Le portail bioinformatique du genre *Helianthus*: Heliagene

***Thibaut Hourlier*[1]**, David Rengel[1], Nicolas Langlade[1], Patrick Vincourt[1], Jérôme Gouzy[1], Sébastien Carrere[1]

[1] Laboratoire des Interactions Plantes-Microorganismes INRA/CNRS

Thibaut.Hourlier@toulouse.inra.fr, Sebastien.Carrere@toulouse.inra.fr, Jerome.Gouzy@toulouse.inra.fr

http://www.heliagene.org

Le portail bioinformatique Heliagene permet de rapidement visualiser les caractéristiques d'un cluster d'ESTs, d'explorer les fonctions des gènes, d'analyser les gènes et les familles protéines et de rechercher des SNPs potentiels à partir de polymorphismes intra et inter spécifiques. Il utilise et propose des web-services et workflows BioMoby.

Résumé

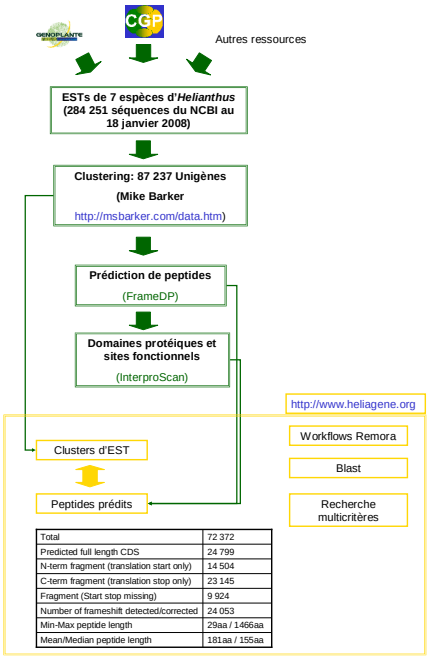
Grâce d'une part à sa capacité d'adaptation aux environnements pauvres en eau des régions du sud de l'Europe et d'autre part à son potentiel de production de matériel pour les bio-carburants, l'espèce de tournesol *Helianthus annuus* est amenée à occuper une place de plus en plus importante parmi les plantes cultivées pour la production de bio-carburants de première génération. A l'heure actuelle, la séquence du génome n'est pas encore connue mais une quantité conséquente d'EST de sept espèces *Helianthus* est disponible dans les banques publiques (284,251 au NCBI à la date du 18 Janvier 2008).

Afin de permettre l'exploitation de ces premières ressources de séquences, nous avons développé le portail HeliaGene dont le système de navigation permet (i) de rapidement visualiser les caractéristiques des clusters d'EST, (ii) d'explorer les fonctions des gènes, (iii) d'analyser les gènes et les familles de protéines, (iv) de rechercher des SNP potentiels à partir de polymorphisme intra et inter spécifique. Aussi, la première analyse à grande échelle des données disponibles a été la prédiction de régions codantes à partir des clusters d'EST; prédiction rendue délicate du fait d'une part de l'hétérogénéité de ces derniers en terme de profondeur de couverture et d'autre part à cause du polymorphisme induit par l'utilisation des séquences de 7 espèces pour la génération des séquences consensus. Cet assemblage « multi-espèces » sert de référence à la communauté car il a été utilisé comme matrice pour la génération de la puce affymetrix « tournesol ». C'est pourquoi nous avons utilisé le programme FrameDP[1] particulièrement adapté à la prédiction de CDS sur des données bruitées. Ainsi, à partir de 72372 séquences consensus de clusters d'EST, FrameDP a prédit 24799 CDS pleine taille, mais également 47573 peptides correspondants à des fragments de protéines. InterProScan a été utilisé pour analyser l'ensemble de ces peptides afin d'en déterminer la composition en domaines et sites fonctionnels. A partir de cette analyse factuelle et en se basant sur la hiérarchie de la base de données InterPro, nous avons déterminé le plus petit ancêtre commun aux différents domaines détectés afin de proposer une fonction pour le peptide ainsi qu'une classification fonctionnelle basée sur la « Gene Ontology ». L'annotation des peptides a été propagée au niveau du cluster d'EST pour compléter l'annotation fonctionnelle automatique.

Le portail héberge les résultats obtenus pour différents projets actuels (par exemple, Sunyfuel, Oléosol,…). Une authentification par login/mot de passe modifie l'affichage des onglets et par conséquent les informations disponibles pour l'utilisateur. Différents outils ayant une licence non libre, tel que POPPI (Pipeline for Orthology-based Primer PIcking, http://www.heliagene.org/cgi/POPPI.cgi) sont accessibles depuis les pages protégées pour les utiliser directement sur les données du genre *Helianthus*.

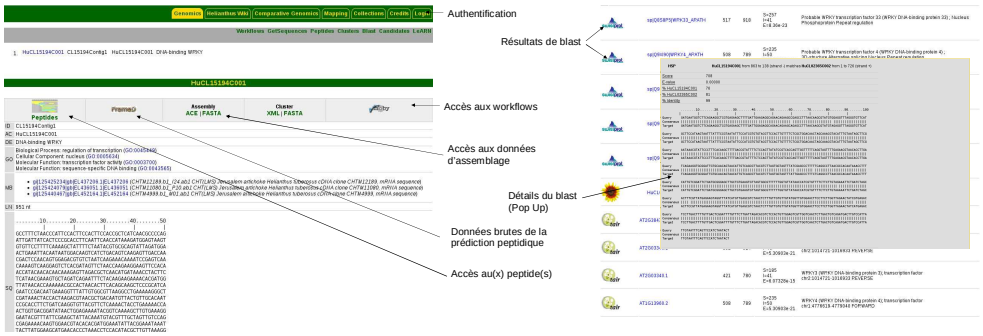
Ressources du portail

Préparation des données

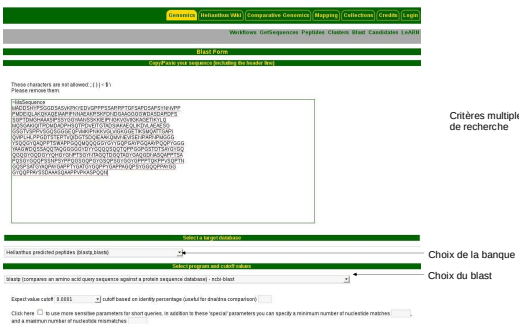


Total	72 372
Predicted full length CDS	24 799
N-term fragment (translation start only)	14 504
C-term fragment (translation stop only)	23 145
Fragment (Start stop missing)	9 924
Number of frameshift detected/corrected	24 053
Min-Max peptide length	29aa / 1466aa
Mean/Median peptide length	181aa / 155aa

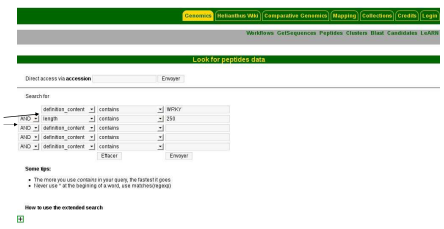
Fiche d'un cluster



Formulaire Blast



Formulaire de recherche



Stockage et accès aux données

Résultats au format XML

Les données (clusters d'EST et peptides) ainsi que les analyses (blast, iprscan) sont structurées au format XML. Nous utilisons le moteur d'indexation CLucene (implémentation C++ de Lucene) afin d'indexer et stocker les fichiers XML. Pour ce faire, nous avons développé une suite de programmes Perl qui permettent à partir d'une description au format XPath d'indexer n'importe quel fichier XML, de requêter ces index et de générer un site web dynamique intégrant un formulaire de recherche multicritères. Nous avons packagé ces scripts sous le nom de EZLucene (http://lipm-bioinfo.toulouse.inra.fr/download/EZLucene) et étendu les fonctionnalités afin d'être capable d'analyser n'importe quel fichier texte à partir d'une description basée sur des expressions régulières Perl.

Résultats bruts

Les résultats d'analyses bruts, tels que les images de FrameD, sont quant à eux compressés et stockés dans une BerkeleyDB afin d'une part de limiter le nombre et la taille des fichiers à gérer sur le système de fichier et d'autre part de permettre un accès direct aux fichiers de données.

Web-services BioMoby

Nous avons déployé des web-services d'accès aux ressources publiques de HeliaGene en utilisant le framework que nous avons développé, PlayMoby. Ces web-services BioMoby[2] sont accessibles via le portail Mobyle[3] du LIPM en plus des outils tels que Remora[4], mais aussi Taverna[5]. En effet, nous avons conçu PlayMoby afin de rapidement déployer des web-services BioMoby à partir de programmes utilisés en ligne de commande. Pour ce faire, nous nous appuyons sur une description des programmes au format Mobyle qui est alors utilisé comme format pivot. De plus, chaque web-service instancié par PlayMoby est déployé avec un jeu de test qui permet au programme de surveillance de contrôler la disponibilité et la stabilité des résultats (« tests fonctionnels »), avec in fine la production de rapports XML et RSS mais surtout émission d'emails lors d'une non-conformité avec le résultat attendu.

Workflows

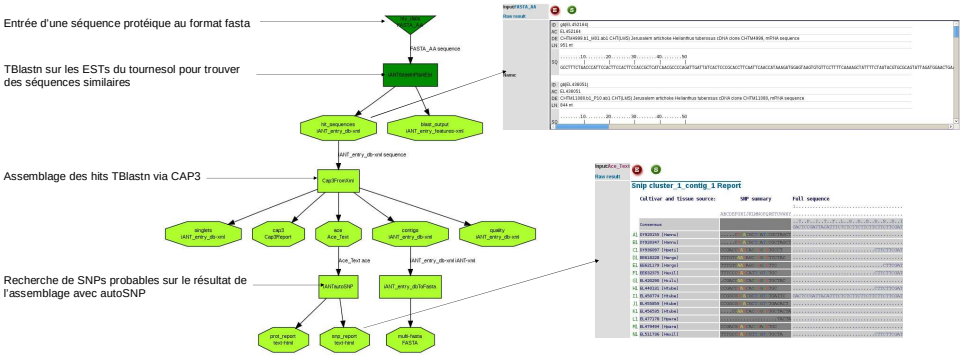
Worflows présents sur Heliagene

Les workflows ou chaînes de traitement automatisé sont des outils collaboratifs à plusieurs niveaux. Premièrement, ils s'avèrent être le support de communication naturel pour la description et la réalisation d'une demande d'un biologiste. De plus, les « workflows » peuvent être vus comme un protocole expérimental permettant au biologiste non seulement d'assurer la sauvegarde de son analyse pour garantir la traçabilité, mais aussi sa ré-exécution et sa personnalisation. Ils peuvent également être partagés entre biologistes travaillant sur les mêmes problématiques et ainsi profiter de l'expérience de la communauté. Lorsque ces « workflows » sont construits sur la technologie des web-services, il devient alors possible d'intégrer des données ou des outils hébergés par d'autres laboratoires. Cela permet une fois encore de profiter de l'expertise et de ressources bioinformatiques extérieures.

Chacun des « workflows » proposé permet de répondre à une question posée par un biologiste:

- Recherche et alignement de protéines similaires dans d'autres plantes à partir de la séquence consensus d'un cluster d'EST du tournesol
- Recherche de motifs protéiques à partir de la traduction dans les 6 phases de lecture d'une séquence nucléique
- Transfert d'information à partir des données de transcriptome de la plante modèle *Arabidopsis thaliana*
- Recherche du « reciprocal best hit » chez le tournesol à partir d'une protéine d'intérêt chez *A.Thaliana*
- Recherche de SNPs dans les régions codantes (intra ou inter spécifique) a partir d'une protéine d'intérêt utilisée pour trouver des ESTs homologues chez *Helianthus*

Workflow Remora de recherche de SNPs



Bibliographie

[1] Gouzy J, Carrere S, Schiex T.,FrameDP: sensitive peptide detection on noisy matured sequences., Bioinformatics. 2009 Jan 19.
[2] BioMoby Consortium, Interoperability with Moby 1.0--it's better than sharing your toothbrush! Brief Bioinform. 2008 May;9(3):220-31.
[3] Bertrand Néron, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Pierre Tufféry, Catherine Letondal, Mobyle: a new full web bioinformatics framework., (2008). Bio Open Source Conference (BOSC), Toronto.
[4] Carrere S. and Gouzy J. (2006). REMORA: a pilot in the ocean of BioMoby web-services *Bioinformatics* 22(7):900-1.
[5] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., and Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20, 3045-3054.