P.O.O. en Python

Projet GO & GOSlim

(Programme GOLIAT - Gene Ontology Link Analysis Tool)

François Bonnardel, Samaneh Kamyabiazar, Aurélien Birer, Yoann Pageaud.

Encadrants : Mme. Yolande Diaz, M. Franck Samson

23 Mars 2016

Table des matières :

Introduction :

Dans le cadre de notre projet d'étude, nous étudions les changements dans les fonctions biologiques exprimées dans le génome d'Arabidopsis thaliana (A. thaliana) due à la duplication de gènes. Pour déterminer quelles fonctions sont affectées par les duplications, une liste de gènes dupliqués a été sélectionnée.

Grace à la base de données Gene Ontology et au logiciel Cytoscape, nous avons développé un outil en Python capable de retrouver les GOSlims les plus représentés à partir d'une liste de Gene Ontologies (GOs).

Les GOSlims, qui sont eux aussi des GOs, mais à un niveau élevé de l'arbre,possèdent des termes très généraux. Ils sont utilisés pour regrouper les GOsde niveaux inférieurs afin de pouvoir comparer des GOSlims (c'est-à-dire des groupes de GO) entre eux.

À partir de la liste de gènes on recherche les GOSlims présents en plus grandes quantités afin de voir les fonctions surreprésentées.

Déterminer des groupes de gènes spécifiques de la condition différenciée permettrait afin d'avancer dans la compréhension des fonctions liés à la duplication des gènes.

TAIR database

La base de données TAIR regroupe toutes les informations liées au génome d'A. thaliana.

Ontology

Une ontologie représente une vue spécifique des données. Elle est définie par 2 parties :

La Conceptualisation : elle permet d'identifier :

- Les concepts-clés du domaine,
- Leurs propriétés et leurs relations,
- D'identifier les termes du langage naturel
- De structurer le savoir du domaine.

La Structuration d'un domaine d'intérêt : elle possède :

- Un concept (gènes, macromolécule),
- Des relations (IS-A, PART-OF, etc.),
- Des attributs/rôles (a_pour_fonction, a_pour_produit),
- Des contraintes (mâle ou femelle mais pas les 2),
- Des objets (instances des concepts),
- Des valeurs (le produit du gène trpA est trytophan-synthetase),

Gene ontology

Le projet GO propose un vocabulaire contrôlé de termes définis représentant les propriétés du produit des gènes. Ceci dans 3 domaines 'cellular component', 'molecular function', 'biological process'.

La GO ontologie est structurée comme un graphe acyclique orienté, où chaque terme à des relations définies avec un ou plusieurs autres termes du même domaine, et parfois dans d'autres domaines. Le vocabulaire GO est adapté pour une utilisation inter-espèces et inclus des termes applicables aux eucaryotes et aux procaryotes, aux organismes unicellulaires et pluri cellulaires.

Les 3 ontologies :

- « Cellular Component » : ces termes décrivent un composant d'une cellule qui fait partie d'un plus grand objet, comme une structure anatomique ou un groupe formé de produit de gènes.
- « Biological Process » : ces termes décrivent une série d'évènements accomplis par un ou plusieurs ensembles de fonctions moléculaires, organisées.
- « Molecular Function » : ces termes décrivent des activités qui ont lieu au niveau moléculaire comme par exemple l'activité catalytique ou l'activité de fixation.

On travaille sur les « Biological process » : série d'événements effectués par un ou plusieurs assemblages ordonnés de fonctions moléculaires (exemple : termes oxydative phosphorylation).

Les types de relations

Les relations entre les termes GO sont de différents types exprimant des interactions biologiquement distinctes :

- is a
- part of
- regulates
- positively regulates
- negatively regulates
- has part
-

La relation « IS A » est particulière car elle exprime une appartenance totale et peux être considérée comme relation primaire tandis que les autres relations peuvent être considérées comme secondaire.

Si un chemin utilisant plusieurs relations est créé, le chemin reste de type « IS A » tant qu'aucune autre relation n'est utilisée.

Dans le cas contraire le chemin deviens du type de la relation secondaire utilisée et ne peux pas contenir d'autres types de relations secondaires.

Matériel et méthodes :

2.1 Les données

Un fichier « go.obo » contenant toutes les informations pour chaque terme GO. Il peut être récupéré directement sur le site du consortium et respecte un format spécifique.

Un fichier contenant les ID des gènes dupliqués d'A. thaliana.

Un fichier provenant de la base de données TAIR fournissant les annotations GO associées à chaque gène d'A. thaliana.

Un nouveau fichier « liste1 » a été créé et contient les gènes dupliqués accompagnés de leur annotation GO. Ce fichier a été réalisé à l'aide du module5 addGO.

Un fichier contenant une liste de GO pouvant être importée en tant que liste de GOSlims supplémentaires. Un ID de GOSLIM_user spécifique à chaque GO est alors ajouté.

2.2 Graphical Navigateur AmiGO

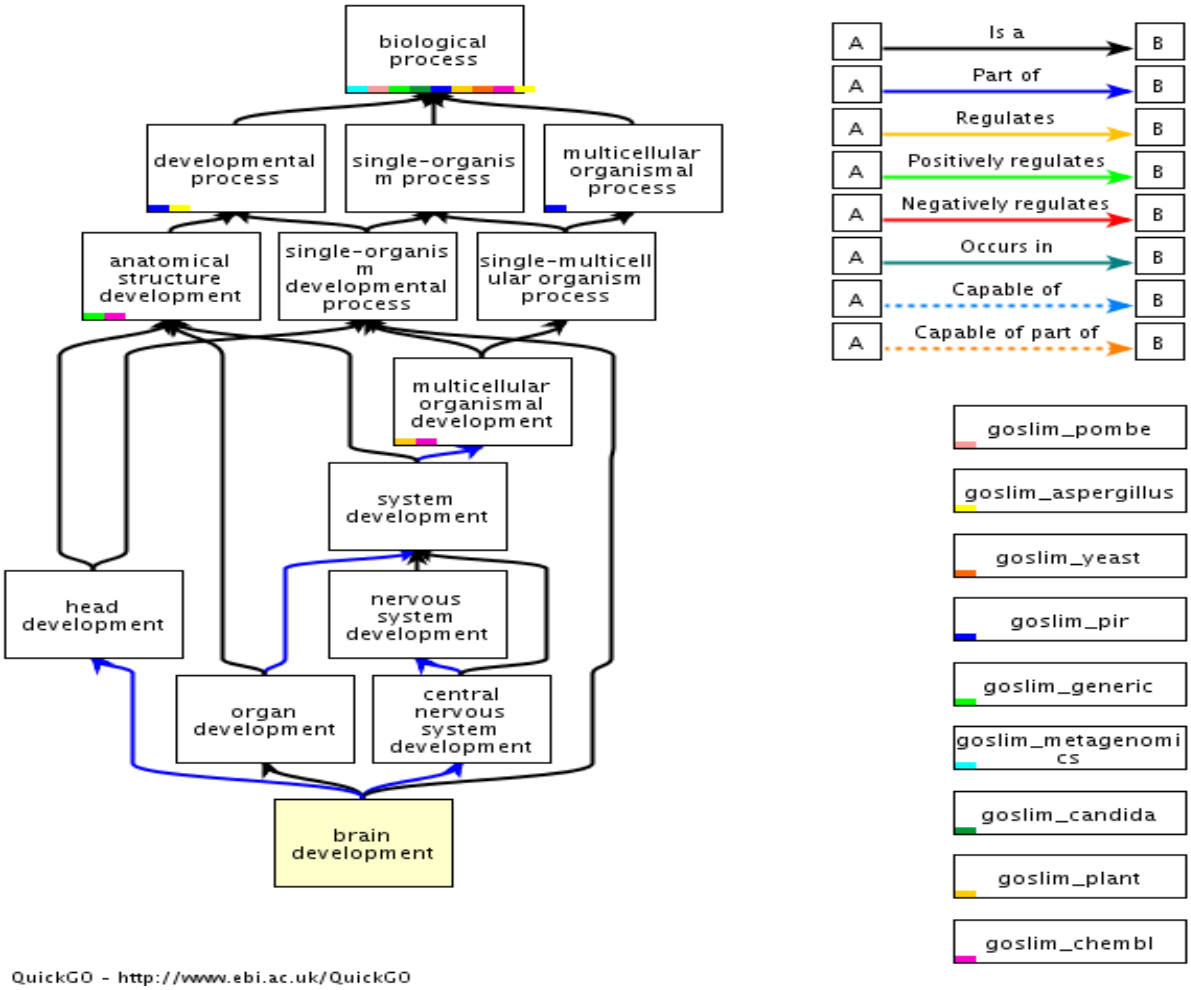


Figure 1 :*Exemple de diagramme de relationspour le terme GO « brain development », les relations rencontrées pour remonter dans des termes GO plus généraux. Les GOSlims rencontrés sont de couleurs différentes.*

Cet outil a permis de vérifier pour des GO sélectionnés le bon fonctionnement de l'outil GOLIAT.

2.3Workflow de GOLIAT

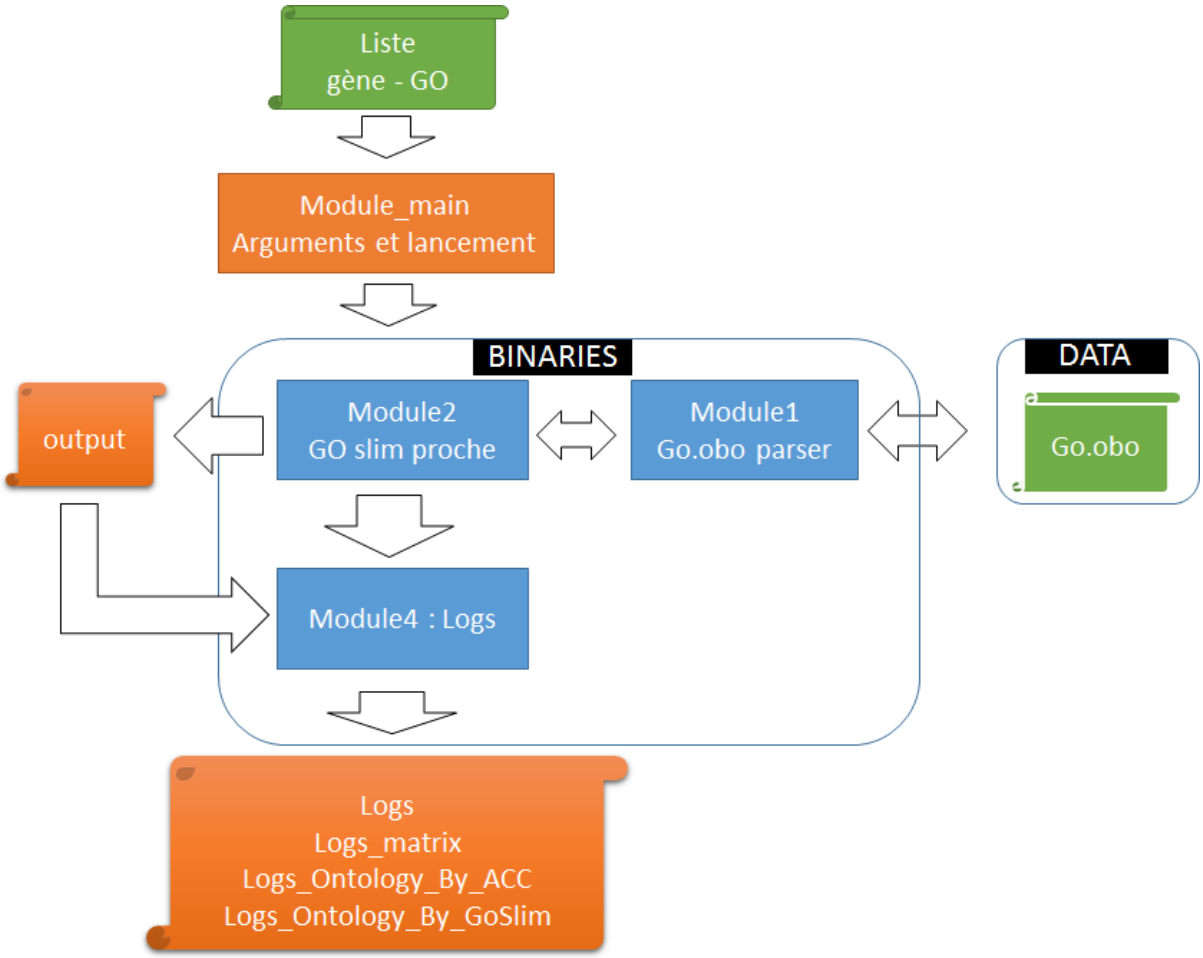


Figure 2 *: Schéma du fonctionnement (Workflow) de l'outil GOLIAT.*

Une liste de gènes est importée par module_main qui vérifie les différentes options et arguments et lance le module2. Le module2 commence par utiliser les fonctions de module1 pour créer un objet « dico_go ». L'objet est ensuite rempli. Pour cela le fichier « go.obo » est parcouru. Pour chaque GO :

1) Récupération des informations du GO, si un marqueur GOslim est présent.
2) Récupération des types de relations que le GO a avec les autres GO (GO Parents et GO Fils)

Une fois l'objet chargé module2 réalise l'analyse en récupérant pour chaque GO tous les GOSlims parents.

Il créé ensuite un fichier de sortie avec les GOs et GOslims associés.

Module4 est ensuite appellé et ouvre le fichier de sortie des GOSlims pour créer des fichiers logs avec, pour chaque gène, pour tous les GOslims rencontrés ceux qui sont utilisés ou non par ce gène.

2.4 SGD Gene Ontology Slim Mapper

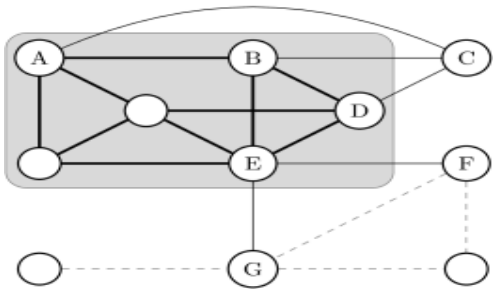
Pour trouver les relations entre les GOs on utilise la méthode « gomapper ».

2.5Testsstatistiques :

K-means.

Analyse et résultats

On utilise l'application Cluster ONE (Clustering with Overlapping Neighborhood Expansion) pour trouver le minimum et maximum des edges entre les clusters et pour trouver la densité. Cluster ONE permet également le chevauchement des sous-graphes (clusters). Puisqu'un gène peut prendre à part plus d'un module fonctionnel unique.



Cluster ONE s'installe dans le menu Plugins de Cytoscape sous un sous-menu nommé Cluster ONE.

Dans Cytoscape :

1. Cliquer sur "Apps" et puis choisir le cluster ONE. Le panneau de cluster ONE se trouve sur un onglet séparé dans le panneau de Cytoscape.

Les paramètres sont regroupés avec ceux de base et avancés.

Vous pouvez soit utiliser les valeurs par défaut, soit modifier les paramètres souhaités.

Pour la similarité choisir le model "simpson coficient"(une technique de calcul numérique d'une intégrale) puis, choisir un seuil de 0.05 quoi ? pourcents ? .

2. Cliquer sur "Start".

La densité du graphe est la somme des poids des edges divisée par le nombre de edges théoriquement possibles.

1. Les clusters sont rangés dans l'ordre croissant des p-value. Après le processus de regroupement, les nœuds du réseau sont colorés en fonction du nombre de grappes qu'ils participant.

Les nœuds :

- correspondant à un seul cluster sont en rouge.
- avec plusieurs clusters sont en jaune.
- dont les valeurs sont aberrantes (nœuds qui ne se retrouvent dans aucun des groupes) sont en gris.

Le résultat est :

- Significatif s'il s'affiche en jaune (entre 0.05 et 0.1)
- Non-significatif s'il s'affiche en rouge (<0.05)

On examine aussi nos données avec l'algorithme MCODE. L'algorithme MCODE est une méthode automatisée bien connue pour trouver sous-graphes fortement interconnectés et trouver des meullière nœuds pour clustering en calculant de score. Clusters dans le réseau peuvent être considérés comme des complexes de protéines et de modules fonctionnels qui peuvent être identifiés comme étant des sous-graphes fortement interconnectés.

Cet algorithme est mis en œuvre par certains plugins Cytoscape tels que AllegroMCODE , MCODE et clusterMaker .

Conclusion

Un outil GOLIAT permettant de récupérer à partir d'une liste de gènes accompagnées de leur annotation GO, les GOSLIM associés a été développé. Cet outil a pour spécificité de prendre en entrée le fichier go.obo de gène ontology et peux donc fonctionné sur des informations actualisées. De plus il est capable d'utiliser tous les types de relations présentent dans gene ontology et ne

se limite pas uniquement à l'utilisation de la relation IS A. Il peut ainsi n'utiliser que les relations voulus par l'utilisateur.

Dans un second temps l'outil récupère tous les goslim rencontré et génère une sortie permettant pour chaque gènes de savoir quels goslim sont rencontrés ou non afin de pouvoir ensuite charger ce fichier pour des études ultérieurs à l'aide de R ou d'autres outils statistiques. A partir des GOSLIM les gènes peuvent être regroupés par GOSLIM afin de déterminer ceux qui sont surreprésentés.

Enfin l'outil GOLIAT propose un module de chargement des résultats GOSLIM dans le programme Cytoscape afin de pouvoir visualiser les GOSLIM et leurs relations.

Références

1. Nat Genet. 2000 May; 25(1):25-9.
   Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
   Ashburner M[1], Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.

2. Genome Res. 2003 Nov; 13(11):2498-504.
   Cytoscape: a software environment for integrated models of biomolecular interaction networks.
   Shannon P[1], Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T.

3. Python 3.5.1; Release Date: 2015-12-07. **Python 3.5.1 was released on December 6th, 2015**.https://www.python.org/downloads/release/python-351/

4. Bioinformatics. 2012 Aug 15; 28(16):2209-10. doi: 10.1093/bioinformatics/bts366. Epub 2012 Jun 27.
   GO-Elite: a flexible solution for pathway and ontology over-representation.
   Zambon AC[1], Gaj S, Ho I, Hanspers K, Vranizan K, Evelo CT, Conklin BR, Pico AR, Salomonis N.

5. BMC Bioinformatics.2003 Jan 13;4:2. Epub 2003 Jan 13.
   An automated method for finding molecular complexes in large protein interaction networks.
   Bader GD1, Hogue CW.

6. Eur Rev Med Pharmacol Sci. 2013 Mar;17(5):618-23.
   The PPI network and cluster ONE analysis to explain the mechanism of bladder cancer.
   Wan FC[1], Cui YP, Wu JT, Wang JM, -Z Liu Q, Gao ZL.

7. **Py2cytoscape is in beta and is installable from PyPI repository:** https://pypi.python.org/pypi/py2cytoscape

8. **Py2cytoscape source code**:https://github.com/idekerlab/py2cytoscape

9. **Gene Ontology Consortium** : http://geneontology.org/