Cyanorak v2 : système d'information pour l'annotation d'orthogroupes de cyanobactéries

http://abims.sb-roscoff.fr/cyanorak

Loraine GUÉGUEN[1], Gregory FARRANT[1,2], Gildas LE CORGUILLÉ[1], Erwan CORRE[1], Wilfrid CARRÉ[1], Laurence GARCZAREK[2], Frédéric PARTENSKY[2], Christophe CARON[1] and Mark HOEBEKE[1]

[1] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France, {loraine.gueguen, gildas.le-corguille, erwan.corre, wilfrid.carre, christophe.caron, mark.hoebeke}@sb-roscoff.fr
[2] Equipe Procaryotes Photosynthétiques Marins, UMR7144 CNRS et UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France, {gregory.farrant, laurence.garczarek, frederic.partensky}@sb-roscoff.fr

Contexte : **Cyanorak v2 est un système d'information dédié à l'annotation de génomes de cyanobactéries marines. En plus des fonctionnalités d'annotation courantes, il permet l'automatisation et l'assistance à la curation de *clusters* de protéines orthologues, et est capable de retracer de manière très détaillée les modifications apportées à la base de données. Il a été conçu pour importer de nouveaux génomes et réaliser une réaffectation automatique des annotations manuelles des *clusters* de la version précédente en utilisant un jeu de règles tenant compte du degré d'inclusion des nouvelles séquences dans les *clusters* précédents.**

1. Présentation de Cyanorak v2

Description	Outil destiné à l'annotation d'orthogroupes de cyanobactéries marines (succède à Cyanorak v1 [1])
Données	- Séquences et annotations de 33 souches de cyanobactéries marines. - Ensemble de *clusters* d'orthologues regroupant les protéines en fonction de leur similarité de séquence.
Fonctionnalités principales	- Consultation et recherche - Extraction de données - Exportation de groupes de séquences au format FASTA, exportation des séquences annotées des contigs au format GenBank - Avec contrôle d'accès : édition des données (modification des annotations des *clusters*, redistribution manuelle du contenu des *clusters*)
Problématiques	- Transfert automatique des informations associées aux *clusters* (contenu en gènes et annotations) lors de l'ajout de nouveaux génomes. - Conservation de l'historique détaillé des modifications (accès public à des « instantanés » / accès restreint pour modifications par la communauté des curateurs).

2. Architecture du système d'information

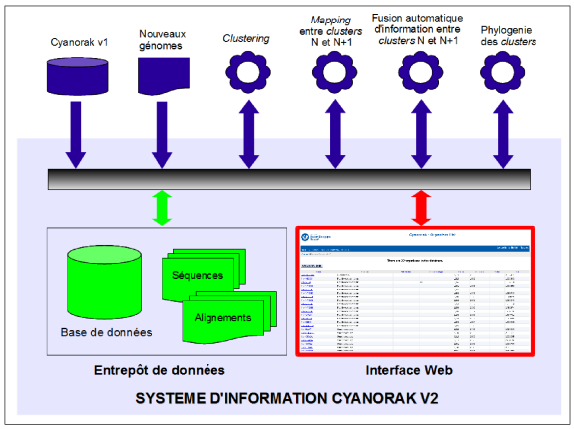
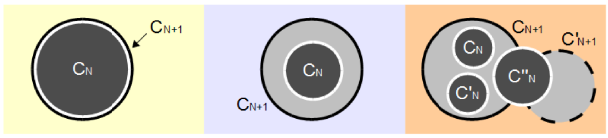


Schéma de l'architecture du système d'information Cyanorak v2 :
- une base de données relationnelles et un entrepôt non structuré,
- des modules complémentaires dédiés aux différentes étapes d'importation de données et de calcul,
- une application Web.

3. Clustering et transfert des annotations

Passage d'une version de Cyanorak à la suivante (ici de Cyanorak v1, 14 génomes, à Cyanorak v2, 33 génomes) et **transposition des informations** (annotations) :

- Étape 1 : Construction d'un nouvel ensemble de *clusters* d'orthologues avec OrthoMCL [2].
- Étape 2 : **Classification des liens entre *clusters* produits par deux *clusterings* successifs**, basée sur le recouvrement relatif de leurs contenus en gènes.
- Étape 3 : Définition d'un jeu de règles pour **mettre en correspondance les *clusters* entre deux versions de *clustering*** (N et N+1).



SINGLE EXACT-MATCH : le *cluster* N+1 est identique au *cluster* N.

Seul le *cluster* N est conservé avec ses annotations

SINGLE SUPERSET : le *cluster* N+1 est un sur-ensemble d'un unique *cluster* N.

Les annotations sont transférées de N vers N+1 avec un niveau de confiance élevé.

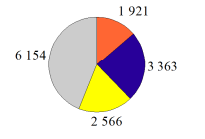
MIX : les protéines sont partagées entre plusieurs *clusters* N et N+1.

Le transfert automatique d'annotations n'est pas possible. Les protéines sont toutes regroupées dans un même *cluster*.

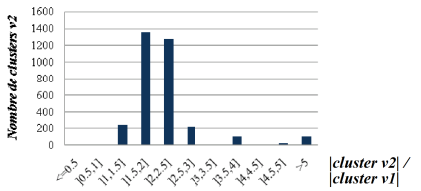
Le curateur, aidé de l'arbre phylogénétique des protéines choisira au cas par cas la validation ou l'éclatement du *cluster*, ainsi que la réaffectation ou non des annotations des *clusters* N.

4. Premiers résultats

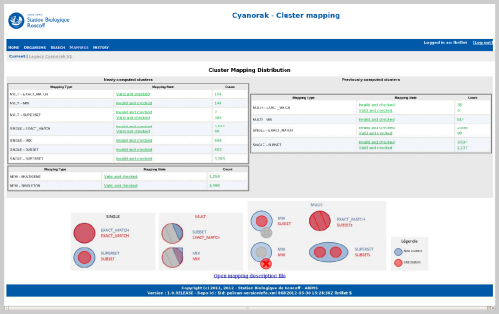
Clustering des 33 génomes pour Cyanorak v2 (14 004 clusters au total)



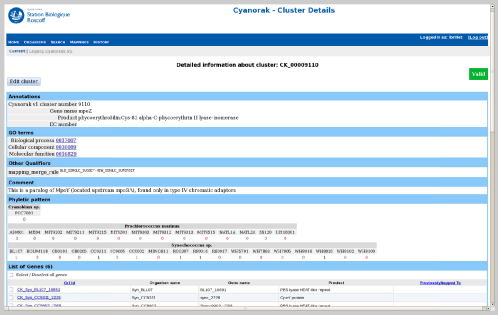
Distribution du rapport (nb gènes v2 / nb gènes v1) pour les clusters de type SINGLE SUPERSET



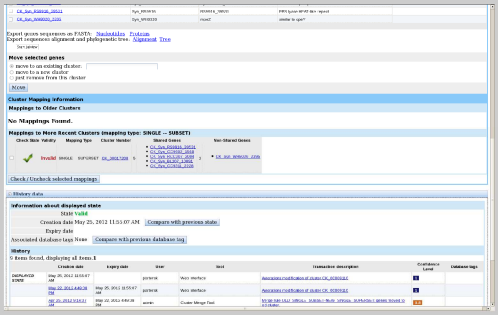
La majorité des clusters v2 contiennent 1.5 à 2.5 fois plus de gènes que les clusters v1. Cette information est importante pour estimer la validité du transfert des annotations des clusters v1 vers les clusters v2.



Page décrivant la classification des liens entre les clusters v1 et v2 qui a permis de définir des règles de réaffectation des annotations entre versions.



Exemple de page décrivant un cluster (haut).



Exemple de page décrivant un cluster (bas).

Conclusion :

Cyanorak v2 constitue un outil précieux pour l'annotation des génomes de cyanobactéries marines. Il permet d'une part la gestion des annotations des gènes et des groupes d'orthologues préexistants, tout en facilitant l'import et l'annotation de nouveaux génomes. Par ailleurs, il inclut plusieurs outils d'aide à la décision pour la gestion des groupes d'orthologues, tels que des règles de *mapping* entre versions et des arbres phylogénétiques représentant les séquences de chaque orthogroupe.

Perspectives :

- Ajout de 23 nouveaux génomes de cyanobactéries marines (projet METASYN, Genoscope)
- Application Web : représentation des gènes dans leur contexte génomique
- Plugin pour la fusion d'information lors de l'import de fichiers GenBank

[1] A. Dufresne, M. Ostrowski, D. J. Scanlan, L. Garczarek, S. Mazard, B. P. Palenik, I. T. Paulsen, N. T. de Marsac, P. Wincker, C. Dossat, S. Ferriera, J. Johnson, A. F. Post, W. R. Hess, F. Partensky. Unraveling the genomic mosaic of a ubiquitous genus of cyanobacteria. *Genome Biology, 9:R90, 2008.*

[2] C. Habib, E. Le Corguillé, E. Corre, L. Brillet, M. Hoebeke, W. Carré, L. Garczarek, F. Partensky, C. Caron. PELICAN: Orthologous Groups and Gene Lateral Transfers for Comparative Genomic Analysis of Marine Cyanobacteria. *Actes JOBIM 2011.*

Le développement de Cyanorak v2 se fait dans le cadre de l'ANR génomique microbienne PELICAN : ANR-PCS-09-GENM-200.