Génomique Comparative
« NARCISSE: a mirror view of conserved syntenies »

Thomas Faraut[1] & Jérôme Gouzy[2]; Yoann Beausse (IE CDD Genopole Toulouse puis EADGENE)

[1] Laboratoire de Génétique Cellulaire INRA - Dep^t de Génétique Animale
[2] Laboratoire des Interactions Plantes-Microorganismes INRA/CNRS - Dep^t Santé des Plantes et Environnement
Thomas.Faraut@toulouse.inra.fr, Jerome.Gouzy@toulouse.inra.fr
http://bioinfo.genopole-toulouse.prd.fr/narcisse/

Introduction: la démarche comparée, qui est depuis l'origine au centre du processus de découverte en biologie, se nourrit désormais de la connaissance des séquences de génomes modèles ou d'intérêts. En accompagnement de ces projets génomes, il est devenu indispensable de maîtriser ou de mettre au point des méthodes de comparaison des séquences génomiques. Cette maîtrise des méthodes doit s'accompagner de la mise à disposition d'outils bioinformatiques permettant aux biologistes de structurer et de représenter les informations de conservation. C'est avec cet objectif que nous avons récemment initié le développement d'une plate-forme multi-génomes, appelé NARCISSE, d'intégration et de représentation des conservations. L'ambition de ce projet est de proposer, à partir de la plate-forme bio-informatique du Génopole Toulouse-Midi Pyrénées, des méthodes et des outils de comparaison de génomes adaptés à la génomique, qu'elle soit végétale, animale ou bactérienne .

Comparaison des génomes ou comment explorer le niveau de conservation des génomes

① La comparaison des séquences génomiques d'espèces apparentées permet d'identifier les remaniements chromosomiques à travers les différentes étapes de l'évolution.

Translocation réciproque

Fusion/Fission

Fusion+Inversion

Transposition

② Dotplot: au niveau du nucléotide la similitude peut se représenter par un graphique d'une fonction indicatrice

$$f(x,y) = \begin{cases} 1 & \text{si la propriété } P_{x,y} \text{ est vraie} \\ 0 & \text{sinon} \end{cases}$$

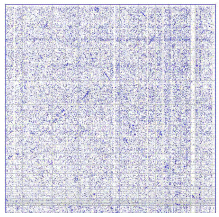
	G	A	T	C	A	A	C
G	•						
T			•				
T			•				
C				•			•
A		•			•	•	
G	•						
C				•			•

③ Cette mesure de similitude au niveau du nucléotide peut se généraliser au niveau d'un segment d'ADN de taille quelconque et l'on peut ainsi envisager à terme :

- d'établir une mesure de similitude ou de dissimilitude (distance) entre génomes
- de calculer cette distance
- de proposer un scénario évolutif

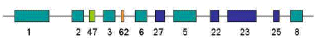
Afin de calculer les segments similaires nous utilisons le programme Pash (Kalafus KJ et al, *Genome Research* 2004) qui permet de calculer en temps « raisonnable » (7jours sur 20 processeurs pour la comparaison G.gallus/H.sapiens) le dotplot des similarités entre deux génomes complètement séquencés.

④ Dotplot de comparaison de *G. Gallus* (axe Y) vs *H. sapiens* (axe X)



⑤ L'information de la matrice de « dotplot » peut, pour chaque chromosome d'une espèce que l'on va utiliser comme cible, être recodée sous forme d'une liste chaînée (figure ⑥) ou les éléments de conservation sont caractérisés par un numéro unique représentant le segment et par un deuxième attribut qui représente le chromosome de l'espèce de référence portant ce segment (cet attribut est représenté par une couleur sur la figure ⑥)

Figure ⑥

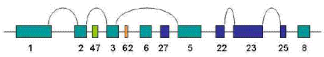


⑦ L'analyse de cette liste chaînée permet d'identifier la structure de conservation sur plusieurs niveaux.

Afin d'analyser cette liste nous utilisons le principe de plus grande sous-séquence croissante (LIS) qui permet d'identifier les plus longs segments conservés. Ainsi sur la figure ⑧ on peut identifier le segment 1,2,3, 5 partagé avec le chromosome « vert » et le segment 22,23,25 partagé avec le chromosome « bleu ».

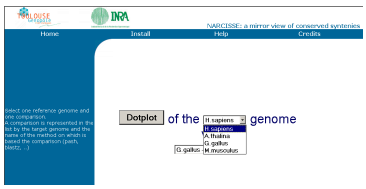
Cette démarche nous permet de détecter simplement inversions et transpositions. Ainsi en fonction du niveau de zoom auquel on se place on pourra observer les remaniements internes d'un segment conservé.

Figure ⑧

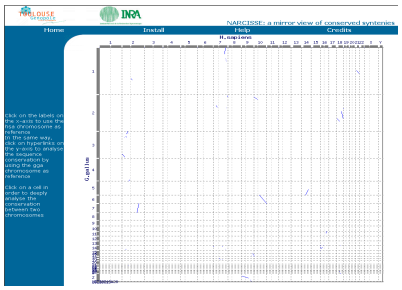


L'environnement de navigation et d'intégration « NARCISSE »

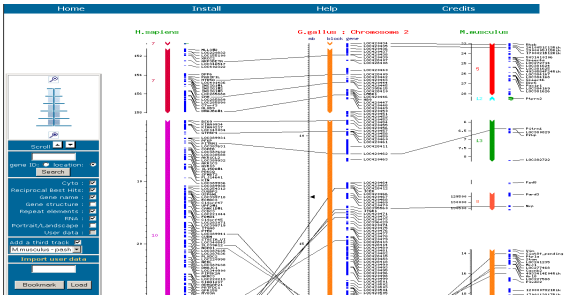
① http://bioinfo.genopole-toulouse.prd.fr/narcisse/



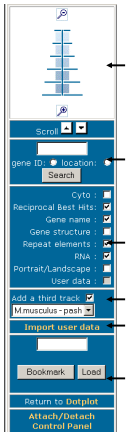
② Une fois la sélection des deux organismes effectuée, le Dotplot présente une vision macroscopique des conservations (ici entre *Hsa* et *Gga*)



③ Le système de navigation proposé permet de cliquer/zoomer sur les zones d'intérêts, de comparer une séquence de référence à 2 autres génomes, de représenter les « reciprocal best hits » , de moduler les objets à représenter en fonction du niveau de zoom, de changer de référence, etc…



Détail de la « télécommande » de Narcisse



5 Niveaux de zoom prédéfinis et paramétrables (chromosome, bras, segment de synténie, gène, structure des gènes)

Accès direct par position ou par identifiant (sur l'organisme de référence)

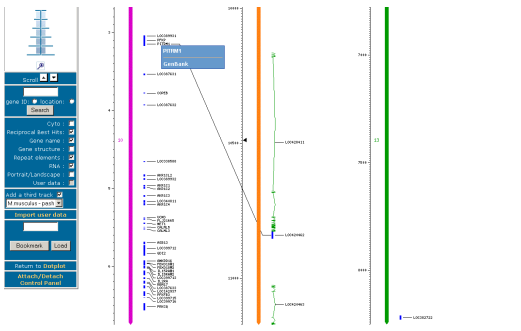
Panneau de contrôle permettant de sélectionner les éléments à afficher. Les valeurs par défaut dépendent du niveau de zoom considéré. Ces valeurs sont paramétrables via un fichier de configuration et l'on pourra ainsi facilement paramétrer l'outil afin de représenter aussi bien des comparaisons de l'ordre du gigabases (homme, porc, poulet, souris, etc..) que des comparaisons de génomes bactériens.

Liste de sélection de la 3eme piste. La comparaison cible est identifiée par un nom d'espèce (ici *Mms*) et de la méthode utilisée pour la comparaison (ici pash) . Cela signifie que l'on peut présenter simultanément les résultats obtenus sur la même cible avec 2 méthodes différentes (Pash ou blastz par exemple)

Possibilité d'importer des données utilisateurs afin de superposer des données privées aux données de narcisse (via un fichier tabulé)

Possibilité de sauvegarder des « vues » via un système de signets.

④ Le niveau actuel de granularité le plus fin permet d'accéder à la structure intron/exon des gènes ainsi qu'aux liens hypertextes vers les banques de données proposant des informations plus précises sur le gène considéré (ci dessous un lien vers GenBank)



Perspectives de développement

- représenter les différents niveaux de conservation (figure 1)
- développer le système de gestion et d'interrogation (via le système SRS installé sur la plate-forme du Génopole Toulouse-Midi Pyrénées)
- mettre en place les outils pour faciliter l'analyse de familles d'intérêts (ex: gènes de résistance)
- intégrer les données d'expression (séquences de type EST ainsi que les données issues des micro/macro arrays)
- développement d'un pipeline de détection des ARN non codant pour des protéines et intégration dans Narcisse
  Coll. C. Gaspin (Unité de Biométrie et d'Intelligence Artificielle, INRA Toulouse)
- développer/paramétrer la version "bactérie" ainsi qu'une version permettant d'exploiter les génomes en cours de séquençage.
- mettre à disposition, sur la plate-forme du Génopole, l'ensemble des comparaisons de génomes pertinentes pour les chercheurs de l'institut

Figure 1: prototype de représentation simultanée de différents niveaux de conservation