde étape, nous appliquons notre algorithme de modélisation et visualisons la contextualisation du POI par rapport au POR. Cet aspect de la contextualisation est réalisé en utilisant les centiles [59] comme outil statistique majeur permettant de positionner un patient par rapport aux autres patients. Un système de visualisation de la courbe de croissance est proposé en modélisant la trajectoire individuelle des données biologiques (paramètres rénaux évalués lors de routine en clinique) en fonction du suivi post-greffe, par rapport à la sous-population POR. De même, Corson et al. [20] fournissent un exemple de pointe qui utilise des centiles pour construire des courbes de croissance de porcs à utiliser dans des études biomédicales.

Le projet KITAPP est un prototype relayant les bases de différentes collaborations de recherche clinique recrutant actuellement des patients, notamment dans deux projets internationaux KTD-Innov [74] et EU-TRAIN [35]. Jusqu'à présent, les centiles sont calculés de manière centralisée, un seul site rassemblant toutes les données.

Au-delà des problèmes techniques, les architectures distribuées permettent des stratégies et des processus de gouvernance des données plus flexibles en les libérant des contraintes de centralisation [16]. La gestion des données des entités de santé s'annonce ainsi facilitée.

Afin de déployer ce type de services médicaux dans un contexte plus large, comme des collaborations nationales ou internationales, des systèmes distribués et des algorithmes pour la médecine de précision doivent être fournis. La contextualisation doit alors être effectuée par rapport à des bases de données médicales distribuées à grande échelle qui sont maintenues sur différents sites. Les schémas de centralisation simples ne sont pas utiles dans ce contexte car les données nécessaires peuvent ne pas être partagées pour des raisons juridiques, des problèmes de sécurité, de confidentialité et de ceux de performances. De plus, les calculs distribués approximatifs ne correspondent pas non plus aux exigences de la médecine de précision.

Afin de contextualiser l'état d'un patient par rapport à une base de données distribuée, le développement des analyses entièrement distribuées est donc très intéressant, si il répond aux exigences d'évolutivité, de sécurité et de confidentialité [13]. Scheel et al. [109] discutent de l'importance de la disponibilité, du droit à la vie privée et de l'accessibilité des données dans la recherche biomédicale distribuée. Ces critères sont difficiles à satisfaire car la signification statistique et la précision des analyses dépendent souvent directement du nombre de cas ou d'individus inclus dans la base de données.

Dans le contexte des systèmes et des algorithmes distribués, le partage et l'analyse des données sont généralement difficiles pour des raisons scientifiques, techniques, réglementaires et de gouvernance. Les politiques de protection des données, telles que le Règlement général sur la protection des données (RGPD) de l'UE, imposent des restrictions bien fondées mais fortes sur le partage des données [42]. Ainsi, les analyses ne sont souvent possibles que « sur place ». De plus, les chercheurs et les institutions peuvent être réticents à perdre le contrôle à la fois des données et des utilisations de celles-ci. Par ailleurs, d'énormes volumes de données sont intrinsèquement difficiles à partager ou à transférer. Par exemple, en raison des coûts associés à l'utilisation des ressources, telles que les ressources de calcul, de stockage et de réseau.

Une solution à ces problèmes peut être trouvée si une analyse distribuée est effectuée séparément dans les locaux des partenaires en ce qui concerne les données sensibles et si le résultat global de l'analyse peut ensuite être calculé à partir de données agrégées ou anonymisées. Ce principe est similaire aux algorithmes traditionnels. Mais il peut nécessiter des algorithmes plus complexes pour qu'il puisse être distribué en termes de systèmes Master-Worker.

Dans ce chapitre :

— Nous motivons et définissons les exigences pour les algorithmes distribués pour la contextualisation dans le cadre du projet KITAPP dans la section 4.2.

— Nous présentons un nouvel algorithme de centile distribué pour la contextualisation des données sensibles en médecine de précision dans la section 4.3.

4.2 Motivation des analyses distribuées pour l'application KITAPP

Les exigences d'échange de données pour les collaborations médicales multicentres/multi-partenaires ont conduit à un changement de paradigme dans le système de partage de données médicales. Traditionnellement, des infrastructures centralisées ont été utilisées pour stocker, traiter ou archiver des informations. Depuis la promulgation des réglementations qui limitent le partage des données de santé y compris la loi du RGPD, ces structures ne sont souvent plus adaptées aux projets collaboratifs de santé en raison des restrictions d'accès aux données sensibles. Pour relever le défi consistant à exploiter les données médicales, tout en conservant les données sensibles sur site, où à assurer une protection renforcée de données en cas de leurs déplacement, les calculs sont souvent effectués aujourd'hui sur des bases de données distribuées. Celles-ci sont liées à un intégrateur de calcul, qui permet à un centre d'interagir avec certaines données, et d'y accéder depuis des sites distants. Chaque centre clinique collecte, stocke et contrôle les données de ses propres patients. Le principe fondateur est qu'aucune donnée des individus ne circule en dehors des centres. Cependant, ce paradigme de partage est très restrictif et inhibe une large gamme d'analyses potentielles, soit parce que les données sensibles ne peuvent pas être protégées de manière appropriée, soit parce que l'analyse ne peut pas être effectuée de manière suffisamment efficace.

Pour ces raisons, l'équipe biomédicale a proposé une extension d'une version distribuée des analyses de KITAPP qui sont détaillé dans la section 3.1. Nous travaillons sur des architectures et des implémentations d'analyses distribuées plus générales qui facilitent la collaboration dans le cadre de projets de recherche multicentriques, où chaque centre peut contrôler et rendre compte de l'utilisation des données de ses propres patients, même s'il est situé à distance.

4.3 Contextualisation distribuée des données biomédicales

Dans le cadre du projet KITAPP, nous avons travaillé sur la contextualisation distribuée des paramètres de l'activité rénale. Dans ce qui suit, nous présentons un algorithme distribué pour calculer les informations de centile sur des données réparties entre différents sites qui ne partagent que de petites quantités de données et aucune donnée sensible entre ces sites.

4.3.1 Définition de l'algorithme

Le rang d'un centile est défini comme suit :

$$k = P/100 * N \tag{4.1}$$

où N est le nombre de valeurs dans l'ensemble de données, P le centile et k le rang d'une valeur donnée. Le calcul de la valeur définissant un centile nécessite donc de calculer d'abord k puis d'identifier la valeur correspondante. L'identification de la valeur k^{th} est difficile et coûteuse dans un tableau non trié mais simple et efficace dans un tableau ordonné. En outre, le fait de rassembler tous les éléments pertinents nécessaires pour déterminer un centile donné ou de classer d'abord les valeurs disponibles sur tous les sites peut entraîner le déplacement d'une grande partie de toutes les valeurs entre les sites. Les deux types d'approches aussi simples peuvent donc souffrir de sérieux problèmes du point de vue de l'efficacité et de la protection des données.

Nous avons donc étudié des algorithmes nécessitant peu d'échanges et de partage de données. Nous nous appuyons sur l'algorithme QuickSelect bien connu (non distribué) (développé par Hoare en 1961 [56]) qui permet de sélectionner le k^{th} élément le plus grand dans un élément non trié tableau en ordonnant partiellement ce tableau. Semblable à QuickSort, QuickSelect partitionne un tableau en fonction d'un élément pivot qui peut être choisi arbitrairement. Contrairement à l'algorithme de tri, QuickSelect ordonne et recherche récursivement mais partiellement une seule des partitions : la partition qui contient la valeur cible. QuickSelect atteint une complexité moyenne de $O(n)$ et $O(n^2)$ dans le pire des cas d'un choix malchanceux de l'élément pivot.

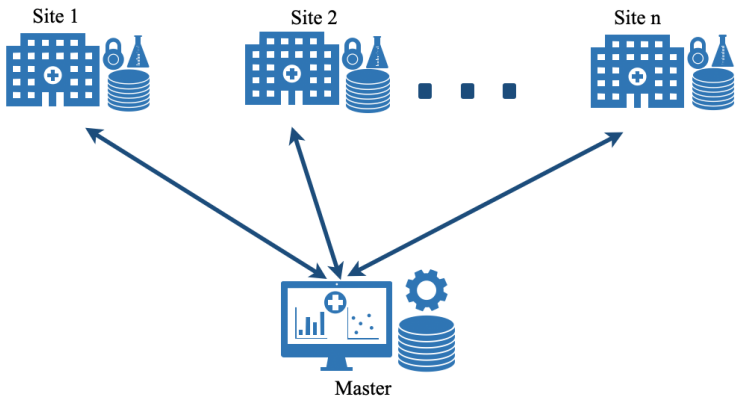


FIGURE 4.1 – Exigences de la collaboration.

Dans le cadre de notre coopération médicale, nous sommes intéressés à une structure de collaboration comme illustrée dans la Fig. 4.1. L'analyse globale est effectuée sur plusieurs sites. Chacun fonctionnant sous la coordination d'un site Master. Cependant, contrairement à de nombreux schémas Master-Worker entièrement parallèles aux données, notre analyse doit observer les dépendances entre les nœuds de travail. Saukas et Song [104] ont proposé un algorithme parallèle pour la sélection du plus petit élément dans un ensemble d'éléments. Cet algorithme nécessite peu de communication et partage également peu de données. Nous avons repris le même principe algorithmique et l'avons adapté aux environnements distribués à grande échelle en mettant, notamment, en place des échanges et partages de données.

4.3.2 Implémentation distribuée

Le pseudocode 1 et l'organigramme montré dans la figure 4.2 présentent notre nouvel algorithme distribué. Le Master calcule d'abord le rang k du centile par rapport à N, la taille totale des données sur tous les hôpitaux. Chaque Worker commence par envoyer la taille de ses données N_i ainsi que la médiane locale m_i. Ensuite, le maître calcule la médiane pondérée M et diffuse M à tous les Worker. Chaque Worker calcule alors l_i, e_i, g_i qui correspondent respec-

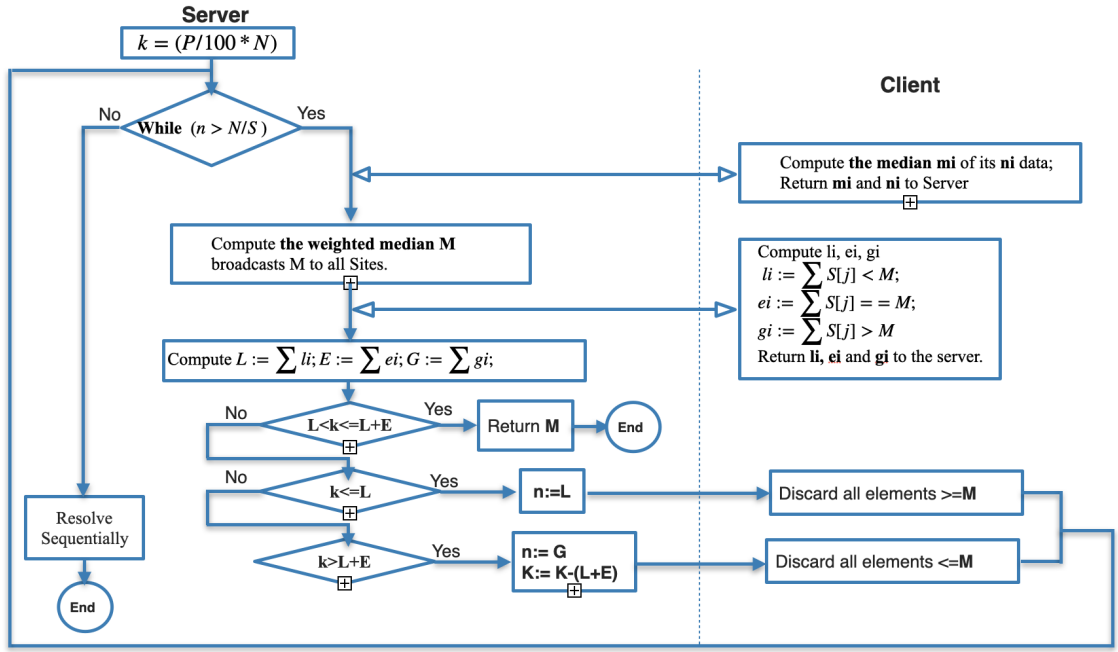


FIGURE 4.2 – Organigramme de l'algorithme des centiles distribué.

tivement au nombre de valeurs inférieures, égales et supérieures à M et les renvoie au serveur.

Après quoi, le serveur calcule les sommes respectives L, E et G. Si $(L < k <= L + E)$. Le résultat correspond, alors, à E. Si $(k <= L)$ alors le maître demande à chaque Worker de rejeter toutes les valeurs inférieures à E, n est fixé à L. Si $(k > L + E)$, alors, le Master demande à chaque Worker d'écarter toutes les valeurs supérieures à M et n est mis à G, k à $k - (L + E)$. Ce processus est répété jusqu'à $(n <= N/S)$. Enfin, si nous n'arrivons pas à la solution après cette analyse parallèle, le serveur commence la même analyse séquentiellement sur le reste des données n.

4.4 Expérimentations

Nous avons implémenté notre algorithme et un programme pilote pour le déployer et l'exécuter dans un environnement basé sur une grille dont les nœuds peuvent être constitués de machines individuelles mais aussi de clusters à part entière. A noter que cet environnement reproduit fidèlement l'architecture de véritables coopérations médicales dont les différents sites partenaires forment une grille et les partenaires individuels peuvent disposer eux-mêmes de clusters informatiques. L'algorithme distribué a été implémenté à l'aide de Python en 840 lignes de code. Son déploiement et son exécution à l'aide d'un nombre, en principe arbitraire, de nœuds de travail ne nécessitent que huit commandes sur l'infrastructure Grid'5000 (voir la section 4.4.1). Nous avons testé notre algorithme de manière approfondie et avec succès en comparant son résultat à sa contrepartie séquentielle.

Nous avons appliqué l'analyse de contextualisation des centiles à des données réelles de transplantation disponibles dans la base de données française DIVAT [25]. Afin de soutenir le projet KITAPP, nous avons exploré la contextualisation (distribuée) des niveaux de créatinine chez les patients. Nous avons donc pris `creatD` comme variable de suivi pour la contextualisation du taux de créatinine chez les patients. Après extraction des données pertinentes pour cette analyse, nous avons obtenu un fichier de 11 028 valeurs. Nous avons divisé ce fichier en plusieurs fichiers de tailles différentes pour une distribution sur différents nombres de sites de travail.

4.4.1 Déploiement sur une infrastructure géo-distribuée

Pour implémenter notre expérimentation, nous avons utilisé la plateforme Grid'5000 [7] en tant qu'environnement d'expérimentation réel basé sur une grille qui contient plusieurs clusters placés dans des dizaines de sites. Grid'5000 est une infrastructure dédiée à la recherche

Algorithm 1: Algorithme des centiles distribués

Input : Set Data of N elements distributed among the S worker sites, each site i with N_i elements;
k : the rank of a specific percentile P;
n : number of remaining data on all sites;
Output: the k^{th} value of N distributed data, of rank k and of P percentile : ($k^{th} = N, P, k$)

$k := P/100 * N$;
while $n > N/S$ **do**
  foreach *site* $i \in S$ **do**
    compute the median m_i of its n_i data;
    return m_i and N_i to Master;
  end
  Master computes the weighted median M;
  Master broadcasts M to all other sites;
  foreach *site* $i \in S$ **do**
    $l_i := \sum_{j=1}^{n} S_i[j] < M$;
    $e_i := \sum_{j=1}^{n} S_i[j] == M$;
    $g_i := \sum_{j=1}^{n} S_i[j] > M$;
    return l_i, e_i, g_i to Master ;
  end
  Master computes $L = \sum_{i=1}^{S} l_i, E = \sum_{i=1}^{S} e_i, G = \sum_{i=1}^{S} g_i$;
  if $L < k \leq L + E$ **then**
    return M and stop;
  else
    if $k \leq L$ **then**
      Master requests each site S to discard all elements $>= M$;
      $n := L$;
    else
      if $k > L + E$ **then**
        Master requests each site S to discard all elements $<= M$;
        $n := G$;
        $k := k - (L + E)$;
      end
    end
  end
end
All the remaining n data are sent to the Master;
Master solves the remaining problem sequentially;
return ;

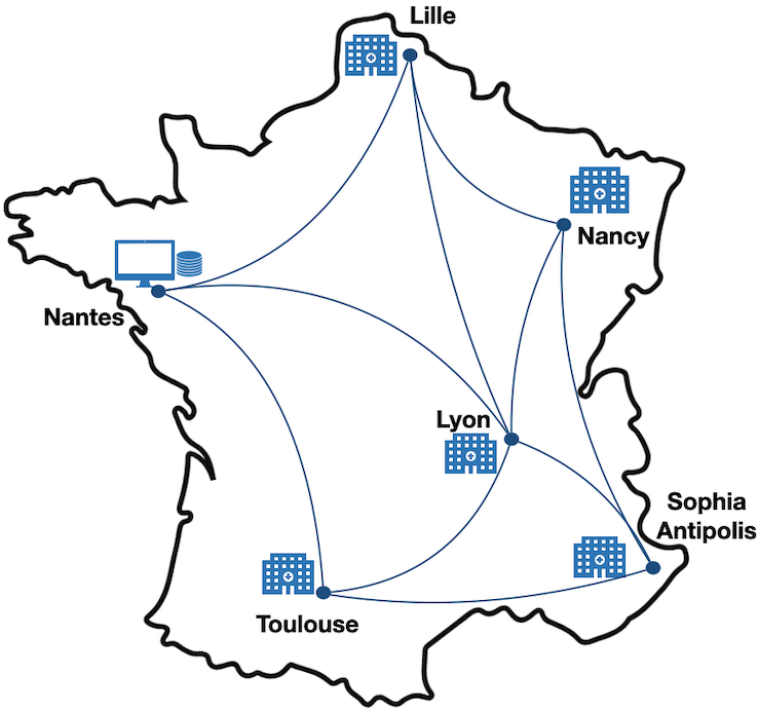


FIGURE 4.3 – Expérimentation avec la plateforme Grid'5000.

informatique dans les domaines des systèmes distribués à grande échelle, de calcul haute performance et des réseaux. Pour chaque expérimentation de notre algorithme, nous avons réservé des machines qui sont utilisées comme Worker comme le montre la figure. 4.3, dans cinq villes différentes. Ces machines contiennent les données stockées sous forme de fichiers CSV ainsi que le code du Worker. Nous avons également réservé une machine en tant que Master (serveur en termes de calcul) qui effectue l'analyse en communiquant avec les Worker à l'aide de la communication par socket sans avoir accès aux données des fichiers CSV et en générant le résultat en centile à la fin de l'analyse, permettant ainsi la visualisation des taux de créatinine d'un patient donné par rapport aux autres.

La figure 4.4 montre nos résultats pour la contextualisation du niveau de créatinine des patients calculés sur cinq sites de travail. Cette figure montre l'état du niveau de créatinine chez les patients par rapport aux autres sur la base des résultats à cinq centiles (3^{rd}, $10^{ème}$, $25^{ème}$, $50^{ème} et 97^{ème}$ centile). Par exemple, les centiles $25^{ème}$ et $50^{ème}$ correspondent à des personnes qui ont un taux de créatinine, respectivement de $61(mol/l)$ et $77(mol/l)$ par rapport à la population répartie sur les cinq sites.

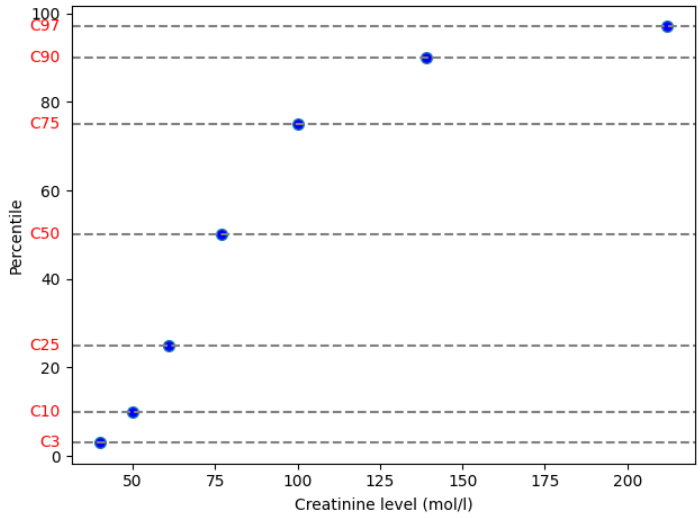


FIGURE 4.4 – Contextualisation du niveau de la créatinine (mol/l).

4.4.2 Évaluation des performances et de la confidentialité

Nous avons évalué notre système par rapport à deux types de paramètres : les paramètres liés aux performances (différents nombres de sites, différentes tailles de données, temps d'exécution) et les paramètres liés au partage des données (proportion de données partagées par centile, taille des données et nombre de sites).

Temps d'exécution par centile.

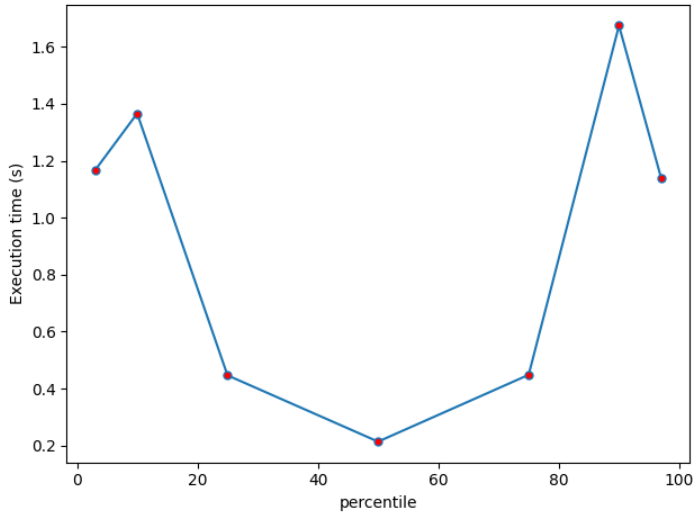


FIGURE 4.5 – Temps d'exécution par 5 sites.

Nous avons mesuré le temps d'exécution pour chaque centile effectué. La figure 4.5 présente le temps d'exécution de chaque centile pour le cas de 5 sites. les abscisses montrent les centiles calculés et les ordonnées montrent le temps nécessaire pour calculer chaque centile. La durée d'exécution du $50^{ème}$ centile est plus court par rapport autres centiles (moins de $0,2s$) car on peut parfois, notamment dans ce cas, ne pas passer à la phase séquentielle de l'algorithme. Le comportement de la courbe s'applique également à d'autres expérimentations avec un nombre de sites plus élevé. Le cas du temps d'exécution pour le calcul des centiles sans passer par la phase séquentielle est toujours plus court. Cela montre l'intérêt des analyses distribuées non seulement pour des raisons de confidentialité et de respect de la vie privée mais aussi pour des raisons de performance. Grâce à notre méthode, le temps d'exécution, même dans les cas où l'on doit passer par la phase séquentielle reste dans les normes (par exemple dans notre cas le plus long est $1,7s$ pour le calcul du $75^{ème}$ qui est court).

Temps d'exécution par nombre de sites.

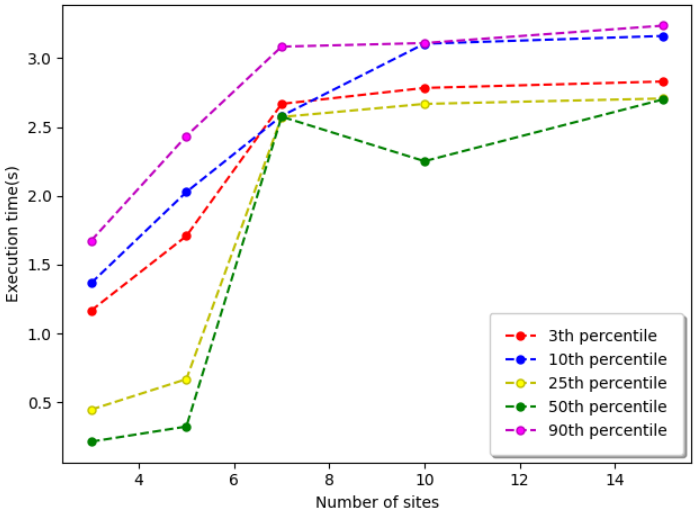


FIGURE 4.6 – Temps d'exécution par nombre de clients.

Nous avons également évalué le temps d'exécution de notre approche avec un nombre de Workers différents, en partitionnant un ensemble de données fixe sur les sites $3, 5, 7, 10$ et 15. La figure 4.6 présente le temps d'exécution de 3^{rd}, $10^{ème}$, $25^{ème}$, $50^{ème}$, $75^{ème}$ et $90^{ème}$ percentiles pour les données partitionnées sur des sites $3, 5, 7, 10$ et 15.

La figure montre que le temps de calcul du percentile augmente avec l'augmentation du nombre de sites. Mais il reste toujours très faible et ne dépasse pas $3.2s$ pour le calcul de $90^{ème}$ centile sur 15 sites pour un jeu de données de taille 11028.

Si on augmente le nombre de sites, le comportement de la courbe reste toujours stable. En effet, même si le temps de communication entre les sites augmente avec l'augmentation du nombre des sites, le temps de calcul par site diminue tout en gardant la même taille des données pour toutes les analyses.

Évolutivité par taille de données.

Nous avons également évalué l'efficacité de notre algorithme proposé sur un grand nombre de données et une grande taille de données par rapport à l'ensemble mentionné précédemment. Le tableau 4.1 montre le nombre de données et les tailles de données utilisées dans ces expériences.

TABLE 4.1 – Le nombre des échantillons pour chaque taille de données

Taille global des données (MB)	**Nombre globale des échantillons**
0.131	11028
0.363	110280
3.6	1102800
36.3	11028000
363	110280000

La figure 4.7 montre le temps d'exécution de notre approche pour calculer le $50^{ème}$ centile pour le cas de 3 sites avec différentes tailles de données globales. Nous remarquons que le temps d'exécution de notre algorithme augmente linéairement relativement à la taille des données tout en donnant des résultats aussi exacts qu'un calcul centralisé.

Confidentialité par pourcentage de données partagées.

Un autre paramètre important dans le contexte médical est la quantité de données partagées, notamment dans le cadre de la phase séquentielle de notre algorithme. Nous avons mené deux expérimentations afin d'évaluer le pourcentage de données partagées en augmentant la taille globale des données ou le nombre de sites participant à l'analyse.

Les résultats sont présentés dans les figures 4.8 et 4.9 sous forme de diagrammes à boîte et à moustaches qui fournissent une vision détaillée de la distribution des données partagées, rendant explicites les médianes, les quartiles inférieures et supérieurs (sous forme de boîtes) et les valeurs aberrantes (sous forme d'étoiles vertes). Les chiffres montrent que le pourcentage de données partagées ne dépasse jamais 0.89% par rapport à l'ensemble des données de tous les sites, ce qui est négligeable d'autant plus que les données partagées ne sont constituées que de valeurs agrégées (médianes) qui ne correspondent à aucune donnée sensible permettant

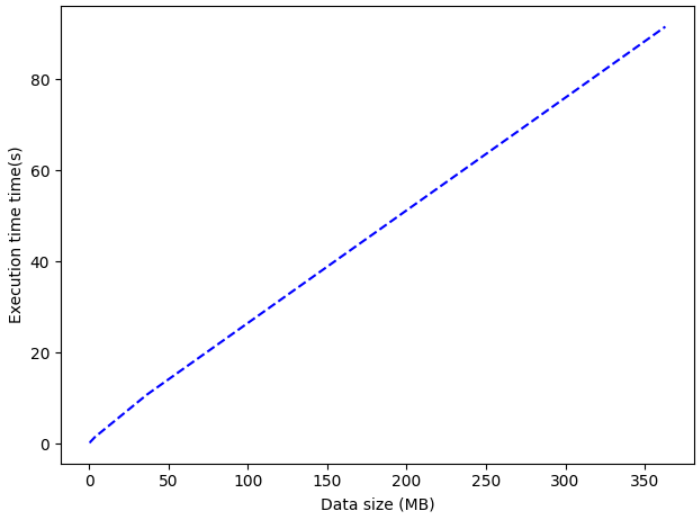


FIGURE 4.7 – Temps d'exécution du $50^{ème}$ centile en fonction de la taille des données.

d'identifier un patient. Ceci sous condition qu'on ne peut pas avoir d'informations personnelles par le taux des créatines des échantillons, ce qui rend négligeable le risque de partage des données sensibles. La médiane de la proportion de données partagées présente dans les deux figures est proche de 0, ce qui explique que, dans la plupart des cas, l'algorithme n'effectue que des calculs parallèles sans nécessiter le recours à des étapes de résolution séquentielle et donc sans partage des données.

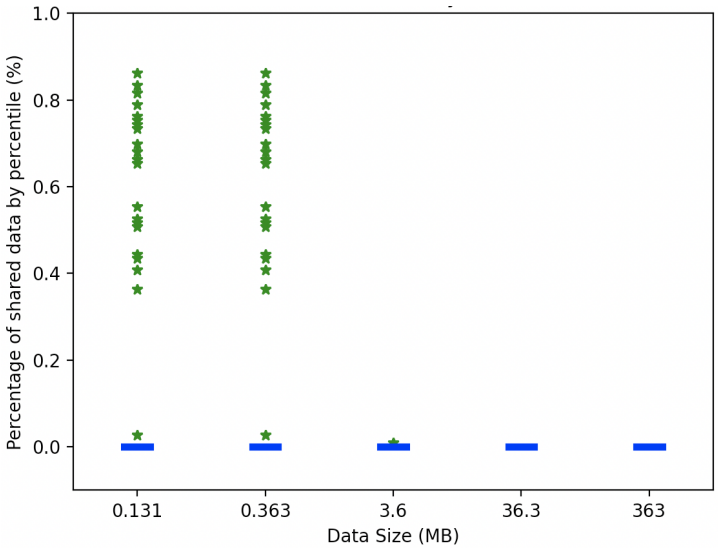


FIGURE 4.8 – Pourcentage des données partagées en fonction de la taille globale des données.

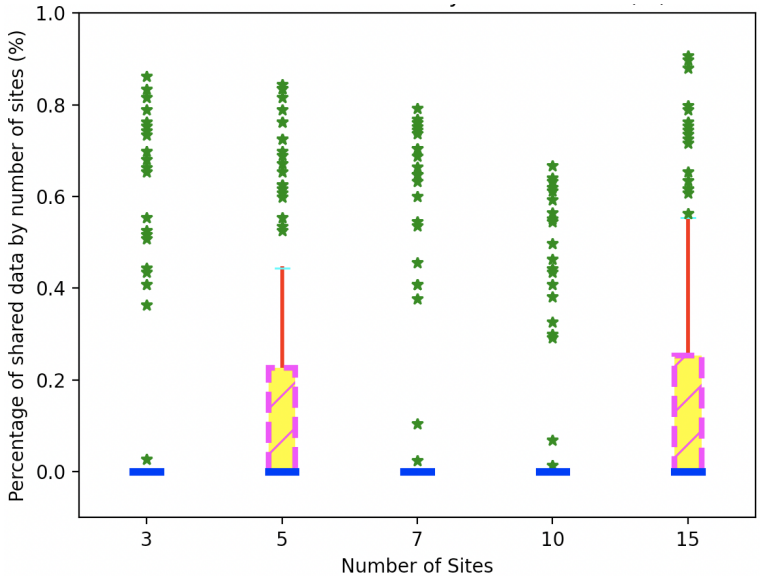


FIGURE 4.9 – Pourcentage des données partagées en fonction du nombre de sites.

Les deux chiffres montrent des valeurs aberrantes comprises entre 0.3% et 0.89%, en particulier lorsque le nombre de sites participant à l'analyse est augmenté. De plus, la figure 4.8 montre que, si nous augmentons la taille des données, notre algorithme sera encore plus efficace en terme de pourcentage de données partagées qui diminue jusqu'à ce qu'il soit trés proche 0. La précision de notre algorithme est préservée même si nous augmentons le nombre de sites et la taille de données ; tout en gardant la confidentialité des données avec le modèle distribué en gardant un pourcentage des données partagées qui ne dépasse pas $0,89\%$ par rapport à données globales.

4.5 Conclusion

Dans ce chapitre, nous avons considéré le problème important de la distribution de la contextualisation. Nous avons motivé notre approche dans le contexte de la contextualisation du monde réel dans le domaine des transplantations médicales. Afin de respecter la confidentialité des données et réaliser ce type d'analyses à grande échelle, nous avons proposé dans ce chapitre un nouvel algorithme distribué de contextualisation des centiles. Cet algorithme ne nécessite que peu de partage de données entre sites. Nous avons réalisé des expériences réalistes sur des données réelles dans un environnement distribué basé sur un cluster de grille réaliste. Nos expérimentations ont montré des résultats précis par rapport à une analyse de contextualisation centralisée

avec de très bons résultats de performances et d'évolutivité en termes de taille du système et de temps d'exécution. Enfin, nous avons montré que notre algorithme ne nécessite que très peu de partage de données qui n'implique pas le partage de données sensibles individuelles.

CHAPITRE 5

ANALYSE FACTORIELLE DE DONNÉES MIXTES DISTRIBUÉE

Contents

Les travaux présentés dans ce chapitre ont donné lieu à une publication dans la *35-th International Conference on Advanced Information Networking and Applications (AINA-2021)* [108].

5.1 Introduction

Les techniques d'analyse de données volumineuses sont de plus en plus populaires pour extraire de nouvelles informations à partir de quantités massives de données. Ces informations servent à améliorer la prise de décision, notamment dans le secteur médical. Un défi majeur pour les cliniciens consiste à prendre en toute sécurité des décisions de traitement correctes basées sur des quantités toujours croissantes de données sur les patients. Ce problème appelle

de nouvelles techniques d'analyse et de nouveaux algorithmes, en particulier pour la médecine de précision.

Comme présenté lors de la motivation de notre thèse dans la section 1.1, la médecine de précision représente une profonde révolution dans les soins de santé. Pour que cette (r)évolution devienne une réalité, il faut apporter la preuve de l'efficacité et de la rentabilité de la médecine de précision. Ceci, à son tour, nécessite que les décisions concernant les patients soient justifiés par une analyse de données de référence à grande échelle, pertinentes pour l'évaluation de leur situation personnelle par rapport à celle des autres [49].

La réduction de dimension est une technique majeure pour transformer de grands espaces de données multidimensionnels en sous-espaces de dimension inférieure. Ceci tout en préservant certaines caractéristiques significatives des données d'origine. Parmi les méthodes de réduction de dimension, celle, la plus courante, consistant en l'analyse en composantes principales (ACP) [67], qui permet la réduction de dimension pour les variables de données quantitatives. D'autre méthodes consistent en l'analyse factorielle de données mixtes (FAMD) [96]. Celle-ci effectue une réduction de dimension pour les variables de données mixtes (quantitatives et qualitatives), et l'apprentissage par dictionnaire (DL) [111], l'une des méthodes les plus puissantes d'extraire des caractéristiques à partir de données.

L'analyse FAMD fournit des représentations simplifiées d'espaces de données multidimensionnels sous la forme d'un nuage de points dans un sous-espace vectoriel de composants principaux. Si deux points sont proches l'un de l'autre dans ce nuage, une forte similarité globale existe entre eux par rapport aux composantes principales sélectionnées. Dans le domaine biomédical, ce type d'analyse est fréquemment utilisé pour présenter des groupes de patients de manière simplifiée et visuelle pour une large gamme de données cliniques complexes ; englobant des données quantitatives (par exemple obtenues à partir d'examens biologiques) et des données qualitatives (par exemple des informations sur le sexe) . Il en résulte des représentations exploitables des caractéristiques individuelles de chaque patient par rapport à celle des autres.

Dans le cadre du partenariat public-privé français KTD-innov et du projet européen H2020 EU-train, FAMD a été utilisée pour la réduction de dimension dans le cadre du système d'aide à la décision clinique KITAPP (l'application de greffe de rein) [53]. Cette application web de médecine de précision calcule des scores prédictifs et représente les distributions des variables des patients dans un sous-groupe de patients de référence après une transplantation rénale.

L'application est conçue pour relayer l'intuition et l'expérience des cliniciens au moyen de calculs à la demande et de représentations graphiques.

Une des fonctionnalités clés de KITAPP consiste dans la "contextualisation" des patients par rapport à une population de référence (POR). À cette fin, il utilise d'abord FAMD pour la réduction de dimension, puis applique un algorithme de modélisation statistique de centile [106] et, enfin, visualise les relations des patients au POR.

Comme mentionné dans les chapitres précédents, les études médicales impliquent souvent de grandes collaborations nationales ou internationales (comme nos projets KTD-innov et EU-Train). Les schémas de centralisation simples pour le placement des données et la réalisation des calculs ne sont souvent pas applicables dans ce contexte, car les données et les calculs peuvent ne pas être partagés pour des raisons juridiques, ou pour des problèmes de sécurité/de confidentialité ou pour ceux liés à la performances. Pour déployer ces type de services médicaux dans des contextes plus larges, des systèmes et algorithmes distribués pour la médecine de précision doivent être fournis. L'un des principaux axes de recherche autour des projets KTD-innov et EU-TRAIN est la mise en place d'une base de données de référence intégrée dans une infrastructure informatique distribuée ; permettant un accès sécurisé aux données tout en respectant le règlement européen RGPD sur la protection des données [42] .

Le partage des données et le placement des analyses sont généralement difficiles pour des raisons de gouvernance, de réglementation, ainsi que pour des raisons scientifiques et techniques (coût, stockage...) cités dans le chapitre précédent de la section 4.1. De plus, les chercheurs et les institutions peuvent être réticents, car ils craignent la perte de contrôle de l'utilisation des données.

Comme les architectures distribuées permettent la mise en place de stratégies de gouvernance des données et de processus d'analyse plus flexibles en les libérant des contraintes de centralisation [16] et en effectuant des calculs locaux sur les données des patients, sans qu'aucune donnée individuelle ne circule en dehors des centres cliniques qui génèrent les données, on peut y recourir pour réaliser les calculs statistiques distribués.

Des analyses entièrement distribuées ont été proposées dans la section 5.2, pour contextualiser l'état d'un patient par rapport aux données POR stockées dans une base de données distribuée. Ces algorithmes doivent répondre à des exigences d'évolutivité, de sécurité et de confidentialité [13], ainsi qu'à des propriétés de disponibilité et au droit à la confidentialité [109]. Ces critères sont cependant difficiles à satisfaire, car la signification statistique et la précision des analyses dépendent souvent directement du nombre de cas ou d'individus inclus dans la base de données.

Une solution à ces problèmes peut être de se baser sur l'exploitation distribuée des analyses qui manipulent les données sensibles directement dans les locaux de leurs propriétaires respec-

tifs et d'exploiter les calculs distribués s'il s'agit de données sensibles, agrégées, résumées ou anonymisées.

Dans ce chapitre :

— Nous présentons une étude bibliographique sur les calculs distribués des algorithmes de réduction de dimension qui sont basés sur l'analyse en composantes principales (ACP) et qui constitue une bonne partie de l'analyse factorielle des données mixtes (FAMD).

— Nous motivons et définissons les exigences des algorithmes distribués pour la réduction de dimension dans le cadre du projet KITAPP.

— Nous présentons un nouvel algorithme FAMD distribué pour la réduction de dimension en présence de données sensibles en médecine de précision et l'appliquons à notre contexte.

5.2 État de l'art

D'autres versions d'algorithmes de réduction de dimension ACP ont déjà été proposées. Liang *et al.* [78] proposent un système client-serveur et envoient des vecteurs singuliers et des valeurs singulières $U\Sigma V$ (les matrices qui constituent le calcul du ACP) du client vers le serveur. Feldman *et al.* [39] ont montré comment calculer les ACP en envoyant des matrices plus petites $U\Sigma$ au lieu d'envoyer toutes les matrices d'une décomposition en valeurs singulières, améliorant ainsi le coût de communication. Wu *et al.* [124] ont introduit un algorithme qui améliore les exigences de stockage et de traitement des données et exploite le cloud computing pour la réduction de la dimension ACP. Ces propositions envoient des matrices de données contenant des données synthétisées à partir des données originales et non des données réelles. Ceci est très intéressant pour les analyses biomédicales afin de garantir la confidentialité des données des patients. Imtiaz *et al.* [60] ont amélioré la proposition de Feldman *et al.s'* en ajoutant des garanties de confidentialité à l'aide de la confidentialité différentielle.

A notre connaissance, aucun algorithme FAMD distribué n'a encore été proposé. Dans ce chapitre, nous proposons un FAMD distribué sur la base d'un algorithme ACP distribué. Notre algorithme est structuré en deux parties (similaire à Pagès [96]) :

1. Transformer les données qualitatives en données quantitatives à l'aide de tableaux disjonctifs complets [8], transformant ainsi le problème original de réduction de dimension FAMD en un problème de ACP.
2. Effectuer la réduction de dimension de l'ACP distribuée sur la base de la proposition de Feldman *et al.* [39].

5.3 Motivation d'une analyse factorielle de données mixtes distribuée pour KITAPP

KITAPP permet d'exploiter un algorithme de contextualisation personnalisé pour comparer les trajectoires de données d'un patient donné (POI) à une sous-population aux caractéristiques similaires (POR) sélectionnée par des filtres ou des mesures de distance. Les informations relatives à une greffe sont sélectionnées parmi des cas similaires au moment de la greffe. Avec l'aide de cliniciens et de la connaissance du corpus de recherche existant, nous avons défini un ensemble de variables pour sélectionner la sous-population de référence.

Nous présentons trois techniques d'analyses sur lesquelles l'algorithme de contextualisation de population est basé. Ces techniques consistent à comparer les données d'un patient donné à des POR avec

1. Des caractéristiques similaires sélectionnées par des filtres ou des approches basées sur une analyse statistique,
2. La méthode du plus proche voisin,
3. La méthode du cluster.

Avec notre approche de filtre (1), le POR est défini en fonction de filtres sélectionnés mis à la disposition du clinicien, tels que l'âge, le sexe et l'indice de masse corporelle (IMC).

Les méthodes (2) et (3) sont basées sur les résultats d'un FAMD.

Suite à cette analyse, on peut alors sélectionner un POR par méthode du voisin proche (2) : en sélectionnant les N individus les plus similaires à un POI. Il est possible de sélectionner un POR par clustering (3) : en sélectionnant les individus du même cluster que notre POI.

Nous avons l'intention d'exploiter l'application KITAPP dans le cadre de coopérations à grande échelle avec de nombreux partenaires (nationaux et internationaux). Pour relever le défi consistant à exploiter les données médicales tout en conservant les données sensibles sur site ou à assurer une protection renforcée des données en cas de déplacement de celles-ci. Aujourd'hui, les calculs sont souvent effectués sur des bases de données distribuées liées à un intégrateur de calcul permettant à un centre d'interagir avec certaines données et d'y accéder depuis des sites distants. Chaque centre clinique collecte, stocke et contrôle les données de ses propres patients. Le principe fondateur de l'architecture est qu'aucune donnée des individus ne circule en dehors des centres.

Cependant, ce paradigme de partage est très restrictif et inhibe une large gamme d'analyses potentielles : soit parce que les données sensibles ne peuvent pas être protégées de manière

appropriée, soit parce que l'analyse ne peut pas être effectuée de manière suffisamment efficace.

Le besoin de stockage local et de distribution des données de référence est motivé par la valeur exploitable. Il offre la possibilité de contrôler localement qui a accédé aux données, quelles sont les utilisations des données locales et comment les limiter en cas de besoin. L'utilisation d'une infrastructure distribuée est un élément central de la gouvernance des données multi-acteurs.

Nous travaillons donc sur des architectures et des implémentations d'analyses distribuées plus générales qui facilitent la collaboration dans le cadre de projets de recherche multicentriques, où chaque centre peut contrôler et rendre compte de l'utilisation des données de ses propres patients, même s'il est situé à distance. La contextualisation doit alors être effectuée par rapport à des bases de données médicales distribuées à grande échelle qui sont maintenues sur différents sites.

5.4 FAMD Distribuée

Dans ce qui suit, nous donnons d'abord un aperçu de l'architecture et des propriétés de notre algorithme avant de le définir en détail.

5.4.1 Description de l'algorithme

L'analyse factorielle de données mixtes (FAMD) [96] est une méthode de réduction dimensionnelle de variables comprenant des données quantitatives et qualitatives mixtes en moins de composants pour des raisons de synthèse d'informations. L'utilité de cette analyse est de permettre la préservation de certaines propriétés de l'espace de données original. Cette analyse peut être définie, par exemple, à l'aide d'opérations matricielles, comme suit :

$$FAMD = ACP + ACM \tag{5.1}$$

où ACP est une réduction de dimension d'analyse en composantes principales pour les variables quantitatives et ACM est une réduction de dimension d'analyse des correspondances multiples pour les variables qualitatives. L'objectif de l'ACP est de minimiser la redondance et de maximiser la variance pour mieux exprimer les données. Pour ce faire, il trouve les vecteurs propres associés à la matrice de covariance des points de données. Les données sont ensuite projetées sur le nouveau système de coordonnées couvert par ces vecteurs propres.

Globalement, notre algorithme fonctionne comme suit. Dans un premier temps, nous transformons les variables qualitatives en variables quantitatives à l'aide d'un codage disjonctif complet (CDC) [88] effectué localement sur chaque site.

Ensuite, nous effectuons une réduction de dimension au moyen d'une ACP distribuée afin d'obtenir un algorithme FAMD sécurisé et distribué.

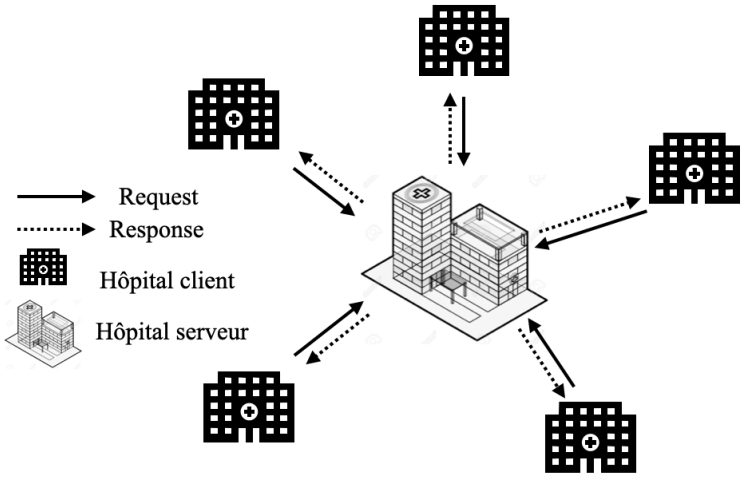


FIGURE 5.1 – Exigences de la collaboration

Nous exploitons l'architecture de coopération distribuée illustrée dans la figure 5.1. La transformation des données et l'analyse de la réduction des dimensions sont effectuées localement en parallèle sur plusieurs sites. La coordination entre les sites est assurée par un site agrégateur, qui reçoit les données synthétiques et effectue la réduction de dimension globale.

Imtiaz *et al.* [60] ont proposé un algorithme sécurisé et distribué pour la réduction de dimension ACP. Cet algorithme utilise la confidentialité différentielle comme technique de sécurité et des données synthétiques pour la communication entre les nœuds.

L'algorithme y résultant a deux propriétés importantes :

— *Faible coût de communication :* Le coût de la communication des algorithmes FAMD parallèles et distribués dépend essentiellement de la taille des matrices transférées entre sites. De nombreux algorithmes de réduction de dimension nécessitent l'envoi de matrices de taille $D \times D$, où D est le nombre d'éléments de données à analyser (c'est-à-dire les données de transplantation dans notre cas). En revanche, l'algorithme d'Imtiaz *et al.* [60] nécessite l'envoi de matrices de type $D \times R$ où R est le nombre de variables avec (typiquement) $R \ll D$.

— *Sensibilisation à la sécurité/à la confidentialité :* Notre algorithme satisfait aux deux

caractéristiques intéressantes : (1) la confidentialité différentielle est utilisée pour la protection des données et (2) la communication entre les sites et l'agrégateur qui implique uniquement des données synthétisées P_S et non les données d'origine, ce qui minimise les possibilités de vol de données et assure la protection des données.

Nous exploitons le même principe et les mêmes propriétés tout en fournissant deux nouvelles contributions : (1) une transformation de variables qualitatives en variables quantitatives afin d'obtenir un algorithme sécurisé et distribué pour la réduction de dimension FAMD et (2) une implémentation distribuée évolutive en utilisant un environnement réellement géo-distribué et des données biomédicales réelles.

5.4.2 Implémentation distribuée

L'algorithme 2 présente notre algorithme FAMD sécurisé et distribué. La première étape consiste à transformer le problème de FAMD complet en un problème ACP (qualitatif) (voir lignes 1–16). Pour chaque variable qualitative, un Codage Disjonctif Complet utilisant le package ade4 du langage R est appliqué afin de transformer les variables qualitatives en variables quantitatives, suivi du calcul des modalités N_k et des proportions $p_k = N_k/N$ pour calculer la fonction de pondération de l'indicateur $X_{i,k}$. Ensuite, chaque site commence par calculer, pour chaque variable quantitative, la moyenne correspondante μ_k, l'écart type σ_k et la fonction de centrage et de réduction $X_{i,k} = \frac{1}{\sigma_k}(x_{i,k} - \mu_k)$.

La deuxième étape est la réduction de dimension, (lignes 17–23). Chaque site calcule la matrice (du second moment) $A_s = \frac{1}{NS} X_s X_s^T$. L'application du schéma de confidentialité différentielle (suivant la proposition de Dwork *et al.* [28]) est effectuée en générant la matrice de bruit E de taille $D \times D$ et la matrice de confidentialité différentielle estimée $A_s = A_s + E$ en ligne 19. Chaque site effectue ensuite la décomposition en valeurs singulières ($SVD(A_s)$) de la matrice A_s pour calculer la matrice ($P_s = U\Sigma^{1/2}$) et la diffuser à l'agrégateur.

Au niveau du site agrégateur, le serveur calcule, voir les lignes 24–26, la matrice $A = \frac{1}{s}\sum_{s=1}^{s} P_s P_s^T$ de tous les sites. Il effectue ensuite la décomposition globale en valeurs singulières $SVD(A) = V\Lambda V^T$.

5.5 Expérimentations

Dans cette section, nous présentons les expérimentations impliquant des analyses sur des données médicales réelles que nous avons menées sur une véritable infrastructure de grille hé-

Algorithm 2: FAMD Distribué

Input : Data matrix $X_s \in \mathbb{R}^{D \times N_s}$ *for* $s \in [S]$ of N elements and P variables, with C quantitative variables and M qualitative variables ; ϵ, δ : privacy parameters ; j : reduced dimension;

Output: V_j : Matrix of eigenvectors on top j

foreach *site* $s \in S$ **do**
  foreach *element* $i \in N$ **do**
    foreach *element* $k \in C$ **do**
      Compute the mean of the variable μ_k;
      Compute the standard deviation of the variable σ_k;
      Compute the Centering and Reduction Function $X_{i,k} = \frac{1}{\sigma_k}(x_{i,k} - \mu_k)$;
      return $X_{i,k}$;
    end
    foreach *element* $k \in M$ **do**
      Apply the Complete Disjunctive Coding using (ade4 package on R);
      Compute the effective of the modality N_k;
      Compute the proportion $p_k = N_k/N$;
      Compute the Indicator Weighting Function $X_{i,k} = \frac{x_{i,k}}{\sqrt{p}_{,k}}$;
      return $X_{i,k}$;
    end
  end
  Compute $A_s = \frac{1}{N_s} X_s X_s^T$;
  Generate $D \times D$ symmetric Matrix E where $E_{i,j} : i \in [D], j \leq i$ drawn i.i.d. from $N(0, \Delta^2_{\epsilon,\delta})$ where $\Delta_{\epsilon,\delta} = \frac{1}{N s_\epsilon}\sqrt{2log(\frac{1.25}{\delta})}$, $E_{i,j} = E_{j,i}$;
  Compute $A_s = A_s + E$;
  Perform $SVD(A_s) = U\Sigma U^T$;
  Compute $P_s = U\Sigma^{1/2}$;
  Send P_s to the aggregator;
end
Compute $A = \frac{1}{s}\sum_{s=1}^{s} P_s P_s^T$;
Perform $SVD(A) = V\Lambda V^T$;
Send V_j to all sites ;
return V_j;

térogène à grande échelle. Nous rapportons notre configuration et évaluons notre mise en œuvre selon trois critères.

5.5.1 Déploiement sur une infrastructure géo-distribuée

Nos expériences ont été menées sur les même données de transplantation rénale utilisées dans le chapitre précèdent qui sont disponibles dans la base de données européenne Divat [25]. Afin de comparer avec les résultats du projet KITAPP, nous avons appliqué notre algorithme distribué à ses analyses sur 11 163 données de transplantation. Nous avons commencé par diviser le fichier de données avant transfert et analyse sur les différents sites.

Nous avons implémenté notre algorithme distribué et l'avons exécuté dans un environnement basé sur une grille avec différentes architectures distribuées, allant du placement de tous les clients sur différentes machines (géo-distribuées) à leur placement en tant que cluster sur une seule machine. Cet environnement distribué constitue une architecture réaliste d'une collaboration médicale impliquant les centres de recherche et cliniques, partenaires du projet KITAPP. Nous avons implémenté notre algorithme distribué en utilisant le langage de programmation Python et R en utilisant 860 lignes de code. L'ensemble du système distribué peut être déployé et exécuté sur un nombre arbitraire de sites sur l'infrastructure Grid'5000 présentée précédemment dans la section 4.4.1 à l'aide d'un petit script de seulement huit commandes.

Pour notre expérience, nous avons réservé une machine en tant que serveur (agrégateur) exécutant un programme Python pour gérer l'analyse, les interactions client et la génération du résultat final. Pour créer plusieurs sites clients nous avons réservé des machines réparties sur cinq sites différents en France.

5.5.2 Résultats et évaluation des performances

L'analyse de réduction de la dimension FAMD est motivée par KITAPP que nous avons utilisée comme une application pour notre contribution et qui a été exécutée sur la base de 11 163 échantillons de transplantation caractérisées à l'aide de 27 variables qualitatives et quantitatives réparties sur cinq sites. Nous avons défini le paramètre de réduction de dimension $j = 2$ sur le serveur pour avoir une figure à 2 dimensions. La figure 5.2 montre le sous-espace bidimensionnel résultant après application de notre analyse FAMD distribuée.

Afin de répartir la sélection POR pour la contextualisation des POI entre les sites, nous avons appliqué la technique de regroupement non supervisée K-means au résultat de l'analyse de réduction de dimension FAMD distribué. La FAMD et le regroupement permettent de regrouper

les données des patients en fonction de leur similarité et de leur proximité par rapport aux composants principaux.

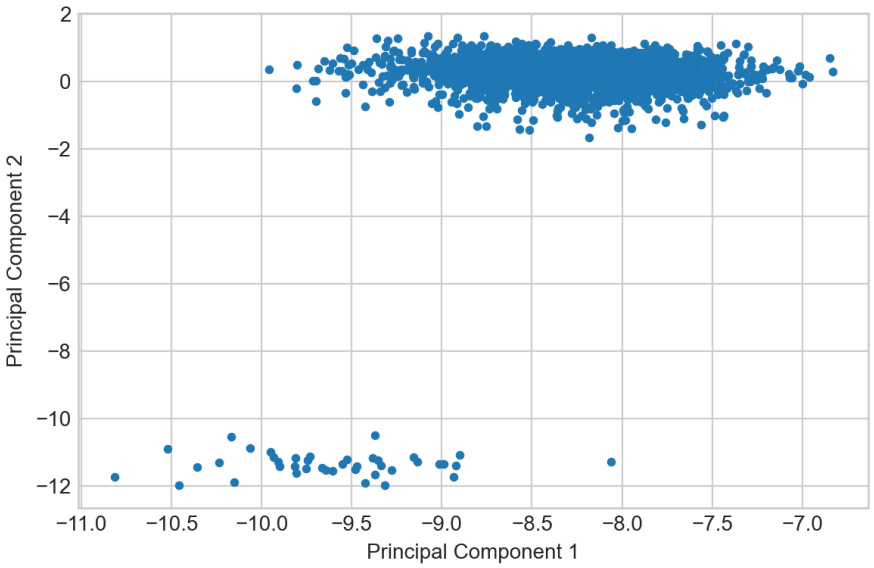


FIGURE 5.2 – FAMD distribuée sur 5 sites.

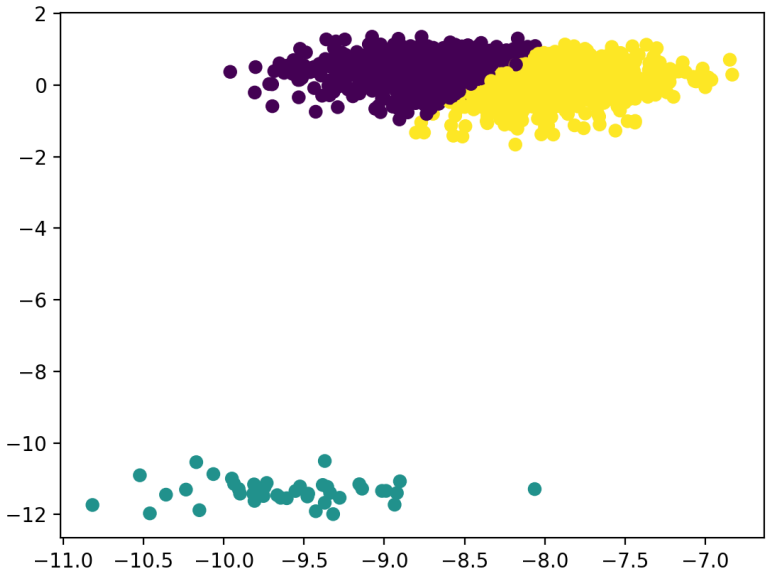


FIGURE 5.3 – Regroupement du résultat de l'algorithme FAMD distribué.

La figure 5.3 présente le résultat du clustering k-means, trois clusters de données indépendants qui correspondent exactement au résultat de l'algorithme séquentiel (centralisé) utilisé dans le cadre du projet KITAPP. Chaque cluster est caractérisé par des combinaisons de variables spécifiques. Le groupe vert correspond aux donneurs vivants. Le groupe jaune correspond aux donneurs décédés et le groupe violet aux donneurs décédés avec des critères étendus, tels que l'âge > 50 ans, le fait d'être sujets à l'hypertension ou à des taux de créatinine $\geq 133 \mu mol/L$.

Notez que nous obtenons toujours les mêmes clusters indépendamment du nombre de sites qui participent à l'analyse FAMD distribuée si nous l'utilisons avec les mêmes données, ce qui montre un fort potentiel de mise à l'échelle de l'algorithme que nous proposons.

Dans ce qui suit nous évaluons trois propriétés de notre implémentation :

— La protection des données sensibles par la technique de confidentialité différentielle en utilisant une notion d'énergie capturée,

— Le temps d'exécution,

— Le coût de la communication.

À des fins d'évaluation, nous considérons trois architectures : notre technique de réduction FAMD distribuée (notée "DPdis" ci-dessous), une version plus centralisée où toutes les matrices de second moment A_s de chaque client sont agrégées au niveau du serveur (notée "fulldis"), et une version FAMD entièrement centralisée (notée "pooled").

Énergie capturée/utilité : Suivant Imtiaz *et al.*, l'énergie capturée q est utilisée pour évaluer la qualité des directions principales Eigen vecteurs V_j en fonction de la différence d'utilité de l'information dans le cas où toutes les données sont centralisées q_{pooled}. Toutes les matrices secondes moment A_s de chaque site est distribuée $q_{fulldis}$ et l'algorithme FAMD distribué sécurisé proposé q_{DPdis} par taille de données et nombre de sites.

L'énergie capturée est définie comme la multiplication matricielle $q = tr(V_j(A)^T A V_j(A)$ mesurant la quantité de valeurs propres optimales capturées dans le sous-espace FAMD. Pour tout autre sous-espace sous-optimal, la valeur doit être inférieure à la valeur optimale.

— *Énergie capturée en fonction du nombre de sites :* Nous avons fait varier le nombre de sites S qui participent à cette analyse en gardant le nombre total d'échantillons $N = 11163$ (c'est-à-dire que nous avons diminué la taille N_s de chaque site). La figure 5.4 montre une détérioration des performances de $q_{fulldis}$ et q_{DPdis} pour un nombre croissant de sites. Cette baisse de performance s'explique par la diminution du nombre d'éléments par site. De plus, la présence d'un bruit de forte variance dégrade le nombre de directions propres plus fortes que le bruit qui est détecté par l'ACP au lieu de capturer toutes les

directions j qui présentent les données.

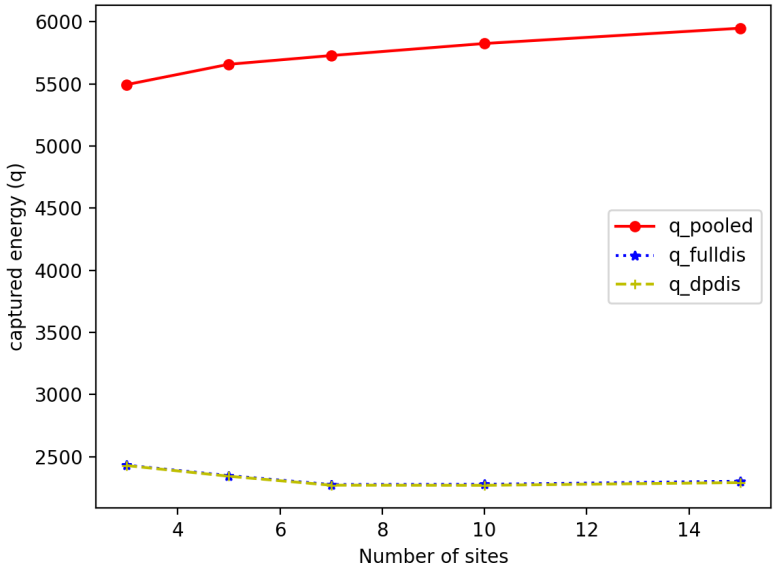


FIGURE 5.4 – L'énergie captée (q) en fonction du nombre de sites.

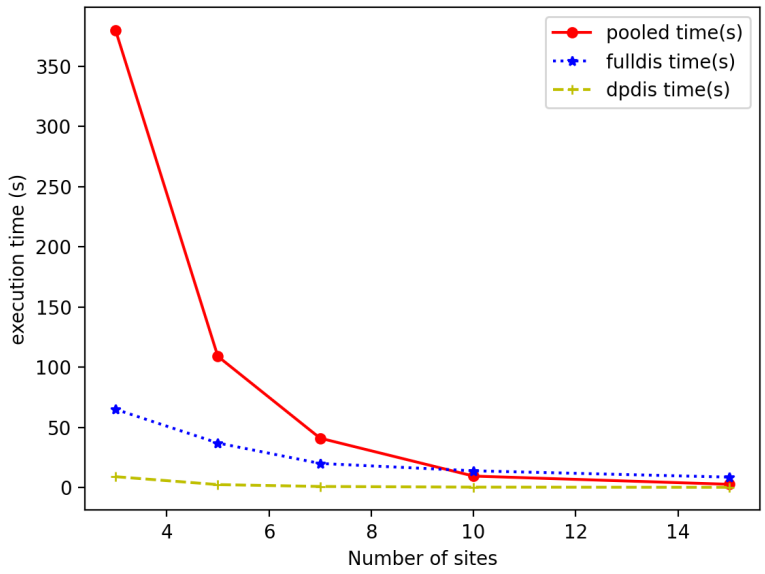


FIGURE 5.5 – Temps d'exécution en fonction du nombre de sites.

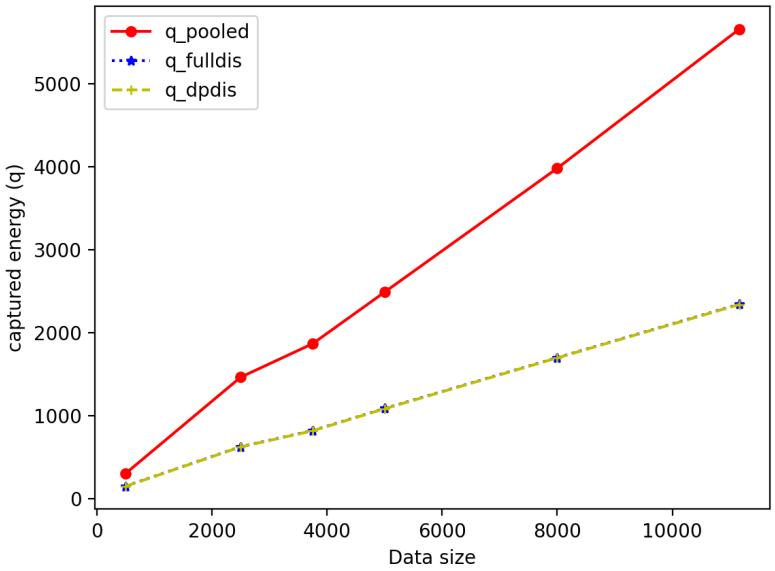


FIGURE 5.6 – L'énergie captée (q) en fonction de la taille des données.

— *Énergie capturée en fonction de la taille de données :* La figure 5.6 montre une augmentation des performances de l'énergie captée q en fonction du nombre d'éléments par site. $q_{fulldis}$ et q_{DPdis} ont presque les mêmes performances en énergie captée. Pour les deux variations de nombre de sites et de taille de données, q_{pooled} conserve toujours de meilleures performances.

Temps d'exécution : Nous avons fait varier le nombre de sites S en gardant le nombre global d'échantillons $N = 11163$ (c'est-à-dire que nous avons diminué la taille N_s de chaque site) et nous avons mesuré le temps d'exécution. La figure 5.5 montre que le temps d'exécution diminue lorsque le nombre de sites augmente. L'approche que nous proposons, $dpdis$ a le temps d'exécution le plus bas. Cela est dû à un faible coût de communication, comme expliqué dans la section 5.4.1.

Coût de la communication/ partage de données : Le moindre coût de communication de l'algorithme que nous proposons, et que nous avons introduit dans la section précédente permet de minimiser la quantité de données à partager (les matrices Ps sont partagées et non A_s). Dans le cas de nos expériences de distribution sur 5 sites avec des échantillons globaux égaux à $11,163$ et 27 variables, la quantité de données partagées par tous les clients est égale à $11,163 \times 27 = 301.401$ valeurs au lieu de $sqr(11,163) = 124.612.569$ valeurs.

5.6 Conclusion

La réduction de dimension FAMD est un outil important pour transformer des données complexes en sous-espaces de dimension inférieure tout en préservant les caractéristiques importantes des données d'origine. Cette technique est généralement utile pour réduire la complexité et soutenir la prise de décision. Dans ce chapitre, nous avons motivé l'utilisation de la réduction de dimension pour les collaborations biomédicales géo-distribuées qui nécessitent des modèles distribués et des implémentations d'algorithmes biomédicaux avec une architecture distribuée. L'évaluation de notre contribution sur une véritable infrastructure de grille géo-distribuée utilisant des données réelles permet de valider ses propriétés d'efficacité, de mise à l'échelle et de protection de la vie privée.

CHAPITRE 6

LA DISTRIBUTION DE LA BASE DE DONNÉES *HLA* EN HISTOCOMPATIBILITÉ

Contents

Les travaux présentés dans ce chapitre ont donné lieu à une publication dans le journal *Exploration of Immunology* [107].

6.1 Introduction

Avec l'augmentation massive de la quantité de données biomédicales de nos jours, l'extraction d'informations pour l'amélioration de la prise de décision clinique nécessite le développement de nouvelles techniques d'analyse. L'augmentation importante du volume de données sur les patients fait de la prise de décisions de traitement efficaces et sûres un grand défi pour les cliniciens. C'est le cas notamment des analyses d'antigènes leucocytaires humains (*HLA*), où le nombre de donneurs volontaires et d'allèles *HLA* augmentent de manière exponentielle. Un allèle est une version alternative du même gène, qui se distingue par des variations dans sa séquence nucléotidique. Les allèles *HLA* sont définis par séquençage de l'ADN et associés à un nom pour identifier leur séquence. Pour chaque gène *HLA*, chaque individu possède au maximum 2 allèles par locus : un de son père et un de sa mère. Les fréquences de l'allèle *HLA* dans

les populations constituent une information très importante pour la recherche et les applications cliniques.

L'une des plus grandes bases de données de phénotypes *HLA* de la World Marrow Donor Association (WMDA) comprend plus de 38 millions de donneurs enregistrés et d'unités de sang de cordon provenant de 55 pays différents [123]. Ce « livre », centralisé sur le site Bone Marrow Donors Worldwide (BMDW), est utilisé dans la pratique clinique pour la recherche de sources appropriées de cellules souches hématopoïétiques non apparentées. BMDW est l'illustration des efforts continus pour collecter les phénotypes *HLA* des donneurs volontaires de cellules souches et des unités de sang de cordon d'une manière centralisée et contrôlée.

Cette base de donnée pourrait être facilement distribuée, selon le besoin des médecins, en transférant le contrôle d'accès aux propriétaire des données et en facilitant les opérations de mise à jour. Un autre grand système pourrait être traité similairement : la base de données centralisée Immuno Polymorphism (IPD). Celle-ci a été développée pour l'étude du polymorphisme dans les gènes du système immunitaire. Elle contient actuellement plus de 30 000 allèles *HLA* dans le cadre de la base de données IPD-IMGT/*HLA* [99].

Toutes ces bases de données fournissent un dictionnaire des allèles *HLA* et de leurs fréquences dans différentes populations. L'intérêt de collecter toutes ces données auprès de différentes populations est de savoir globalement si les personnes d'une population donnée ont des *HLA* de certaines fréquences. Cela aide beaucoup à résoudre des problèmes d'histocompatibilité tenue à expliquer volontairement de trouver un donneur correspondant au *HLA* d'un receveur. La collecte de toutes ces données auprès de plusieurs pays n'est pas aisée en raison de la sensibilité de ces données et des restrictions correspondantes.

Dans la recherche pharmacogénétique et la pratique clinique, on s'intéresse de plus en plus à la compréhension des distributions mondiales des allèles *HLA* pour le profilage des risques liés à une maladie spécifique. En effet, les allèles *HLA* sont associés à diverses maladies auto-immunes et infectieuses, comme la sclérose en plaques [79].

Les données des populations saines dans l'Allele Frequency Net Database (AFND) [47] peuvent être d'un grand apport pour évaluer ces fréquences [1]. Cette base est une ressource disponible gratuitement pour le stockage des données de fréquence sur les polymorphismes de plusieurs gènes liés à l'immunité, y compris le *HLA*, sur les récepteurs tueurs de type immunoglobuline (KIR), la séquence liée aux polypeptides (MIC) du CMH de classe I et plusieurs polymorphismes de gènes de cytokines (ex., IL4, TGFB1, TNF). Ces gènes, connus pour être parmi les régions les plus polymorphes du génome humain, jouent un rôle important dans la réponse du système immunitaire. Les données de population *HLA* de l'AFND sous-tendent également

souvent des études anthropologiques, ainsi que des analyses in silico pour le développement de vaccins basé sur la prédiction d'épitopes, parmi de nombreuses autres applications. La base de données AFND a résolu le problème d'accès aux fréquences *HLA*. Techniquement, toutes les bases de données ont toujours vu la centralisation comme la seule architecture possible. Nous proposons ici une architecture distribuée, différente de stockage et d'interrogation des données.

L'architecture centralisée de la base de données facilite la collecte de données de différentes populations dans une base de données commune. Cliniciens et chercheurs peuvent accéder à un large ensemble de données afin d'élargir leurs analyses.

Historiquement, l'infrastructure de données centralisée de la plate-forme AFND a montré sa robustesse pour des analyses à grande échelle. Avec le renforcement des lois sur la gouvernance des données et des questions de propriété intellectuelle au cours des dernières années, les infrastructures de données distribuées ont gagné en popularité. Cela est dû aux garanties de sécurité et de confidentialité des données qui doivent être assurées pour minimiser les risques d'atteinte à la vie privée des participants. Les architectures distribuées visées permettent aussi de contrôler l'usage (qui accède à quoi, pour faire quoi) des infrastructures distribuées.

L'accroissement de données de santé électroniques, comme détaillé précédemment dans la motivation de la thèse, a créé un énorme défi en matière de gouvernance des données. La gestion d'un volume croissant de données et la fusion de jeux de données hétérogènes sont des enjeux complexes. En Europe, des politiques de protection des données de santé, imposant des restrictions fondées mais fortes quant au partage des données, ont été définies en 2016 par le Règlement général sur la protection des données (RGPD) [42]. Aux États-Unis, la loi HIPAA (Health Insurance Portability and Accountability Act) couvre la sécurité et la confidentialité des informations médicales ou, dans le langage HIPAA, les informations de santé protégées (PHI). Selon la loi, les «entités couvertes», principalement les hôpitaux, les assurés et les organisations qui traitent les RPS pour eux, ont la responsabilité légale d'assurer la protection des informations de santé [55].

La gouvernance des données offre aux établissements de santé une méthode de partage des données médicales à la fois standardisée et structurée, afin de délivrer des soins de la plus haute qualité à chaque patient. L'infrastructure distribuée, l'analyse collaborative et le partage de données ont le potentiel de transformer profondément le domaine de la santé pour le mieux. Cependant, pour y parvenir, les établissements de santé (universités, hôpitaux, centres de recherche et entreprises technologiques) doivent coopérer et faire émerger des analyses distribuées sécurisées ; où les données agrégées peuvent circuler librement et en toute sécurité dans tout le système de santé.

Dans le cadre de notre projet, nous avons pris en compte les contraintes de protection de données sur les données génétiques, comme imposées par HIPAA et RGPD, et nous avons assuré la flexibilité des analyses biomédicales sécurisées et distribuées à grande échelle. Nous proposons un modèle distribué fonctionnant à différents niveaux, dans le cadre de l'AFND, permettant une analyse distribuée sécurisée entre sites sans partage de données confidentielles entre eux.

Dans ce chapitre, nous appliquons la distribution des bases de données et des calculs aux ensembles de données *HLA*, offrant une alternative à la méthode de centralisation historique dans l'AFND. Nous décrivons également les différents niveaux d'analyse sur AFND centralisé et les enjeux d'une analyse distribuée. Il présente également le modèle distribué proposé pour l'AFND distribué.

6.2 Méthode

La base de données AFND a été créée en 2003 avec des fréquences d'allèles *HLA*/ lignées alléliques. Elle s'est enrichie au fil des années de nouveaux outils intégrés dans une nouvelle version majeure en 2015 [98, 46]. En 2020, une nouvelle mise à jour a été effectuée sur les jeux de données disponibles et permet maintenant la soumission et le partage de données en utilisant un critère de classification de la qualité des données (GSB : or, argent, bronze) [47].

La base de données de l'AFND a été consultée par plus de 100 000 utilisateurs différents de 186 pays au cours des dernières années. Cette référence contient des informations provenant de plus de 10 millions d'individus sains et de plus de 1600 populations. Ces données démographiques proviennent de 141 pays du monde entier avec une couverture de population variée. Avec ces données, les utilisateurs peuvent effectuer des analyses sur les fréquences alléliques, sur les gènes, les génotypes ou les haplotypes pour *HLA*, KIR, MIC ainsi que sur cytokines.

Les données du site de l'AFND sont issues (1) de publications à comité de lecture, (2) de données démographiques émanant d'ateliers internationaux *HLA* et immunogénétique (IHWS), (3) de soumissions de laboratoires du monde entier et (4) de rapports de publication (SPR) en collaboration avec la revue «Human Immunology».

Pour la soumission des données, l'AFND impose certaines exigences telles que la validation du nom de l'allèle selon les lignes directrices officielles de la nomenclature IPD-IMGT/*HLA* et IPD-KIR, l'homogénéisation de la dénomination de la population, l'attribution de la région géographique à laquelle appartient la population et la validation des données de fréquence par l'outil de téléchargement des soumissions SPR de l'AFND.

La base de données AFND est disponible pour stocker des données de fréquence sur les polymorphismes de plusieurs gènes de l'immunité. Cette ressource donne une idée des distributions mondiales des allèles *HLA*. Les utilisateurs intègrent les résultats de leur travail dans une base de données commune (centralisée) et recherchent dans la base de données les informations déjà disponibles. Ce référentiel contient actuellement des données dans les formats d'allèle, d'haplotype et de génotype. La plupart des données gérées sont des tables de fréquences alléliques pour un gène. Celles-ci sont évaluées par comptage des données individuelles de génotypage ou par sommation marginale sur le tableau de fréquence des haplotypes.

En tant que référentiel de données majeur, l'AFND s'appuie sur la centralisation des ressources de données, c'est-à-dire le transfert réel de données agrégées (fréquences d'allèles ou d'haplotypes), plutôt que sur les données de génotype *HLA* individuelles d'origine. Ces trois niveaux d'information génétique font l'objet de différents enjeux de gouvernance de données comme présenté dans le tableau 6.1. La version distribuée de l'AFND que nous proposons répond à ces 3 défis.

TABLE 6.1 – Les différents niveaux d'AFND et les défis pour AFND distribué

Niveau analytique	**AFND**	**Défis pour la gouvernance des données**		
		Risque de confidentialité	Contrôle d'usage	Exigence de calcul
1	Fréquence allélique		X	
2	Fréquence des haplotypes		X	X
3	Données génotypiques individuelles	X	X	X

Afin de respecter la confidentialité des données sensibles des patients et d'éviter la centralisation des transferts de données de l'établissement propriétaire des données vers l'établissement qui utilise les données, les données individuelles sont souvent confinées à l'établissement propriétaire des données. La centralisation implique un transfert de la violation de la vie privée, un risque de perte de données et une perte implicite de contrôle des usages.

Nous présentons maintenant les 3 niveaux :

Niveau 1 : Distribution de la fréquence des allèles :

Actuellement, toutes les données de fréquence allélique sont stockées dans une base de données commune (centrale). Cette base de données contient tous les allèles et les fréquences

alléliques correspondantes de plusieurs populations. Pour visualiser ou analyser ces données, l'AFND propose aux utilisateurs un outil d'interrogation permettant d'explorer ces fréquences alléliques dans une ou plusieurs populations ; en fonction de critères donnés.

Notre version distribuée, de ce niveau, offre les mêmes outils que l'AFND (classique/centralisé) aux utilisateurs ; pour explorer les fréquences alléliques. De plus, cela permet aux centres de conserver leurs données sur leurs sites d'origine ; sans avoir besoin de les soumettre ou de les partager et de préserver les connaissances qui utilisent leurs données.

Niveau 2 : Distribution des données de fréquence des haplotypes :

À ce niveau, les fréquences des haplotypes sont traitées de la même manière que le fréquences alléliques. Notre version distribuée, pour l'analyse des fréquences d'haplotypes, permet aux utilisateurs de consulter un haplotype particulier dans un ensemble de populations avec deux gènes ou plus comme le cas centralisé. Notre approche permet de réaliser des analyses distribuées sans obliger les centres à partager leurs données dans une base de données centrale. Notre nouvelle version d'AFND distribuée, offre non seulement le contrôle de qui fait quoi avec les données, mais permet également d'effectuer des calculs localement et d'envoyer simplement les résultats agrégés au serveur. Celle-ci effectue un calcul global des fréquences d'haplotypes à la fin.

Niveau 3 : Distribution des données génotypiques individuelles :

Enfin, au niveau le plus sensible, la version actuelle de l'AFND collecte les données génotypiques individuelles via des rapports de population. L'utilisateur, à ce niveau, peut visualiser les fréquences génotypiques d'un profil donné. La sensibilité des données démographiques nous motive à proposer notre nouvelle version distribuée de l'AFND. Notre méthode n'a pas besoin de partager de données de génotype individuelles. Elle permet aux utilisateurs d'effectuer des calculs de fréquence d'allèles sur les sites des fournisseurs de données et simplement de partager ces fréquences avec le serveur qui effectue ensuite le calcul global. A ce niveau, notre nouvelle approche permet aux fournisseurs de données de savoir qui fait quoi avec leurs données, d'effectuer des calculs en local, de ne pas surcharger le serveur par des calculs et de garantir la confidentialité des données individuelles des patients.

Architecture centralisée et architecture distribuée

Actuellement, tous les sites participant à l'AFND envoient leurs fréquences alléliques, leurs fréquences haplotypiques et leurs données génotypiques aux serveurs. Les utilisateurs qui souhaitent analyser des données, envoient leurs requêtes à une plateforme reliée au serveur central hébergeant toutes les données. Le serveur exécute la requête sur les données soumises et pré-

sente l'ensemble de données résultantes en fonction de critères bien déterminés (voir figure 6.1-a)).

Afin de prendre en charge des analyses distribuées à grande échelle et d'assurer la gouvernance des données, plusieurs modèles/architectures informatiques distribués peuvent être exploités. Cela permet de distribuer les analyses sans partager les données entre les sites. L'utilisation de telles infrastructures distribuées est un élément central de notre gouvernance des données multi-acteurs. Nous nous efforçons de relever le défit qui consiste à activer l'exploration de données ; tout en conservant les données sensibles sur site et en garantissant une protection renforcée des données lorsqu'elles sont déplacées.

Notre approche est basée sur l'architecture Master/Worker. Dans ce modèle, les calculs sont effectués sur des sites distribués (clients), reliés à un agrégateur de calculs (serveur). Celui-ci permet à un site d'interagir et d'accéder à certaines données depuis des sites distants. Chaque centre collecte, stocke, analyse et contrôle les données de ses propres patients. Lorsqu'un utilisateur envoie une requête au serveur, celui-ci envoie les requêtes appropriées aux sites concernés afin qu'ils puissent effectuer leurs calculs localement. Les sites répondent par des résultats de calculs ou des paramètres demandés. Ensuite, le serveur collecte ces résultats, effectue l'agrégation localement et fournit le résultat à l'utilisateur.

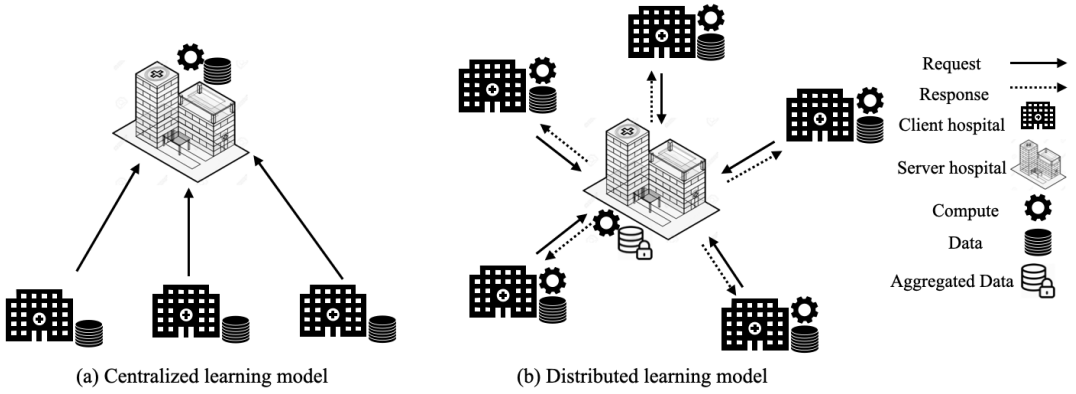


FIGURE 6.1 – Visualisation du modèle centralisé et du modèle distribué pour l'analyse des données appliquée à l'AFND

La figure 6.1 montre deux modèles différents de fonctionnement des calculs du site AFND.net. Le modèle (a) montre le modèle de calcul centralisé actuel, où le serveur collecte toutes les données des sites participant au calcul et effectue des calculs collaboratifs sur le serveur de manière

centralisée. Le modèle (b) présente notre nouveau modèle distribué proposé pour AFND.net qui permettrait à chaque site, participant à l'analyse collaborative, d'effectuer les calculs localement ; sans déplacer ses données vers le serveur. Après cela, ils envoient des résultats de calcul qui permettent au serveur d'effectuer en toute sécurité des calculs agrégés et de générer le résultat.

Le principe fondateur de notre architecture est qu'aucune donnée individuelle ne circule en dehors des centres (sites). Cependant, ce paradigme de partage offre la possibilité de contrôler localement qui accède aux données, quelles sont les utilisations de ces données et également de garantir la confidentialité de celles-ci.

La fréquence allélique y de A est définie dans le cas centralisé comme suit :

$$y_A = \frac{1}{2N} * \sum_{j=1}^{N} [g_i = A] \tag{6.1}$$

où N est le nombre d'individus et g sont les génotypes des individus.

Pour le cas distribué, la fréquence allélique est calculée comme suit :

$$y_A = \sum_{i=1}^{s} (\frac{1}{2n_i} * \sum_{j=1}^{n_i} [g_j = A]) \tag{6.2}$$

où s est le nombre de sites et n_i est le nombre d'individus sur chaque site.

6.3 Évaluation

Dans cette section, nous présentons les différentes étapes de déploiement de nos analyses sur une infrastructure distribuée. Nous décrivons un exemple de scénario d'un calcul distribué. Nous détaillons aussi une évaluation des performances en terme de précision des résultats, de temps d'exécution et de la taille des échantillons analysés.

6.3.1 Déploiement des analyses dans un environnement distribué

Pour implémenter les différents niveaux de nos calculs génétiques distribués, nous avons utilisé un ensemble de données de 1000 échantillons (allèles) disponibles sur le site «allelefrequencies.net». Ces données HLA proviennent de la population de l'Irlande du Nord. Ils ont été déposés fin 2019 [2]. Afin d'effectuer des analyses à grande échelle et de mesurer la précision et le temps d'exécution pour évaluer la stabilité de notre nouveau modèle distribué, nous avons

également utilisé un autre ensemble de données HLA public [62] qui contient 2 504 individus avec 16 979 fréquences alléliques pour 1000 génomes.

Nous avons utilisé la plateforme Grid'5000 pour réaliser nos expérimentations sur un environnement distribué. Grid'5000 [7] est une plateforme européenne, regroupant des clusters sur huit sites différents, pour la recherche dans le domaine des systèmes distribués à grande échelle et du calcul haute performance.

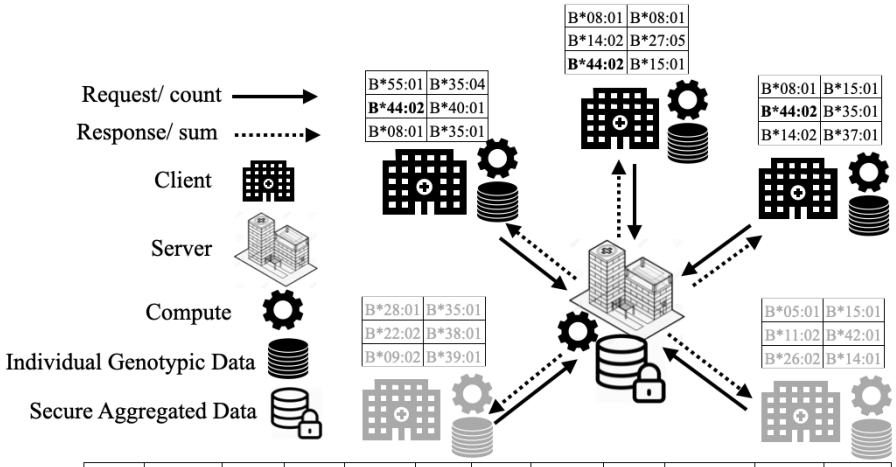
Pour notre expérimentation, nous avons réservé une machine sur un serveur exécutant un programme Python pour gérer l'analyse, les interactions avec les clients et la génération du résultat. Nous avons également utilisé des machines réparties sur différents sites pour fonctionner comme des machines clientes pour notre modèle. Notre modèle et nos algorithmes sont disponibles sur un site Gitlab [23].

6.3.2 Aperçu d'un scénario de calcul distribué

Pour évaluer notre approche, nous avons réalisé des expériences pour chaque niveau. La figure 6.2 montre un exemple de distribution de niveau 3 de données génotypiques individuelles en montrant les données génotypiques sur chaque site, le modèle de distribution et les données de fréquence d'allèle résultantes.

La figure 6.2 montre le scénario de distribution des allèles $HLA - B$. Au lieu de partager des données avec le serveur, les clients effectuent des calculs sur leurs données localement, puis retournent les résultats des calculs qui leur permettent d'agréger les résultats côté serveur sans accéder aux données locales de chaque site. Concrètement, une demande de calcul est distribuée de l'allèle $B * 44 : 02$ entre les différents sites de données. Les sites hospitaliers et leurs données en noir montrent les sites qui contiennent l'allèle $B*44 : 02$. Les sites hospitaliers et les données en gris ne contiennent pas d'allèle pertinent pour le calcul. Le tableau montre les résultats d'un calcul distribué de tous les allèles en gris et les résultats de l'allèle $B * 44 : 02$ qui correspond à notre requête en noir. Cette analyse est effectuée de manière distribuée sans déplacer les données de leurs sites. Un utilisateur d'un site envoie une demande de comptage sur un allèle affecté. Le serveur effectue une sommation sur les fréquences alléliques reçues des autres sites et diffuse le résultat final.

Le principe illustré dans la figure 6.2 s'applique à tous les allèles de tous les loci pour les fréquences d'allèles (niveau 1), les fréquences d'haplotypes (niveau 2) ainsi que les fréquences de génotypes (niveau 3).



Allele	Allele freq	Frq indiv	Allele	Allele freq	Frq indiv	Allele	Allele freq	Frq indiv	Allele	Allele freq	Frq indiv
A*01:01	0.277	0.555	B*08:01	0,222	0,444	C*04:01	0.222	0.444	DRB1*04:01	0.222	0.444
A*02:01	0.277	0.555	**B*44:02**	**0.166**	**0.333**	C*07:01	0.222	0.444	DRB1*03:01	0.166	0.333
A*11:01	0.166	0.333	B*14:02	0.111	0.222	C*05:01	0.166	0.333	DRB1*07:01	0.111	0.222
A*24:02	0.111	0.222	B*15:01	0.111	0.222	C*08:02	0.111	0.222	DRB1*15:01	0.111	0.222
A*32:01	0.055	0.111	B*35:01	0.111	0.222	C*03:03	0.111	0.222	DRB1*01:01	0.055	0.111
A*03:01	0.055	0.111	B*55:01	0.055	0.111	C*02:02	0.055	0.111	DRB1*01:03	0.055	0.111
A*66:01	0.055	0.111	B*27:05	0.055	0.111	C*03:04	0.055	0.111	DRB1*04:04	0.055	0.111
			B*37:01	0.055	0.111	C*06:02	0.055	0.111	DRB1*11:01	0.055	0.111
			B*40:01	0.055	0.111				DRB1*11:04	0.055	0.111
			B*35:04	0.055	0.111				DRB1*13:01	0.055	0.111
									DRB1*13:02	0.055	0.111

FIGURE 6.2 – Exemple de données génotypiques individuelles distribuées de $HLA-B*44:02$.

6.3.3 Évaluation des performances

Nous avons comparé le modèle de calcul centralisé classique de l'AFND à notre nouveau modèle distribué relatif à la précision des résultats et aux différents critères de performance.

Comparaison de la précision des résultats entre l'AFND et l'AFND distribué :

Nous avons utilisé la précision comme premier critère de comparaison entre les cas centralisés et distribués. La figure 6.3 montre que les points forment une droite linéaire pour comparer les estimations de fréquence d'allèles pour plusieurs cas de méthodes centralisées et distribuées. Chaque point représente une expérience d'un allèle. Cette figure est une visualisation de la méthode 2 (cas distribué) en fonction de la méthode 1 (cas centralisé). Comme attendu, il n'y a pas de dispersion des données. Les valeurs sont concordantes. Nous obtenons les mêmes valeurs pour les deux méthodes avec un coefficient de corrélation r^2 =1. La figure 6.3 montre le résultat de la corrélation d'une requête de fréquence allélique de 5 allèles pris comme exemple pour les gènes $HLA-A$, $HLA-B$, $HLA-C$, $HLA-DQB1$ et $HLA-DRB1$ entre le cas centralisé et le cas distribué. Ceci est représentatif de tous les allèles avec un coefficient de corrélation r^2 = 1.

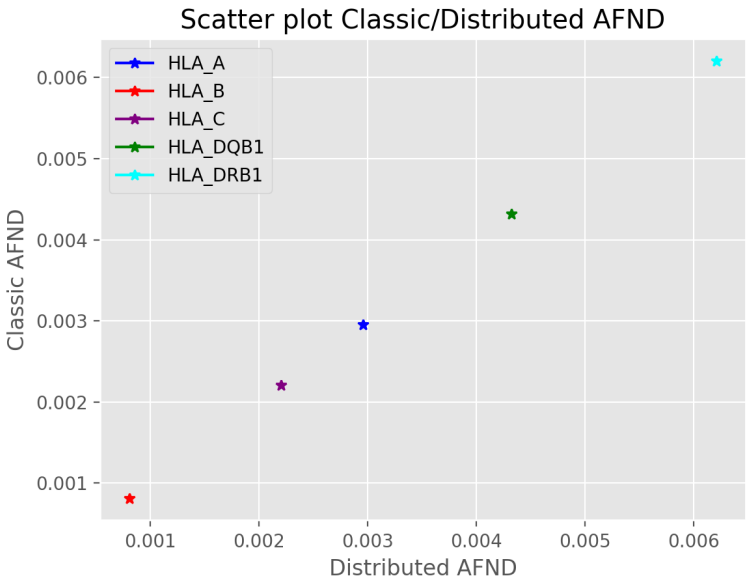


FIGURE 6.3 – Précision de l'estimation de la fréquence des allèles.

Comparaison de temps d'exécution en fonction du nombre de sites :

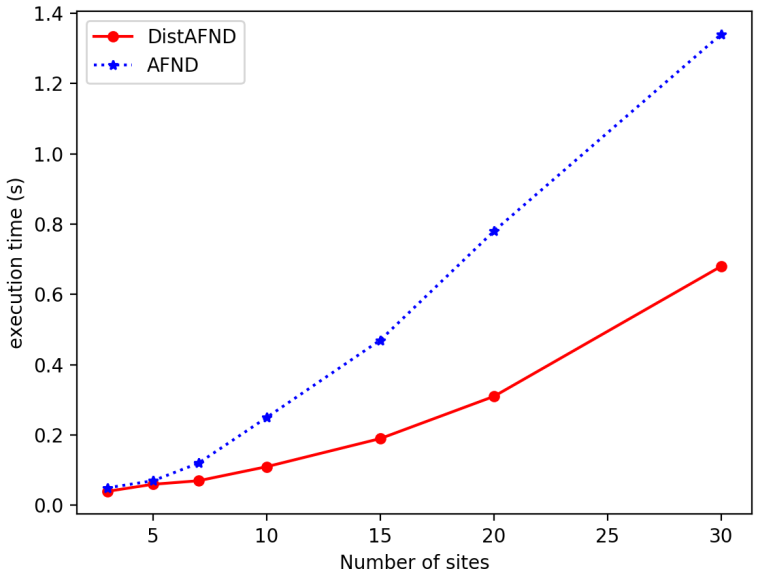


FIGURE 6.4 – Temps d'exécution par nombre de sites.

Nous avons également évalué le temps d'exécution par nombre de sites. La figure 6.4 montre

le temps d'exécution en fonction du nombre de sites. Chaque point représente un test pour le même nombre de sites pour les cas centralisés et distribués. Cette figure montre que plus le nombre de sites est grand, plus le calcul de la fréquence allélique est rapide dans le cas distribué. Cela est dû au fait que nous augmentons le nombre de sites participant au calcul pour une même taille de données globales. Le calcul sera effectué plus rapidement. La répartition des calculs sur plusieurs sites permet d'effectuer des calculs de la part des données sur chaque site séparément ; avec une agrégation des résultats des calculs sur le serveur à la fin. Ceux-ci accélèrent le temps d'exécution.

Comparaison de temps d'exécution en fonction de la taille de l'échantillon par sites : Nous avons également évalué le temps d'exécution en fonction de la taille de l'échantillon dans les cas distribués et centralisés. Chaque point représente une expérience pour les deux cas pour une taille d'échantillon définie. La figure 6.5 montre que l'estimation de la fréquence allélique devient plus rapide dans le cas distribué après une certaine taille d'échantillon (30 000 dans notre cas).

Pour les deux figures 6.4 et 6.5, l'algorithme distribué est plus rapide que l'algorithme centralisé. L'algorithme distribué est beaucoup plus rapide dans le cas où on augmente la taille de données, que dans celui où on augmente le nombre de sites participants à l'analyse. Ceci s'explique, car dans le cas où on augmente le nombre de sites, on augmente aussi le temps de communications entre les sites en plus de temps de calcul.

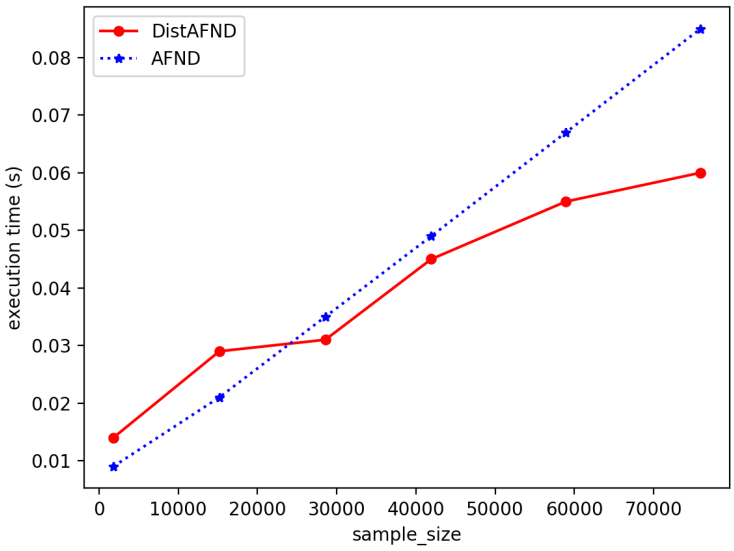


FIGURE 6.5 – Temps d'exécution par taille d'échantillons.

6.4 Discussion

Notre évaluation a montré que notre approche donne une précision équivalente par rapport à l'AFND centralisée ; ainsi qu'une meilleure performance en terme de temps d'exécution par rapport à l'augmentation de la taille de l'échantillon et du nombre de sites.

Ce nouveau modèle sécurisé et distribué de données et d'analyses sensibles apporte de réels avantages au monde de la médecine de précision. En effet, il permet des calculs précis à grande échelle sans partager les données sensibles des patients, en particulier avec l'émergence de données massives dans le monde. Les calculs distribués ont été considérés comme un outil très important en informatique. Cela est confirmé par le temps d'exécution minimal de nos travaux par rapport à l'AFND centralisée. Notre nouveau modèle de distribution d'analyses et de bases de données HLA offre une grande contribution dans le domaine biomédical en raison de son support de calculs sécurisés et distribués appliqués à la recherche de fréquences alléliques de données HLA. Il est, aussi, d'un apport important pour des collaborations entre plusieurs organisations dans le monde, sans avoir à partager les données HLA à l'aide d'une base de données unique ainsi que le contrôle de l'utilisation des données que celle-ci fournit (utilisateurs et fournisseurs de données savent qui fait quoi avec leurs données).

Notre nouvelle base de données de génétique des populations HLA distribuée s'applique à de nombreuses ressources génétiques des populations, en particulier, la solution AFND. Nous avons utilisé le modèle distribué Master/Worker et l'agrégation comme processus de synthèse des données afin de faciliter l'analyse statistique. Avec cette méthode, nous sommes en mesure de diffuser et d'exploiter des informations agrégées (informations sur tous les patients ou groupes de patients spécifiques que nous combinons afin qu'un patient individuel ne puisse plus être identifié ou mentionné). La méthode d'agrégation pour l'analyse distribuée est un choix pour rester dans un cadre conceptuel et méthodologique plus standardisé et mieux connu dans la recherche biomédicale. L'agrégation de calculs n'est pas la seule solution de synthèse de données utilisable pour ce genre de problème.

Des données synthétiques peuvent être créées à partir de différentes méthodes d'apprentissage automatique, par exemple, en utilisant la méthode d'avatarisation [14] qui a été validée par la CNIL, l'agence française de protection des données, comme méthode d'anonymisation d'un ensemble de données anonymisé. Cette méthode a été utilisée comme méthode de synthèse des données génomiques HLA pour la non identification des données afin de sortir du périmètre RGPD. Ce travail a montré sa bonne robustesse statistique même avec des données qualitatives importantes.

Nous pouvons également utiliser d'autres infrastructures distribuées et sécurisées pour gérer des données biomédicales sur différents sites distants en toute sécurité. La technologie blockchain, par exemple, a été utilisée comme solution aux défis de gouvernance associés au partage de données génomiques [110].

Pour l'analyse distribuée, le nouveau paradigme d'apprentissage fédéré a également montré son efficacité pour assurer un apprentissage distribué et sécurisé à grande échelle. Cette méthode consiste à envoyer le modèle sur des sites distants, à s'entraîner sur des jeux de données locaux à chaque site, puis, à mettre à jour et agréger les paramètres provenant de chaque site et à retransmettre les résultats aux sites.

Cet algorithme a été utilisé, par exemple, pour le développement d'un cadre d'apprentissage fédéré qui permet l'étude des relations structurelles du cerveau entre les maladies et les cohortes cliniques. Ce cadre permet un accès sécurisé et une méta-analyse de toutes les données biomédicales sans partager d'informations individuelles [114].

La solution proposée dans ce chapitre a été appliquée comme solution alternative à la référence AFND. Nos algorithmes distribués ont été testés et évalués sur des analyses appliquées aux données HLA. Notre solution distribuée peut être utilisée pour tout type de données comme KIR et bien d'autres.

6.5 Conclusion

Nous avons développé une base de données génétique distribuée de données HLA. Entre autres, la nouvelle version distribuée s'appliquerait parfaitement à la base de données historique de l'AFND. Ce référentiel permet la recherche et l'analyse des fréquences alléliques, des fréquences haplotypiques et des données génotypiques individuelles. Il s'effectue d'une manière distribuée et sécurisée sans obliger les institutions à partager leurs données dans une base de données commune. Nous avons basé notre algorithme sur le modèle distribué Master/ Worker. Nous l'avons évalué dans un environnement d'expérimentation distribué réaliste (Grid'5000) avec des données réelles. Les expériences ont montré de bons résultats de performance (précision et temps d'exécution) pour notre nouvelle version distribuée.

CHAPITRE 7

ESTIMATION DISTRIBUÉE DE LA FRÉQUENCE DES HAPLOTYPES *HLA*

Contents

Les travaux présentées dans ce chapitre ont donné lieu à une soumission dans *International Journal of Immunogenetics*.

7.1 Introduction

Depuis des années, l'estimation des fréquences des haplotypes est devenue de plus en plus importante pour les cliniciens. Cette analyse a été initialisée en 1975 par Dempster et al. [22] afin d'obtenir les estimations du maximum de vraisemblance dans un modèle comprenant des données manquantes. Dans le cadre du *HLA*, Excoffier et al. [36] ont proposé un nouvel algorithme nommé EM (espérance-maximisation), qui prend en entrée une liste de génotypes *HLA* et permet d'estimer les fréquences des haplotypes dans cette population.

Les méthodes les plus couramment utilisées pour estimer les fréquences des haplotypes *HLA* sont les algorithmes basés sur l'EM [57]. Ceux-ci peuvent accueillir plusieurs loci avec un nombre arbitraire d'allèles pour un grand nombre d'individus avec des haplotypes ambigus [57,

102]. Cependant, ils ont montré des performances limitées avec une petite taille d'échantillon. Ils ne permettent pas la détermination d'haplotype à partir d'un individu unique. Une autre technique d'imputation d'haplotypes *HLA* ; basée sur le maximum de vraisemblance a été validée sur plusieurs ensembles de données pour la recherche de donneurs non apparentés et ce, dans le cadre d'une transplantation de moelle osseuse[48]. Cette méthode calcule la paire d'haplotypes la plus probable à partir des génotypes *HLA* ; sur la base des fréquences des génotypes *HLA* dans les registres de transplantation de donneurs.

Tous ces travaux ont montré leurs succès pour l'estimation des haplotypes sur des bases de données de génotypes centralisés, sans avoir accès aux génotypes des parents.

Depuis l'apparition de la loi RGPD [42] qui contrôle le partage des données des patients, les recherches médicales sont effectués d'une manière centralisée au sein des institutions. Avec l'accroissement important de la taille des données et le besoin d'efficacité de calculs des analyses biomédicales, la centralisation de ces données pseudonymes n'est plus efficace. Ceci est considéré comme une vraie contrainte pour l'évolution de la recherche médicale. Un problème émergent concerne la ré-identification des données pseudonymes et qui peut limiter la protection et la confidentialité des données [19].

Dans ce chapitre, nous proposons une nouvelle version distribuée de l'estimation des fréquences des haplotypes *HLA* à partir des génotypes apparentés en utilisant l'algorithme de l'espérance maximisation (EM).

7.2 EM distribué pour l'estimation des fréquences des haplotypes

7.2.1 Description de l'algorithme

L'algorithme Espérance-Maximisation (EM) est un algorithme itératif (voir notre introduction dans la section 3.4). Il permet de regarder en même temps toutes les phases possibles qui donnent les haplotypes, et de trouver les paramètres du maximum de vraisemblance d'un modèle qui dépend de variables non observables. Il pondère toutes les possibilités pour donner la meilleure estimation des fréquences possibles.

L'estimation des fréquences d'haplotypes, par maximum de vraisemblance dans l'algorithme EM a été réalisée comme suit :

— Une étape de l'évaluation de l'espérance E, où on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées. Dans notre cas d'applica-

tion, on calcule généralement la probabilité de l'échantillon en utilisant des estimations d'haplotype de l'itération précédente.

— Une étape de maximisation M, où on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E. L'estimation de la fréquence des haplotypes est, alors, inspirée d'une procédure de comptage de gènes. Pour chaque génotype, la présence d'un haplotype est comptée par la probabilité de sa phase résultante.

L'aspect itératif de l'algorithme EM et son fonctionnement basé sur deux étapes compliquent la réalisation de sa version distribuée.

Afin de proposer une nouvelle version distribuée de l'algorithme EM, nous avons choisi de distribuer l'étape E sur plusieurs sources et garder le calcul de l'étape M centralisé, comme l'algorithme EM classique au niveau d'un agrégateur.

La figure 7.1 montre une représentation graphique d'un scénario d'estimation des haplotypes, en utilisant l'algorithme EM distribué.

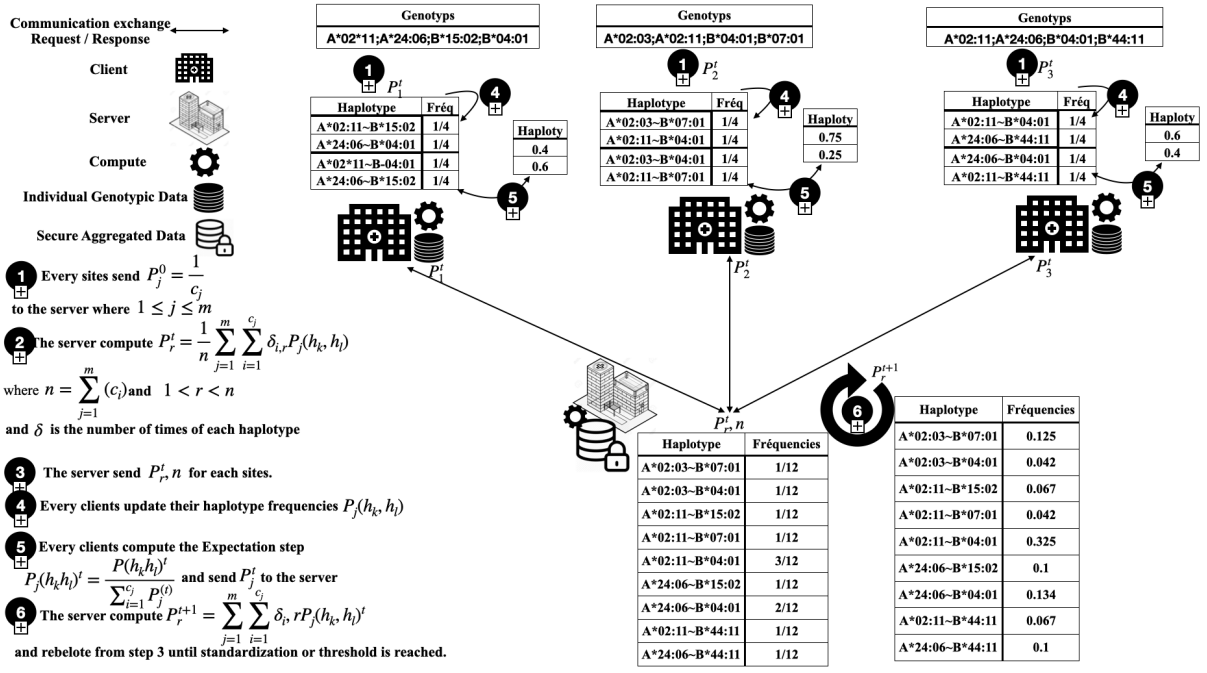


FIGURE 7.1 – Scénario d'une estimation distribuée des haplotypes en utilisant l'algorithme EM distribué.

Dans la légende de la figure, chaque étape du scénario est définie par du pseudo code, qui correspond à une communication entre les entités. Le numéro 1, par exemple, correspond à l'initialisation des fréquences des haplotypes aux niveaux de chaque site. Le numéro 2 cor-

respond à une agrégation des fréquences des haplotypes effectué par le serveur. Le numéro 3 montre une communication entre le serveur et les clients pour envoyer des données. Le numéro 4 réalise une mise à jour des fréquences des haplotypes au niveau de chaque client. Le numéro 5 représente l'étape E au niveau de chaque client. Le numéro 6 montre l'étape de maximisation de vraisemblance distribuée.

7.2.2 Algorithme distribué

Algorithm 3: Estimation distribuée des fréquences d'haplotypes

Input : m is the number of sites, $c_1, c_2, ..., c_m$ is the count of possibilities of haplotypes for each site and n is the total number of haplotypes on all sites;
Output: The distributed estimation of haplotypes frequecies is P_r^{t+1}

1 **foreach** *site* m **do**
2 Initialize $P_j(h_k h_l)^0 = \frac{1}{c_j}$ where $1 \leq j \leq m$;
3 Send P_j to the server
4 **end**
5 The master computes the global estimation of haplotype frequencies : $P_r = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{i,r} P_j(h_k h_l)$ where $1 < r < n$ and δ is the number of times the haplotype is present;
6 **while** $|P_r^{t+1} - P_r^t| \geq threshold$ **do**
7 The master send P_r to the sites;
8 **foreach** *site* m **do**
9 Update the values of $P_j(h_k h_l)$ from P_r received ;
10 Compute the expectation step : $P_j(h_k h_l)^t = \frac{P(h_k h_l)^t}{\sum_{i=1}^{c_j} P_j^t}$;
11 Send P_j^t to the server;
12 **end**
13 The master compute the maximisation step $P_r^{t+1} = \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{i,r} P_j(h_k h_l)^t$;
14 **end**
15 **return** P_r^{t+1};

L'algorithme 3 présente notre nouvel algorithme EM distribué qui permet d'estimer les fréquences des haplotypes *HLA* provenant de plusieurs sources géographiquement distribuées. Les lignes 1–4 implémentent l'étape d'initialisation. Chaque site commence par calculer P_j. Lors de la deuxième étape (ligne 5), le serveur fait une première estimation globale des fréquences des haplotypes P_r à partir des P_j reçu de chaque site et envoie cette quantité à tous les clients. Puis, chaque site met à jour les valeurs de P_j à partir de P_r reçu du serveur pour calculer la

nouvelle valeur de l'espérance E à cette itération. Chaque site envoie, ensuite, cette valeur au serveur, voir les lignes 8–12.

Au niveau du site master, le serveur calcule le maximum de vraisemblance M global à partir des valeurs des Espérances E reçues de chaque site, voir la ligne 13. Le code python de ce pseudocode est disponible sur Gitlab [24].

7.3 Expérimentations

Dans cette section, nous présentons les expérimentations de la mise en place et de l'évaluation des performances de notre contribution. Nous fournissons également, une comparaison de l'efficacité des calculs et du temps d'exécution de notre nouvelle solution par rapport à la version centralisée classique.

7.3.1 Déploiement géo-distribué

Afin de déployer notre solution, nous avons utilisé les mêmes environnements des contributions précédentes. Nous avons récupéré des données de génotypes *HLA* à partir des données du site AFND. Nous avons utilisé un algorithme existant qui permet de générer les possibilités des haplotypes pour chaque génotype. Puis, nous avons groupé chaque ensemble des haplotypes qui correspond à un génotype dans un fichier CSV. Nous avons, ensuite, implémenté notre pseudo code présenté précédemment en deux fichiers Python (un fichier pour la partie serveur et un autre pour les clients). La plate-forme Grid'5000 présentée dans les chapitres précédents, a été utilisée pour la réalisation de nos expérimentations. Nous avons réservé une machine serveur et des machines pour les clients répartis sur plusieurs sites en france. Nous avons déployé sur chaque machine le code Python et le fichier CSV qui contient les haplotypes. En exécutant le code coté serveur, les interactions entre le serveur et les clients et les résultats de chaque itération jusqu'à l'obtention du résultat optimal, sont lancés.

7.3.2 Évaluation des performances

Pour l'évaluation des performances de notre contribution, nous avons utilisé des données de génotype *HLA* pour nos expérimentations. Nous présentons maintenant une comparaison de l'efficacité en temps de calcul et en précision des résultats de notre algorithme relatif à la version centralisée.

Précision de l'estimation des fréquences des haplotypes distribuées

Le tableau 7.1 montre que l'estimation des fréquences des haplotypes à source unique (OSE) et l'estimation à sources multiples (MSE) donnent les mêmes résultats. On obtient exactement les mêmes valeurs dans notre solution distribuée ; quel que soit le nombre de sources d'estimation par rapport à la solution centralisée (OSE). Le nombre maximum de sources correspond au nombre d'individus qui, par construction, définissent un algorithme EM privé/ centralisé. En réalisant plusieurs expérimentations, on observe des valeurs écart type (SD) à zéro dans le tableau 7.1. Ce qui montre la précision maximale des résultats de notre nouvelle solution par rapport à la version centralisée.

TABLE 7.1 – Comparaisons des fréquences des haplotypes et SD entre la version centralisée (OSE) et la version distribuée (MSE)

	Haplotype Frequency (OSE)	**SD (OSE)**	**Haplotype Frequency (MSE)**	**SD (MSE)**
A*02 :11 B*04 :01	0.5	0	0.5	0
A*02 :03 B*07 :01	0.2	0	0.2	0
A*24 :06 B*15 :02	0.15	0	0.15	0
A*24 :06 B*44 :01	0.15	0	0.15	0
A*24 :06 B*04 :01	0	0	0	0
A*02 :11 B*15 :02	0	0	0	0
A*02 :11 B*44 :11	0	0	0	0
A*02 :03 B*04 :01	0	0	0	0
A*02 :11 B*07 :01	0	0	0	0

Comparaison du temps d'exécution de l'estimation des fréquences des haplotypes entre la version OSE et MSE

La figure 7.2 montre que l'estimation des fréquences des haplotypes distribuées est toujours plus rapide que la solution centralisée. Ceci est expliqué par le fait que le calcul est réalisé d'une manière parallèle entre plusieurs sites dans le cas distribué. Les analyses sont réalisées en même temps dans les différents clients (Chaque client effectue les analyses de l'estimation

sur ces propres données parallèlement) ; ce qui permet d'avoir un temps d'exécution global plus rapide que la version centralisée où toutes les étapes sont réalisées d'une façon séquentielle.

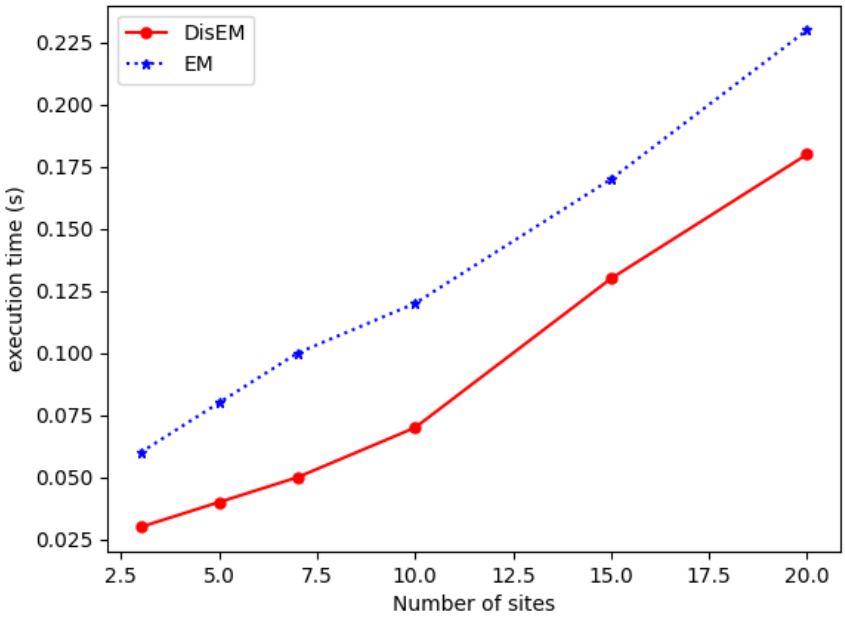


FIGURE 7.2 – Comparaison du temps d'exécution pour l'estimation des fréquences des haplotypes avec EM centralisé et EM distribué.

7.4 Discussion

Pour la première fois, nous décrivons une version distribuée de l'algorithme EM classique utilisé pour estimer la fréquence des haplotypes *HLA* pour un génotype apparenté. Elle est intéressante notamment pour limiter le partage des données sensibles. Notre solution permet de définir une estimation privée de la fréquence des haplotypes, même lorsque le nombre de sources est égal au nombre de patients. Ceci permet à chaque patient de contrôler, autoriser ou non le calcul avec ses données.

Notre approche privée permet de contrôler l'usage des données par leurs propriétaires des données. Des patients rassurés peuvent alors participer aux analyses tout en possédant le plein droit d'autoriser ou non l'utilisation de leurs données.

Les avantages offerts par le modèle distribué en sécurité et en contrôle d'usage des données sensibles encouragent les propriétaires des données à participer à ces analyses. Ceci permet

d'avoir plus de données réelles et plus de participants aux analyses ; pour plus d'efficacité des résultats pour la médecine de précision et pour la gouvernance des données.

Les modèles d'analyses distribués offrent une distribution des calculs entre les participants. Ceci permet d'éviter des attaques d'une base de données centralisées qui peuvent être considérées comme un point de défaillance unique en cas de problème.

7.5 Conclusion

L'estimation des fréquences des haplotypes *HLA* est très intéressante pour des cliniciens dans le domaine de la médecine de précision. Dans ce chapitre, nous avons proposé une nouvelle version distribuée de l'algorithme EM pour l'estimation des fréquences des haploypes ; sans accéder aux génotypes des parents. Cette nouvelle version permet de protéger des données sensibles, tout en offrant le contrôle d'usage des données à leurs propriétaires. Les expérimentations réalisées sur une infrastructure réelle géo-distribuée et les évaluations correspondantes ont montré la précision des résultats et l'efficacité en temps d'exécution des calculs de notre solution distribuée ; comparées à la version centralisée.

CHAPITRE 8

CONCLUSION

Contents

Ce chapitre résume, d'abord, les contributions présentées dans cette thèse, apportant des solutions aux problématiques présentées dans l'introduction. Ensuite, nous discutons des pistes potentielles pour les travaux futurs.

8.1 Réalisations

Cette thèse multidisciplinaire vise à motiver l'utilisation des systèmes et des algorithmes distribués pour faire face aux réglementations de l'utilisation des données ; afin de garantir la confidentialité des patients et aux coûts d'utilisation et de transfert des données ; pour améliorer les analyses en médecine de précision. Pour cela, nous avons concentré nos travaux de recherche sur les technologies qui permettent des analyses distribuées et sécurisées à grande échelle. Pour mener nos expérimentations, nous avons proposé, pour chaque analyse, un nouvel algorithme basé sur l'architecture distribuée Master / Worker, des expériences déployées sur des vraies infrastructures géo-distribuées en utilisant le banc d'essai de recherche Grid'5000 ainsi que des données réalistes fournies par nos partenaires en médecine et des comparaisons des résultats et des performances entre les analyses classiques centralisées et nos nouvelles solutions distribuées et sécurisées proposées.

Les recherches présentées dans ce manuscrit permettent de proposer une nouvelle solution pour chaque application présentée dans l'introduction.

Pour l'application de transplantation rénale KITAPP, nous avons proposé deux nouveaux algorithmes distribués et sécurisés pour une contextualisation des données collaboratives à grande échelle :

-Nous avons présenté dans le chapitre 4 un nouveau modèle distribué et sécurisé pour une

contextualisation collaborative des données des patients en utilisant un nouveau algorithme parallèle pour le calcul des centiles ne nécessitant que très peu de partage de données et qui n'implique pas le partage de données sensibles individuelles. Notre algorithme a montré son efficacité sur les résultats de calculs centralisés avec de bons résultats de performances et d'évolutivité.

-Notre deuxième contribution pour cette application concerne la préparation des données. Nous avons présenté dans le chapitre 5 une nouvelle méthode pour la réduction des dimensions distribuées des données quantitatives et qualitatives. Notre contribution est basée sur un travail publié d'une version distribuée pour l'analyse en composantes principales. Nous avons intégré la phase de la transformation des données qualitatives en données quantitatives dans le modèle distribué. Nous avons obtenu, à la fin, une nouvelle méthode distribuée et sécurisée pour l'analyse factorielle des données mixtes. Notre solution nous a permis de transformer les données complexes provenant de plusieurs sites en un sous espace de dimensions inférieures ; tout en préservant les caractéristiques des données originales. Les expérimentations présentées dans le chapitre 5 ont montré que notre solution a validé les propriétés d'efficacité, d'évolutivité et de protection de la vie privé au même niveau que la version centralisée.

Comme solution alternative à la ressource librement disponible Allele Fréquency Net Database (AFND), qui impose la centralisation des données, nous avons proposé ,dans le chapitre 6, un nouveau modèle distribué et sécurisé pour le calcul des fréquences alléliques, des fréquences haplotypiques et des données génotypiques individuelles sans partage de données entre les sites. Notre solution a, non seulement montré des résultats équivalents aux résultats de la solution centralisé avec un temps d'exécution plus rapide ,mais a garantit aussi le contrôle d'utilisation et la confidentialité des données.

Notre dernière contribution décrite dans le chapitre 7 présente une nouvelle version distribuée pour l'estimation des fréquences des haplotypes HLA ; en se basant sur l'algorithme EM distribué sans avoir accès au génotypes des parents. En plus de la garantie de la privatisation des données génotypiques individuelles, notre solution a montré son efficacité en résultats et en temps de calcul par rapport à la solution centralisée dans les expérimentations.

Les travaux présentés dans cette thèse ont donné lieu à la publication d'articles dans deux conférences internationales et dans une revue.

8.2 Perspectives

Toutes nos contributions sont basées sur le modèle distribué Master / Worker. Pour chaque application, nous avons travaillé sur la distribution des calculs sans révéler d'informations sur les données d'origine lors des communications entre le serveur et les clients. Notre modèle utilisé peut être utile pour distribuer n'importe quelle nouvelle analyse. Il nécessite tout simplement un peu d'effort pour les calculs au niveau du serveur et des clients. L'intégration des modules de calcul à notre modèle, pour qu'il soit adapté à toute nouvelle application, pourra être une bonne extension à notre travail. Cette intégration ne nécessitera qu'une petite configuration ; sans besoin d'un développeur pour chaque nouvelle analyse.

Pour chaque nouveau modèle proposé, nous avons supposé que les données au niveau des clients ont les mêmes structures et caractéristiques. Une extension de nos algorithmes pourrait s'effectuer par l'amélioration des calculs pour qu'ils fonctionnent sur des données de structures différentes.

BIBLIOGRAPHIE

[1] *AFND.NET*, Online ; accessed 07 Mars 2022, URL : http://www.allelefrequencies.net/.

[2] *AFND.NET/submit*, Online ; accessed 07 Mars 2022, URL : http://www.allelefrequencie net/submit/Default.aspx..

[3] C. AGBO CORNELIUS, H. MAHMOUD QUSAY et J. MIKAEL EKLUND, *Blockchain Technology in Healthcare : A Systematic Review*, 2019, DOI : 10.3390/healthcare7020056.

[4] T. AHRAM et al., « Blockchain technology innovations », in : *2017 IEEE Technology Engineering Management Conference (TEMSCON)*, 2017, p. 137-141, DOI : 10.1109/TEMSCON.2017.7998367.

[5] S. ANGRAAL, HM. KRUMHOLZ et WL. SCHULZ, « Blockchain Technology : Applications in Health Care », in : (2017), DOI : 10.1161/CIRCOUTCOMES.117.003800.

[6] A. AZARIA et al., « MedRec : Using Blockchain for Medical Data Access and Permission Management », in : *2016 2nd International Conference on Open and Big Data (OBD)*, Los Alamitos, CA, USA : IEEE Computer Society, 2016, p. 25-30, DOI : 10.1109/OBD.2016.11, URL : https://doi.ieeecomputersociety.org/10.1109/OBD.2016.11.

[7] Daniel BALOUEK et al., « Adding Virtualization Capabilities to Grid'5000 », in : t. 367, juil. 2012, DOI : 10.1007/978-3-319-04519-1_1.

[8] Jörg BLASIUS et Michael GREENACRE, « Multiple Correspondence Analysis and Related Methods », in : *Multiple Correspondence Analysis and Related Methods* (juin 2006), DOI : 10.1201/9781420011319.ch1.

[9] Michael BLOT et al., « GoSGD : Distributed Optimization for Deep Learning with Gossip Exchange », in : *Neurocomputing* (nov. 2018), DOI : 10.1016/j.neucom.2018.11.002.

[10] Thomas BOCEK et al., « Blockchains everywhere - a use-case of blockchains in the pharma supply-chain », in : *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)* (2017), p. 772-777.

[11] Keith BONAWITZ et al., « Practical Secure Aggregation for Privacy-Preserving Machine Learning », in : *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, Dallas, Texas, USA : Association for Computing Machinery, 2017, p. 1175-1191, ISBN : 9781450349468, DOI : 10.1145/3133956.3133982, URL : https://doi.org/10.1145/3133956.3133982.

[12] Keith BONAWITZ et al., « Towards Federated Learning at Scale : System Design », in : (fév. 2019).

[13] Fatima-zahra BOUJDAD et al., « On distributed collaboration for biomedical analyses », in : *CCGrid-Life 2019 Workshop on Clusters, Clouds and Grids for Life Sciences*, Larnaca, Cyprus : IEEE, mai 2019, p. 1-10, DOI : 10.1109/CCGRID.2019.00079, URL : https://hal.archives-ouvertes.fr/hal-02080463.

[14] *Brevet*, Online ; accessed 07 Mars 2022, URL : https://bases-brevets.inpi.fr/fr/document/FR3091602/publications.html?p=5%5C&s=1594642475255%5C&cHash=462efb7d021bce0c34a691b065b05a1d..

[15] Theodora BRISIMI et al., « Federated learning of predictive models from federated Electronic Health Records », in : *International Journal of Medical Informatics* 112 (jan. 2018), DOI : 10.1016/j.ijmedinf.2018.01.007.

[16] Paul BROUS, Marijn JANSSEN et Riikka VILMINKO-HEIKKINEN, « Coordinating Decision-Making in Data Management Activities : A Systematic Review of Data Governance Principles », in : sept. 2016, p. 115-125, ISBN : 978-3-319-44420-8, DOI : 10.1007/978-3-319-44421-5_9.

[17] Giuseppe CATTANEO et al., « MapReduce in Computational Biology - A Synopsis », in : avr. 2017, p. 53-64, ISBN : 978-3-319-57710-4, DOI : 10.1007/978-3-319-57711-1_5.

[18] *Cloud computing*, Online ; accessed 22 August 2022, URL : https://fr.wikipedia.org/wiki/Cloud_computing.

[19] CNIL, *Le G29 publie un avis sur les techniques d'anonymisation*, URL : https://www.cnil.fr/fr/le-g29-publie-un-avis-sur-les-techniques-danonymisation (visité le 11/07/2022).

[20] A.M. CORSON et al., « Percentile growth charts for biomedical studies using a porcine model », in : *Animal* 2.*12* (2008), p. 1795-1801, ISSN : 1751-7311, DOI : https://doi.org/10.1017/S1751731108002966, URL : https://www.sciencedirect.com/science/article/pii/S1751731108002966.

[21] Jeff DAILY et al., *GossipGraD : Scalable Deep Learning using Gossip Communication based Asynchronous Gradient Descent*, 2018, arXiv : 1803.05880 [cs.DC].

[22] A. P. DEMPSTER, N. M. LAIRD et D. B. RUBIN, « Maximum Likelihood from Incomplete Data Via the *EM* Algorithm », in : *Journal of the Royal Statistical Society : Series B (Methodological)* 39.*1* (sept. 1977), Publisher : Journal of the Royal Statistical Society : Series B (Methodological), p. 1-22, ISSN : 00359246, DOI : 10.1111/j.2517-6161.1977.tb01600.x, URL : http://doi.wiley.com/10.1111/j.2517-6161.1977.tb01600.x (visité le 23/10/2020).

[23] *Distributed AFND*, Online ; accessed 07 Mars 2022, URL : https://gitlab.inria.fr/ssayadi/distributed_afnd.

[24] *Distributed EM*, Online ; accessed 15 octobre 2022, URL : https://gitlab.inria.fr/ssayadi/distributed_em.git.

[25] *divatfrance*, Online ; accessed 21 January 2020, URL : www.divat.fr.

[26] Christos DOULKERIDIS et Kjetil NØRVÅG, « A Survey of Large-Scale Analytical Query Processing in MapReduce », in : *The VLDB Journal* (déc. 2013), DOI : 10.1007/s00778-013-0319-9.

[27] *Dropox*, Online ; accessed 22 August 2022, URL : https://www.dropbox.com/.

[28] Cynthia DWORK et al., « Calibrating Noise to Sensitivity in Private Data Analysis », in : *Theory of Cryptography*, sous la dir. de Shai HALEVI et Tal RABIN, Berlin, Heidelberg : Springer Berlin Heidelberg, 2006, p. 265-284, ISBN : 978-3-540-32732-5.

[29] Cynthia DWORK et al., « Calibrating Noise to Sensitivity in Private Data Analysis », in : t. Vol. 3876, jan. 2006, p. 265-284, DOI : 10.1007/11681878_14.

[30] Khaled EMAM, « Seven Ways to Evaluate the Utility of Synthetic Data », in : *IEEE Security & Privacy* 18 (juil. 2020), p. 56-59, DOI : 10.1109/MSEC.2020.2992821.

[31] Khaled EMAM, Lucy MOSQUERA et Jason BASS, « Evaluating Identity Disclosure Risk in Fully Synthetic Health Data : Model Development and Validation », in : *Journal of Medical Internet Research* 22 (nov. 2020), e23139, DOI : 10.2196/23139.

[32] *EMBL-EBI, IPD-IMGT/HLA Statistics, octobre 2020, https ://www.ebi.ac.uk/ipd/imgt/hla/stats.html*, URL : https://www.ebi.ac.uk/ipd/imgt/hla/stats.html (visité le 26/10/2020).

[33] EMMA BLUEMKE, *Privacy-Preserving AI in Medical Imaging : Federated Learning, Diffetrental Privacy, and Encrypted Computation*, Online ; accessed 21 January 2020, URL : https://blog.openmined.org/federated-learning-differential-privacy-and-encrypted-computation-for-medical-imaging/.

[34] Mark A. ENGELHARDT, « Hitching Healthcare to the Chain : An Introduction to Blockchain Technology in the Healthcare Sector », in : *Technology Innovation Management Review* 7 (2017), p. 22-34, ISSN : 1927-0321, DOI : http://doi.org/10.22215/timreview/1111, URL : http://timreview.ca/article/1111.

[35] *EU-TRAIN*, Online ; accessed 21 January 2020, URL : eu-train-project.eu.

[36] L. EXCOFFIER et M. SLATKIN, « Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population », in : 12.*5* (sept. 1995), Publisher : Molecular Biology and Evolution, p. 921-927, ISSN : 0737-4038, DOI : 10.1093/oxfordjournals.molbev.a040269.

[37] Kai FAN et al., « MedBlock : Efficient and Secure Medical Data Sharing Via Blockchain », in : *Journal of Medical Systems* (2018), DOI : https://doi.org/10.1007/s10916-018-0993-7.

[38] FATE, *An Industrial Grade Federated Learning Framework*, Online ; accessed 21 January 2020, URL : https://fate.fedai.org.

[39] Dan FELDMAN, Melanie SCHMIDT et Christian SOHLER, « Turning Big Data into Tiny Data : Constant-Size Coresets for k-Means, PCA and Projective Clustering », in : SODA '13, New Orleans, Louisiana : Society for Industrial et Applied Mathematics, 2013, p. 1434-1453, ISBN : 9781611972511.

[40] tensor FLOW, *Federated Learning : Machine Learning on Decentralized Data (Google I/O'19)*, Online ; accessed 21 January 2020, URL : https://www.youtube.com/watch?v=89BGjQYA0uE&t=251s.

[41] *Gagner du temps avec le cloud*, Online ; accessed 22 August 2022, URL : https://partitio.com/gagner-du-temps-avec-le-cloud/.

[42] *GDPR*, Online ; accessed 21 January 2020, URL : https://eur-lex.europa.eu/legal-content/FR/%20TXT/HTML/?uri=CELEX:32016R0679.

[43] Estelle GEFFARD, « Développement d'outils de médecine de précision pour accompagner la prise de décision médicale en transplantation », 2020NANT1035, thèse de doct., 2020, URL : http://www.theses.fr/2020NANT1035/document.

[44] Robin GEYER, Tassilo KLEIN et Moin NABI, « Differentially Private Federated Learning : A Client Level Perspective », in : (déc. 2017).

[45] Badih GHAZI, Rasmus PAGH et Ameya VELINGKER, « Scalable and Differentially Private Distributed Aggregation in the Shuffled Model », in : (juin 2019).

[46] Faviel F. GONZALEZ-GALARZA et al., « Allele frequency net : a database and online repository for immune gene frequencies in worldwide populations », in : *Nucleic Acids Research* 39 (2011), p. D913-D919.

[47] Faviel.F GONZALEZ-GALARZA et al., « Allele frequency net database (AFND) 2020 update : gold-standard data classification, open access genotype data and new query tools », in : *Nucleic Acids Research* 48.*D1* (nov. 2019), p. D783-D788, ISSN : 0305-1048, DOI : 10.1093/nar/gkz1029, eprint : https://academic.oup.com/nar/article-pdf/48/D1/D783/31697448/gkz1029.pdf, URL : https://doi.org/10.1093/nar/gkz1029.

[48] Pierre-Antoine GOURRAUD et al., « Inferred HLA Haplotype Information for Donors From Hematopoietic Stem Cells Donor Registries », in : *Human immunology* 66 (juin 2005), p. 563-70, DOI : 10.1016/j.humimm.2005.01.011ISTEX.

[49] Pierre-Antoine GOURRAUD et al., « Precision Medicine in Chronic Disease Management : The Multiple Sclerosis BioScreen », in : *Annals of Neurology* 76 (nov. 2014), DOI : 10.1002/ana.24282.

[50] Morgan GUILLAUDEUX et al., *Patient-centric synthetic data generation, no reason to risk re-identification in the analysis of biomedical pseudonymised data*, mai 2022, DOI : 10.21203/rs.3.rs-1674043/v1.

[51] Runxin GUO et al., « Bioinformatics applications on Apache Spark », in : *GigaScience* 7 (août 2018), DOI : 10.1093/gigascience/giy098.

[52] Stephen HARDY et al., *Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption*, 2017, arXiv : 1711.10677 [cs.LG].

[53] Corentin HERVÉ et al., *The Kidney transplantation application (KiTapp) : A visualization and contextualization tool in a kidney graft patients' cohort*, EFI 2017, Poster, mai 2017, URL : https://www.hal.inserm.fr/inserm-02161749.

[54] Nathan R. HILL et al., « Global Prevalence of Chronic Kidney Disease – A Systematic Review and Meta-Analysis », in : *PLOS ONE* 11 (juil. 2016), p. 1-18, DOI : 10.1371/journal.pone.0158765, URL : https://doi.org/10.1371/journal.pone.0158765.

[55] *HIPAA*, Online ; accessed 07 Mars 2022, URL : https://www.hipaaguide.net/hipaa-for-dummies/.

[56] C. A. R. HOARE, « Algorithm 65 : Find », in : *Commun. ACM* 4.7 (1961), p. 321-322, ISSN : 0001-0782, DOI : 10.1145/366622.366647, URL : https://doi.org/10.1145/366622.366647.

[57] Eberhard HP et al., « Comparative validation of computer programs for haplotype frequency estimation from donor registry data », in : *Tissue Antigens* (août 2013), p. 93-105, DOI : 10.1111/tan.12160ISTEXISTEXISTEXISTEXISTEXISTEXISTEXISTEXISTEXI

[58] Ahmed Faeq HUSSEIN et al., « A medical records managing and securing blockchain based system supported by a Genetic Algorithm and Discrete Wavelet Transform », in : *Cognitive Systems Research* 52 (2018), p. 1-11.

[59] Rob J. HYNDMAN et Yanan FAN, « Sample Quantiles in Statistical Packages », in : *The American Statistician* 50.4 (1996), p. 361-365, DOI : 10.1080/00031305.1996.10473566, eprint : https://www.tandfonline.com/doi/pdf/10.1080/00031305.1996.10473566, URL : https://www.tandfonline.com/doi/abs/10.1080/00031305.1996.10473566.

[60] Hafiz IMTIAZ et Anand D. SARWATE, « Differentially Private Distributed Principal Component Analysis », in : *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, p. 2206-2210, DOI : 10.1109/ICASSP.2018.8462519.

[61] INTEL, *Federated Learning for Medical Imaging*, Online ; accessed 21 January 2020, URL : https://www.intel.ai/federated-learning-for-medical-imaging/#gs.r4zqwu/.

[62] *International Genome*, Online ; accessed 07 Mars 2022, URL : https://www.internationalgenome.org/1000-genomes-summary..

[63] *IntraCranial ANeurysms : From familial forms to pathophysiological mechanisms – I-CAN*, Online ; accessed 22 August 2022, URL : https://anr.fr//Project-ANR-15-CE17-0008.

[64] Andrzej JARYNOWSKI et al., *Attempt to understand public health relevant social dimensions of COVID-19 outbreak in Poland*, avr. 2020, DOI : 10.31234/osf.io/dxkc3.

[65] Shan JIANG et al., « BlocHIE : A BLOCkchain-Based Platform for Healthcare Information Exchange », in : *2018 IEEE International Conference on Smart Computing (SMARTCOMP)* (2018), p. 49-56.

[66] Zhanhong JIANG et al., « Collaborative Deep Learning in Fixed Topology Networks », in : (juin 2017).

[67] I. T. JOLLIFFE, *Principal Component Analysis*, Springer Series in Statistics, New York : Springer-Verlag, 2002, ISBN : 0-387-95442-2, DOI : 10.1007/b98835, URL : http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95442-4.

[68] Amit JUNEJA et Michael Mahmoud MAREFAT, « Leveraging blockchain for retraining deep learning architecture in patient-specific arrhythmia classification », English (US), in : *2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018*, t. 2018-January, Institute of Electrical et Electronics Engineers Inc., avr. 2018, p. 393-397, DOI : 10.1109/BHI.2018.8333451.

[69] A. E. KENNEDY, U. OZBEK et M. T. DORAK, « What has GWAS done for HLA and disease associations ? », in : *International Journal of Immunogenetics* 44.*5* (2017), p. 195-211, ISSN : 1744313X, DOI : 10.1111/iji.12332.

[70] Craig KOLLMAN et al., « Estimation of HLA-A, -B, -DRB1 Haplotype Frequencies Using Mixed Resolution Data from a National Registry with Selective Retyping of Volunteers », in : 68.*12* (1er déc. 2007), Publisher : Human Immunology, p. 950-958, ISSN : 0198-8859, DOI : 10.1016/j.humimm.2007.10.009ISTEXISTEX, URL : http://www.sciencedirect.com/science/article/pii/S0198885907004612 (visité le 23/10/2020).

[71] Jakub KONEČNÝ, Brendan MCMAHAN et Daniel RAMAGE, « Federated Optimization :Distributed Optimization Beyond the Datacenter », in : (nov. 2015).

[72] Jakub KONEČNÝ et al., « Federated Learning : Strategies for Improving Communication Efficiency », in : (oct. 2016).

[73] Jakub KONEČNÝ et al., « Federated Optimization : Distributed Machine Learning for On-Device Intelligence », in : (oct. 2016).

[74] *KTD-Innov*, Online ; accessed 21 January 2020, URL : www.ktdinnov.fr.

[75] Do LE et al., « SGX-PySpark : Secure Distributed Data Analytics », in : mai 2019, p. 3564-3563, DOI : 10.1145/3308558.3314129.

[76] Qinbin LI, Zeyi WEN et Bingsheng HE, « A Survey On Federated Learning Systems : Vision, Hype and Reality for Data Privacy and Protection », in : (juil. 2019).

[77] Tian LI, *Federated Learning : Challenges, Methods, and Future Directions*, Online ; accessed 21 January 2020, URL : https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/.

[78] Yingyu LIANG, Maria-Florina BALCAN et Vandana KANCHANAPALLY, « Distributed PCA and k-Means Clustering », in : 2013.

[79] Jenny LINK et al., « Importance of Human Leukocyte Antigen (HLA) Class I and II Alleles on the Risk of Multiple Sclerosis », in : *PLOS ONE* 7.5 (mai 2012), p. 1-11, DOI : 10.1371/journal.pone.0036779, URL : https://doi.org/10.1371/journal.pone.0036779.

[80] Jennifer LISTGARTEN et al., « Statistical Resolution of Ambiguous HLA Typing Data », en, in : *PLOS Computational Biology* 4.2 (fév. 2008), Publisher : Public Library of Science, e1000016, ISSN : 1553-7358, DOI : 10.1371/journal.pcbi.1000016, URL : https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000016 (visité le 03/11/2020).

[81] Jian LIU et al., « Oblivious Neural Network Predictions via MiniONN Transformations », in : oct. 2017, p. 619-631, ISBN : 978-1-4503-4946-8, DOI : 10.1145/3133956.3134056.

[82] Benchoufi M, Porcher R et Ravaud P, « Blockchain protocols in clinical trials : Transparency and traceability of consent », in : (2017).

[83] MCCORBY, *Federated Learning : Client application doing classification of images and local training. Works better with the Parameter Server at Federated Learning : Client application doing classification of images and local training. Works better with the Parameter Server at https ://github.com/mccorby/PhotoLabellerServer*, Online ; accessed 21 January 2020, URL : https://github.com/mccorby/PhotoLabeller.

[84] H. MCMAHAN et al., « Federated Learning of Deep Networks using Model Averaging », in : (fév. 2016).

[85] H. Brendan MCMAHAN et al., *Communication-Efficient Learning of Deep Networks from Decentralized Data*, 2016, arXiv : 1602.05629 [cs.LG].

[86] H. Brendan MCMAHAN et al., *Learning Differentially Private Recurrent Language Models*, 2017, arXiv : 1710.06963 [cs.LG].

[87] Ishan MEENA et al., « Healthcare Analysis Using Hadoop Framework », in : 4 (oct. 2018), p. 300-306.

[88] Michel MELLINGER, « Correspondence analysis in the study of lithogeochemical data : General strategy and the usefulness of various data-coding schemes », in : *Journal of Geochemical Exploration* 21 (1984), p. 455-469.

[89] M. METTLER, « Blockchain technology in healthcare : The revolution starts here », in : *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2016, p. 1-3, DOI : 10.1109/HealthCom.2016.7749510.

[90] Diogo MEYER et al., « A genomic perspective on HLA evolution », in : *Immunogenetics* 70.*1* (2018), p. 5-27, ISSN : 0093-7711, DOI : 10.1007/s00251-017-1017-3, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5748415/ (visité le 26/10/2020).

[91] P. MYTIS-GKOMETH et al., « Notarization of Knowledge Retrieval from Biomedical Repositories Using Blockchain Technology », in : (2018), sous la dir. de Nicos MAGLAVERAS, Ioanna CHOUVARDA et Paulo de CARVALHO, p. 69-73.

[92] Satoshi NAKAMOTO, « Bitcoin : A peer-to-peer electronic cash system », in : (2009), URL : http://www.bitcoin.org/bitcoin.pdf.

[93] NVIDIA, *NVIDIA and King's College London Debut First Privacy-Preserving Federated Learning System for Medical Imaging*, Online ; accessed 21 January 2020, URL : https://news.developer.nvidia.com/first-privacy-preserving-federated-learning-system/?ncid=so-yout-39336&linkId=100000008539607#cid=organicSocial_en-us_YouTube_NVIDIA-Research-Research-NR01.

[94] Aisling O' DRISCOLL, Jurate DAUGELAITE et Roy SLEATOR, « 'Big Data', Hadoop and Cloud Computing in Genomics. », in : *Journal of biomedical informatics* 46 (juil. 2013), DOI : 10.1016/j.jbi.2013.07.001.

[95] *Overview of Precision Medicine*, Online ; accessed 16 August 2022, URL : https://www.thermofisher.com/fr/fr/home/clinical/precision-medicine/precision-medicine-learning-center/precision-medicine-resource-library/precision-medicine-articles/overview-precision-medicine.html.

[96] Jérôme PAGÈS, *Multiple Factor Analysis by Example Using R*, nov. 2014, p. 1-253, ISBN : 9780429171086, DOI : 10.1201/b17700.

[97] Le PHONG et al., « Privacy-Preserving Deep Learning via Additively Homomorphic Encryption », in : *IEEE Transactions on Information Forensics and Security* PP (déc. 2017), p. 1-1, DOI : 10.1109/TIFS.2017.2787987.

[98] James ROBINSON et al., « IPD-IMGT/HLA Database », in : *Nucleic acids research* 48 (oct. 2019), DOI : 10.1093/nar/gkz950.

[99] James ROBINSON et al., « The IPD and IMGT/HLA database : allele variant databases », in : *Nucleic acids research* 43 (nov. 2014), DOI : 10.1093/nar/gku1161.

[100] James ROBINSON et al., « The IPD and IMGT/HLA database : allele variant databases », en, in : *Nucleic Acids Research* 43.*D1* (jan. 2015), Publisher : Oxford Academic, p. D423-D431, ISSN : 0305-1048, DOI : 10.1093/nar/gku1161, URL : https://academic.oup.com/nar/article/43/D1/D423/2438496 (visité le 26/10/2020).

[101] Theo RYFFEL et al., *A generic framework for privacy preserving deep learning*, 2018, arXiv : 1811.04017 [cs.LG].

[102] Rany M. SALEM, Jennifer WESSEL et Nicholas J. SCHORK, « A comprehensive literature review of haplotyping software and methods for use with unrelated individuals », in : *Human Genomics* 2 (2005), p. 39-66.

[103] Salman SALLOUM et al., « Big data analytics on Apache Spark », in : *International Journal of Data Science and Analytics* (oct. 2016), DOI : 10.1007/s41060-016-0027-9.

[104] E. L. G. SAUKAS et S. W. SONG, « A Note on Parallel Selection on Coarse-Grained Multicomputers », in : *Algorithmica* 24.*3-4* (1999), p. 371-380, DOI : 10.1007/PL00008268ISTEX.

[105] S. SAYADI, S. BEN REJEB et Z. CHOUKAIR, « Blockchain Challenges and Security Schemes : A Survey », in : (2018), p. 1-7, DOI : 10.1109/COMNET.2018.8621944.

[106] Sirine SAYADI et al., « Distributed Contextualization of Biomedical Data : A Case Study in Precision Medicine », in : *17th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2020, Antalya, Turkey, November 2-5, 2020*, IEEE, 2020, p. 1-6, DOI : 10.1109/AICCSA50499.2020.9316502, URL : https://doi.org/10.1109/AICCSA50499.2020.9316502.

[107] Sirine SAYADI et al., « Distributing HLA database in histocompatibility : a shift in HLA data governance », in : *Exploration of Immunology* (août 2022), URL : https://hal.archives-ouvertes.fr/hal-03747555.

[108] Sirine SAYADI et al., « Secure Distribution of Factor Analysis of Mixed Data (FAMD) and Its Application to Personalized Medicine of Transplanted Patients », in : *Advanced Information Networking and Applications*, Springer International Publishing, 2021, p. 507-518, ISBN : 978-3-030-75100-5.

[109] Heiko SCHEEL et al., « A Privacy Preserving Approach to Feasibility Analyses on Distributed Data Sources in Biomedical Research », in : *Studies in health technology and informatics* 267 (sept. 2019), p. 254-261, DOI : 10.3233/SHTI190835.

[110] Mahsa SHABANI, « Blockchain-based platforms for genomic data sharing : a de-centralized approach in response to the governance problems ? », in : *Journal of the American Medical Informatics Association* 26.*1* (nov. 2018), p. 76-80, ISSN : 1527-974X, DOI : 10.1093/jamia/ocy149, eprint : https://academic.oup.com/jamia/

article-pdf/26/1/76/34151251/ocy149.pdf, URL : https://doi.org/10.1093/jamia/ocy149.

[111] Zahra SHAKERI, Anand SARWATE et Waheed BAJWA, « Sample Complexity Bounds for Dictionary Learning from Vector- and Tensor-Valued Data », in : jan. 2021, p. 134-162, ISBN : 9781108427135, DOI : 10.1017/9781108616799.006.

[112] Micah SHELLER et al., « Multi-Institutional Deep Learning Modeling Without Sharing Patient Data : A Feasibility Study on Brain Tumor Segmentation », in : (oct. 2018).

[113] Reza SHOKRI et Vitaly SHMATIKOV, « Privacy-preserving deep learning », in : sept. 2015, p. 909-910, ISBN : 9781509018246, DOI : 10.1109/ALLERTON.2015.7447103.

[114] Santiago SILVA et al., « Federated learning in Distributed Medical Databases : Meta-Analysis of Large-Scale Subcortical Brain Data (Supplementary Material) », in : (oct. 2018), URL : https://hal.inria.fr/hal-01895800.

[115] Dimitris STRIPELIS et al., « A Scalable Data Integration and Analysis Architecture for Sensor Data of Pediatric Asthma », in : t. 2017, avr. 2017, p. 1407-1408, DOI : 10.1109/ICDE.2017.198.

[116] Nugent T, Upton D et Cimpoesu M, « Improving data transparency in clinical trials using blockchain smart contracts », in : (2016).

[117] Ronald TAYLOR, « An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics », in : *BMC bioinformatics* 11 Suppl 12 (déc. 2010), S1, DOI : 10.1186/1471-2105-11-S12-S1.

[118] Aurelle TCHAGNA KOUANOU et al., « An optimal big data workflow for biomedical image analysis », in : *Informatics in Medicine Unlocked* 11 (mai 2018), DOI : 10.1016/j.imu.2018.05.001.

[119] Paul VANHAESEBROUCK, Aurélien BELLET et Marc TOMMASI, « Decentralized Collaborative Learning of Personalized Models over Networks », in : (oct. 2016).

[120] Shiqiang WANG et al., *Adaptive Federated Learning in Resource Constrained Edge Computing Systems*, 2018, arXiv : 1804.05271 [cs.DC].

[121] Martin WEISS et al., « Blockchain as an enabler for public mHealth solutions in South Africa », in : *2017 IST-Africa Week Conference (IST-Africa)* (2017), p. 1-8.

[122] *What is precision medicine ?*, Online ; accessed 16 August 2022, URL : https://medlineplus.gov/genetics/understanding/precisionmedicine/definition/.

[123] *WMDA*, Online ; accessed 07 Mars 2022, URL : https://statistics.wmda.info/.

[124] Zebin WU et al., « Parallel and Distributed Dimensionality Reduction of Hyperspectral Data on Cloud Computing Architectures », in : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.6 (2016), p. 2270-2278, DOI : 10.1109/JSTARS.2016.2542193.

[125] Qi XIA et al., « MeDShare : Trust-less Medical Data Sharing Among Cloud Service Providers Via Blockchain », in : *IEEE Access* PP (juil. 2017), p. 1-1, DOI : 10.1109/ACCESS.2017.2730843.

[126] N. YADAV, Eswara B et K. SRINIVASA, « Cloud-Based Healthcare Monitoring System Using Storm and Kafka », in : jan. 2018, p. 99-106, DOI : 10.1007/978-981-13-2348-5_8.

[127] Timothy YANG et al., *Applied Federated Learning : Improving Google Keyboard Query Suggestions*, 2018, arXiv : 1812.02903 [cs.LG].

[128] Matei ZAHARIA et al., « Spark : Cluster Computing with Working Sets », in : *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing* 10 (juil. 2010), p. 10-10.

[129] Aiqing ZHANG et Xiaodong LIN, « Towards Secure and Privacy-Preserving Data Sharing in e-Health Systems via Consortium Blockchain », in : *J. Med. Syst.* 42.8 (août 2018), p. 1-18, ISSN : 0148-5598, DOI : 10.1007/s10916-018-0995-5, URL : https://doi.org/10.1007/s10916-018-0995-5.

[130] Peng ZHANG et al., « FHIRChain : Applying Blockchain to Securely and Scalably Share Clinical Data », in : *Computational and Structural Biotechnology Journal* 16 (juil. 2018).

[131] Guy ZYSKIND et al., « Decentralizing Privacy : Using Blockchain to Protect Personal Data », in : mai 2015, p. 180-184.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Architectures réparties et conteneurs logiciels sécurisés pour coopérations médicales multi-sites

Mot clés : analyse de données biomédicales, analyses distribués multi-site et multi-partenaires, protection et sécurisation des données, médecine de précision, contextualisation distribuée

Résumé : L'analyse de larges quantités de données biomédicales permet d'obtenir une meilleure précision des résultats et aide les chercheurs et les cliniciens à établir de bonnes diagnostiques et prendre les bonnes décisions. Les collaborations nationales et internationales qui explorent la médecine de précision se fondent sur des analyses sur des données provenant de populations variées et nombreuses. Les analyses biomédicales actuelles sont sujets à des problèmes en terme de protection de données.

Dans cette thèse multidisciplinaire, nous contribuons à la résolution de ces problèmes par le développement d'analyses de données biomédicales distribuées ainsi que par leur déploiement dans des infrastructures informatiques et leur intégration dans des outils médicaux réels. Dans le cadre d'une application de transplantation rénale, nous proposons une nouvelle solution de contextualisation distribuée qui aide les cliniciens à évaluer les problèmes rénaux de patients. Pour la même application nous proposons également une nouvelle version distribuée de l'analyse factorielle des données mixtes (FAMD) pour la réduction des dimensions. Dans le cadre des données *HLA*, nous proposons un nouveau modèle distribué et sécurisé pour l'estimation de fréquences appliquées sur des bases de données *HLA*. Nous proposons également un nouveau algorithme pour l'estimation distribuée de la fréquence d'haplotypes *HLA* en utilisant l'algorithme EM.

Toutes nos contributions ont été déployées sur une infrastructure réelle de nuage géo-distribués et évaluées sur des données réelles ou réalistes. Les expérimentations ont montré de (très) bons résultats en terme de performances comparés aux solutions centralisées.

Title: Distributed architectures and secure software containers for multi-site medical cooperation

Keywords: precision medicine, data analysis, distributed analysis, aggregate computing, data security, distributed contextualization

Abstract: The analysis of large quantities of biomedical data makes it possible to obtain better results and helps researchers and clinicians to establish good diagnoses and make the right decisions. National and international collaborations that explore precision medicine are based on analyzes of data from diverse and numerous populations. Current biomedical analyzes are subject to problems in terms of data protection.

In this multidisciplinary thesis, we contribute to the resolution of these problems through the development of distributed biomedical data analyses as well as their deployment in IT infrastructures and their integration into real medical tools. As part of a kidney transplant application, we provide a new distributed contextualization solution that helps clinicians assess patients' kidney problems. For the same application, we also propose a new distributed version of Factor Analysis of Mixed Data (FAMD) for dimension reduction. In the context of *HLA* database, we propose a new distributed and secure model for frequency estimation applied to *HLA* databases. We also propose a new algorithm for the distributed estimation of *HLA* haplotype frequency using the EM algorithm.

All our contributions have been deployed on a real geo-distributed cloud infrastructure and evaluated on real or realistic data. Experiments have shown (very) good results in terms of performance compared to centralized solutions.