Projet :

En biologie, l'émergence des méthodes de séquençage a sans aucun doute permis de faire des avancées considérables dans le domaine de la régulation. L'un des acteurs de ces avancées est la puce à ADN. Pouvant contenir des millions de cDNA d'un organisme, il est possible de déterminer le niveau d'expression de milliers de gènes d'une cellule dans des conditions données en hybridant l'ARN rétrotranscrit de cette cellule à celui de la puce.

Ainsi, en observant des profils d'expression dans différentes conditions, on peut établir un réseau : mettre en relation chaque gène et les gènes qu'il régule. Les acteurs au coeur de la régulation sont les facteurs de transcription et comprendre leur fonctionement demande d'avoir au préalable identifié les séquences qu'ils lient.

Les expériences de ChIpSeq (Immunoprécipitation de chromatine) ou de SELEX (évolution systématique de ligands par enrichissement exponentiel) permettent l'identification de ces séquences. On peut alors construire des matrices PFM (matrice position fréquence), qui donnent des scores aux séquences d'ADN en fonction des nucléotides vus par la matrice à chaque position. Les scores étant liés à l'affinité de liaison, l'expérience de SELEX peut permettre de prédire les séquences qui sont le plus succeptibles d'êtres liées au sein d'un génome. Si on connait au préalable les régions d'ADN lié, il est possible de confronter l'affinité de liaison théorique donnée par la matrice (qui ne prend en compte que les intéractions ADN facteur de transcription) et la liaison véritable qui fait souvent intervenir d'autres paramètres : état de la chromatine, méthylation de l'ADN, liaison en tant que dimère... En intégrant ces paramètres à un modèle basé sur la PFM, on peut créer un modèle qui se rapproche de la réalité, permettant d'évaluer l'importance de ces paramètres dans la liaison.

Certaines bases enregistrent les PFM d'un grand panel de facteur de transcription. C'est le cas de JASPAR, une base de données qui recense les PFM d'un grand nombre de facteur de transcription dans plusieurs organismes. L'accès gratuit, les normes de qualités exigées pour la construction des PFM et l'abscence de doublon pour un même facteur de transcription font de JASPAR un outil unique et incontournable pour l'étude des séquences cis-régulés. La base est actualisée tous les deux ans et un grand nombre y est à chaque fois incorporé

Le LPCV alimente la sous partie de la base concernant les plantes à chaque occasion. Les dernières données à ajouter vont représenter un gros échantillon ; en 2016 une équipe a inventé une nouvelle technique de séquençage effectué sur de l'ADN nu appelé DAPSeq, pour trouver les sites de liaison d'un facteur de transcription. Avec cette méthode, ils ont étudié 529 facteurs de transcription dans A. thaliana. À titre de comparaison, 496 facteurs de transcription au total (toutes les espèces) ont été rajoutés à la base lors de sa dernière actualisation.

Ce travail est donc d'une grande ampleur et constituerait le coeur du projet. Étant étudiant en thèse de bioinformatique dans l'équipe Régulation du Développement Floral au Laboratoire de Physiologie Cellulaire et Végétale (LPCV), les question relatives aux sites de liaison me concernent directement. Le LPCV ne dispose pas réellement d'expertise bioinformatique et il m'est difficile de savoir si les méthodes que j'emploie sont les méthodes types. Le mathelier-lab ayant un grand savoir-faire bioinformatique dans les sites de liaisons, je pourrai me former correctement autour de ces questions. De plus, Anthony Mathelier a été le principal acteur dans la complétion de la base JASPAR en 2014 et en 2016.

En deuxième lieu, nous pourrions étendre des analyses à des sets de données beaucoup plus grands : les analyses que je fais sur quelques facteurs de transcription pourraient êtres généralisées à l'ensemble des facteurs si je disposais des méthodes nécessaires.

Ma thèse porte sur le facteur de transcription LEAFY, qui déclenche à lui seul le développement des fleurs. Certaines expériences qui montrent la régulation de LEAFY par les facteurs de réponse à l'auxine (ARF) orientent également ma thèse vers les sites de liaisons et la spécificité des ARFs.

Ainsi, les sites de liaison des ARF sont presque similaires alors que les régions qu'ils lient peuvent êtres radicalement différentes. Dans une certaine mesure, nous avons été capables d'expliquer les différences entre les sites de liaisons des différents ARFs par le fait qu'ils se lient en tant que dimères : les configuration de dimères qu'ils préfèrent (espacement entre sites, orientation des sites...) peuvent être différentes malgré des sites de liaison identiques.

Regarder si les sites de liaison d'un facteur de transcription ont tendance à être espacés n'est pas une question qui se limite aux ARFs. Il serait extrêment intéressant de mener cette étude au niveau de tous les facteurs de transcriptions dont on connait les régions liées.

Enfin, Anthony mathelier est spécialiste de la “shape” de l'ADN. Certaines études menées au LPCV pour expliquer la spécificité des ARF ont tenté d'incorporer la “shape” sans succès. Peut-être serait-il possible de les mener à bien lors de mon séjour.