Reconnaissance automatique des anomalies chromosomiques afin d'accélérer le diagnostic pronostic dans la prise en charge des cancers du sang (Part I)

Ludivine, Erwan, Céline, Patricia

Août 2021

Microsoft IA/Simplon

Sommaire :

Introduction

Enjeux de la reconnaissance des chromosomes

A l'heure actuelle, les cancers du sang, hémopathies malignes, représentent 30% des cancers de l'enfant et 5% des cancers chez l'adulte. Afin de pouvoir classifier ces cancers, il est nécessaire de prendre en compte le tissu d'origine, la morphologie, les marqueurs protéiques, ainsi que les données chromosomiques de ces cancers.
Les cancers du sang sont diagnostiqués grâce au caryotype, car ils sont dus à une translocation de gènes, dans laquelle les chromosomes 9 et 22 ont échangé réciproquement des fragments de chromosomes.

Pour pouvoir mettre en évidence ce type d'anomalie chromosomique, il est donc nécessaire de **classer les chromosomes afin d'établir le caryotype**, puis de l'analyser.
Ce projet est composé de deux parties, la première se charge de la segmentation et de la classification (Part I), et la deuxième se consacre à la détection d'anomalies (Part II).
Dans ce document, nous allons traiter la première partie.

Un caryotype est réalisé à partir d'une photographie d'une cellule en vue microscopique lors de la métaphase de la mitose. A cette étape, la chromatine est condensée ce qui rend les chromosomes visibles.
Dans un caryotype, les chromosomes sont classés par paire, par taille et en fonction de la position du centromère. Il y a 23 paires de chromosomes autosomes (de 1 à 22) et une paire de gonosomes (XX pour les femmes et XY pour les hommes).

Les logiciels de réalisation des caryotypes actuels sont semi-automatiques, et nécessitent encore l'intervention des techniciens de laboratoire qui savent reconnaître les chromosomes.

L'objectif de ce projet est d'automatiser entièrement ce processus.
Pour ce faire, il est nécessaire de procéder en deux étapes. Il s'agit d'abord d'isoler chacun des chromosomes présents dans les vues microscopiques des cellules en phase de mitose (que nous appellerons "mélange"), puis de les classifier, afin d'établir le caryotype final.

Qu’est-ce qu’un caryotype ?

Les chromosomes

Les chromosomes sont visibles uniquement au moment des divisions cellulaires (mitose ou méiose). Ils contiennent essentiellement les molécules d'A.D.N. porteuses de l'information génétique et des protéines telles les histones qui maintiennent la structure des chromosomes.

Pour reconnaître spécifiquement chaque paire chromosomique, on utilise donc des techniques de marquage particulières qui permettent d'obtenir une coloration inhomogène des chromosomes par le Giemsa et l'apparition de bandes.
C'est la succession de bandes sombres et claires le long d'un chromosome, identique chez tous les individus pour un chromosome donné,

Les chromosomes métaphasiques sont constitués d'un bras court (noté p) et d'un bras long (noté q), reliés entre eux par le centromère qui correspond à un étranglement situé à un niveau variable du chromosome et qui sert de point d'attache au fuseau de division pendant la division cellulaire.

Plusieurs critères vont permettre de reconnaître et de classer les chromosomes :

- la taille : par convention, les chromosomes sont classés du plus grand au plus petit chromosome,
- l'index centromérique, c'est-à-dire le rapport entre la taille du bras court et la taille totale du chromosome? Cet index permet de reconnaître trois familles de chromosomes :
    - - Les chromosomes métacentriques dont les deux bras ont une taille à peu près équivalente.
    - - Les chromosomes submétacentriques qui ont un bras franchement plus petit que le bras long
    - - Les chromosomes acrocentriques dont le bras court est quasi inexistant (on ne trouve sur ces bras courts que les gènes codant pour les ribosomes; ces gènes étant présents à plusieurs centaines d'exemplaires double génome, la perte du bras court d'un chromosome acrocentrique n'a pas de conséquence clinique)

Le nombre de bandes visibles est variable d'une mitose à l'autre et dépend du niveau de condensation du chromosome.
Plus les chromosomes sont condensés, moins on peut observer de bandes et moins l'analyse permet de dépister des anomalies de petite taille. Le nombre de bandes par lot haploïde (c'est-à-dire pour 23 chromosomes) permet de définir la résolution de l'analyse cytogénétique ; un caryotype standard a une résolution de 300 à 550

bandes ; certaines techniques dites de haute résolution permettent d'augmenter le nombre de bandes visualisées en bloquant les chromosomes au tout début de leur condensation : on peut ainsi obtenir 800 ou même 1000 bandes par lot haploïde. Ces techniques de haute résolution sont de réalisation et d'interprétations plus délicates que le caryotype standard, mais permettent la mise en évidence d'anomalies de taille beaucoup plus réduite.

Anomalies

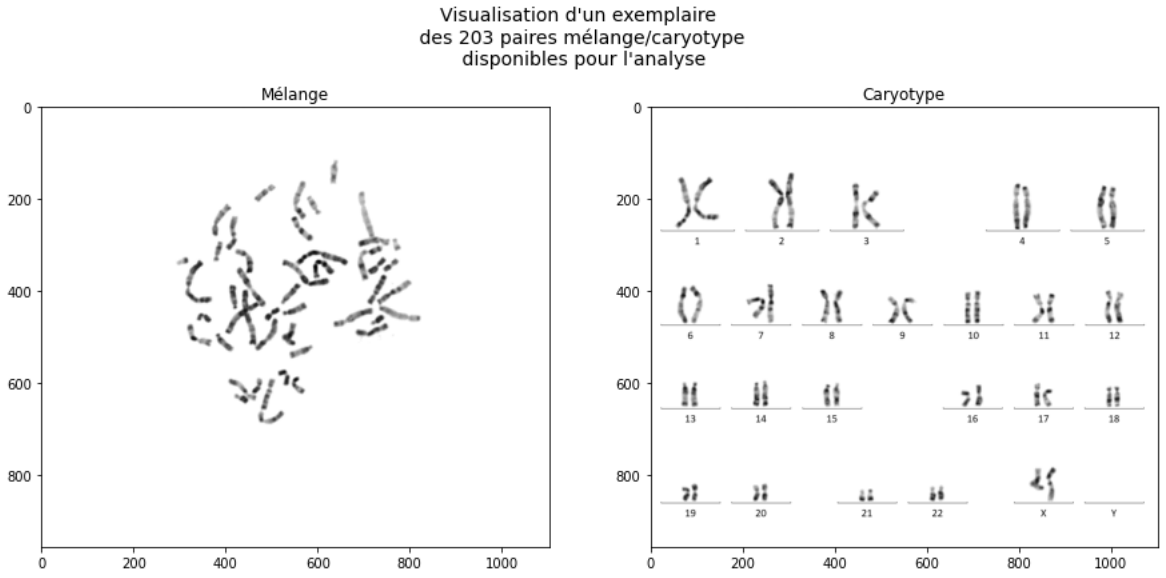
Les anomalies chromosomiques sont de deux types :

- Anomalies de nombres : aneuploïdies , les plus fréquentes (95%)
    - Trisomies : 3 exemplaires d'un chromosome
    - Monosomies : 1 seul exemplaire d'un chromosome
    - Triploïdies : 3 exemplaires de chacun des chromosomes.
- Anomalies de structures : Mode majeur de l'évolution phylogénique mais peuvent avoir des conséquences en pathologie humaine
    - 1 seul chromosome impliqué : délétions , inversions ,duplications
    - 2 chromosomes impliqués : translocations réciproques, translocations robertsoniennes, insertions

Les cancers du sang (hémopathies malignes), que nous chercherons à détecter automatiquement dans la seconde partie de ce projet, font partie de ces anomalies de structure à translocation réciproques, entre les chromosomes 9 et 22.

Données disponibles

203 couples d'images mélange/caryotype on été mise à disposition par le CHRU pour la réalisation de ce projet

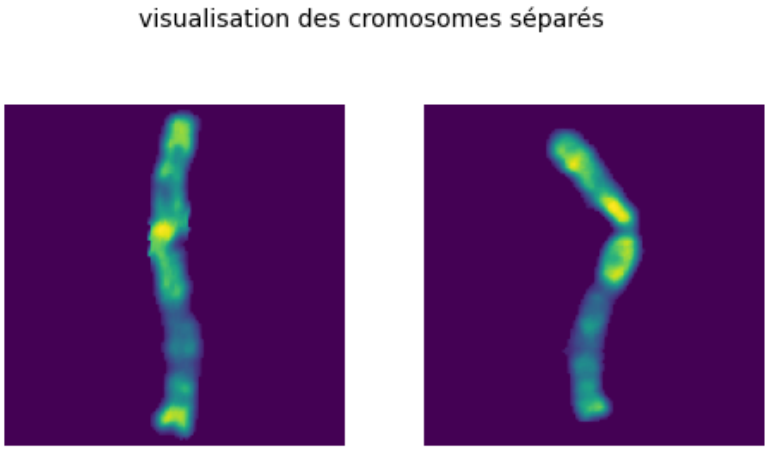


Classification des images des chromosomes

Le travail de classification se fait en utilisant uniquement les caryotypes.

En premier lieu, un travail est réalisé pour extraire des images de chromosomes de chacun des 203 caryotypes. On obtient donc environ 23x203=4669 images de chromosomes (plus ou moins, en fonction des anomalies qui ont pu êtres rencontrées : trisomies, monosomies, ...).

Les images obtenues, extraites et enregistrées accompagnées de leur label (leur n° chromosomique), permettront d'entrainer un modèle de classification.
Ces images sont similaires à celles visibles ci-dessous :



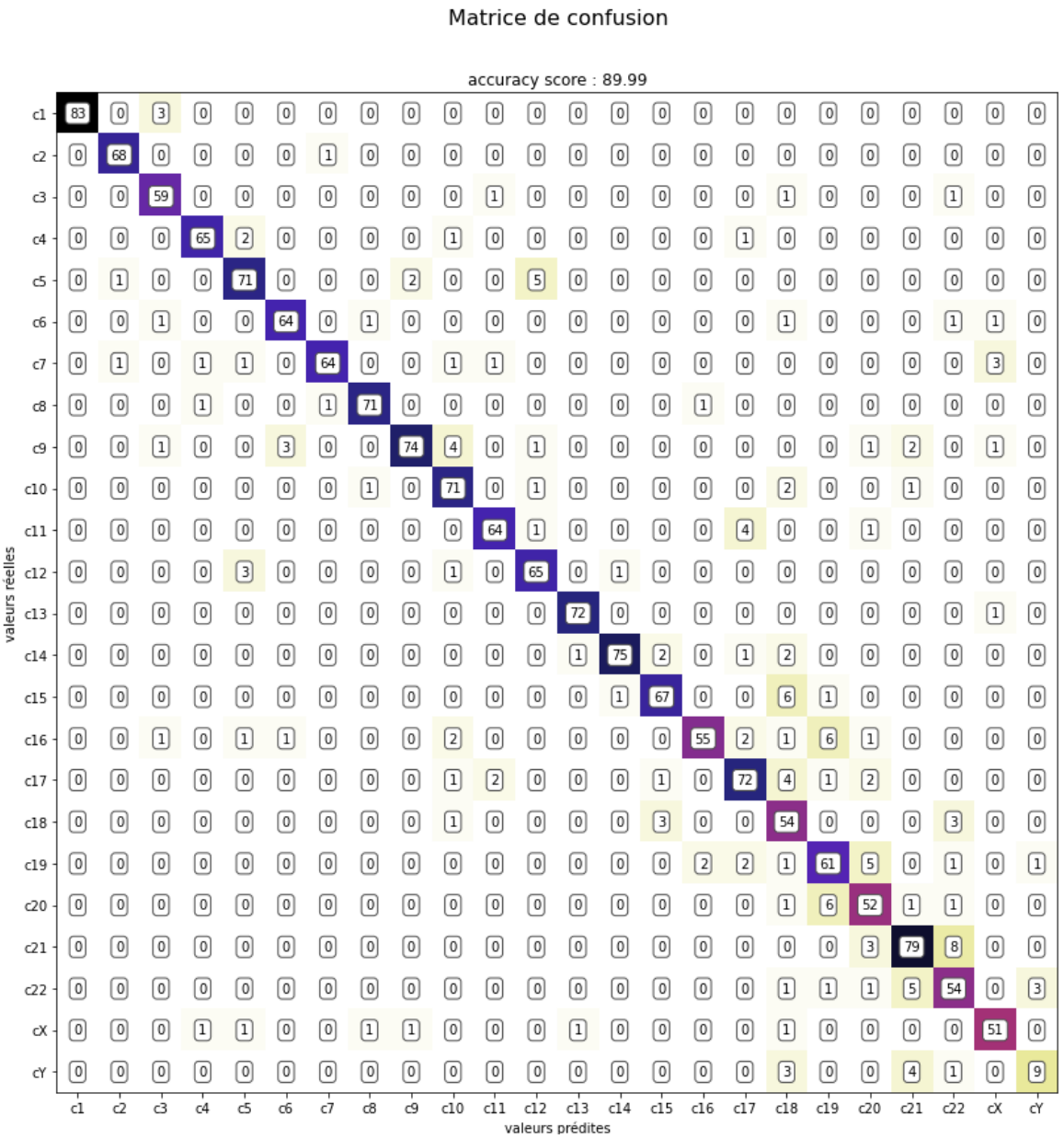
Le code pour obtenir les images et leur label est décrit dans le notebook chomosomes_extraction.ipynb

Grâce à ces images, on peut ensuite entraîner un modèle de classification efficace.

Une data-augmentation a été réalisée pour donner une rotation aléatoire à chaque image, les chromosomes détectés dans les mélanges n'étant pas dirigés tous dans le même sens.

Un modèle séquentiel de CNN convolutif est ensuite entraîné sur les images.

Les résultats en sortie du modèle sont plutôt bons, avec une accuracy finale de 90% sur la base de test.



Résultats

Les résultats de la classification sont plutôt bons, excepté pour le chromosome Y. Cela vient probablement du fait que le chromosome Y est sous-représenté dans la base de données. En effet, un être humain possède 22 paires de chromosomes autosomes, et 1 paire de chromosomes gonosomes, soit XX pour une femme, et XY pour un homme. Donc, en admettant que les hommes et les femmes soient également représentés dans la base de caryotypes, si N1 est le nombre de chromosomes “1” dans la base, N2, et nombre de chromosomes 2, etc., on a environ N1 = N = … = N22 = 4*NX/3 = 4*NY.

En pratique, dans notre dataset, et comme on peut de voir au début du notebook de classification, nous avons environ 364 exemplaires de chaque type de chromosomes autosomes, 280 exemplaires du chromosome X, et 82 exemplaires du chromosome Y. Ce déséquilibre dans les classes entraine la maivaise reconnaissance du chromosome Y.

Segmentation des images des chromosomes

Premières approches

Différentes approches ont été testées pour la segmentation des vues microscopiques des cellules en phase de mitose ("mélanges").

Le défi dans cette segmentation est de séparer les chromosomes qui présentent un chevauchement. Pour les chromosomes isolés, une segmentation avec des outils simples comme les méthodes déjà implémentées dans scikit-image conviennent (voir Label Image Regions).

Nous avons d'abord tenté une segmentation sémantique multiclasse en créant nos propres chevauchements de chromosomes avec leurs masques de segmentation. Malheureusement, ces mélanges créés de toutes pièces ne ressemblaient pas assez aux mélanges réels, et les résultats sur les mélanges réels n'étaient pas bons.

Nous avons ensuite essayé une segmentation par instance avec pytorch et détectron, en labellisant manuellement les mélanges réels existants. Les résultats, légèrement meilleurs que les précédents, n'étaient pas non plus convaincants.

Approche fonctionnelle avec UNet++

Finalement, la solution résidait dans l'article scientifique de Cao X et Al, *ChromSeg: Two-Stage Framework for Overlapping Chromosome Segmentation and Reconstruction*[1].

Lien vers l'article : http://www.bio8.cs.hku.hk/pdf/chromseg.pdf

LIen vers le code (github) : https://github.com/HKU-BAL/ChromSeg

Cette équipe de recherche est parvenue à segmenter les chevauchements de chromosomes en utilisant un réseau UNet++, qui extrait séparément les croisements et chacun des chromosomes.

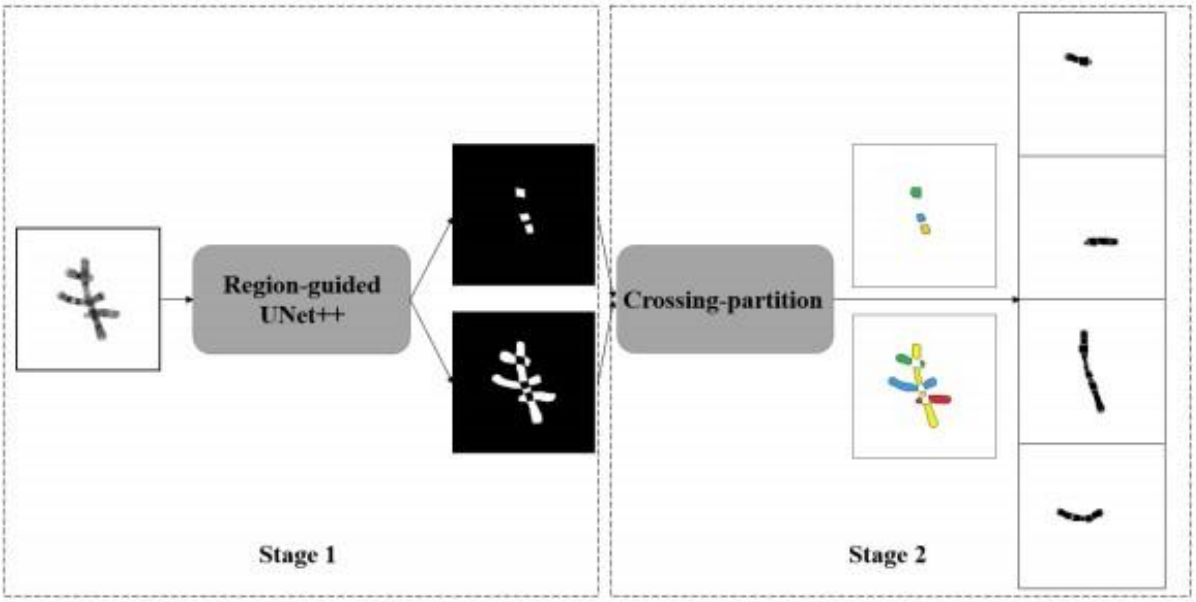
[1] X. Cao, F. Lan, C. -M. Liu, T. -W. Lam and R. Luo, "ChromSeg: Two-Stage Framework for Overlapping Chromosome Segmentation and Reconstruction," *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 2335-2342, doi: 10.1109/BIBM49941.2020.9313458.

Approche générale

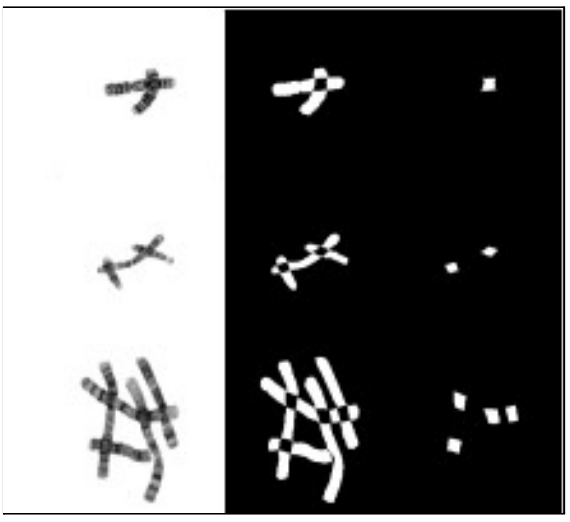
Leurs approche est composé de deux partie :

- La première (ci-dessous, “stage 1”), qui est le modèle UNet++ qui permet de séparer, d’un côté, les chromosomes sans les régions de chevauchement, de l’autre, les zones de chevauchement,
- et la seconde (ci-dessous, “stage 2”), qui est un script de reconstitution des chromosomes à partir des images obtenues en sortie de la phase 1.

Le schéma général de la solution :

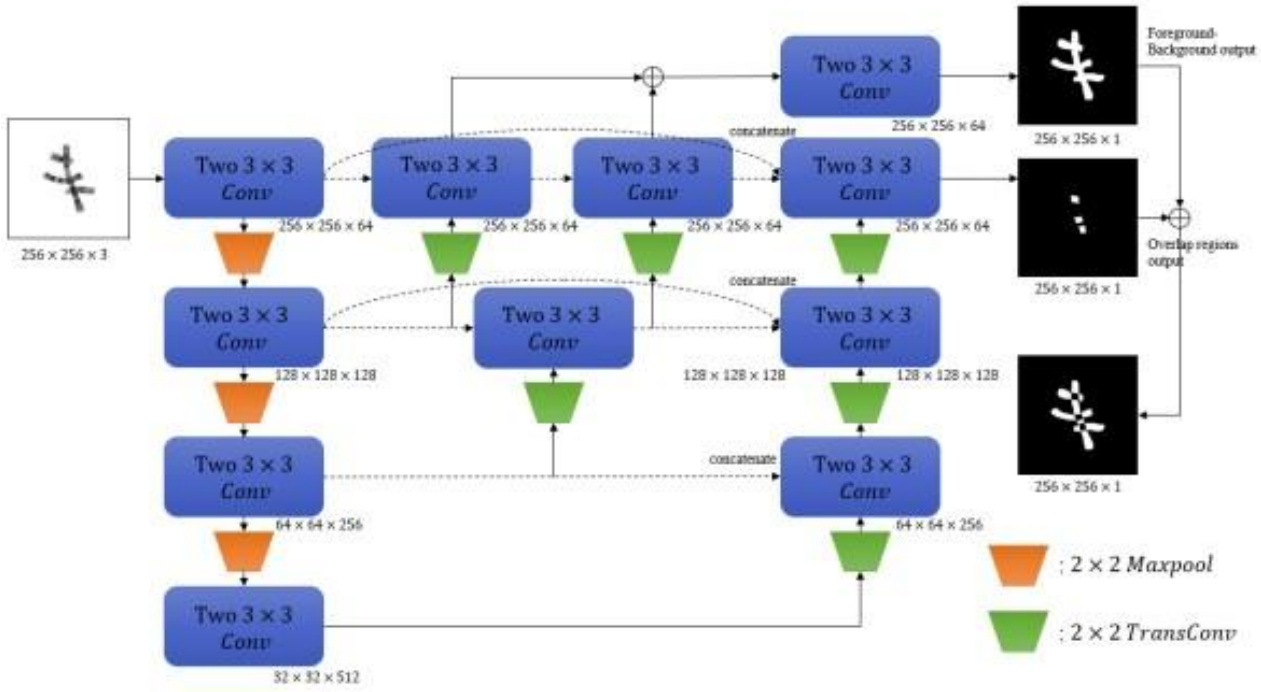


Exemples de masques obtenus en sortie de la phase 1 :



Détails du réseau de deep-learning utilisé

Voici, plus en détail, le schéma Unet++ utilisé :



Il s'agit d'un réseau de segmentation sémantique pour séparer les régions de chevauchement et de non-chevauchement des images d'entrée.
L'idée de l'UNet++ guidé par région a été inspiré par UNet++. Celui-ci est un réseau de neurones convolutif dédié à la segmentation d'images, qui ajoute des modules d'agrégation par rapport à son prédécesseur U-Net, ce qui permet de mieux fusionner les informations.

Dans cet UNet++ guidé par région, UNet++ a été étendu pour qu'il devienne un prédicteur de région multibranche.
L'entrée de l'UNet++ guidé par région est une image 256 × 256 RVB à 3 canaux.
Les auteurs ont utilisé un U-Net peu profond comme encodeur-décodeur et ajouté trois couches de convolution médianes au chemin de connexion de saut.

Le modèle, dont le code téléchargeable sur le lien github mis à disposition par les auteurs, a été entraîné avec 345 images en 256 x 256 toutes annoté avec deux images binaires représentant les régions de chevauchement et les régions de non-chevauchement.

Utilisation des deux script python

Pour utiliser ce modèle il faut au préalable avoir installé pytroch, ou utiliser Google Colab. Pour l'utilisation avec Colab, il faudra modifier la partie `argparse.ArgumentParser()` et dans tous les cas supprimer la partie `ContourAttention()` dans le fichier `Unet_plus.py`.

Une fois tout cela fait, il ne restera plus qu'à fournir une image en 256 x 256 au modèle pour qu'il retourne un masque, les régions de chevauchement et les régions de non-chevauchement. Ensuite, une fois en possession de toutes les images, il suffit de donner l'image de base, le masque, et les régions de chevauchement au deuxième script pour qu'il sépare les chromosomes.