Inférence de réseaux chez Arabidopsis thaliana

Marie-Laure Martin-Magniette

1er février 2011

Arabidopsis thaliana est la plante modèle des Brassicacées, son génome séquencé en 2000 contient environ 25 000 gènes codants des protéines. La plante modèle est étudiée dans de nombreux laboratoires et de nombreuses données transcriptomes ont été générées principalement sur la puce Affymetrix ATH1. Depuis quelques années, quelques articles sont sortis sur l'inférence de réseaux de gènes chez Arabidopsis thaliana. Pour cela ils exploitent les données Affymétrix déposées à GEO ou à Arrayexpress. La plupart des articles utilisent des méthodes fondées sur la corrélation de Pearson pour mettre en évidence de la co-expression car il est supposé que les gènes co-exprimés sont de bons candidats pour être des gènes co-régulés. Seuls, Ma et al. (2007) proposent d'étudier le réseau de régulation directement à l'aide d'un modèle graphique gaussien. Ils utilisent la méthode de Schäffer et Strimmer (2005) pour identifier un réseau de plus de 22 000 gènes à l'aide de 3000 expériences Affymétrix représentant environ 150 conditions expérimentales différentes.

Au début des années 2000, grâce à un projet européen, une puce à ADN nommée CATMA (Complet Arabidopsis Thaliana MicroArray) a pu être élaborée. Les sondes de cette puce ont été dessinées pour être le plus spécifique possible afin d'éviter de la cross-hybridation. Elles ont été définies à partir de l'annotation officielle du TAIR et de l'annotation fournie par le logiciel Eugene. Grâce à cette particularité, les puces CATMA permettent d'étudier l'expression d'environ 6000 gènes non représentés sur la puce Affymetrix ATH1 (5000 gènes TAIR non représentés sur Affymetrix, 500 gènes Eugene et 500 prédictions de miRNA). La puce CATMA est exploitée à l'URGV (Unité de Recherche en Génomique Végétale) depuis 2003 sur la plate-forme transcriptome. Toutes les données sont produites avec le même protocole expérimental et analysées avec la même routine de base pour la normalisation et l'analyse différentielle. De plus toutes les données sont stockées dans la base de données CATdb qui est reliée à la base de données FLAGdb++ pour gérer les liens entre les sondes de la puce CATMA et l'annotation structurale d'Arabidopsis thaliana. La ressource CATMA est actuellement en cours de transfert vers la base Genevestigator pour que toute la communauté Arabidopsis puisse la consulter.

Jusqu'à maintenant les principales demandes des biologistes sur la plate-forme de l'URGV étaient de fournir la liste des gènes différentiellement exprimés entre 2 conditions. Mais leurs demandes évoluent car ils veulent maintenant étudier le comportement des gènes différentiellement exprimés dans d'autres conditions afin de formuler de nouvelles hypothèses biologiques. Implicitement, leur objectif final est de mettre en évidence le réseau de gènes (quel gène agit sur quel gène) et de grouper les expériences par similarité de réponses. Le projet ci-dessous se concentre donc sur cette nouvelle demande et la manière dont elle peut être déclinée et étudiée.

De mon expérience à l'URGV et d'une review récente de Usadel et al. (2009), j'ai identifié 4 manières pour appréhender les réseaux de gènes

1. Pour un ensemble de conditions données, on souhaite identifier les gènes participant dans un processus biologique et dire comment ils collaborent. Ces travaux sont effectués dans le cadre du projet SONATA dont l'objectif est d'identifier tous les gènes impliqués dans le stress (biotique et abiotique) et de caractériser leur collaboration. Le projet est porté par Sébastien Aubourg, responsable de l'équipe bioinformatique pour la génomique prédictive de l'URGV et est une collaboration avec l'équipe Génomique Fonctionnelle d'Arabidopsis de l'URGV, Gilles Celeux (INRIA) e, Cathy Maugis (INSA Toulouse), Caroline Meynet et Christine Keribin (univ. Orsay), Sylvie Huet, Brigitte Schaeffer et Nicolas Verzelen (MIA

INRA) et Christophe Giraud (Ecole polytechnique). Le projet est financé par l'INRA pour 3 ans (projet AllEnvi 2010-2013 et financement MIA et GAP en 2010).

2. Pour un ensemble de gènes donnés, on souhaite regarder leur comportement transcriptomique sur l'ensemble des conditions disponibles dans CATdb pour déterminer dans quelles conditions ils sont transcrits et aussi pour inférer un réseau. Les biologistes partent généralement de la liste de gènes différentiellement exprimés au moins une fois dans leur projet, puis vont explorer Genevestigator ou utiliser des outils tels que ceux disponibles sur le site de l'université de Toronto (http://bar.utoronto.ca/eapop/cgi-bin/ntools_expression_angler.cgi) pour essayer de mieux les caractériser ou de compléter la liste des gènes à étudier.
3. Pour un réseau en cascade avec un nombre fini et raisonnable de gènes identifiés. On est capable de muter chaque gène du réseau, d'étudier alors le transcriptome de chacun des mutants en le comparant aux transcriptomes d'individus sauvages ou en regardant le mutant dans des conditions qui permettent de voir l'impact de la mutation. L'objectif est de construire le réseau. L'équipe de Héribert Hirt « Protein Kinases in Signal Transduction » de l'URGV étudie actuellement un tel réseau.
4. On dispose de données transcriptomes sur des lignées recombinantes de deux écotypes d'Arabidopsis ainsi que de marqueurs génétiques et l'on souhaite étudier le réseaux de gènes quand la plante est soumise à un stress hydrique. Ceci implique l'URGV pour la création et le stockage des données transcriptomes et Olivier Loudet (INRA Versailles) pour l'analyse des eQTLs.

Toutes ces approches sont complémentaires et il me semble qu'il est possible de travailler dessus parallèlement. Mon idée est de partir de projets biologiques identifiés, caractéristiques des points 2, 3 et 4 et d'en formuler des problèmes méthodologiques à résoudre. Le point 1 concernant le projet SONATA est écarté de la demande car le projet est déjà en cours et fait à lui seul l'objet d'une demande de renouvellement de réseau MIA. SONATA est en effet un projet qui aborde de nombreuses questions biologiques et où sont déployées une variété importante de méthodologie (mélange gaussien avec sélection de variable, biclustering, modèle graphique gaussien, gestion des données manquantes, pénalité de type Lasso). L'objectif de SONATA-stat est de travailler sur toutes les méthodologies nécessaires pour exploiter un jeu de données, alors que la demande concernant le réseau « inférence de réseau » est dans un premier temps plus en amont pour identifier les verrous méthodologiques lorsqu'on souhaite inférer des réseaux de régulation de gènes chez un organisme qui possède plus de 25 000 gènes codant des protéines.

Les premières questions sont les suivantes :

- Quels réseaux est-il possible d'inférer à partir des données transcriptomes, quelles méthodes pourraient être utilisées ?
  - Quels types de résultats sont envisageables ?
  - Quels types de réponse biologiques ?
- D'autres informations de nature très différentes existent par ailleurs (interactome, transcriptome Affymetrix, chIP-chip). Pourrait-on les utiliser pour inférer des réseaux ?
- Si on opte pour une démarche de construction de réseau à partir d'un sous-ensemble de gènes (point 2), on va alors avoir une collections de petits réseaux. Peut-on inférer un grand réseau qui inclurait tous les gènes étudiés dans au moins un petit réseau ?
- Comment et quelles données seraient nécessaires pour identifier dans des plantes cultivées des réseaux de gènes (module fonctionnel simple) connus chez Arabidopsis thaliana.

Pour la partie biologie, ce projet sera une collaboration avec l'URGV en particulier avec

- l'équipe « Bioinformatique pour la génomique prédictive » de l'URGV en charge de la partie bioinformatique de la plate-forme et qui développe en parallèle des projets de recherche sur l'exploitation des données « omics »
- l'équipe « Genomique Fonctionnelle d'Arabidopsis » qui gère la plate-forme transcriptome de l'URGV. J'ai en particulier sollicité Richard Berthomé qui travaille actuellement sur un tel projet pour les gènes impliqués dans la carence azotée. Il a réalisé les expériences transcriptomes sur CATMA, identifié 4 sous-ensembles de gènes qui l'intéressent et cherchent maintenant des méthodes pour aller plus loin. Je travaille actuellement avec lui pour étudier la co-expression mais il est très intéressé par l'inférence de réseau. Lui et des collègues de l'INRA de Versailles seraient prêts à travailler avec notre groupe pour formuler correctement les questions biologiques et pour nous faire un retour lors de l'application de méthodes sur leurs données. Etienne Delannoy, responsable de la plate-forme transcriptome est également très intéressé par le développement de méthodes pour analyser de manière approfondie la ressource CATdb. Il est également prêt à s'impliquer dans ce réseau MIA.
- l'équipe « Protéines kinases » dirigée par Héribert Hirt qui s'intéresse aux rôles des protéines kinases dans la réponse de la plante aux stress.
- L'équipe « génomique des plantes cultivées » dirigée par Abdelhafid Bendahmane pour les questions sur les transferts des connaissances des plantes modèles aux plantes cultivées et pour l'exploitation d'une collection de 100 000 mutuants d'Arabidopsis.

L'analyse utilisant des lignées recombinantes sera effectuées dans le cadre d'une collaboration avec Olivier Loudet de l 'équipe « Variation et tolérance aux stress abiotiques » de l'INRA de Versailles.

Pour le planning, l'idée est d'amorcer les discussions entre biologistes, bioinformaticiens et méthodologistes pour bien comprendre les questions biologiques, les formuler d'un point de vue mathématique et de monter s'il y a matière un projet méthodologique autour de l'inférence de réseaux chez Arabidopsis d'ici un an.