Discovery of metabolic gene mutations causing intellectual delay

Wyeth W. Wasserman

David Aubert
Ahmed Rafik
University of Montpellier
Bio-informatique

05 jan 2015

Aujourd'hui il existe quelques domaines qui occupent une place importante du monde de la recherche, à cause de leurs importances et de leurs compléxités et qui font régulièrement appel aux centres de calculs haute performances. L'étude du génome en fait partie.
Aujourd'hui beaucoup des maladies difficilement curables sont dûes à des problèmes de mutations et de génétique (Trisomie, cancer etc...). Ces problèmes trouvent leurs sources dans l'ADN et/ou son expression.
A l'heure actuelle, le séquencage du génome pose des problèmes à la recherche et ralenti son avancée.
Le Dr. Wasserman est à la tête d'un groupe travaillant sur la partie calcul et analyse du génome. Il travaille sur des nouvelles méthodes de calculs et des outils pour identifier et implementer des bases de données pour améliorer la recherche. Le Dr. Wasserman essaie d'avancer la comprehension sur les facteurs de transcriptions qui jouent un rôle important dans l'expression du génome et dans ce but, il est venu nous présenter ses méthodes d'analyses et de prédictions des sites de transcriptions.

Il sera presenté dans cet article l'outil MeSH, qui est une base de données spécialisée dans le domaine de la recherche médical, ensuite la détection des transcriptions si importantes dans le fonctionnement. Ensuite une fois ces transcriptions détectées comment elles sont utilisés, et comment elles s'expriment mais dans un premier temps l'intêret de l'étude du génome, la portée ainsi que les méthodes utilisées.

1 Etude du génome

1.1 Pourquoi cette étude

Un laboratoire a réussi à mettre en évidence le lien entre des maladies entrainant un retard intellectuelle et des mutation génétique, ce qui a poussé le docteur Wasserman et son équipe à se pencher sur le sujet. Il commence par décrire les différents symptomes que l'on peut observer chez une fratrie de nouveau nées comme par exemple, à la naissance,

des difficultés respiratoires qui réapparaissent à 2 ans et demi ainsi qu'à 3 ans et demi. Rapidement, ils ont pu écarté de nombreuses thèses et se concentrer sur les pistes qui nous intéressent.

1.2 WGS VS WES

Défintitions simplifiées The Whole genome sequencing : c'est un procédé qui permet de déterminer l'ensemble des séquences d'ADN du génome d'un organisme donné. The Whole exome sequencing : c'est un procédé qui va selectionner l'ensemble des séquences d'ADN qui encode des protéines, et ensuite va séquencer celles-ci.

Le procédé WGS est donc plus lourd a utilisé que WES (WGS = WESx6 bp)

1.3 le gène CA5A

D'abord, nous devons définir ce qu'est l'anhydrase carbonique en général qui l'enzyme secrété par CA5A: C'est une enzyme présente à la surface plasmique intracellulaire des globules rouges qui permet transforme le gaz carbonique CO_2 en H_2CO_3. le CA5A est donc un gène de la famille des Anhydrases Carboniques. Il encode l'anhydrase carbonique dans les cellules mitochondrial.

Malgré le rôle de CA-VA dans le métabolisme intermédiaire extérieur des crises mortelles, on peut observer des phénotypes légers (léger retard dans le developement mental ou dans la croissance par exemple) c'est peut s'expliquer éventuellement de la manière suivante : Il est possible qu'il y ait un chevauchement fonctionnel de la production de bicarbonate avec l'enzyme mitochondriale CA5B. CA-VA étant une enzyme néonatale, elle devient moins important avec l'âge.

1.4 L'approche de l'équipe

L'équipe a eu une approche très classique pour aborder ce probleme qui leur donna les résultat suivants : un premier séquensage avec WES sur 67 familles révéla que 3 n'était pas génétique tandis que 64 l'était. Parmis ceux-ci, ils ont pu identifier des mutations chez 52 familles et pour les 12 restantes, il fallu utiliser WGS par manque de résultat.

2 Medical Subject Heading Over–representation Profiles

Après avoir vu comment était étudier le génome, les différents méthodes utilisés et certains gènes ciblés maintenant interessons nous à une approche plus bibliographique, pour faire des recherches sur des sujets déjà expréimentés, rendu possible grâce à MeSH.

2.1 Définition

MeSHOP est une application très complète.Celle-ci permet de centraliser les recherches et d'indexer les publications scientifiques. Celli ci englobe des thématiques allant de la recherche génétique à la chimie en passant par les maladies, ce qui interesse l'auteur.

Figure 2.1: *exemple d'un nuage de mot*

2.2 fonctionnement

Le moteur de cette application repose sur une base de données MEDLINE et le moteur de recherche PubMed.

2.2.1 Annotations

Chaque article devra être annoté avec un vocabulaire particulier. En effet, c'est cette annotation qui permettra une recherche efficace. Le système MeSH a une liste de vocabulaire (*PMIDs*) qui servira d'index pour le fichier.

2.3 Editer un article

Après l'écriture d'un article, il est nécessaire de le publier. Et c'est à ce moment là que va servir le système d'annotation. En effet une lecture par un robot de l'article sera fait et permettra, en fonction du vocabulaire utilisé, de classer l'article dans des catégories et de l'indexer pour des recherches.

Chaque mot sera classé en fonction de son nombre d'ocurence dans les *PMIDs*

2.4 Recherche

Un point important est la thématique recherchée. En effet, comme vu précédemment, les sujets d'articles englobent un grand nombre de thématique (génétique, maladies, chimie etc...) et donc le contexte peut changer pour des mêmes mots ou combinaisons de mots.

3 Détection de transcription

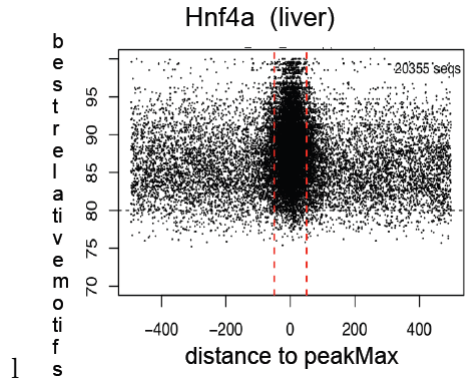

Figure 3.1: *Motif Hnf4a au milieu du bruit*

3.1 Les facteurs de transcription

Le facteur de transcription est une proteine qui a pour but lire et transcrire l'ADN. Pour cela, elle parcours le brin d'ADN jusqu'à trouver le site dont la forme des nucléotides lui correspond.

3.2 JASPAR 2014

JASPAR est la plus grande base de données en open source de nucleotides stoqué sous forme de matrice décrivant la liaison de facteurs de transcription sur plusieurs espèces.

3.3 ChIP-Seq

ChIP-Seq = Chromatin ImmunoPrecipitation Sequencing
C'est une méthode utilisée pour analyser les interactions entre les protéines et l'ADN. Cette technologie combine l'immunoprécipitation de la chromatine (ChIP), en utilisant des anticorps spécifiques d'une protéine d'intérêt et le séquençage haut débit.
Le principe consiste à fixer de façon covalente les protéines liées à l'ADN par un traitement chimique, de fragmenter la chromatine, d'immunoprécipiter les fragments en présence de l'anticorps d'intérêt et après purification et élimination des protéines, de séquencer les fragments d'ADN obtenus.
Il est ainsi possible de cartographier tous les sites de liaison sur l'ADN de cette protéine à l'échelle du génome.

3.4 Les Zingers

Lorsque l'on utilise les ChIP-Seq pour transcrire un brin d'ADN, nous repérons plus facilement les motifs au milieu du bruit comme on le voit sur le motif Hnf4a ci-dessous.

Mais certain motifs apparaissent toujours peut importe la méthode utiliser sans que l'on comprennent réellement ce qu'il font là.

Ces motifs sont appelés “Zingers”.

3.5 Les annotations de regions (sequences) de régulation

Définition :
Une region de regulation est une zone du gène qui lui permet reguler la production de protéine trancrite par ce gène.

L’annotation de ces regions est une méthodes qui consiste à etudier ces zones afin de savoir si le gène est actif, mutant, ou autre ...

4 Site de liaison d’allèle spécifique

Dans cette partie sera présenté une manière plus affiné de détecter les sites de liaisons.

4.1 Définition

Pour commencer un transcription, une protéïne est nécessaire. Celle-ci doit se fixer sur l’hélice. Une liaison d’allèle spécifique est ce même phénomène mais avec une protéine qui se liera d’avantages sur les allèles récessives.

4.2 allèles préférées

Une des premières observations faites est que sur les individus hétérozygotes (ayant deux allèles différentes) on remarque que certaines allèlles créent plus de sites de connections que une autre, alors que il était possible de s’attendre à une distribution equi-probable.

Mais comme les allèles fortes s’expriment plus que les allèles récessives celà peut être dû à plusieurs raisons mais la plus commune est que, par selection naturelle, un gène subira une mutation, et aura un nombre supérieur de site de transcriptions que le précédent et ainsi pourra se transmettre plus facilement de génération en génération. De ce fait l’étude à de meilleurs chances de résultats si il est éffectué sur des gènes forts.

Pour trouver les places de liaisons de transcription, une unité a été crée: **PWM**: La matrice de position de poids.
Il s’agit d’une matrice où sont marquées les probabilités de chaque nucléotide d’apparaitre dans le site de liaison pour la transcription, ainsi que sa position.

Ceci a une importance particulière car grâce à ces matrices, il a été possible de remarquer que les liaisons ce faisaient plus sur les allèles récessives que les autres.
Mais ces travaux sont ralentis par le fait qu’il peut exister de multiples facteurs qui modifient le taux de transcription d’un site comme les mutations, l’hérédité et l’epigenetique.

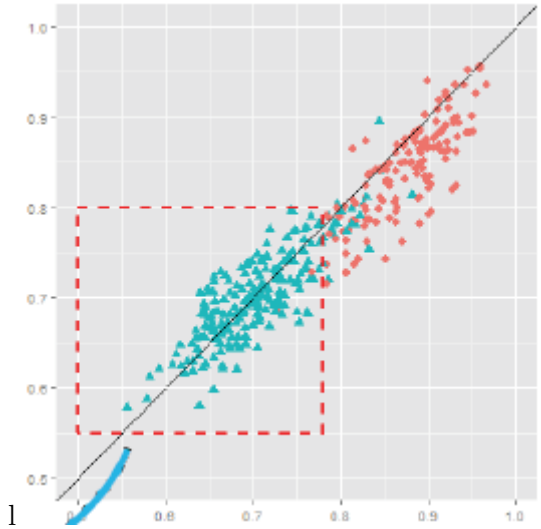

Figure 4.1: *Allèle score*

4.3 Recheche de site de liaison pour la transcription pour les lymphomes

C'est ici que la présentation recoupe avec les travaux du Dr. Wasserman.
En effet, grâce aux outils présentés précédemment, beaucoup de données ont été recueillis, notamment des échantillons d'ADN, d'ARN de malades atteints de lymphomes.

Ensuite certaines zones du génomes ont été ciblés, et une des premières remarques est que les sites de transcriptions ont eu des taux de mutations plus élevé comparés aux séquences saines.

La mort des cellules, à un rôle dans ce fonctionnement. En effet, la mort des cellules est régulé d'une manière naturel et saine mais si cette régulation est perturbée celà peut causer des problèmes, dans le cas d'une diminution et dans le cas d'une augmentation un lymphome, par exemple. Une augmentation des cellules (ou plutôt une non décroissance normale du nombre de cellules) augmentent le nombre de site de liaison augmente.

5 Finalité

Comme expliqué, les sites de transcriptions sont importants dans l'expression du génome. Des malades attendent des avancées médicales mais la technologie bloque sur le séquencage ADN. La recherche est mobilisée. Malgré l'aspect financier, il existe des initiatives qui permettent de partager des résultats, des expériences et d'aider la recherche (MeSH).

Mais malgré cela la recherche bloque sur le séquencage du génome, pour se subsituer à cet obstacle de nouvelles idées émergent et au lieu de faire une recherche brute, des méthodes plus affinées sont implémentés par le Dr. Wasserman et dans le cas où le problème ne serait pas entièrement génétique mais dans son expression, via les sites de transcriptions.

5.1 perspectives