Application du système GenFam à la réponse au stress des plantes : intégration de l'identification d'éléments cis spécifiques

Jonathan Lorenzo, Delphine Larivière, Jean-François Dufayard, Dominique This, Stéphanie Bocs

Application du système GenFam à la réponse au stress des plantes : intégration de l'identification d'éléments cis spécifiques

Jonathan Lorenzo*, Delphine Larivière, Jean-François Dufayard**, Dominique This***, et Stéphanie Sidibe-Bocs****

* Master STIC pour la santé, spécialité "Bioinformatique, Connaissances et Données"
** CIRAD, UMR AGAP ; *** Montpellier SupAgro, UMR AGAP:
TA-A 108/03. Avenue Agropolis. 34398 Montpellier cedex, France
jonathan.lorenzo@cirad.fr, delphine.lariviere@cirad.fr, jean-francois.dufayard@cirad.fr, dominique.this@supagro.fr, stephanie.sidibe-bocs@cirad.fr

JOBIM - juillet 2015

Résumé - *GenFam est un système intégratif d'analyse de familles de gènes. Ce système permet (i) de créer des familles de gènes de génomes complets, (ii) d'exécuter une analyse phylogénétique de cette famille à travers le gestionnaire de workflows Galaxy afin de définir les relations d'homologie, (iii) d'étudier des événements évolutifs à partir de blocs de synténie précalculées avec le workflow SynMap de la plateforme de génomique comparative (CoGe) et (iv) d'intégrer ces résultats dans l'interface de visualisation synthétique. La première application de GenFam est d'identifier des gènes candidats pour la tolérance aux stress environnementaux. Il nécessite de mettre en évidence la présence de séquences régulatrices cis spécifiques de la réponse aux stress (de type ABRE, DRE). Dans ce contexte, nous avons besoin d'intégrer de nouveaux outils afin de découvrir et chercher des sites de fixation de facteurs de transcription (Transcription Factor Binding Sites, TFBS) dans les séquences promotrices des gènes membre de la famille étudiée. Ce workflow Galaxy va, d'une part, sélectionner les régions flanquantes en 5' ou en 3' des gènes d'intérets selon le choix de l'utilisateur. D'autre part, les régions flanquantes sont analysées afin de découvrir et rechercher les motifs de séquences régulatrices cis spécifiques de la réponse aux stress avec des méthodes complémentaires comme MEME, STIF, PHYME. Ces résultats ainsi que l'annotation fonctionnelle des gènes étiquetés comme étant impliqués dans la réponse au stress seront intégrés dans l'interface de visualisation. Ce travail doit permettre une réflexion sur la notion d'orthologie fonctionnelle et effectuer une recherche translationnelle depuis les espèces modèles jusqu'aux espèces d'intérêt agronomique (i.e identifier des gènes candidats pour la réponse au stress du caféier à partir d'informations fonctionnelles connues chez Arabidopsis).*

Contact:

- jonath.lorenzo@gmail.com
- delphine.lariviere@cirad.fr

Liste des espèces disponibles
Arabidopsis thaliana
Brachypodium distachyon
Glycine max
Gossypium raimondii
Lotus japonicus
Medicago truncatula
Musa acuminata
Oryza sativa ssp. Indica
Oryza sativa ssp. Japonica
Populus trichocarpa
Solanum lycopersicum
Sorghum bicolor
Theobroma cacao
Vitis vinifera
Zea mays
Malus X domestica
Manihot esculenta
Ricinus communis
Setaria italica
Solanum tuberosum
Coffea canephora

Composition et analyse de familles de gènes - Premier *workflow* genfam.southgreen.fr

Petit jeu de séquences (<20 séquences)

Recherche de séquences d'intérêt - gènes connus d'une famille
Petit nombre de séquences, soigneusement choisies

Base de données Chado, Familles GreenPhyl automatiques

Panier de séquences ajout/suppression de séquences

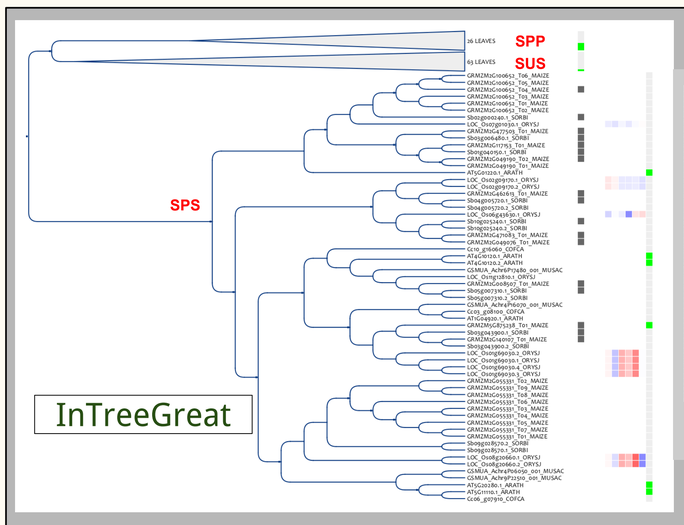
Recherche des gènes homologues dans un ensemble d' espèces
HMMer : HMM Build / HMM Search
(e.g. Coffea + Grapevine + Rice …)

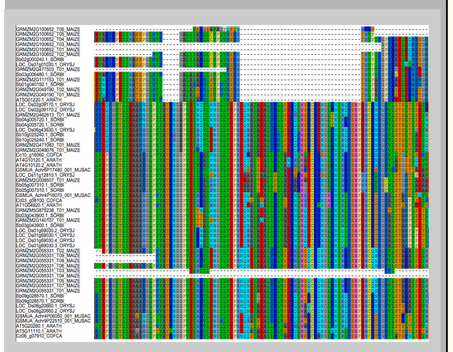
Séquences CoGe (Blast, Synténies)

Fichier de séquences (Toutes les familles)

Analyses Galaxy

Focus sur l'analyse des régions promotrices





Alignement en SVG

Intégration des données autour des familles de gènes

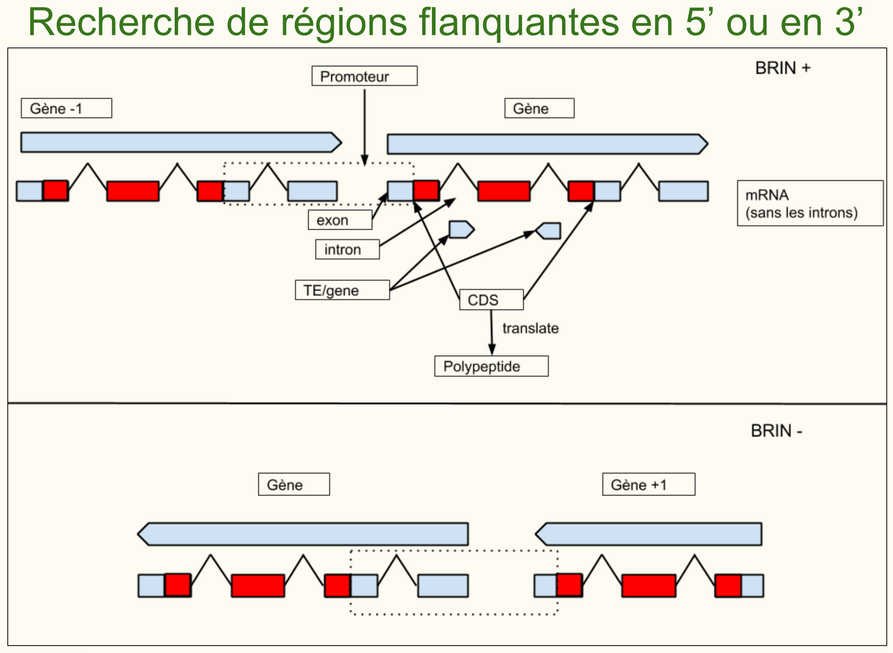
Visualisation synthétique

Exemple de la famille SPS

Génère une image SVG à partir d'un fichier CSV contenant la description des TFBS (position sur la région flanquante)

"Protein Domain Drawer"
Auteur: Jean-François DUFAYARD

Recherche de régions flanquantes en 5' ou en 3'



Sélection de la région flanquante
Développement spécifique en Perl et Python (multi_gff2fna.pl et multi_gff2fna.py)

Liste des cas particuliers:

Brin +
- ❑ premier gène du chromosome
- ❑ fin de l'élément précédent identique au début de notre gène
- ❑ gène à l'intérieur d'un gène ou d'un élément transposable
- ❑ gène(s) précédent(s) identique(s) et avec un cas normal
- ❑ gène(s) précédent(s) identique(s) et avec le cas du premier gène du chromosome
- ❑ gène(s) précédent(s) identique(s) et avec le cas du gène à l'intérieur d'un gène ou d'un élément transposable

Brin -
- ❑ dernier gène du chromosome
- ❑ gène à l'intérieur d'un gène ou d'un élément transposable précédent et dernier gène
- ❑ gène à l'intérieur d'un gène ou d'un élément transposable suivant
- ❑ gène à l'intérieur d'un gène ou d'un élément transposable précédent
- ❑ gène chevauchant l'élément précédent et suivant
- ❑ gène chevauchant l'élément précédent et fin du/des élément(s) suivant(s) inférieure à la fin de notre gène
- ❑ gène chevauchant l'élément précédent et cas normal pour l'élément suivant.

Utilisation des GFF3 et gènome des espèces présentes

=> BEDTOOLS

Séquences au format FASTA des régions flanquantes 3' ou 5' et de taille variable selon le choix de l'utilisateur.

Si:
taille du promoteur > intervalle entre le gène d'intérêt et le gène précédent ou suivant
alors:
l'outil récupère l'intervalle maximal entre ces derniers

Matrice de motif de TFBS (matrice poids/position, PWM)
Base de données: JASPAR, PLACE

La première ligne représente la position de chaque base du motif.
La valeur dans le tableau représente le nombre de fois où la base est présente dans un jeu de séquence.

Position	1	2	3	4	5	6	7	8
A	3	21	25	0	0	24	1	0
C	13	1	0	0	5	0	0	0
G	4	0	0	0	0	1	0	2
T	5	3	0	25	20	0	24	23

Exemple: HAT5
Source: JASPAR

Détection des sites de fixations des facteur de transcription (TFBS) (e.g. PlantPAN)

Développement spécifique en Perl et C++
(pwm_genome_map.pl)
Utilisation de la librairie MOODS

Fichier de sortie:
fichier CSV contenant les positions de chaque élément cis

Conclusion

Le système GenFam permet une histoire de l'évolution des familles de gènes (duplications et évènements de spéciation) par des analyses plylogénétiques, des études de synténies, et leur mise en relation avec des indices fonctionnels (annotations, profiles d'expression, structure des promoteur, etc). Ces analyses sont réalisées par l'intégration de diverses sources de données, ainsi que le développement de pipelines d'analyses destinées à l'analyse de familles de gènes. Les résultats des pipelines d'analyse sont représentés grâce à une visualisation synthétique permettant une vision globale des informations disponibles pour une famille. L'analyse de l'histoire évolutive d'une famille permet au chercheur d'identifier des gènes candidats dans des espèces non modèles pour des gènes impliqués dans la tolérance aux stress, utilisant l'identification d'orthologies fonctionnelles. L'identification des séquences promotrices et terminatrices présentée dans ce poster s'intègre dans l'outil GenFam et permet d'apporter des indices fonctionnels complémentaires. Suite à l'implémentation de ces outils, l'ajout d'outils différents permettront de compléter ce travail tels que des méthodes de recherches de motif utilisant des HMM comme le programme STIF mais aussi des méthodes de découvertes comme le programme MEME.

Références

J. Korhonen, P. Martinmäki, C. Pizzi, P. Rastas and E. Ukkonen. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* 25(23), pages 3181-3182. (2009)

C. Pizzi, P. Rastas and E. Ukkonen: Finding Significant Matches of Position Weight Matrices in Linear Time. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 8(1), pages 69 - 79. (2011)

Quinlan, Aaron R., and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26.6 (2010): 841-842.

MATYS, Vea, FRICKE, Ellen, GEFFERS, R., *et al.*TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 2003, vol. 31, no 1, p. 374-378.

PORTALES-CASAMAR, Elodie, THONGJUEA, Supat, KWON, Andrew T., *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research*, 2009, p. gkp950.

Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research*,*34*(suppl 2), W369-W373.

Naika, M., K. Shameer, O. K. Mathew, R. Gowda, and R. Sowdhamini. "STIFDB2: An Updated Version of Plant Stress-Responsive TranscrIption Factor DataBase with Additional Stress Signals, Stress-Responsive Transcription Factor Binding Sites and Stress-Responsive Genes in Arabidopsis and Rice." *Plant and Cell Physiology* 54, no. 2 (12, 2013): E8. doi:10.1093/pcp/pcs185.

CHANG, Wen-Chi, LEE, Tzong-Yi, HUANG, Hsien-Da, *et al.* PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. *BMC genomics*, 2008, vol. 9, no 1, p. 561.

Sundar, Ambika Shyam, Susan Mary Varghese, Khader Shameer, Nataraja Karaba, Makarla Udayakumar, and Ramanathan Sowdhamini. "STIF: Identification of Stress-upregulated Transcription Factor Binding Sites in Arabidopsis Thaliana." *Bioinformation* 2, no. 10 (12, 2008): 431-37. doi:10.6026 /97320630002431.

Higo, K., Y. Ugawa, M. Iwamoto and T. Korenaga (1999) Plant cis-acting regulatory DNA elements (PLACE) database. Nucleic Acids Res. 27 (1): 297-300.

Participation financée par le Labex NUMEV
http://www.lirmm.fr/numev/

Master Sciences & Numérique pour la Santé
- maurice.hayot@umontpellier.fr
- pascal.poncelet@umontpellier.fr
- claudie.fabry@umontpellier.fr
- http://www.telecom-montpellier.fr/

Spécialité« Bioinformatique, Connaissances, Données »
- alban.mancheron@umontpellier.fr
- anna-sophie.fiston-lavier@umontpellier.fr
- nicolas.molinari@umontpellier.fr
- http://www.lirmm.fr/BCD/