# Intelligence artificielle et apprentissage profond appliqués à la reconnaissance d'images en cytogénétique

Alexandra Petit

▶ **To cite this version:**

Alexandra Petit. Intelligence artificielle et apprentissage profond appliqués à la reconnaissance d'images en cytogénétique. Médecine humaine et pathologie. 2018. dumas-02100870

**HAL Id: dumas-02100870**

**https://dumas.ccsd.cnrs.fr/dumas-02100870**

Submitted on 16 Apr 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

*Unité de Formation et de Recherche de Médecine*
*d'Amiens*
*3, rue des Louvels*
*80036 Amiens Cedex 1*

**MÉMOIRE POUR LE DIPLÔME D'ÉTUDES SPÉCIALISÉES DE BIOLOGIE MÉDICALE**
**tenant lieu conformément au décret du 23 janvier 2003 de**
**THÈSE D'ÉTAT DE DOCTEUR EN MÉDECINE**
**N°2018-119**
**Mention Spécialité**

**Mardi 25 Septembre 2018**

**Madame Alexandra PETIT**

TITRE DE LA THÈSE :

# INTELLIGENCE ARTIFICIELLE ET APPRENTISSAGE PROFOND APPLIQUÉS À LA RECONNAISSANCE D'IMAGES EN CYTOGÉNÉTIQUE

Vu : les Membres de Jury

Le Président de Jury,

**Monsieur le Professeur Henri COPIN**

Les Juges,

**Monsieur le Professeur Gilles DUVERLIE**

**Madame le Docteur Christine TERRE**

**Monsieur Michel TERRE**

**Monsieur le Docteur Yann FERRET**

Le directeur de thèse,

**Monsieur le Professeur Loïc GARÇON**

# Remerciements

**À mon président de thèse,**

**Monsieur le Professeur Henri COPIN**
*Professeur des Universités – Praticien Hospitalier*
*(Histologie-Embryologie-Cytogénétique)*
*Chef du Service de Médecine et Biologie de la Reproduction et de Cytogénétique et CECOS de Picardie*
*Pôle "Femme-Couple-Enfant"*
*Chevalier dans l'Ordre des Palmes Académiques*

Pour m'avoir fait l'honneur d'accepter de présider cette thèse, et plus largement pour tout ce que vous avez fait pour moi depuis ces cours de première année. Si j'ai choisi la cytogénétique comme spécialité, c'est en grande partie grâce à vous. Soyez assuré de mon profond respect.

**À mes juges,**

**Monsieur le Professeur Gilles DUVERLIE**
*Professeur des Universités – Praticien Hospitalier*
*(Virologie-Biologie)*
*Chef du Laboratoire de Virologie*
*Pôle "Biologie, Pharmacie et Santé des Populations"*
*Doyen, Directeur de l'Unité de Formation et de Recherche de Pharmacie d'Amiens*
*Chevalier dans l'Ordre des Palmes Académiques*

Pour avoir accepté de juger ce travail, et pour votre soutien dans mes projets universitaires. Veuillez trouver ici l'expression de ma reconnaissance.

**Madame le Docteur Christine TERRÉ**
*Praticien Hospitalier*
*(Cytogénétique)*
*Laboratoire de Cytogénétique*
*Centre Hospitalier de Versailles – Hôpital André Mignot*

Sans votre généreuse ouverture des archives, ce travail serait probablement resté à l'état d'ébauche. Merci pour votre accueil chaleureux, pour votre aide continue, pour votre disponibilité et pour avoir accepté de faire partie de ce jury.

**Monsieur Michel TERRÉ**
*Professeur des Universités*
*Directeur de l'École d'Ingénieurs du Conservatoire National des Arts et Métiers*
*Département Ingénierie des Systèmes Mécaniques et Électrotechniques*

Pour vos conseils avisés, votre expertise, et votre engagement dans ce projet. La confiance que vous m'avez accordée est pour moi un honneur. Soyez assuré de toute ma gratitude.

**Monsieur le Docteur Yann FERRET**
*Assistant Hospitalier*
*Service de Médecine et Biologie de la Reproduction et de Cytogénétique et CECOS de Picardie*
*Pôle "Femme-Couple-Enfant"*

Pour y avoir cru, tout d'abord ; pour avoir rassemblé les membres de l'équipe et organisé nos discussions, pour avoir encadré mon travail et relu avec attention les multiples versions de cette thèse, jusqu'au dernier moment. Merci pour tout.

**A mon directeur de thèse**,

**Monsieur le Professeur Loïc GARÇON**
*Professeur des Universités – Praticien Hospitalier*
*(Hématologie-Transfusion)*
*Chef du Laboratoire d'Hématologie*
*Pôle "Biologie, Pharmacie et Santé des Populations"*

Pour m'avoir fait l'honneur d'accepter d'être mon directeur de thèse, et pour m'avoir soutenu dans ma spécialisation. Soyez assuré de ma sincère considération et de ma reconnaissance.

**A toutes les personnes qui ont participé à ce projet :**

Monsieur le Professeur Luc Fety, pour votre implication dans ce projet, pour vos conseils, votre attention et votre gentillesse. Travailler avec vous a été un privilège.

Les technicien(ne)s et biologistes du laboratoire de cytogénétique d'Amiens, pour leur soutien dans les premières phases de ce projet et pour tous ces bons moments passés ensemble.

Les technicien(ne)s et biologistes du laboratoire de cytogénétique de Versailles. La base de données qui a été exploitée pour ce travail n'existerait pas sans vous.

Les technicien(ne)s et biologistes du laboratoire de cytogénétique de Lille, pour m'avoir accueillie à bras ouverts, pour m'avoir permis d'utiliser vos archives, et pour tout ce que vous m'avez appris. Merci pour avoir fait de cet inter-CHU un des meilleurs semestres de mon internat.

**Je tiens également à remercier** :

Les technicien(ne)s et biologistes du laboratoire de biologie moléculaire d'Amiens, et en particulier Mme le Docteur Estelle Cadet, pour leur soutien et leur amitié.

Mes co-internes, merci de m'avoir accueillie parmi vous. Merci à Morgane, pour tout ce que tu as fait pour nous en général, et en particulier pour avoir facilité ma transition vers la biologie. Merci à toi et à Camille pour ces longues discussions passionnées au déjeuner. Merci à Samy, qui le premier m'a fait découvrir l'apprentissage profond. Sans ton enthousiasme, cette thèse n'aurait simplement jamais existé. Merci à Mathilde, pour ta gentillesse et ton optimisme. Merci à tous.

Elann et Nestor, pour leur aide pour la conception de Jera et Mannaz. Anaïs et Remy, pour leur relecture des résumés, *thank you* ! Tout le groupe du WriYe, pour leur soutien inconditionnel. Que l'inspiration soit toujours avec vous.

Anne-Lore, si tu n'es pas là pour ma soutenance, c'est pour la meilleure des raisons. Pour avoir toujours été la meilleure des amies, tout simplement.

Mes parents, pour m'avoir soutenu tout le long du chemin, et pour avoir cru en moi.

# Table des matières

**Abréviations....8**
**Introduction....9**
Histoire de la cytogénétique....9
Evolution des techniques....10
Cytogénétique constitutionnelle et caryotype conventionnel....13
Cytogénétique onco-hématologique....14
Place du caryotype conventionnel en cytogénétique onco-hématologique....15
Automatisation de la cytogénétique conventionnelle....17
Réalisation d'un caryotype conventionnel....17
Automatisation des techniques....18
Informatisation....19
Intelligence artificielle et apprentissage profond....21
Architecture des réseaux neuronaux....22
Utilisation d'un réseau neuronal en recherche....24
Objectif de cette étude....26
**Matériel et méthodes....27**
Base de données utilisée....27
Traitement des images....28
Réseaux neuronaux utilisés....29
Inception v3....29
MobileNet v1....29
Réseau neuronal "maison" : Mannaz....29
Tests préliminaires....30
Evolution de la précision en fonction du nombre d'itérations....30
Effet des hyper-paramètres sur les résultats....31
Tests effectués....31
Analyse statistique....33
**Résultats....34**
Comparaison des types d'images et reconnaissance de chromosomes séparés....34
Reconnaissance de sous-groupes de chromosomes....34
Reconnaissance des acrocentriques....34
Reconnaissance des groupes de la classification de Denver....35
Reconnaissance des groupes de la classification de Denver selon leur statut acrocentrique ou non-acrocentrique....35
Reconnaissance des chromosomes au sein de chaque groupe de la classification de Denver…36

Comparaison aux logiciels existants....37
Reconnaissance des chromosomes anormaux....37
Reconnaissance des chromosomes dérivés de la translocation t(9;22)(q34;q11)....37
Reconnaissance d'une délétion des bras longs du chromosome 5....39
**Discussion....40**
Comparaison des types d'images et reconnaissance des chromosomes séparés....40
Reconnaissance des chromosomes par la classification de Denver (sous-groupes)....41
Comparaison aux logiciels existants....42
Reconnaissance des anomalies de structure....42
Perspectives....43
**Conclusion....46**
**Réalisation des travaux....47**
**Bibliographie....48**

# Abréviations

ADN : acide désoxyribonucléique

BCR-ABL : *breakpoint cluster region – Abelson proto-oncogene 1 non-receptor tyrosine kinase*

CBF : *core-binding factor*

CGH : hybridation génomique comparative

CH : Centre Hospitalier

CNAM : Conservatoire National des Arts et Métiers

FAB (classification) : franco-américano-britannique

FISH : hybridation *in situ* en fluorescence

LAL : leucémie aiguë lymphoïde

LAM : leucémie aiguë myéloïde

LMC : leucémie myéloïde chronique

OMS : Organisation Mondiale de la Santé

PCR : réaction en chaîne par polymérase

RAM : *random access memory* (mémoire vive)

RT-MLPA : amplification multiplexe de sondes dépendant d'une ligation, par transcriptase inverse

SNP : polymorphisme d'un seul nucléotide

# Introduction

## Histoire de la cytogénétique

La cytogénétique est la spécialité scientifique et médicale qui étudie les chromosomes, porteurs du génome humain. Si l'on peut considérer qu'elle apparaît dès 1882, avec la découverte des chromosomes humains par le cytologiste autrichien Walther Flemming, avec l'invention du mot chromosome en 1888 par Heinrich W. Waldeyer, ou avec l'invention du mot cytogénétique en 1900 par Walter Sutton qui développa une théorie chromosomique de l'hérédité, la cytogénétique ne prit réellement son essor que plus de cinquante ans plus tard[1].

En effet, le développement continu des techniques d'analyse durant la première moitié du vingtième siècle permit une précision toujours plus importante dans l'analyse des chromosomes, donnant naissance au caryotype. Celui-ci est une représentation de l'ensemble des chromosomes d'une cellule, classés par paires et selon leur taille, à partir d'une prise de vue microscopique (figure 1). L'arrangement du caryotype est spécifique à l'espèce étudiée, et constitue l'outil de base de l'étude cytogénétique.

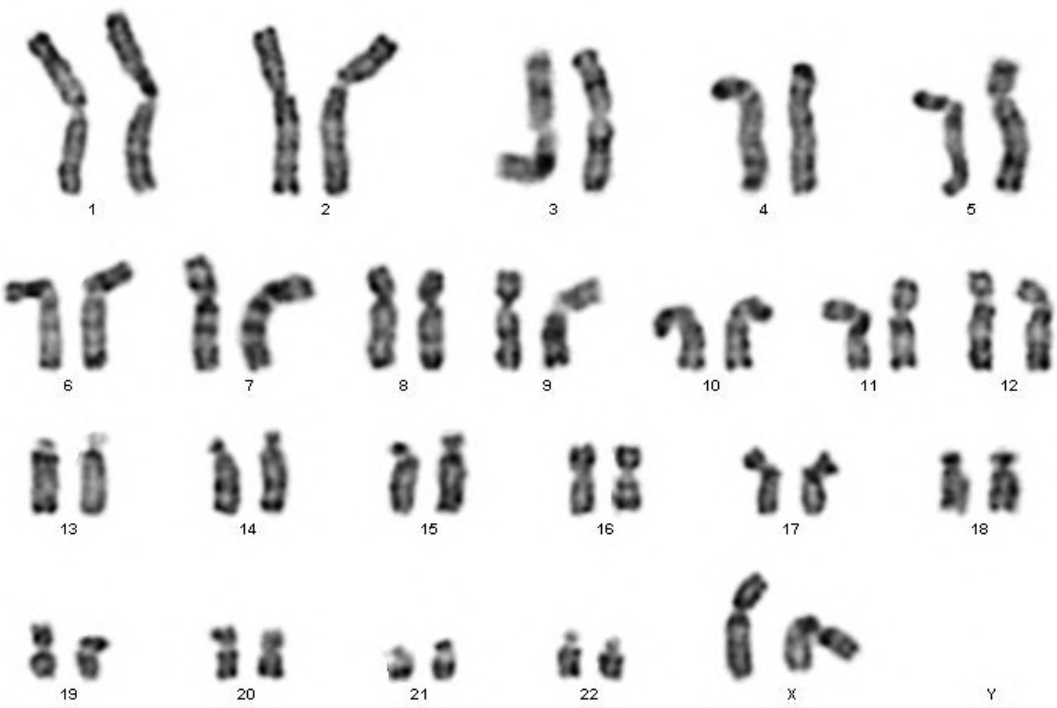


*Figure 1 : exemple de caryotype*

L'avènement de la culture cellulaire en 1907[2] puis la découverte de l'effet de la colchicine sur la division cellulaire en 1937[2] facilitèrent les premières études chromosomiques mais il fallut attendre 1953 et la découverte accidentelle du choc hypotonique par Tao-Chiuh Hsu[3] pour permettre de séparer suffisamment les chromosomes pour pouvoir les compter aisément au microscope. Cette technique a donné le signal de départ de la cytogénétique médicale en tant que telle en permettant trois ans plus tard la détermination du nombre correct de chromosomes humains[4]. Les découvertes par la suite s'enchaînèrent assez rapidement, avec tout d'abord la description en 1959 par Jérôme Lejeune et ses collègues d'un petit chromosome surnuméraire dans le syndrome de Down, anomalie maintenant plus connue sous le nom de trisomie 21[5]. La même année, Charles Ford *et al.* rapportèrent l'absence du chromosome Y chez les femmes atteintes de syndrome de Turner[6], Patricia Jacobs et John A. Strong indiquèrent la présence d'un chromosome surnuméraire du groupe du chromosome X chez les hommes atteints du syndrome de Klinefelter[7], et Jacobs décrivit également une patiente infertile porteuse de 47 chromosomes, considérée comme une possible trisomie X[8], sans certitude à l'époque. L'année suivante vit la description de deux autres syndromes majeurs de la cytogénétique prénatale, la trisomie 13 par Klaus Patau *et al*[9], et la trisomie 18 par John H. Edwards *et al*[10].

Ces observations eurent des répercussions profondes sur la connaissance du génome humain, élucidant notamment la présence du corps de Barr (chromosome X inactivé) dans les cellules, et la détermination du sexe chez l'humain (présence d'un chromosome Y ou non).
Les premières anomalies de structure furent décrites rapidement, avec par exemple le syndrome du cri du chat décrit par Jérôme Lejeune en 1963, lié à une délétion du bras court du chromosome 5[11], ou le syndrome de l'X fragile, décrit pour la première fois par Herbert Lubs en 1969[12]. Dans les années qui suivirent, d'autres anomalies de structure furent rapportées, notamment les instabilités liées au syndrome de Bloom[13] et de Fanconi[14].

# Evolution des techniques

A cette époque, il était difficile de distinguer les chromosomes autrement que par leur taille et la position de leur centromère, composant une silhouette générale commune à plusieurs chromosomes. La classification prévalant était celle de ***Denver*** (présentée dans le tableau 1) composée de groupes allant de A à G, et permettant les premières catégorisations de pathologies. Le groupe A correspondait aux grands chromosomes métacentriques (1 à 3), le groupe B aux grands sub-

métacentriques (4 et 5), le groupe C aux chromosomes sub-métacentriques de taille moyenne (6 à 12 et X), le groupe D aux chromosomes acrocentriques de taille moyenne (13 à 15), le groupe E aux plus petits des chromosomes de taille moyenne (le 16 étant métacentrique et le 17 et le 18 sub-métacentriques), le groupe F aux chromosomes métacentriques de petite taille (19 et 20), et le groupe G aux acrocentriques de petite taille (21, 22 et Y)[1].

*Tableau 1 : classification de Denver*

| Groupe | Chromosomes | Taille | Index centromérique |
|---|---|---|---|
| A | 1,2,3 | Grande | Métacentrique |
| B | 4,5 | Grande | Submétacentrique |
| C | 6,7,8,9,10,11,12,X | Moyenne | Submétacentrique |
| D | 13,14,15 | Moyenne | Acrocentrique |
| E | 16,17,18 | Assez petite | Submétacentrique (sauf 16 : métacentrique) |
| F | 19,20 | Petite | Métacentrique |
| G | 21,22,Y | Petite | Acrocentrique |

Une nouvelle ère s'ouvrit pour la cytogénétique en 1970 avec l'invention par Torbjörn Caspersson *et al.* du ***marquage des bandes chromosomiques***, ou banding, utilisant la coloration à la quinacrine[15]. La cytogénétique sortit alors du domaine restreint d'une poignée d'anomalies du développement. En effet, la méthode à la moutarde de quinacrine permettait de marquer les chromosomes, faisant apparaître de manière fluorescente intense les régions riches en bases A et T du génome, et créant ainsi des bandes spécifiques de chaque chromosome. Leur identification précise devint donc possible. La méthode à la quinacrine, difficile à mettre en place à l'époque car liée à la fluorescence, fut rapidement concurrencée par l'apparition des bandes dites G, correspondant à une digestion enzymatique des préparations suivie d'un marquage au Giemsa, facilitant l'application de ces nouvelles méthodes pour tous les laboratoires[16]. Les bandes G sont ainsi superposables aux bandes fluorescentes de la méthode à la quinacrine (bandes Q). Une dernière méthode, basée sur le traitement à la chaleur des lames, permit de faire apparaître des bandes dites R (pour *reverse*, car elles sont l'inverse des bandes G) qui ont l'avantage de rendre bien visibles l'extrémité des chromosomes et donc de faciliter la description d'anomalies touchant ces régions[17]. Le marquage en bandes R, découvert par une équipe française, restera par la suite le plus utilisé en France.

Dix ans plus tôt, Peter Nowell décrivait une méthode d'induction de la prolifération des lymphocytes T par l'utilisation de phytohémagglutinine, permettant d'obtenir suffisamment de lymphocytes pour une étude cytogénétique en 48 à 72h[18]. Cette technique transforma une méthode

lente et difficile en un test diagnostique relativement simple et rapide qui, associé au banding, rendait enfin possible l'identification d'anomalies chromosomiques en routine au laboratoire. Le marquage des bandes chromosomiques donna le coup d'envoi d'une décade riche en découverte d'anomalies cytogénétiques, qu'il s'agisse d'aneuploïdies plus rares, de microdélétions ou duplications, de modifications de structures telles que les inversions, les insertions, les translocations, etc.

L'identification de régions de plus en plus petites du caryotype permit une cartographie plus précise du génome et notamment de la position des gènes sur les chromosomes. Cette tendance à la recherche d'une précision toujours plus grande fut renforcée par l'apparition des sondes d'hybridation fluorescentes (hybridation in situ en fluorescence, ou FISH) dans les années 1980 qui donnèrent un essor à la ***cytogénétique moléculaire***[19]. En effet, cette nouvelle technique permettait, sans les inconvénients des sondes radioactives auparavant utilisées, la localisation précise des loci et la caractérisation d'anomalies chromosomiques auparavant trop petites pour être vues en caryotype conventionnel, élargissant ainsi à nouveau le champ d'étude de la cytogénétique. La cytogénétique moléculaire évolua de nouveau à la fin des années 1990 avec l'apparition de l'hybridation génomique comparative (CGH), puis de la CGH array (hybridation sur puce, voir figure 2)[20].

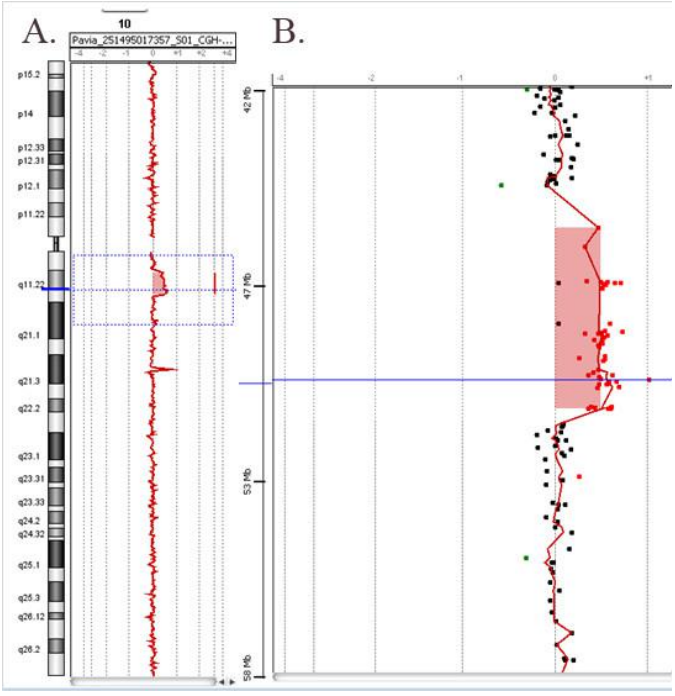


*Figure 2 : profil de CGH-array du chromosome 10 indiquant une duplication interstitielle*[21]

Celle-ci, par la comparaison entre l'ADN (acide désoxyribonucléique) du patient et celui d'un témoin, permit de décrire des insertions et des délétions avec une résolution 100 à 500 fois

supérieure à celle du caryotype conventionnel (anomalies de taille minimale comprise entre 5 et 10 kb), mais en perdant la possibilité de détecter les translocations équilibrées, car celles-ci ne provoquent ni perte ni gain de matériel chromosomique. La SNP array[22] (single nucleotide polymorphism : polymorphisme d'un seul nucléotide), technologie également basée sur les puces à ADN mais adaptée à la recherche des pertes d'hétérozygotie, fut développée dans les années suivantes, offrant notamment la possibilité de détecter les disomies uniparentales.

Parallèlement à l'évolution de la cytogénétique, une autre discipline de la génétique se développait : la ***biologie moléculaire***. A la suite de la découverte de la structure de l'ADN en 1953[23], les techniques de séquençage ont été développées avec notamment la méthode Sanger, qui est restée la méthode de référence des années 80 aux années 2000[24]. C'est ainsi cette méthode, couplée à la technique de la réaction en chaîne par polymérase (PCR) découverte en 1983[25], qui permit le séquençage complet du génome humain en 2001, étape historique majeure dans notre compréhension de la génétique humaine[26].
A partir de la fin des années 2000, la généralisation de l'utilisation du séquençage haut débit, permettant à la fois de réaliser le séquençage d'un grand nombre de gènes ciblés ou d'un génome entier en moins d'une semaine, mais également de caractériser les anomalies chromosomiques (délétions, duplications, inversions), permit d'envisager une précision de celles-ci à la base près, résultats jusqu'ici inaccessibles en routine de manière pangénomique[27]. L'étude fonctionnelle des chromosomes, et notamment des domaines d'activation de la transcription, permit également d'envisager le fonctionnement des chromosomes sous un nouvel angle, avec des interactions entre ADN codant et non-codant en trois dimensions, et de manière dynamique dans le temps[28]. On a donc assisté à un rapprochement des deux disciplines, qui rend possible d'affiner nos connaissances sur le génome, ainsi que de comprendre plus aisément la manière dont les différentes parties de celui-ci interagissent. A chaque évolution technique, la cytogénétique évolue ainsi en affinant son analyse du génome qui devient de plus en plus résolutive.

## *Cytogénétique constitutionnelle et caryotype conventionnel*

Les applications de la cytogénétique constitutionnelle sont maintenant très nombreuses et la plupart relèvent de la pratique de routine. Ainsi, les femmes enceintes peuvent facilement bénéficier d'une amniocentèse en cas d'anomalies fœtales vues à l'échographie, ou de risque augmenté pour une pathologie fœtale, notamment pour la trisomie 21. Les bilans de première intention des enfants

présentant des malformations multiples ou des difficultés notamment d'apprentissage, tout comme des couples infertiles, comportent de manière systématique une étude cytogénétique. Le rôle de la cytogénétique reste essentiel dans la prise en charge de nombreuses pathologies, et cela est d'autant plus vrai avec l'évolution constante des techniques. A l'heure actuelle, la CGH-array est ainsi l'examen le plus utilisé pour le bilan initial, le caryotype conventionnel étant surtout utilisé dans la détection des anomalies de structure équilibrées, et dans la documentation et le contrôle des anomalies détectées sur puce à ADN[29].

# Cytogénétique onco-hématologique

Parmi les nombreuses applications cliniques possibles, la cytogénétique fut très tôt employée dans l'étude des pathologies cancéreuses, essentiellement hématologiques puis plus tardivement également des tumeurs solides. Ainsi, la première description d'une anomalie cytogénétique dans une pathologie hématologique remonte au tout début de la cytogénétique humaine, en 1960. C'est avec la découverte du ***chromosome Philadelphie*** (figure 3) par David Hungerford et Peter Nowell[30] que se crée une nouvelle spécialité, la cytogénétique onco-hématologique.

*Figure 3 : paires de chromosomes 9 et 22 avec translocation t(9;22)(34;q11) visible (chromosomes dérivés de la translocation à droite) – le chromosome 22 dérivé est aussi connu sous le nom de chromosome Philadelphie*

Le cancer est une pathologie génétique acquise, qui peut se développer à partir d'une mutation prédisposante constitutionnelle avec mutations somatiques acquises par la suite, ou par l'accumulation de mutations somatiques. Il existe de très nombreuses altérations génétiques impliquées dans l'oncogenèse, et certaines intéressent plus particulièrement la cytogénétique, comme les aneuploïdies, les délétions ou les duplications de matériel chromosomique, les translocations ou les amplifications génétiques.

Si certaines tumeurs solides sont étudiées en cytogénétique, l'intérêt du caryotype dans ces pathologies est limité à quelques rares anomalies spécifiques qui sont souvent également détectables

par d'autres techniques, notamment de biologie moléculaire. Pour la plupart des cancers solides, le caryotype est très complexe et n'apporte pas d'informations sur le diagnostic ou le pronostic. De plus, il est techniquement difficile à réaliser et de ce fait très peu pratiqué en routine. En revanche, les pathologies hématologiques sont plus intéressantes du point de vue cytogénétique : environ 50% d'entre elles présentent des anomalies chromosomiques (proportion variant fortement en fonction du type de pathologie), et nombre d'entre elles sont spécifiques et orientent le diagnostic et/ou le pronostic, voire le traitement[31].

Ainsi, des translocations récurrentes dans certaines pathologies hématologiques ont permis d'améliorer la compréhension des mécanismes génétiques impliqués dans l'oncogenèse, et même de développer des traitements ciblés spécifiques. L'exemple le plus célèbre est, à la suite de la découverte du chromosome Philadelphie alors identifié en chromosome minute par David Hungerford et Peter Nowell, la description par Janet Rowley en 1973 de son origine cytogénétique, la translocation t(9;22)(q34;q11)[32]. Cette translocation est responsable de l'apparition de la protéine chimérique BCR-ABL1 (*breakpoint cluster region – Abelson proto-oncogene non-receptor tyrosine kinase 1*). Cette découverte permit le tout premier traitement par inhibiteurs de tyrosinase kinase dans les leucémies myéloïdes chroniques (LMC), transformant le pronostic de cette pathologie naguère très sombre[33].

Si la classification historique (dite franco-américano-britannique ou FAB) des néoplasies hématologiques, encore partiellement utilisée, est basée sur la cytologie et son analyse morphologique, dès 1995 l'Organisation Mondiale de la Santé (OMS) a développé une nouvelle classification incorporant les données de cytogénétique et de biologie moléculaire. L'importance de ces données n'a fait que croître depuis, ainsi la classification OMS 2016 met-elle particulièrement en valeur les anomalies cytogénétiques dans la définition même de nombreuses pathologies[34][35].

## *Place du caryotype conventionnel en cytogénétique onco-hématologique*

A l'heure actuelle, et ce parallèlement à l'importance toujours croissante de la biologie moléculaire dans le diagnostic et le suivi des pathologies hématologiques, notamment grâce à l'arrivée de panels de séquençage haut débit, le caryotype conventionnel occupe toujours une place prépondérante dans l'étude de ces pathologies. En effet, la classification OMS des pathologies hématologiques utilise explicitement celui-ci pour classifier les pathologies, se basant sur des considérations à la fois

diagnostiques et pronostiques. Le caryotype peut ainsi poser le diagnostic pour certaines pathologies, par exemple la présence d'une translocation t(8;21)(q22;q22) impliquant le gène *RUNX1* définit une des leucémies aiguës myéloïdes (LAM) dites *core-binding factor* (CBF) et ce même si le taux de blastes dans la moelle est inférieur à 20%, seuil correspondant à la définition cytologique de la LAM. Mais il est également indispensable à la stratification pronostique de la plupart des hémopathies malignes (LAM[36][37], leucémies aiguës lymphoïdes ou LAL[31], syndromes myélodysplasiques[38][39], syndromes myéloprolifératifs[31], lymphomes[31], etc.). Il permet encore d'orienter la thérapeutique dans de nombreux cas, ainsi la translocation t(15;17)(q22;q12) indique le traitement par acide tout-trans rétinoïque.

La particularité de la cytogénétique hématologique est de travailler sur des ***clones***, souvent présents en faible nombre au sein des cellules du patient. De plus, les anomalies étudiées sont fréquemment des ***translocations équilibrées*** avec des points de cassure qui peuvent être très variables. Le caryotype conventionnel est ici la méthode privilégiée car il s'agit encore actuellement de la seule méthode qui soit à la fois pangénomique et suffisamment sensible. La CGH et la SNP-array sont des méthodes plus récentes, mais elles sont moins performantes quant à la limite de détection car elles analysent l'ensemble des cellules en même temps, contrairement à la FISH ou au caryotype conventionnel qui analysent chaque cellule séparément : certains clones présents en faible quantité peuvent donc être invisibles par ces techniques. Elles sont donc privilégiées en cas d'envahissement important par les cellules tumorales, et notamment au diagnostic. Une autre limite de ces techniques est qu'elles ne détectent pas les translocations équilibrées.

D'autre part, des techniques de biologie moléculaire (RT-MLPA : amplification multiplexe de sondes dépendant d'une ligation, par transcriptase inverse) permettent maintenant la détection des translocations équilibrées avec une sensibilité supérieure à celle du caryotype par la recherche de transcrits de fusion[40]. Mais cette détection n'est pas exhaustive et elle ne peut pas s'appliquer aux translocations à points de cassure variables. De plus, pour les anomalies détectables à la fois en biologie moléculaire et en cytogénétique conventionnelle, cette dernière bénéficie en pratique d'une plus grande rapidité permettant des analyses en urgence.

Ces techniques sont donc une aide précieuse, mais elles ne peuvent pas remplacer le caryotype conventionnel dans la prise en charge des hémopathies malignes.

# Automatisation de la cytogénétique conventionnelle

## *Réalisation d'un caryotype conventionnel*

Comme nous l'avons vu, les innovations techniques ont été à l'origine d'évolutions majeures dans l'histoire de la cytogénétique, et dans notre compréhension des mécanismes qui sous-tendent l'organisation du génome. Les progrès technologiques et notamment informatiques ont eux aussi accompagné le développement du caryotype conventionnel. Pour mieux le comprendre, il est nécessaire de revenir sur les différentes étapes de la réalisation du caryotype.

On effectue un premier lieu un prélèvement sur le patient, le plus souvent de sang ou de moelle osseuse en cytogénétique hématologique, mais de nombreux autres tissus peuvent être analysés. Des cellules sont mises en culture *in vitro,* dans un milieu de culture favorisant la division du type cellulaire d'intérêt. Après un temps de culture variant selon la pathologie suspectée, on introduit de la colchicine dans le milieu de culture, ce qui va paralyser les fuseaux mitotiques et ainsi bloquer les cellules en métaphase de la mitose. On prélève alors les cellules et on les place dans un milieu hypotonique qui va les faire gonfler par différence de pression osmotique, fragilisant les membranes cellulaires. On étale ensuite la suspension cellulaire traitée par un fixateur sur une lame de verre. Cette phase d'étalement est critique car d'elle dépend en grande partie la résolution finale des images obtenues, le choc permettant l'éclatement des membranes et libérant les chromosomes qui sont alors répartis plus largement sur la lame. Une fois les lames étalées, on marque les bandes chromosomiques, la méthode variant selon les bandes désirées (le plus souvent bandes R ou G). La lame peut alors être étudiée au microscope.

On recherche d'abord sur la lame des métaphases qui peuvent être analysées, c'est-à-dire celles dont les chromosomes ont été suffisamment espacés les uns des autres par les étapes de choc et d'étalement pour être identifiables (figure 4). On prend également en compte la résolution de l'image en fonction du nombre de bandes que l'on peut distinguer sur chaque chromosome (figure 5)[29]. On choisit alors les métaphases que l'on désire analyser (jusqu'à 20 le plus souvent en cytogénétique hématologique), et pour chaque métaphase on sépare chaque chromosome des chromosomes avoisinant, notamment s'ils se chevauchent, une étape appelée ***segmentation***. Enfin, on classe les chromosomes selon leur taille, la position de leur centromère, et les bandes caractéristiques visibles.

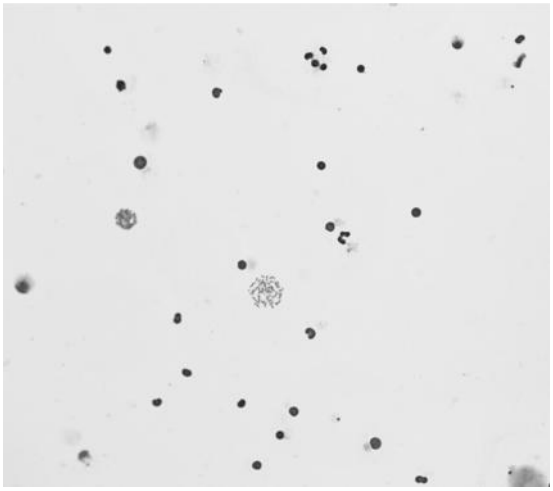

*Figure 4 : vue au microscope d'une lame (grossissement x10), une métaphase visible au centre*

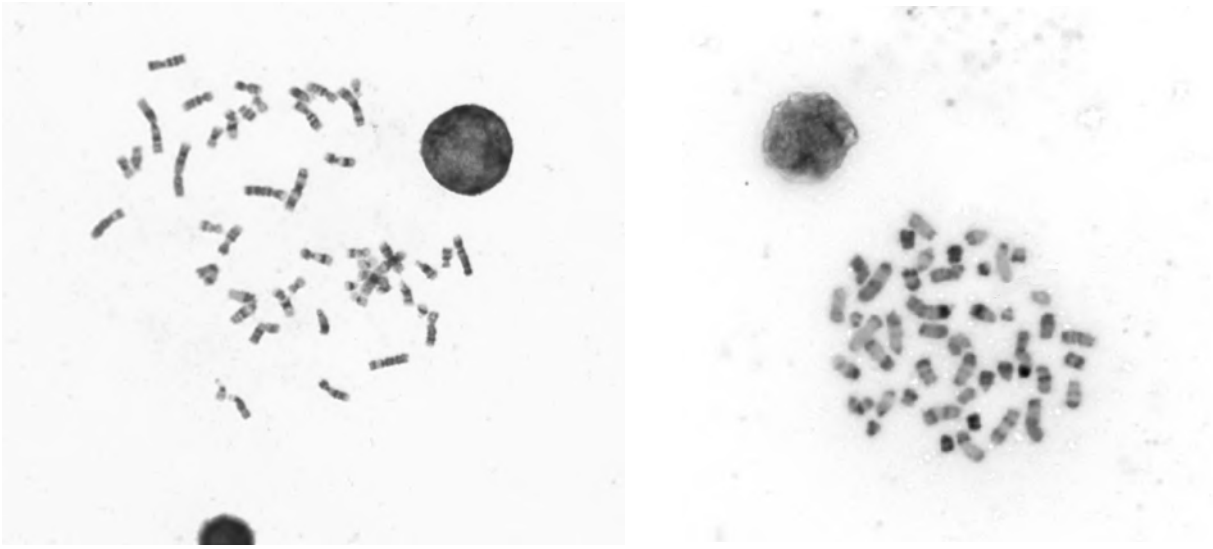

*Figure 5 : métaphase de bonne qualité (haute résolution) à gauche, de qualité médiocre (faible résolution) à droite*

## *Automatisation des techniques*

Il s'agit donc d'une technique encore très manuelle, même si certaines étapes peuvent être automatisées, comme les arrêts de culture ou l'étalement. Mais c'est surtout à partir de la phase de lecture que l'automatisation s'est développé au cours des dernières années. Ainsi si dans les années 60 les images étaient obtenues par photographie, les microscopes ont été rapidement raccordés à des caméras permettant d'enregistrer les images obtenues, puis à des ordinateurs pour traiter ces images. Récemment, les chercheurs automatiques de métaphases (figure 6) ont fait leur apparition : il s'agit de microscopes qui lisent en continu les lames préparées et en prennent des photos à faible

grossissement. Un logiciel intégré reconnaît les métaphases sur ces images et le microscope prend alors des photographies à fort grossissement de chaque métaphase. Cela permet d'éviter aux technicien(ne)s la recherche des métaphases sur la lame, étape qui peut être très chronophage, notamment quand la culture est pauvre. De plus, le chercheur automatique étudiant l'intégralité de la lame, il est plus facile de sélectionner dans l'ensemble des métaphases photographiées les plus informatives pour les analyser.

*Figure 6 : un chercheur automatique de métaphases*[41]

# *Informatisation*

De même, les logiciels de classification de chromosomes se sont rapidement répandus, qu'il s'agisse de logiciels propriétaires liés aux microscopes, comme par exemple Cytovision (Leica), ou de logiciels libres comme par exemple Lucky. Ces systèmes sont généralement constituées de quatre sous-parties : l'amélioration de la qualité des images (ou pré-traitement), la segmentation des chromosomes, l'extraction des caractéristiques structurelles, et enfin le classement à partir de ces caractéristiques[42]. Le ***pré-traitement*** correspond à l'élaboration d'algorithmes permettant d'augmenter le rapport signal/bruit des images, et également d'améliorer la qualité des images à faible contraste[43]. La ***segmentation*** repose sur d'autres algorithmes pour séparer les chromosomes isolés (souvent par seuillage dans les logiciels commerciaux), ceux qui se touchent ou se croisent[44], et faire la différence entre ces deux situations[45]. Les ***caractéristiques structurelles*** extraites pour permettre par la suite le classement sont le plus souvent la longueur des chromosomes, leur surface, leur périmètre, le ratio centromérique, le tracé des bandes en suivant l'axe médian[46].

La méthode utilisée pour ce classement varie d'un logiciel à un autre : par exemple, Lucky calcule des vecteurs pour chaque chromosome en fonction de la taille et de la position du centromère et classe ensuite ces vecteurs. Ce classement est de qualité variable selon le logiciel utilisé, et ne fait pas de différence entre chromosome de structure normale et anormale. Il permet néanmoins un gain de temps important et simplifie la tâche des technicien(ne)s qui peuvent se concentrer sur des tâches de classement plus complexes, tels que la reconnaissance d'anomalies.

En dehors des logiciels de caryotypage commercialisés qui relèvent donc plutôt d'une semi-automatisation, de nombreuses études informatiques ont été réalisées afin de tendre vers une automatisation complète. Des recherches ont ainsi été effectuées sur des moyens d'améliorer les résultats en facilitant l'analyse, par exemple pour redresser les chromosomes déviés[47] ou sélectionner les métaphases de meilleure qualité au sein d'un ensemble d'images[48]. Enfin, de très nombreuses méthodes, mathématiques, géométriques, statistiques ou relevant de l'intelligence artificielle (ces dernières seront vues plus en détail ci-dessous), ont été appliquées au problème du classement, souvent avec de très bons résultats (précision du classement entre 67 et 97%), mais souvent également sur uniquement certains chromosomes ou groupes de chromosomes, ou sur une partie uniquement du set d'images[42].

En parallèle de ces avancées, des ***bases de données*** ont été développées pour faciliter le développement et la comparaison des nouveaux algorithmes de classement : Copenhague, Édinbourg et Philadelphie en bandes G (images de qualité variable issues de sang périphérique ou de villosités choriales)[49], BioImLab en bandes Q (base dédiée à la segmentation avec de nombreux croisements de chromosomes, voir figure 7)[46] et MFISH en images multi-FISH[42]. Une seule base de données concerne la cytogénétique onco-hématologique avec des images extraites de dossiers de patients atteints de leucémies, la base de données Lisbon K1[50].

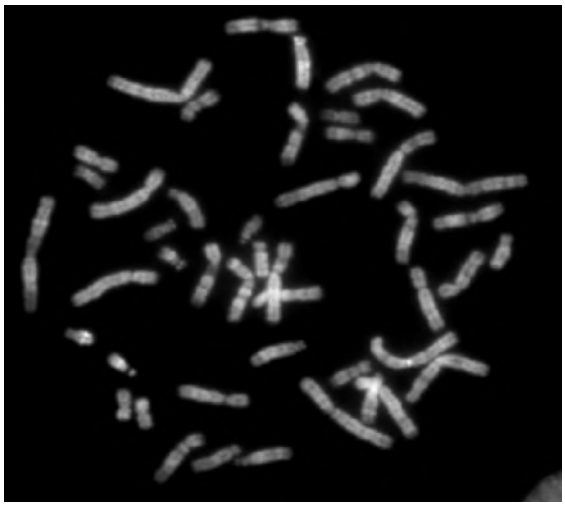

*Figure 7 : exemple de métaphase en bandes Q de la base de données BioImLab*[51]

# Intelligence artificielle et apprentissage profond

L'étude présentée ici est centrée sur l'utilisation d'une innovation technologique informatique encore peu connue dans le domaine médical et en particulier en cytogénétique : l'intelligence artificielle. En effet, depuis le début des années 2010, le développement d'un sous-type particulier d'intelligence artificielle cherchant à s'inspirer du fonctionnement du cerveau humain, les ***réseaux neuronaux***, a connu une expansion fulgurante qui a notamment permis l'amélioration de la reconnaissance vocale ou de texte écrit, la prédiction des habitudes des consommateurs, et même la conduite autonome d'une voiture. Dans le champ médical, cette technologie s'est traduite par la création de logiciels capables d'identifier des tumeurs cutanées malignes[52], ou de classer des fonds d'œil[53], avec une précision rivalisant avec celle d'un médecin expérimenté. Les réseaux neuronaux ont également été appliqués à la biologie moléculaire, avec notamment la prédiction de l'effet de variants[54] ou des interactions entre ADN et protéines[55].

Dans le domaine de la cytogénétique, les réseaux neuronaux ont été appliqués à plusieurs étapes du processus d'analyse du caryotype. En amont du classement en lui-même, un système a par exemple été développé pour trier les métaphases photographiées par un chercheur automatique, afin de ne sélectionner que celles considérées comme facilement caryotypables (bonne résolution, peu de chromosomes croisés, etc.)[56]. L'appariement des paires de chromosomes homologues a également été étudié, notamment en cytogénétique hématologique où la petite taille des chromosomes et la résolution inférieure du marquage chromosomique obtenues rend le classement en 24 catégories (une par chromosome) plus difficile[50]. La structure de réseau neuronal la plus utilisée pour le

classement des chromosomes est un système à deux étages reposant sur la classification de Denver. Un premier réseau effectue un tri des chromosomes parmi les 7 groupes de la classification de Denver, et 7 autres réseaux effectuent le tri des chromosomes dans ces sous-groupes[49]. Avec ce type de réseau, de très bons résultats ont été obtenus : sur un set d'images issues du laboratoire de génétique de l'université des sciences de la santé d'Oklahoma, en bandes G et sur des échantillons de sang périphérique de patients sains, l'équipe de Xingwei Wang obtient un taux de reconnaissance compris entre 67 et 98% selon les sous-groupes (le taux le plus faible correspondant au groupe C, qui comprend le plus grand nombre de chromosomes)[49]. Sur une base de données publiques (Copenhague, un set d'images de très bonne qualité sur bandes G, sur prélèvement de sang périphérique), Siamak Delshadpour décrit des résultats entre 69 et 88% de bonnes reconnaissances[57]. Sur une autre base de données publiques, BioImLab, en bandes Q, l'équipe d'Enea Poletti a atteint des résultats compris entre 92 et 95% avec un système complet comprenant le pré-traitement des images, la segmentation, l'extraction des caractéristiques chromosomiques, le classement par réseau neuronal à double étage et un algorithme simple supplémentaire de classement contraint. Celui-ci permet d'améliorer les résultats en forçant le classement à suivre les règles d'un caryotype humain sans particularités : deux chromosomes maximum par classe d'autosomes, deux XX ou un X et Y[51]. On peut également noter une étude récente sur un système basé sur des séparateurs à vaste marge (un type de technique d'apprentissage proche des réseaux neuronaux), développé pour identifier les chromosomes dicentriques, chromosomes qui à la suite d'une exposition radioactive présentent deux centromères au lieu d'un, qui obtient une reconnaissance correcte à 92% du premier centromère et à 70% du second[58].

## *Architecture des réseaux neuronaux*

Ces résultats sont liés en grande partie à la structure des réseaux neuronaux utilisés. Les réseaux neuronaux ont la capacité d'apprendre seuls, sans supervision humaine, et d'extraire d'une très grande quantité de données les valeurs qui leur sont nécessaires à remplir correctement leur tâche, souvent une tâche de classification ou de prédiction de résultat. Ils sont basés sur la structure de base du cerveau humain, le neurone, qui est représenté par une structure mathématique simple (une fonction, par exemple). On donne à ce neurone des informations en amont, il effectue une opération sur l'information donnée, et la répercute sur un autre neurone, qui à son tour effectue une autre opération sur l'information transmise (figure 8). Les neurones sont ainsi disposés en couches successives, et la connexion entre deux neurones est appelée une synapse, tout comme dans le

cerveau. On considère qu'à partir de trois couches de neurones, celle du milieu est une couche dite "cachée" entre la couche d'entrée de l'information et celle de sortie. Dans ce travail comme dans les études présentées plus haut, des réseaux neuronaux dits profonds ont été utilisés. Ceux-ci sont définis par le fait qu'ils présentent plus de trois couches de neurones, ou plus d'une couche cachée de neurones (figure 9). L'étude de ce type de réseau est donc le domaine de l'***apprentissage profond***[59].

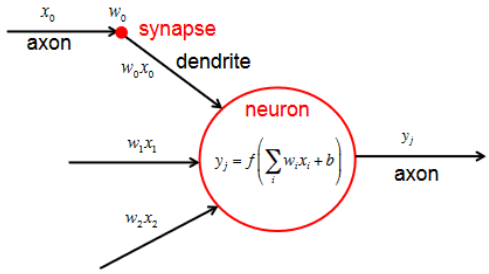


*Figure 8 : fonctionnement d'un neurone artificiel*[55]

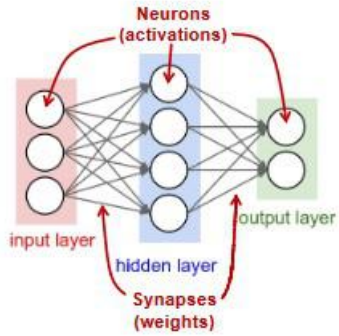


*Figure 9 : fonctionnement d'un réseau neuronal profond avec couche d'entrée, couche cachée et couche de sortie, et les synapses où s'appliquent les poids entre chaque*[55]

Une caractéristique clé du neurone artificiel est sa capacité à attribuer un poids à l'information qu'il reçoit, et donc à l'échelle du réseau de neurone, à déterminer quelles informations sont plus importantes que les autres pour aboutir au résultat souhaité. Ainsi, l'apprentissage d'un réseau neuronal est défini comme l'ajustement des poids en réponse à la stimulation initiale. Seuls les poids changent lorsque le réseau neuronal apprend une nouvelle tâche, l'organisation globale du réseau restant exactement la même. Il est donc possible d'utiliser un modèle qui a été entraîné sur un set de données particulier, comme par exemple une reconnaissance de chiffres manuscrits, et de lui donner comme nouvelle tâche la reconnaissance de lettres manuscrites. Le réseau ré-utilisera les bases qu'il a appris pour la première tâche (identification de traits droits, de courbes...) afin de parvenir beaucoup plus rapidement au résultat souhaité. Cette capacité particulière des réseaux

profonds, appelée ***ré-entraînement***, a permis au cours des cinq dernières années la mise à disposition du public de réseaux neuronaux pré-entraînés sur de très larges bases de données, que l'utilisateur peut ensuite adapter à ses propres besoins, sans avoir besoin de ressources informatiques importantes en matière de mémoire ou de puissance d'analyse du processeur. On peut citer comme exemple Inception, une série de réseaux neuronaux conçus par les équipes de Google dédiées à l'intelligence artificielle, et distribués librement et gratuitement[60]. Plus la base de données est importante et pertinente, meilleur sera le résultat de l'apprentissage[59][61].

Dans le cadre de la reconnaissance d'images, il est important que l'algorithme puisse reconnaître un objet quelle que soit sa position dans l'image. Pour cela, des réseaux dits convolutifs sont utilisés, dont le principe est de diviser l'image analysée en multiples images de très petite taille (souvent de 2*2 ou 3*3 pixels) qui se recoupent partiellement. Cela permet une simplification de l'image et la mise en valeur des caractéristiques les plus discriminantes comme les angles, les courbes, les variations de couleur, etc. A chaque couche convolutive, le réseau apprend ainsi des caractéristiques de plus en plus fines de l'image, améliorant la précision de la reconnaissance[59].

A l'issue de la phase d'apprentissage, les valeurs des poids et les opérations réalisées afin d'accomplir la tâche demandée sont fixes (on considère que le réseau a fini son apprentissage) et on obtient donc une simple fonction appelée algorithme qui exécutera la tâche que le réseau a apprise. Comme il s'agit d'une fonction mathématique, elle ne prend que très peu de place et de ressources, et peut ainsi être facilement déployée sur des architectures légères comme des téléphones portables[62].

## *Utilisation d'un réseau neuronal en recherche*

Le principe d'utilisation de ces réseaux est assez simple : on crée des catégories d'images que l'algorithme utilisé doit savoir reconnaître, par exemple différents types de lésions cutanées, différentes pathologies visibles sur un fond d'œil, ou dans ce cas différents chromosomes. Pour chaque catégorie, on sélectionne des images représentatives en grande quantité, de préférence les plus variées possible, afin de permettre à l'algorithme de généraliser les traits spécifiques de chaque catégorie. Le réseau sélectionne ensuite aléatoirement une partie de ces images pour en faire un ***set de test*** (ce sont des images "non vues" par l'algorithme, qui va les utiliser pour vérifier sa performance en fin d'apprentissage), un ***set d'entraînement***, qui constitue la plus grande partie des images (permettant l'apprentissage proprement dit), et à l'intérieur de ce dernier un ***set de validation*** (utilisé pour tester le réseau au fur et à mesure de son apprentissage mais sans incidence directe sur l'apprentissage : il s'agit d'un indicateur de performance). Pour chaque image du set

d’entraînement, l'algorithme propose une classification en fonction des données dont il dispose à ce stade – ainsi au tout début de l'apprentissage, l'algorithme choisit au hasard une classification. Selon que la réponse donnée est bonne ou non, le réseau fait varier les poids de sa structure pour favoriser les neurones qui ont donné les informations qui ont permis de faire le bon choix. On définit par ***itération*** le passage d’un certain nombre d’images (que l’on peut faire varier). A chaque itération, le réseau se teste donc et adapte son comportement pour obtenir de meilleurs résultats à l'image suivante. Des tests réguliers sur le set de validation permettent au chercheur d’analyser les performances du réseau en cours d’apprentissage[63].

Après un certain nombre d’itérations, les résultats sur le set de validation adoptent une progression logarithmique et ont donc tendance à atteindre un plateau. A ce stade, on peut alors tester le réseau sur le set de test, qui permet de vérifier que le réseau a bien généralisé les informations pour reconnaître les images, et est donc capable de reconnaître des images nouvelles. Chez l'humain, ce serait faire la différence entre un élève qui a compris une leçon et sait l'appliquer à un problème proche de ceux vus en classe mais non identique, et un élève qui a appris tous les cas étudiés en cours par cœur sans comprendre, et aura de mauvais résultats face à un problème inédit (ce qu’on appelle en intelligence artificielle du ***sur-apprentissage***). Pour chaque image, le réseau donne un score de classement de cette image dans toutes les catégories, avec un score d’autant plus haut que le réseau considère ce classement comme fiable. L’image est alors classée dans la catégorie obtenant le plus haut score.
Si les résultats sur les deux sets (c’est-à-dire la proportion d’images correctement classées) sont satisfaisants, on peut arrêter l'apprentissage car celui-ci ne progressera plus et utiliser l'algorithme directement[64].

# Objectif de cette étude

L'objectif de cette étude était d'évaluer le potentiel de l'utilisation de l'intelligence artificielle, et en particulier de l'apprentissage profond, dans la classification automatique de caryotype. En effet, si le sujet de l'automatisation du classement des chromosomes a été très étudié, l'utilisation de l'intelligence artificielle, et en particulier des réseaux neuronaux profonds, a fait l'objet de moins d'analyses. La cytogénétique hématologique a été particulièrement peu explorée sous cet angle, avec notamment une seule base de données publique[50]. Or cette spécialité est comme nous l'avons vu encore très liée au caryotype conventionnel et il nous paraissait d'autant plus intéressant d'explorer les nouvelles possibilités offertes par les progrès de l'intelligence artificielle. Dans ce but, nous avons préparé une base de données d'images nécessaire pour l'apprentissage. Nous avons ensuite développé et testé plusieurs réseaux neuronaux dédiés au classement automatique de chromosomes déjà individualisés et à la reconnaissance d'anomalies chromosomiques de structure simples.

# Matériel et méthodes

## Base de données utilisée

La base de données initiale à notre disposition était une banque d'images provenant du CH André Mignot de Versailles. Elle représentait l'ensemble des caryotypes analysés et validés de 2005 à 2016 dans le secteur de cytogénétique onco-hématologique du laboratoire de cytogénétique (Dr Christine Terré). La technique de marquage chromosomique utilisée dans ce laboratoire était celle des bandes R. Le choix d'utiliser uniquement des images de caryotypes de cytogénétique hématologique a été fait pour plusieurs raisons. Tout d'abord, les images de ces caryotypes sont de moins bonne résolution que ceux des caryotypes de cytogénétique constitutionnelle. En entraînant un réseau sur ces images, il était probable que la généralisation des caractéristiques des chromosomes serait suffisamment importante pour permettre à l'algorithme de reconnaître aussi bien des images de faible que de haute résolution. D'autre part, et comme on l'a vu en introduction, le caryotype conventionnel est encore la référence en matière de cytogénétique onco-hématologique, alors que son utilisation en cytogénétique constitutionnelle a diminué au cours des dernières années. Nous avons utilisé la base de données de Versailles et non celle d'Amiens car celle-ci est une archive Cytovision et nous n'avions pas accès à des images de chromosomes individuels. De plus, la base de données de Versailles présente une taille très importante et son organisation permet d'extraire facilement les dossiers d'images en fonction des anomalies recherchées.
L'extraction automatisée des images de chromosomes individuels a été réalisée par l'intermédiaire du logiciel Lucky par le Pr Luc Fety (Conservatoire National des Arts et Métiers - CNAM) qui en est le concepteur avec le Pr Michel Terré. En raison de la taille de la base de données, un set restreint de 983 dossiers ont été choisis pour l'extraction, correspondant à l'année 2006.

Un programme en langage informatique Python nommé ***Jera*** a ensuite été développé permettant de placer automatiquement les 169 000 images obtenues dans des dossiers correspondant à leur numéro de chromosome. Pour entraîner les réseaux à reconnaître les chromosomes normaux, un tri de ces dossiers a été effectué par le programme, pour ne retenir que les chromosomes provenant de métaphases sans anomalies clonales de structure décelées. Il n'a pas été effectué de tri sur la résolution des métaphases et celle-ci était donc variable d'une métaphase à l'autre. Dans un second temps, Jera a permis un tri des métaphases présentant des anomalies spécifiques afin d'étudier

celles-ci. Il a donc été construit deux bases de données de travail : l'une contenant uniquement des images de chromosomes sans anomalies de structure décelées (soit 76740 images réparties en 24 catégories équilibrées entre elles), et l'autre comportant en plus des images d'anomalies de structure spécifiques (dont la composition détaillée est précisée plus bas dans la partie *Tests préliminaires*).

# Traitement des images

Trois types de traitement d'images ont été réalisés via le programme Jera :

- Dans le premier cas (figure 10), les images ont été présentées à l'algorithme à l'état brut (c'est-à-dire sans modifications de l'image initiale), toutes les images étant donc de taille différente, mais comprises pour la plupart d'entre elles entre 20*20 pixels et 260*30 pixels.

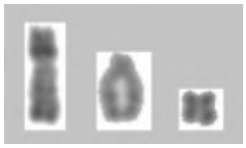

*Figure 10 : images brutes de chromosomes extraits de la base de données (6, 14 et 19)*

- Dans le deuxième cas (figure 11), les images ont été déformées toutes de la même façon, afin de créer un set comparable d'images de taille 128*128 pixels, permettant l'utilisation d'un réseau neuronal "maison" qui nécessitait une taille fixe en entrée.

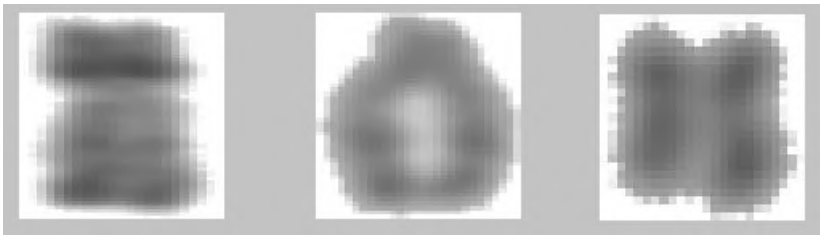

*Figure 11 : images déformées de chromosomes extraits de la base de données (6, 14 et 19)*

- Dans le troisième cas (figure 12), les images ont été réduites à la taille de 128*128 pixels mais sans les déformer : les images correspondant aux chromosomes les plus petits comportent donc beaucoup plus de pixels blancs que ceux des chromosomes les plus grands.

*Figure 12: images réduites de chromosomes extraits de la base de données (6, 14 et 19)*

# Réseaux neuronaux utilisés

## *Inception v3*

Le programme Inception, aussi connu sous le nom de GoogleLeNet, a été introduit en 2014. Il s'agit d'un réseau neuronal convolutionnel développé par Google, qui comprend 42 couches neuronales. Google a permis l'utilisation libre de cet algorithme et notamment son utilisation pour un ré-entraînement. Lors du ré-entraînement, nous avons présenté à l'algorithme les images spécifiques de la tâche demandée, et celui-ci les a classées en fonction de son entraînement initial, basé sur la reconnaissance de 1000 classes d'images du quotidien (ImageNet)[65].

## *MobileNet v1*

MobileNet est une version plus légère du réseau Inception, adaptée au traitement d'images plus petites, plus rapide mais moins précise[65].

## *Réseau neuronal "maison" : Mannaz*

Pour les besoins de cette thèse, un petit réseau neuronal appelé Mannaz a été développé en langage informatique Python. L'architecture a été adaptée sur les conseils des Pr. Luc Fety et Michel Terré, suivant leur expertise de la reconnaissance d'images appliqués à la cytogénétique. Ce réseau est constitué de 4 couches neuronales. Les trois premières couches sont dites convolutionnelles, et permettent le traitement des images en entrée, en filtrant ces images par sections de 3*3 pixels, auxquelles un poids est attribué. La dernière couche est constituée de 128 neurones et c'est à ce niveau que les poids sont recalculés en fonction des résultats sur le set d'entraînement. Ce réseau

n'est pas pré-entraîné et est donc entraîné intégralement sur les images de chromosomes à chaque test.

# Tests préliminaires

Pour permettre la comparabilité entre des réseaux fonctionnant selon des organisations différentes, des tests ont été effectués portant sur la précision maximale, l'évolution de la précision avec le nombre d'itérations effectuées par l'algorithme, et l'utilisation des hyper-paramètres de chaque réseau. Ces termes sont définis ci-dessous. A cette étape, des tests (présentés dans la partie *Résultats*) ont également permis de déterminer le choix du type d'images sur lesquelles s'est effectué l'apprentissage des réseaux neuronaux utilisés dans ce travail pour les différentes tâches étudiées : les images réduites.

## *Evolution de la précision en fonction du nombre d'itérations*

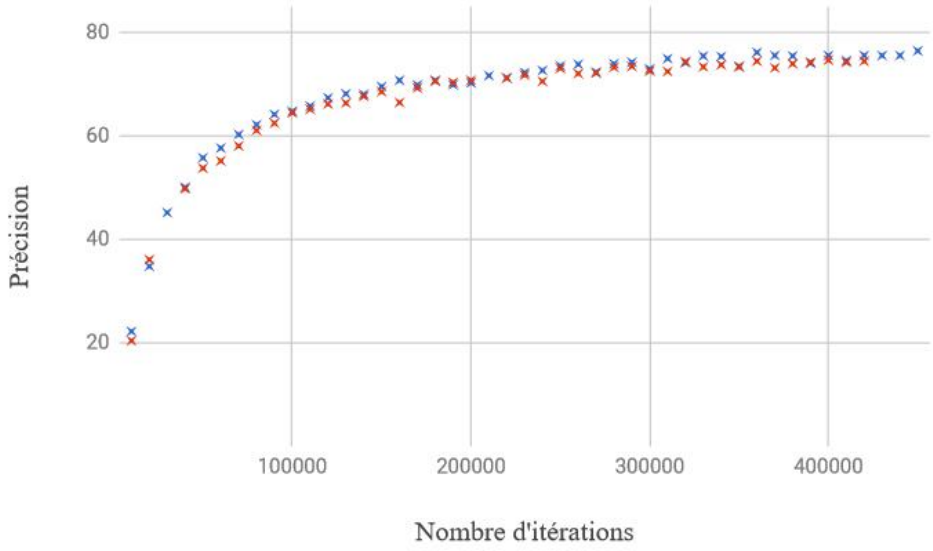


*Figure 13 : évolution de la précision des résultats en fonction du nombre d'itérations (algorithme Mannaz)*

La figure 13 présente un exemple de résultats de deux essais de Mannaz (avec les mêmes données et paramètres). La précision à chaque itération évolue selon une courbe d'allure logarithmique et atteint un plateau (ici à partir d'environ 200 000 itérations) au-delà duquel toute augmentation du

nombre d'itérations ne permet plus d'amélioration significative de la précision. Cette répartition des résultats est la même pour tous les réseaux neuronaux. Les valeurs de précision présentées dans la suite de cette thèse sont toujours des précisions dites maximales, correspondantes au pourcentage de classement exact obtenu sur le set de test lorsque le plateau des résultats sur le set d'apprentissage est atteint. Les graphes d'évolution des résultats ne sont donc pas inclus dans cette thèse. Pour chaque test réalisé, la vérification de la cohérence des résultats entre les sets d'apprentissage et de test a permis de confirmer l'absence de sur-apprentissage.

## *Effet des hyper-paramètres sur les résultats*

Les hyper-paramètres sont des paramètres modificateurs concernant la structure ou le fonctionnement d'un réseau neuronal. Les deux hyper-paramètres étudiés dans cette thèse sont le nombre d'images analysées à chaque itération et le taux d'apprentissage, qui définit la magnitude des changements apportés aux poids à chaque étape de l'apprentissage. Des variations de ces hyper-paramètres ont été testées de façon empirique pour chaque réseau : nombre d'images analysées par itération de 1, 2, 4, 8, 16 et 32 pour Mannaz et de 10, 25, 50 et 100 pour Inception et MobileNet, taux d'apprentissage de 0,0001 / 0,001 / 0,01 / 0,1 et 1 pour les trois réseaux. Ces modifications n'apportant pas d'amélioration, tous les résultats présentés dans cette thèse sont issus des paramètres par défaut de chaque réseau (indiqués dans le tableau 2).

*Tableau 2 : valeurs des hyper-paramètres par défaut pour Inception, MobileNet et Mannaz*

| | Images par itération | Taux d'apprentissage |
|---|---|---|
| Inception | 10 | 0,01 |
| MobileNet | 10 | 0,01 |
| Mannaz | 1 | 0,0001 |

# Tests effectués

Plusieurs tests ont été effectués dans le cadre de cette étude, utilisant les paramètres et le type d'images déterminés lors des tests préliminaires. Pour tous ceux et sauf mention contraire, trois réseaux ont été utilisés : Inception v3, MobileNet v1, et Mannaz. Chaque test a fait l'objet d'un apprentissage dédié. La précision est ici définie comme le pourcentage de chromosomes correctement reconnus sur le set de test.

Les tests réalisés sont :

1) Une comparaison de la précision des réseaux selon le type de traitement d'images.

2) Une comparaison de la précision des réseaux pour la reconnaissance de chromosomes normaux, en classant les chromosomes directement en 24 classes.

3) Une comparaison de la précision pour la reconnaissance des chromosomes normaux en regroupant ceux-ci selon la classification de Denver, c'est-à-dire en utilisant plusieurs réseaux enchaînés pour séparer en sous-groupes puis différencier les chromosomes au sein des sous-groupes. Ceci a permis de déterminer si une méthode orientée par l'expérience humaine, reposant sur plusieurs réseaux au lieu d'un seul, obtenait une meilleure précision.

4) Une comparaison de la précision obtenue avec ceux des logiciels existants. Pour permettre d'évaluer les résultats des algorithmes plus précisément par rapport aux outils déjà existants et disponibles, nous avons étudié deux logiciels utilisés couramment en cytogénétique : Lucky, logiciel libre développé par les Pr Luc Fety et Michel Terré, et Cytovision, logiciel propriétaire du groupe Leica. L'étude des performances de Lucky a été effectuée de manière automatisée, sur une sélection aléatoire de caryotypes représentant plus de 20 000 images issues de la base de données utilisée dans ce travail. L'étude des performances de Cytovision a été réalisée sur une sélection aléatoire de caryotypes représentant 4600 images issues des archives du laboratoire de cytogénétique de Lille (images de résolution similaire, en bandes G). Le nombre total d'images analysées était plus faible pour Cytovision que pour Lucky en raison des contraintes liées à l'utilisation du logiciel (classement manuel et non automatisé).

5) Une comparaison de la précision sur des chromosomes comportant des anomalies de structure : chromosomes dérivés d'une translocation t(9;22)(q34;q11) et chromosomes 5 porteurs d'une délétion des bras longs. Ce choix a été motivé par le fait que ces anomalies sont très fréquemment retrouvées en cytogénétique onco-hématologique, et qu'une quantité importante d'images est cruciale pour un apprentissage de qualité. Un travail statistique a été effectué sur les résultats (formules chromosomiques) des dossiers présents dans la base de données. La translocation t(9;22)(q34;q11) était retrouvée de façon clonale dans 10,45% des dossiers analysés, permettant l'obtention de 329 images de dérivés 9 et 300 images de dérivés 22. La délétion des bras longs du chromosome 5 était présente de façon clonale dans 3,77% des

dossiers analysés, permettant l'obtention de 111 images de chromosomes 5 porteurs de délétion. Les images des chromosomes anormaux ont été extraites grâce au programme Jera. Ces images ont été ajoutées au set d'images utilisé pour le test de reconnaissance des chromosomes séparés pour former une base de données spécifique. L'objectif du test était de privilégier la sensibilité tout en maintenant une spécificité supérieure à 50%. Quand ces analyses nécessitaient des tests en sous-groupes, les sets d'images spécifiques ont été préparés en pondérant par rapport au nombre de chromosomes dans chaque groupe.

# Analyse statistique

Les tests de significativité statistiques ont été effectués par méthode des comparaisons de proportions (loi normale centrée réduite) sur tableur LibreOffice Calc. Le seuil de significativité choisi était $p<0,05$.

# Résultats

Pour tous les résultats présentés dans cette partie, seules les valeurs de précision sur le set de test (images jamais présentées à l'algorithme) sont données.

## Comparaison des types d'images et reconnaissance de chromosomes séparés

Le tableau 3 compare les précisions obtenues pour les différents traitements d'images, avec les trois algorithmes utilisés. La tâche servant à la comparaison est la classification des chromosomes en 24 classes (de 1 à 22, ainsi que X et Y).

*Tableau 3 : précisions pour chaque algorithme en fonction du traitement d'image utilisé*

| | Images brutes | Images déformées | Images réduites | Significativité |
|---|---|---|---|---|
| Inception | 49,3% | 48,3% | 48,1% | p<0,05 |
| MobileNet | 61,6% | 60,5% | 59,9% | p<0,01 |
| Mannaz | N/A | 67,0% | 76,5% | p<0,001 |
| Significativité | p<0,001 | p<0,001 | p<0,001 | |

Le set d'images de test utilisé par les trois réseaux pour le calcul de précision comportait 15356 images. Inception et MobileNet présentent une plus grande précision pour les images brutes ($p<0,05$), alors que Mannaz obtient de meilleurs résultats avec les images réduites, avec une significativité plus importante (76,5% avec $p<0,001$). Il n'a pas été possible de tester Mannaz sur les images brutes car le réseau n'accepte pas les images de taille variable. Par ailleurs, Mannaz obtient des précisions plus élevées sur cette tâche d'identification des 24 chromosomes que les deux autres réseaux avec une précision de 76,5% sur les images réduites, et 67,0% sur les images déformées ($p<0,001$). Les images réduites ont été utilisées pour tous les tests ci-dessous.

## Reconnaissance de sous-groupes de chromosomes

### *Reconnaissance des acrocentriques*

Un test a constitué en la division du stock d'images réduites de chromosomes en deux groupes : chromosomes acrocentriques et non-acrocentriques. Les résultats sont présentés dans le tableau 4.

L'algorithme Mannaz distingue correctement les chromosomes acrocentriques des non-acrocentriques à 92,6%, de manière significativement meilleure que les deux autres réseaux (p<0,001).

*Tableau 4 : précision des trois algorithmes sur la reconnaissance des chromosomes acrocentriques et non-acrocentriques*

| | Précision |
|---|---|
| Inception | 87,7% |
| MobileNet | 88,5% |
| Mannaz | 93,1% |
| Significativité | p<0,001 |

## *Reconnaissance des groupes de la classification de Denver*

Les différents réseaux on été entraînés sur la tâche de classification des chromosomes dans les différents groupes correspondant à la classification de Denver. Les résultats sont présentés dans le tableau 5. L'algorithme Mannaz atteint une précision pour les groupes A, B, C, D, E, F et G de 77,6%, résultat significativement supérieur à ceux des deux autres réseaux (p<0,001).

*Tableau 5 : précision des trois algorithmes sur la reconnaissance des groupes A, B, C, D, E, F et G*

| | Précision |
|---|---|
| Inception | 66,1% |
| MobileNet | 71,3% |
| Mannaz | 77,6% |
| Significativité | p<0,001 |

## *Reconnaissance des groupes de la classification de Denver selon leur statut acrocentrique ou non-acrocentrique*

Les différents réseaux ont été entraînés sur la tâche de classification en sous-groupes à partir des groupes acrocentriques et non-acrocentriques, c'est-à-dire classer les acrocentriques dans les groupes D ou G, et les non-acrocentriques dans les groupes A, B, C, E ou F.

Le tableau 6 regroupe les résultats du test de différenciation entre les groupes D et G. Le tableau 7 regroupe les résultats du test de différenciation entre les groupes A, B, C, E et F. L'algorithme Mannaz distingue correctement les chromosomes appartenant au groupe D et ceux appartenant au

groupe G parmi les acrocentriques à 95,7% (p<0,001), et classe correctement les chromosomes appartenant aux groupes A, B, C, E et F à 85,7% (p<0,001).

*Tableau 6 : précision pour les trois algorithmes pour la reconnaissance des groupes D et G*

| | Précision |
|---|---|
| Inception | 93,5% |
| MobileNet | 94,4% |
| Mannaz | 95,7% |
| Significativité | p<0,001 |

*Tableau 7 : précision pour les trois algorithmes pour la reconnaissance des groupes A, B, C, E et F*

| | Précision |
|---|---|
| Inception | 63,7% |
| MobileNet | 72,1% |
| Mannaz | 85,7% |
| Significativité | p<0,001 |

## *Reconnaissance des chromosomes au sein de chaque groupe de la classification de Denver*

Pour cette dernière étape l'étude a uniquement été réalisée avec l'algorithme Mannaz en raison de ses résultats significativement supérieurs pour tous les tests précédents. Ce réseau a donc été entraîné pour la reconnaissance des chromosomes parmi les sous-groupes, c'est-à-dire la classification correcte des différents chromosomes de chaque sous-groupe (1,2 et 3 pour le groupe A, par exemple). Le tableau 8 présente les résultats obtenus pour les différents sous-groupes. La précision varie de 83,8 % pour le sous-groupe le moins bien reconnu (groupe C) à 94,3 % pour le sous-groupe le mieux reconnu (groupe B).

*Tableau 8 : précision de l'algorithme Mannaz pour la reconnaissance en sous-groupes*

| | Précision |
|---|---|
| Groupe A | 91,7% |
| Groupe B | 94,3% |
| Groupe C | 83,8% |
| Groupe D | 90,5% |
| Groupe E | 90,2% |
| Groupe F | 91,2% |
| Groupe G | 87,8% |

Le tableau 9 présente les résultats comparés de la méthode par identification directe des chromosomes et de la méthode par sous-groupes. La précision de la méthode par sous-groupes a été

calculée en tenant compte des résultats sur le test de distinction entre acrocentriques et non-acrocentriques, de ceux sur la reconnaissance des groupes en fonction de cette distinction, et de ceux sur la reconnaissance des chromosomes au sein de ces groupes. La méthode par identification séparée des chromosomes donne une précision supérieure avec p<0,001.

*Tableau 9 : précision comparée des méthodes séparée et groupée pour l'algorithme Mannaz*

| | Précision |
|---|---|
| Méthode séparée | 76,5% |
| Méthode groupée | 72,5% |
| Significativité | p<0,001 |

# Comparaison aux logiciels existants

Le tableau 10 présente les résultats des précisions comparées de Mannaz et des logiciels Cytovision et Lucky. Les résultats de Mannaz ne sont pas significativement différents de ceux de Lucky (p>0,05).

*Tableau 10 : précisions comparées de Mannaz, Cytovision et Lucky*

| | Précision | Significativité |
|---|---|---|
| Mannaz | 76,5% | N/A |
| Cytovision | 67,2% | p<0,001 |
| Lucky | 76,9% | p>0,05 |

# Reconnaissance des chromosomes anormaux

## *Reconnaissance des chromosomes dérivés de la translocation t(9;22)(q34;q11)*

Le tableau 11 présente les résultats sur une tâche de classification des 24 chromosomes normaux et du dérivé 9 de cette translocation (soit 25 catégories). Le set de test comprenait 12316 images.

*Tableau 11 : précisions des trois algorithmes pour la reconnaissance des 24 chromosomes et du dérivé 9 de translocation t(9;22)(q34;q11)*

| | Précision |
|---|---|
| Inception | 48,5% |
| MobileNet | 58,0% |
| Mannaz | 69,3% |
| Significativité | p<0,001 |

Mannaz obtient une précision pour ce test de 69,3%, significativement supérieure aux résultats des deux autres réseaux (p<0,001). Il est constaté en analysant les données que plus de la moitié des dérivés 9 de translocation t(9;22)(q34;q11) sont identifiés comme des 9 (29 sur 55) et que 24% sont identifiés comme des 10 (13 sur 55). Peu de chromosomes sont identifiés à tort comme des dérivés 9 (3 chromosomes 9 et 1 chromosome 22 : « faux positifs ») mais seuls 3 dérivés 9 sont correctement identifiés, soit 57% de faux positifs et une précision sur les dérivés 9 de 5% (3 sur 55). Le réseau est alors entraîné à différencier les chromosomes 9, 10 et les dérivés 9 dans une analyse en sous-groupe avec un set de test de 196 images (réparties également entre chaque classe). L'algorithme Mannaz obtient une précision de 79,1% pour l'ensemble de l'analyse et de 64% pour les dérivés 9 de translocation t(9;22). Il est finalement obtenu un score de reconnaissance des dérivés 9 en deux étapes de 51%. Le taux de faux positifs est diminué à 36%.

Le tableau 12 présente les résultats sur une tâche de classification des 24 chromosomes normaux et du dérivé 22 de cette translocation (soit 25 catégories). Le set de test comprenait 12298 images.

*Tableau 12 : précisions des trois algorithmes pour la reconnaissance des 24 chromosomes et du dérivé 22 de translocation t(9;22)(q34;q11)*

| | Précision |
|---|---|
| Inception | 48,3% |
| MobileNet | 59,2% |
| Mannaz | 73,0% |
| Significativité | p<0,001 |

Mannaz obtient une précision pour ce test de 73,0%, significativement supérieure aux résultats des deux autres réseaux (p<0,001). Il est constaté en analysant les données que si 29% des dérivés 22 de translocation t(9;22)(q34;q11) sont identifiés correctement (12/45), 44% (20 sur 45) sont identifiés comme des chromosomes 21 normaux et 20% comme des 22 (9 sur 45). L'algorithme classe 4 chromosomes 21, 3 chromosomes 22, et un chromosome Y en dérivés 22, soit un taux de faux positifs à 40%.

Le réseau est alors entraîné à différencier les chromosomes 21, 22 et les dérivés 22 dans une analyse en sous-groupe avec un set de test de 177 images (réparties également entre chaque classe). L'algorithme Mannaz obtient une précision de 67,2% pour l'ensemble de l'analyse et de 60% pour les dérivés 22 de translocation t(9;22). Il est finalement obtenu un score de reconnaissance des dérivés 22 en deux étapes de 54%, avec 40% de faux positifs.

## *Reconnaissance d'une délétion des bras longs du chromosome 5*

Le tableau 13 présente les résultats sur une tâche de classification des 24 chromosomes normaux et des chromosomes 5 anormaux (soit 25 catégories). Le set de test comprenait 12268 images.

*Tableau 13 : précisions des trois algorithmes pour la reconnaissance des 24 chromosomes et de la délétion des bras longs du chromosome 5*

| | Précision |
|---|---|
| Inception | 48,2% |
| MobileNet | 59,6% |
| Mannaz | 74,0% |
| Significativité | p<0,001 |

Mannaz obtient une précision pour ce test de 74,0%, significativement supérieure aux résultats des deux autres réseaux (p<0,001). Cependant aucun des réseaux n'arrive à identifier correctement les chromosomes 5 porteurs de délétion, tous les exemples du set de test (n=18) étant classés dans des catégories différentes, et ce quel que soit le réseau utilisé.

# Discussion

L'objectif de la présente thèse était de développer et d'évaluer le potentiel de différents outils d'apprentissage profond utilisables en cytogénétique pour la reconnaissance des chromosomes. Ce travail est à notre connaissance le premier concernant le classement de chromosomes sur une base de données de cytogénétique hématologique en bandes R. C'est également à notre connaissance le seul s'attachant à évaluer la capacité d'un réseau neuronal profond à reconnaître des anomalies de structure (dérivés de translocation, délétions). Dans le but d'établir des points de comparaison concrets, trois algorithmes différents ont été testés. Les analyses systématiques effectuées sur les trois systèmes retrouvent les données connues de la littérature, qu'il s'agisse de la progression logarithmique des résultats par convergence[59] ou de l'amélioration de ceux-ci par l'augmentation de la taille du set de données d'entraînement[61].

## Comparaison des types d'images et reconnaissance des chromosomes séparés

La comparaison des différents traitements d'images montre de bien meilleurs résultats pour les images réduites sans être déformées (troisième technique) pour l'algorithme Mannaz. En effet, cette méthode permet de garder une information de taille relative sur les chromosomes et de garder la silhouette générale du chromosome intacte. Cela est particulièrement vrai pour les plus petits chromosomes, qui ont tendance à être très déformés par l'agrandissement nécessaire pour obtenir une image en 128*128 avec la deuxième technique. Il n'a pas été possible de tester les images brutes sur le réseau Mannaz car celui-ci ne tolère pas d'images de taille variable. Il est cependant intéressant de noter que les deux réseaux issus d'un pré-entraînement (Inception et MobileNet) obtiennent des résultats assez comparables avec les trois techniques, avec de meilleurs résultats obtenus pour les images brutes.

Il n'a pas été effectué de pré-traitement des images qui sont donc celles issues directement de la capture microscopique, car l'objectif de ce travail était centré sur l'étape de classement des chromosomes. Ceci constitue néanmoins une limite de notre étude et pourrait expliquer partiellement des résultats dans la moyenne basse de ceux retrouvés dans la littérature.

Les performances comparées des trois réseaux sont assez surprenantes au premier abord. En effet, on s'attendrait à obtenir les meilleurs résultats avec Inception, qui est le réseau le plus complexe de tous et est déjà pré-entraîné, et les pires avec Mannaz, qui est un petit réseau et n'est pas pré-entraîné. En réalité, ces résultats semblent refléter une surpuissance d'Inception et de son réseau dérivé MobileNet, qui ont été entraînés sur une base de données de 1000 catégories d'objets du quotidien. Ces images sont variées, en couleur, et globalement comportent plus de différences entre elles (par exemple entre une photo de chien et celle d'un chat) qu'entre deux images de chromosomes. De fait, le pré-entraînement semble agir comme un obstacle qui empêche Inception et MobileNet de s'adapter totalement à la tâche demandée. Une limite de cette étude est qu'il n'a pas été possible de tester un ré-entraînement complet d'Inception ou de MobileNet sur la base de données dont nous disposions car cela nécessitait un matériel beaucoup plus puissant que celui disponible – par exemple, pour entraîner Inception v3, il a fallu deux semaines à l'équipe de recherche de Google sur un ordinateur équipé de 8 processeurs graphiques bénéficiant chacun de 12 Go de mémoire vive (*random access memory*, ou RAM). On ne peut pas non plus exclure qu'avec une base de données plus importante, les performances d'Inception et de MobileNet soient considérablement améliorés, bien que ces réseaux soient capables de très bonnes performances sur des sets d'images de taille comparable[65]. En revanche, Mannaz est peu complexe mais son entraînement exclusif sur des images de chromosome le rend plus adapté à la tâche demandée et pourrait expliquer sa performance supérieure ici.

# Reconnaissance des chromosomes par la classification de Denver (sous-groupes)

La méthode des sous-groupes, principalement utilisée dans la littérature, montre de très bons résultats sur la différenciation des chromosomes acrocentriques et non-acrocentriques quel que soit le réseau utilisé, mais des différences plus marquées pour la différenciation en groupes A, B, C, D, E, F et G et pour les analyses en sous-groupes. La raison pour laquelle ce système est privilégié dans les études déjà publiées est que celles-ci datent des débuts de l'application des réseaux neuronaux en cytogénétique, et qu'un gros réseau avec 24 classes était trop lourd à gérer avec les moyens informatiques de l'époque, les réseaux utilisés ne dépassant pas trois couches neuronales[45]. On constate ici que les résultats en sous-groupes sont inférieurs au résultat de la classification globale. Il est possible que cette différence soit liée aux changements d'architecture des réseaux, permettant une plus forte puissance pour les gros réseaux à l'heure actuelle. Il est

également possible que les images de cytogénétique onco-hématologiques, de plus mauvaise qualité que celles des bases de données de cytogénétique constitutionnelle, ne permettent pas un classement en sous-groupes satisfaisants, car les chromosomes présentent des différences morphologiques moins marquées.

# Comparaison aux logiciels existants

Le choix des logiciels utilisés pour la comparaison aux systèmes existants a été fait en fonction de la disponibilité des logiciels et de leur utilisation large au sein des laboratoires de cytogénétique. L'absence de différence entre les résultats de Lucky et de Mannaz permet de prouver que la méthode par apprentissage profond fonctionne et obtient des résultats comparables à ceux des logiciels utilisés tous les jours dans les laboratoires. Il s'agit donc d'une preuve du potentiel de l'apprentissage profond appliqué à la reconnaissance des chromosomes en cytogénétique. Plusieurs pistes sont possibles pour améliorer encore ces résultats : l'augmentation du nombre d'images dans la base de données et l'utilisation d'ordinateurs plus puissants (RAM en particulier) afin de pouvoir traiter toutes ces images. La complexification de Mannaz, notamment par l'ajout de couches de neurones supplémentaires, pourrait de manière empirique également améliorer les résultats.

# Reconnaissance des anomalies de structure

En ce qui concerne la reconnaissance des anomalies de structure, les résultats sont encourageants puisque Mannaz est capable de reconnaître les dérivés de translocation t(9;22)(q34;q11), une des translocations les plus fréquentes en cytogénétique hématologique, avec une sensibilité de 51% et une spécificité de 64% pour les dérivés du chromosome 9, et une sensibilité de 54% et une spécificité de 60% pour les dérivés du chromosome 22. En revanche, l'algorithme s'est révélé tout à fait incapable d'identifier les chromosomes 5 anormaux, porteurs d'une délétion des bras longs. Si ces résultats montrent qu'il est possible d'automatiser la recherche d'anomalies de structure, ils montrent également que celle-ci est loin d'être optimale pour l'instant. La reconnaissance d'anomalies rares pourrait apporter une vraie aide au diagnostic, mais cela reste encore un défi pour l'apprentissage. Une voie pour améliorer ces résultats est l'exploitation d'une base de données plus large. En effet, il est probable que l'échec d'identification des délétions des bras longs du chromosome 5 soit lié au très faible nombre d'images de cette délétion dans notre set de données,

trois fois moins que d'images des dérivés de translocation t(9;22)(q34;q11). On peut de plus noter que les délétions du chromosome 5 n'ont pas toujours les mêmes points de cassure, complexifiant encore la tâche de reconnaissance.

# Perspectives

Dans le prolongement de ce travail, et sous l'initiative du Pr. Michel Terré, une équipe d'étudiants des Ponts et Chaussées (composée de M. Théo Giraudon, M. Mouad Marrakchi Benazzouz, M. Youval Vanlaer, M. Thomas Chabal et M. Elhoussein El Beghdadi) a travaillé sur le même sujet, avec la même base de données mais en utilisant des approches variées au sein du groupe. Une de ces approches utilise un réseau profond de structure assez voisine de Mannaz, et des modifications des images (par rotation, pivotements, décalages de l'image, etc.). Les images sont par ailleurs réduites à une taille de 30*30 pixels. Le groupe d'étudiants obtient cependant des résultats assez similaires puisqu'ils aboutissent à une précision de 75% pour la reconnaissance des chromosomes de manière séparée, soit à peine inférieure à celle obtenue par Mannaz. Les deux travaux, menés de manière indépendante, renforcent donc l'intérêt de l'étude de cette nouvelle technologie pour la reconnaissance des chromosomes.

Les réseaux neuronaux profonds semblent donc être capables d'identifier et de classer correctement les chromosomes, dans les limites de cette thèse. Plusieurs voies de recherche sont possibles pour améliorer la précision obtenue. Il est tout d'abord raisonnable d'envisager une amélioration de la précision de ces réseaux avec l'augmentation de taille de la base de données, telle que le permettrait l'étude complète de celle du laboratoire de cytogénétique du CH André Mignot de Versailles. Cette base pourrait par la suite être enrichie des images d'autres laboratoires, dans la limite des possibilités offertes par les logiciels propriétaires pour la collecte de données.

Nous pouvons également aborder la question sous un autre angle, qui est celui de l'apprentissage profond non-supervisé, une autre méthode de construction d'algorithme dans laquelle l'algorithme construit lui-même des catégories en fonction des points communs entre les sets de données au lieu d'apprendre des catégories qu'on lui impose. Il est donc possible que ce type d'algorithme puisse établir des groupes de chromosomes plus adaptés pour la reconnaissance d'images que ceux de la classification de Denver, notamment dans le cadre spécifique de la cytogénétique onco-hématologique.

Il est également à noter que nous ne disposons à l'heure actuelle que de peu de données concernant les performances attendues de ce type de réseau, et qu'une comparaison large avec de nombreux réseaux différents permettrait une meilleure estimation statistique des résultats obtenus.

Les réseaux neuronaux semblent capables de reconnaître correctement les chromosomes porteurs d'anomalies de structure, avec une base de données comprenant relativement peu d'images. Mais la méthode de reconnaissance utilisée a une limite assez évidente : le réseau ne reconnaît que ce qu'il connaît, et est donc incapable d'identifier un chromosome anormal autres que ceux présentés dans son set d'apprentissage. De plus, la sensibilité globale diminue à chaque catégorie de classement supplémentaire. Plusieurs méthodes pourraient être testées pour contourner cet obstacle.
Ainsi, un test de l'appariement des chromosomes, déjà étudié dans la littérature et notamment en cytogénétique hématologique[50], pourrait être une voie d'étude intéressante. En effet, le contrôle de l'appariement est une étape importante du classement manuel, permettant de vérifier une éventuelle inversion entre chromosomes visuellement proches (comme les 11 et les 12 par exemple) et de suspecter une anomalie si les deux chromosomes d'une paire sont différents. L'appariement facilite ainsi à la fois le classement et la recherche d'anomalies. Une utilisation possible de ce type de test pourrait être le couplage au classement des chromosomes, afin de corriger celui-ci si nécessaire et donc d'augmenter encore sa précision, sur le modèle des couples algorithme de reconnaissance/algorithme de correction déjà présentés dans la littérature[51].
Une autre possibilité pourrait être d'utiliser les scores de classement de l'algorithme, qui correspondent à la fiabilité estimée du classement. Les chromosomes obtenant un faible score, pour lesquels le classement de l'algorithme est donc douteux, pourraient être regroupés dans une catégorie spécifique qui serait alors recontrôlée par l'opérateur humain, permettant une mise en valeur des chromosomes suspects d'être porteurs d'anomalies de structure.

Dans une optique plus large, il serait intéressant d'intégrer ces réseaux à un système complet de reconnaissance d'images, incluant les étapes de pré-traitement des images, de segmentation, voire d'extraction des caractéristiques structurelles des chromosomes. Les étapes de pré-traitement et d'extraction des caractéristiques pourraient ainsi améliorer la précision du classement. Mais le gain de temps pour les opérateurs qui pourrait résulter de l'automatisation complète du classement est pour l'instant limité par la nécessité d'une intervention manuelle systématique au niveau de l'étape cruciale de segmentation. Celle-ci reste un défi pour l'automatisation et limite les possibilités d'utilisation de systèmes intelligents de classement.

Des algorithmes issus de l'apprentissage profond sont à l'heure actuelle utilisés dans l'audiovisuel pour sélectionner des parties spécifiques d'une image ou d'une vidéo (par exemple un visage). Il est ainsi envisageable d'utiliser également cette technique pour repérer les différents chromosomes au sein d'une image de métaphase et effectuer ainsi une segmentation automatisée.

A plus long terme, il est possible d'imaginer également de coupler un tel système à un chercheur automatique de métaphases. L'algorithme de classement pourrait alors analyser un très grand nombre de métaphases et les classer non seulement par qualité d'images, mais aussi en fonction de la présence ou non d'anomalies détectées. Ce gain de sensibilité s'avérerait d'autant plus intéressant en cytogénétique onco-hématologique qu'il pourrait mettre en évidence des clones de faible fréquence, et cela permettrait également de réduire le taux d'échec.

# Conclusion

Au terme de cette thèse, l'intérêt de l'utilisation de l'intelligence artificielle dans le cadre de la reconnaissance des chromosomes, et en particulier de l'apprentissage profond, a été prouvé. En élargissant les domaines d'application de l'apprentissage profond, il est ainsi possible d'imaginer de nombreuses voies de développement pour cette discipline dans un avenir très proche. Il s'agit en effet d'une méthode assez facile d'accès, ne nécessitant pas forcément de lourds moyens informatiques, très versatile et adaptable à des tâches nombreuses et variées. L'accès libre aux algorithmes, ainsi que la grande qualité de la documentation et des guides disponibles permet à une personne n'ayant aucune connaissance initiale en intelligence artificielle, et uniquement une connaissance de base en programmation, de développer l'outil adapté à ses besoins spécifiques.

De très nombreuses possibilités d'application de l'apprentissage profond s'ouvrent en cytogénétique, de l'amélioration du classement automatique à l'appariement des chromosomes, en passant par la reconnaissance des chromosomes anormaux. La même méthode pourrait tout aussi bien être appliquée au traitement des images de FISH, permettant une détection automatisée des anomalies de nombre, des fusions et des séparations de signaux. Il est même envisageable de poursuivre l'apprentissage des algorithmes dans chaque laboratoire, permettant une adaptation poussée aux conditions et techniques particulières de ceux-ci.

L'intelligence artificielle telle qu'elle est développée actuellement pourrait rapidement devenir une aide précieuse, permettant de gagner du temps et de se concentrer sur l'essentiel : la détection et l'interprétation des anomalies et la précision du diagnostic. Elle serait donc un auxiliaire, sans se substituer pour autant à l'expérience et à l'intelligence humaine. Loin de ne reposer que sur l'intelligence artificielle, le futur de la cytogénétique pourrait donc bien se jouer à trois partenaires : technicien(ne)s, biologistes, et algorithmes.

# Réalisation des travaux

Ce travail n'aurait pas été possible sans la participation des Pr Luc Fety et Michel Terré et des Dr Christine Terré et Yann Ferret.

Le Dr Christine Terré a permis l'accès à la base de données utilisée ainsi qu'à son fichier d'annotations, a réalisé l'étude de la performance de classement de Lucky, et participé à l'organisation des différentes phases du projet.

Le Pr Luc Fety a extrait les images de la base de données, réalisé l'étude de la performance de classement de Lucky, et participé au choix des paramètres architecturaux de Mannaz.

Le Pr Michel Terré a participé à l'organisation des différentes phases du projet ainsi qu'au choix des paramètres architecturaux de Mannaz, soumis le sujet à un groupe d'étudiants des Ponts et Chaussées et encadré leur travail.

Le Dr Yann Ferret a mis les différents intervenants en contact, participé à l'organisation des différentes phases du projet et encadré ce travail.

# Bibliographie

1. Gersen S.L, Keagle M.B. (2005) Principles of Clinical Cytogenetics 2nd edition. *Humana Press*, 611 pages.

2. Alberts B., Johnson A., Lewis J. et al. (2002) Molecular Biology of the Cell. 4th edition. New York: *Garland Science*, 1616 pages.

3. Hsu T.C., & Pomerat, C.M. (1953) Mammalian chromosomes in vitro II: A method for spreading the chromosomes of cells in tissue culture. *Journal of Heredity* 1953, 44”’, 23-29

4. Tjio, J.H. and Levan, A. (1956) The chromosome number of man. *Hereditas* 1956, 42: 1-6.

5. Lejeune J., Gautier M., Turpin R. (1959) Étude des chromosomes somatiques de neuf enfants mongoliens. *CR Acad Sci* 1959;248: 1721–1722.

6. Ford C.E., Jones K.W., Polani P.E., De Almeida J.C., Briggs J.H. (1959) A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* Apr 4;1(7075):711-3.

7. Jacobs PA, Strong JA. (1959) A case of human intersexuality having possible XXY sex determining mechanism. *Nature* 2:164-167.

8. Jacobs P.A., Baikie A.G., Court Brown W.M., MacGregor T.N., Maclean N., Harnden D.G. (1959) Evidence for the existence of the human ‘super female'. *Lancet* 274:423–425.

9. Patau K., Smith D.W., Therman E., Inhorn S.L., Wagner H.P. (1960) Multiple congenital anomaly caused by an extra autosome. *Lancet* 1, 790.

10. Edwards J.H., Harnden D.G., Cameron A.H., Crosse V.M., Wolff O.H. (1960) A new trisomic syndrome. *Lancet* 1:787-90.

11. Lejeune J., Lafourcade J., Berger R., Vialatte J., Boeswillwald M., Seringe P., Turpin R. (1963) Trois cas de délétion partielle du bras court d'un chromosome 5. *CR Acad Sci* (D) 257:3098-3102.

12. Lubs H.A. (1969) A marker X chromosome. *Am. J. Hum. Genet.* 21:231–244.

13. German J., Archibald R., Bloom D. (1965) Chromosomal Breakage in a Rare and Probably Genetically Determined Syndrome of Man. *Science* Vol. 148, Issue 3669, pp. 506-507

14. Schroeder T.M., Anschultz F., Knoff A. (1964) Spontane chromosomenaberrationen bei familiärer Panmyelopathie. *Humangenetik*. 194–196;1

15. Caspersson T., Zech L. and Johansson J. (1970) Differential banding of alkylating fluorochromes in human chromosomes. *Exp. Cell Res.* 60,315-319

16. Seabright M. (1971) A rapid banding technique for human chromosomes. *Lancet* 2:971-2.

17. Dutrillaux B., Lejeune J. (1961) Sur une nouvelle technique d'analyse du caryotype humain. *CR. Acad. Sci.* 272: 2638-2640

18. Nowell P.C. (1960) Phytohemagglutinin : an initiator of mitosis in cultures of normal human leucocytes. *Cancer Res.* 20. 462-466.
19. Langer-Safer P.R., Levine M., Ward D.C. (1982) Immunological method for mapping genes on *Drosophila polytene* chromosomes. *PNAS* 79 (14) 4381-4385
20. Pinkel D., Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo W-L, Chen C, Zhai Y. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics* 20:207–211.
21. Manolakos E., Vetro A., Kefalas K., Rapti S-M, Louizou E. Garas A.et al. (2010). The use of array-CGH in a cohort of Greek children with developmental delay. *Molecular cytogenetics*. 3. 22. 10.1186/1755-8166-3-22.
22. Wang D.G., Fan J.B., Siao C.J., Berno A., Young P., Sapolsky R., Ghandour G., Perkins N., Winchester E., Spencer J., et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* vol. 280 (pg. 1077-1082)
23. Watson J.D. & Crick F.H.C. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* volume 171, pages 737–738
24. Sanger F., Nicklen S., Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA.* 74 (12): 5463–77
25. Mullis K.F., Faloona F., Scharf S., Saiki R., Horn G., Erlich H. (1986) Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposium in Quantitative Biology*, 51:263–273.
26. Venter J.C., Adams M.D., et al. (2001) The sequence of the human genome. *Science* 291 (5507): 1304–51.
27. Goodwin S., McPherson J.D., McCombie W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*. 17 (6): 333–51.
28. Pombo A., Dillon N. (2015) Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology* 16 (4): 245–57.
29. ACLF, GFCH, GFCO. (2014) Guide de Bonnes Pratiques en Cytogénétique version 3
30. Nowell P.C., Hungerford D.A. (1960) A minute chromosome in human chronic granulocytic leukemia. *Science* 132,1497
31. GFCH. (2016) Place de la cytogénétique dans la prise en charge des hémopathies malignes : actualisation par le Groupe francophone de cytogénétique hématologique. *Anal Biol Clin* 74 (5) : 509-595
32. Rowley J.D. (1973) A new consistent chromosomal abnormality in chronic myelogenous leukemia identified by quinacrine fluorescence and Giemsa staining. *Nature* 243,290-293

33. Wapner J. (2014) The Philadelphia Chromosome: A Genetic Mystery, a Lethal Cancer, and the Improbable Invention of a Lifesaving Treatment. *The Experiment*
34. Arber D.A., Orazi A., Hasserjian R., Thiele J., Borowitz M.J., Le Beau M.M., Bloomfield C.D., Cazzola M., Vardiman J.W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood, 127(20), 2391-2405.
35. Swerdlow S.H., Campo E., Pileri S.A., Harris N.L., Stein H., Siebert R., Advani R., Ghielmini M., Salles G.A., Zelenetz A.D., Jaffe E.S. (2016) The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood, 127(20), 2375-2390.
36. Döhner H., Estey E., Grimwade D., Amadori S., Appelbaum F.R., Büchner T., Dombret H., Ebert B.L., Fenaux P., Larson R.A., Levine R.L., Lo-Coco F., Naoe T., Niederwieser D., Ossenkoppele G.J., Sanz M., Sierra J., Tallman M.S., Tien H., Wei A.H., Löwenberg B., Bloomfield C.D. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, 129(4), 424-447.
37. Grimwade D., Hills R.K., Moorman A.V., Walker H., Chatters S., Goldstone A.H., Wheatley K., Harrison C.J., Burnett A.K. (2010). Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. *Blood*, 116(3), 354-365.
38. Schanz J., Tüchler H., Solé F., Mallo M., Luno E., Cervera J., et al. (2012) New Comprehensive cytogenetic scoring system for primary myelodysplastic syndromes (MDS) and oligoblastic acute myeloid leukemia after MDS derived from an international database merge. *J Clin Oncol* 30:820-9.
39. Greenberg P.L., Tuechler H., Schanz J., Sanz G., Garcia-Manero G., Solé F., et al. (2012) Revised international prognostic scoring system for myelodysplastic syndromes. *Blood* 120 : 2454-65
40. Ruminy P., Marchand V., et al. (2016) Multiplexed targeted sequencing of recurrent fusion genes in acute leukaemia. *Leukemia* volume 30, pages 757–760
41. Schunck C., Johannes T., Varga D., Lörch T., Plesch A. (2004) New developments in automated cytogenetic imaging: unattended scoring of dicentric chromosomes, micronuclei, single cell gel electrophoresis, and fluorescence signals. *Cytogenet Genome Res* 104:383–389
42. Munot M.V. (2018) Development of computerized systems for automated chromosome analysis: current status and future prospects. *IJARCS* volume 9, n°1, pp 782-791
43. Wang Y-P., Wu Q., Castleman K.R., Xiong Z. (2001) Image Enhancement Using Multiscale Differential Operators. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol.3, pp 1853-1856

44. Munot M., Mukhopadhyay J., Joshi M. (2013) A Novel Approach for Efficient Extrication of Overlapping Chromosomes in Automated Karyotyping. *Med Biol Eng Comput* 51(12):1325-38.
45. Shunren, X., Weidong X., Yutang S. (2003) Two Intelligent Algorithms Applied To Automatic Chromosome Incision. *IEEE Conference on ASSP*, pp 697-700
46. Grisan E., Poletti E., Ruggeri A. (2009) Automatic segmentation and disentangling of chromosomes in Q-band prometaphase images. *IEEE Transactions on Information Technology in Biomedicine* 13 575–581.
47. Khan S., Dsouza A., Sanches J., Ventura R. (2012) Geometric Correction of Deformed Chromosomes for Automatic Karyotyping. *34th Annual International Conference of the IEEE EMBS,* pp 4438-4441
48. Arora T., Dhir R. (2015) A review of metaphase chromosome image selection techniques for automatic karyotype generation. *Med Biol Eng Comput* 54(8):1147-1157
49. Wang X., Zheng B., Li S., Mulvihill J.J., Wood M.C., Liu H. (2009) Automated classification of metaphase chromosomes: Optimization of an adaptive computerized scheme. *Journal of Biomedical Informatics* 42, pp 22–31
50. Khmelinskii A, Ventura R., Sanches J. (2010) A Novel Metric for Bone Marrow Cells Chromosome Pairing. *IEEE Transactions on Biomedical Engineering.* Volume 57, n°6, pp 1420-1429
51. Poletti E., Grisan E., Ruggeri A. (2011) A modular framework for the automatic classification of chromosomes in Q-band images. *Elsevier journal on computer methods and programs in biomedicine*, pp 1-11
52. Esteva A., Kuprel B., Novoa R.A., Ko J., Swetter S.M., Blau H.M., Thrun S. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* volume 542, pages 115–118
53. Rahimy, E. (2018) Deep learning applications in ophthalmology. *Current Opinion in Ophthalmology* Volume 29 Issue 3 p. 254–260
54. Zhou J., Troyanskaya O.G. (2015) Predicting effects of non-coding variants with deep learning based sequence model. *Nature methods* vol 12 n°10, pp. 931-934
55. Zeng H., Edwards M.D., Liu G., Gifford, D.K. (2016) Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* vol 32 n°12, pp. i121-i127
56. Wang X., Li S., Liu H., Wood M., Chen W.R., Zheng B. (2008) Automated identification of analyzable metaphase chromosomes depicted on microscopic digital images. *Journal of Biomedical Informatics*. Volume 41, issue 2, pages 264-271

57. Delshdapour S. (2003) Reduced size multi layer perceptron neural network for human chromosome classification, pp 2249–2252, in: *The 25th Silver Anniversary International Conference of the IEEE Engineering in Medicine and Biology Society, Cancun (Mx)*
58. LI Y. , Knoll J.H., Wilkins R.C., Flegal F.N., Rogan, P.K. (2016) Automated discrimination of dicentric and monocentric chromosomes by machine learning-based image processing. *Microsc. Res. Tech*., 79: 393-402.
59. Sze V., Chen, Y-H., Yang T-J., Emer J. (2017) Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *ArXiv:1703.09039v2*
60. Szegedy, C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich, A. (2014) Going Deeper with Convolutions. *ArXiv:1409.4842*
61. Joulin A., Van der Maaten L., Jabri A., Vasilache N. (2015) Learning visual features from large weakly supervised data. *ArXiv:1511.02251*
62. Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861*
63. Ripley, B.D. (1996) Pattern Recognition and Neural Networks Cambridge: *Cambridge University Press*, p. 354
64. Prechelt L. (2012) Early Stopping — But When? In: Montavon G., Orr G.B., Müller K.R. (eds) Neural Networks: Tricks of the Trade. *Lecture Notes in Computer Science*, vol 7700. Springer, Berlin, Heidelberg
65. https://ai.googleblog.com/2016/03/train-your-own-image-classifier-with.html

## INTELLIGENCE ARTIFICIELLE ET APPRENTISSAGE PROFOND APPLIQUES A LA RECONNAISSANCE D'IMAGES EN CYTOGÉNÉTIQUE

Introduction. Le récent développement de l'apprentissage profond, branche de l'intelligence artificielle, a permis de nombreuses avancées en reconnaissance d'images médicales. Le but de ce travail était d'étudier les possibilités de cette technologie dans le cadre de la cytogénétique.

Matériel et méthodes. Trois algorithmes (deux publics, Inception et MobileNet, et un, Mannaz, développé dans le cadre de cette étude) ont été testés dans la reconnaissance simple de chromosomes, de manière séparée ou par analyse en sous-groupes, et par rapport aux logiciels commerciaux d'analyse répandus dans les laboratoires, ainsi que dans la reconnaissance d'anomalies de structure. Les images étaient issues d'une base de données de cytogénétique hématologique en bandes R.

Résultats. Mannaz obtient les meilleurs résultats ($p<0,001$) avec une précision de 76,5% pour la reconnaissance des chromosomes de manière séparée et de 72,5% en sous-groupes. Ces performances étaient équivalentes à celles des logiciels commerciaux. Mannaz reconnaît correctement les dérivés 9 et 22 de translocation t(9;22)(q34;q11) dans respectivement 51% et 54% des cas. La reconnaissance des délétions des bras longs du chromosome 5 est un échec.

Discussion. Les résultats du classement des chromosomes segmentés et de la reconnaissance d'anomalies telles que les dérivés de translocation sont encourageants. L'importance d'une base de données d'images multicentrique permettant de renforcer la sensibilité est discutée et d'autres voies de recherche sont présentées.

Conclusion. Cette étude met en évidence le potentiel et l'accessibilité pour le biologiste de l'apprentissage profond dans l'automatisation du caryotype conventionnel.

Mots-clés : *cytogénétique, caryotypage, aberrations des chromosomes, intelligence artificielle, apprentissage machine, réseaux neuronaux (ordinateur)*

## ARTIFICIAL INTELLIGENCE AND DEEP LEARNING APPLIED TO IMAGE RECOGNITION IN CYTOGENETICS

Objectives. Recent progress in deep learning, a branch of artificial intelligence, led to numerous applications in the field of medical imaging. The goal of this study was to examine the potential of this technology in cytogenetics.

Study design and methods. Three algorithms (two of them open-source: Inception and MobileNet, and one developed for this study: Mannaz) were tested for simple chromosome classification, separately or through subgroup analysis, and compared to the results of private software widely used in laboratories. They were also tested for the recognition of structurally abnormal chromosomes. The pictures used were extracted from a hematological cytogenetic database using R-banding.

Results. Mannaz gets the best results ($p<0.001$) with an accuracy of 76.5% for the recognition of chromosomes in a separate manner, and 72.5% with the subgroup method. These results are comparable to the ones of the private software. Mannaz correctly identifies derivative chromosomes 9 and 22 from translocation t(9;22)(q34;q11) in respectively 51% and 54% of cases. Recognition of deletions in the long arm of chromosome 5 failed.

Conclusions. Results from chromosome classification and abnormal chromosomes recognition are encouraging. The interest of a multicentric image database to improve accuracy is discussed and other research approaches are presented. This study highlights the potentials and the accessibility for the biologist of deep learning in the conventional karyotype automation.

Keywords: *cytogenetics, karyotyping, chromosomal abnormalities, artificial intelligence, machine learning, neural network (computer)*