Insyght: Using symbols to visualize homologies, conserved syntenies and genomic insertions across multiple genomes

Thomas Lacroix, Valentin Loux, Annie Gendrault-Jacquemard Gendrault, Mark M. Hoebeke, Jean-François Gibrat

Insyght: Using symbols to visualize homologies, conserved syntenies and genomic insertions across multiple genomes

Thomas Lacroix[1], Valentin Loux[1], Annie Gendrault[1], Mark Hoebeke[2] and Jean-François Gibrat[1]

[1] Mathématique, Informatique et Génome, INRA, Jouy-en-Josas, 78352 France
{thomas.lacroix, valentin.loux, annie.gendrault, jean-francois.gibrat}@jouy.inra.fr

[2] CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France
mark.hoebeke@sb-roscoff.fr

Abstract *Insyght proposes a new way to explore the landscape of conserved and idiosyncratic genomic regions across multiple genomes and their rearrangements throughout evolution. Its unique display consists of a symbolic representation tightly integrated with a proportional view. The symbols highlight a region of interest and provide legibility while the proportional view simultaneously allows grasping genomic locations and complex rearrangements scattered across the genomes and occurring at different scales. A second type of display is dedicated to the analysis of the presence, absence, or multiple copies of a given set of homologs. A functionality based on filters has been implemented to facilitate the retrieval of genes of interest and allow the formulation of relevant biological questions, such as finding niche-specific or core genome genes that match a few particular functions or biological processes. Our public dataset currently consists of 389 prokaryotes genomes. Alternatively, a virtual machine can be downloaded and installed locally to visualize private data. It contains a pre-installed version of the pipeline, database and visualisation tool. Insyght is suitable for a variety of analyses: genome-wide inference of gene function, detection of evolutionary events, phylogenetic profiling and investigation of the core genome or niche-specific genes. It is freely available at http://genome.jouy.inra.fr/Insyght/*

Keywords homology browser, conserved synteny, multi-genome visualisation

1 Introduction

Les génomes subissent des réarrangements de différentes natures lors des processus évolutifs: translocation, duplication, fusion, etc... D'un point de vue de la génomique comparée, chaque génome peut être considéré comme une succession de régions conservées intercalées par des régions idiosyncratiques. Les technologies de séquençage haut débit fournissent une grande quantité de données aux biologistes qui ont besoins d'outils pour les aider à annoter les gènes de manière efficace et rapide à l'échelle du génome. La conservation de l'ordre des gènes peut permettre d'assigner les fonctions d'un ensemble de gènes simultanément ou fournir des indices concernant la fonction des protéines hypothétiques [1,2]. De nombreux outils d'exploration de synténies existent, et les paradigmes de visualisation dans ce domaine sont variés: le dot plot [3,4,5,6], l'idéogramme [7,8], la vue en trapézoïde [9,10,11], ou la représentation symbolique [12,13,14]. Parmi les défis posés par l'analyse des synténies et homologues, on peut noter la visualisation des réarrangements qui sont répartis sur les génomes et se produisent à différentes échelles, ou la navigation parmi une quantité de donnée abondante et multi-dimensionnelle (coordonnée génomique, plusieurs génomes comparés, plusieurs homologues par comparaison,...).

2 Design et fonctionnalités principales

Insyght est un outil de visualisation qui permet d'analyser les homologies, les synténies conservées et les régions génomiques idiosyncratiques à l'échelle de plusieurs organismes. Sa caractéristique principale est l'association d'une représentation symbolique (Figure 1-A) à une représentation proportionnelle (Figure 1-B). Cette combinaison originale de paradigme visuel facilite la navigation exhaustive de données d'homologies complexes. L'utilisateur peut interagir avec divers symboles qui représentent les évènements évolutifs:

homologues, synténies, régions génomiques insérées. Ces symboles sont étroitement intégrés avec une vue proportionnelle où les mêmes évènements sont représentés selon leurs coordonnées génomiques. Dans la vue proportionnelle, les régions génomiques homologues sont jointes par des trapézoïdes. La représentation symbolique améliore la lisibilité parmi le grand nombre d'événements évolutifs et permet de naviguer parmi les multiples copies d'homologues. La représentation proportionnelle permet de localiser les réarrangements complexes dispersés dans le génome et se produisant à différentes échelles. L'utilisateur peut interagir avec les symboles à l'aide d'un menu contextuel pour, par exemple, afficher ou masquer les gènes d'une synténie ou trouver des gènes d'intérêt. Le zoom et la navigation peuvent être synchronisés entre les résultats ce qui permet d'analyser plusieurs génomes en parallèle.

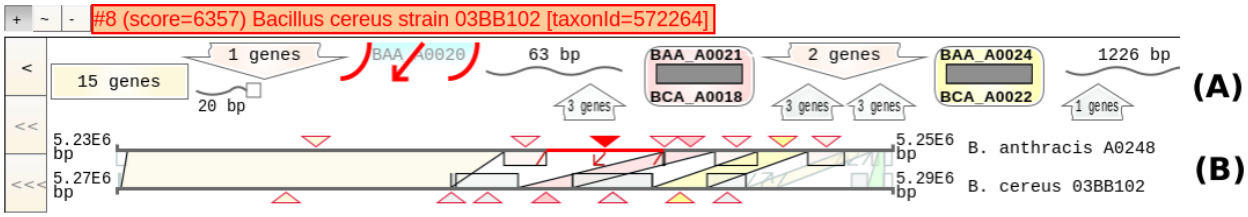

Figure 1. Représentation symbolique (A) et proportionnelle (B) de la vue d'organisation génomique.

Une deuxième vue est consacrée à l'analyse exhaustive des homologies d'un jeu de gènes d'intérêt. Une fonctionnalité de recherche par combinaison de filtres (Figure 2) a été implémentée pour faciliter la constitution de groupes de gènes significatifs d'un point de vue biologique. Les opérateurs booléens logiques d'intersection (AND) ou d'union (OR) permettent de combiner différents types de filtres: présence / absence d'homologues, coordonnées génomiques, identifiants, fonctions, processus biologiques, produits, localisation cellulaire, ou numéro EC. Par exemple, il est possible de formuler des requêtes combinées qui permettent de trouver les gènes spécifiques ou partagés au sein d'une espèce correspondant à un processus biologique particulier. La vue d'analyse des homologues ressemble à un tableau où les colonnes sont les gènes sélectionnés et les lignes sont les espèces comparées (Figure 3). Ainsi l'utilisateur peut visualiser la présence, l'absence, ou les multiples copies d'homologues et détecter par exemple les espèces avec des pertes de fonction ou des familles de gènes abondantes. Les gènes sont colorés en fonction de la synténie à laquelle ils appartiennent. D'autre fonctionnalités ont été implémentées, tel que trier le tableau de résultat selon divers critères et échelles ou visualiser l'emplacement des gènes sur les génomes. Les deux vues, organisation génomique et tableau des homologues, sont interconnectées et il est possible de passer de l'une à l'autre.

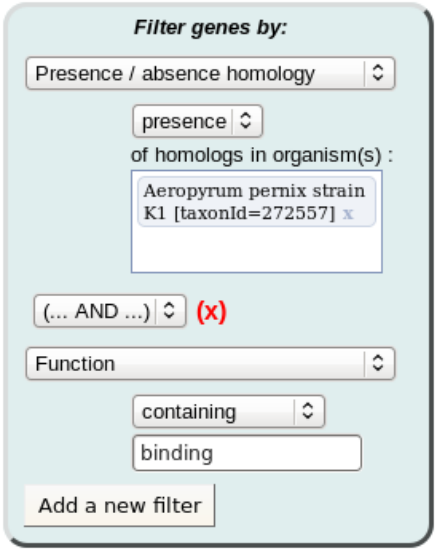

Figure 2. Fonctionnalité de recherche de gènes (combinaison de filtres).

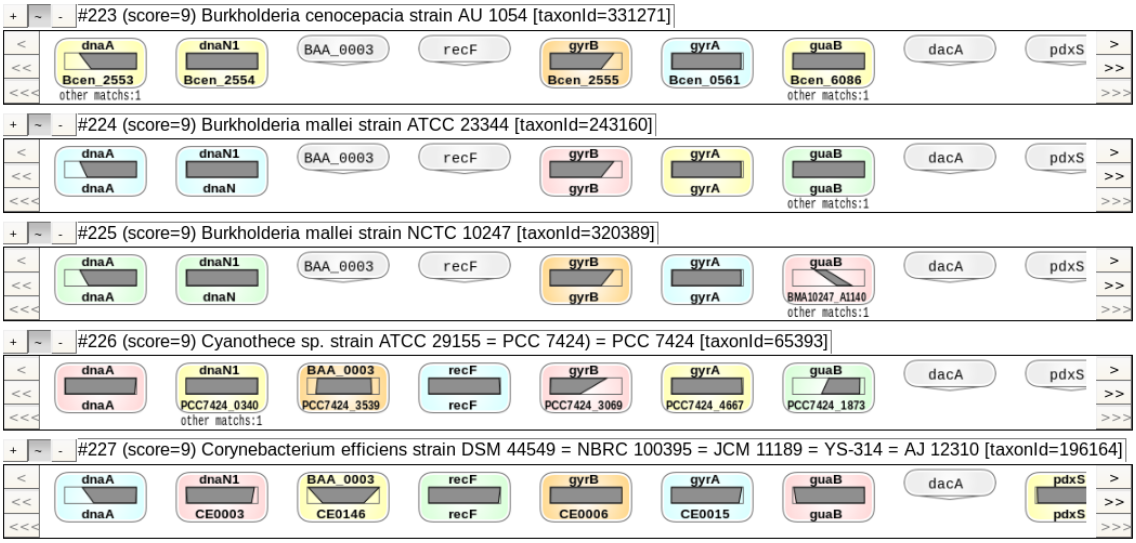

Figure 3. Vue tableau des homologues

389 génomes complets procaryotes ont été intégrés dans la base de données PostgreSQL à ce jour. Le pipeline s'appuie sur Genome Reviews, BLAST, le bi-directional best hit pour inférer l'orthologie, et la programmation dynamique pour déterminer les synténies. Toutes les données et méthodes utilisées par le pipeline sont publiques ou ont été publiées par leurs auteurs. L'interface est une application web développée en GWT et HTML5. Pour analyser des données privées, une machine virtuelle peut être téléchargée qui contient une version pré-installée du pipeline, de la base de données et de l'application web de visualisation.

3 Conclusions

Insyght (http://genome.jouy.inra.fr/Insyght) propose une représentation visuelle et une navigation originale des synténies et homologies qui ouvrent une perspective nouvelle en ce qui concerne diverses analyses biologiques classiques: annotation de la fonction des gènes à l'échelle du génome, détection des évènements d'évolutions (par exemple transfert horizontal), profilage phylogénétique, et analyse de gènes niche-spécifiques ou core-génome. L'analyse des gènes dans le contexte d'espèces proches ou distantes phylogénétiquement est un besoin identifié par les biologistes [15,16]. Insyght permet de constituer un jeu de gènes qui satisfait plusieurs critères hétérogènes et d'analyser les gènes candidats. La vue du tableau d'homologues offre une approche exhaustive et simple pour étudier les homologies abondantes. La vue génomique permet l'identification d'évènements évolutifs et améliore la lisibilité parmi le grand nombre de régions synténiques et idiosyncratiques.

Remerciements et financements

Les auteurs souhaitent remercier Dr Philippe Bessières et Dr Catherine Juste pour leur commentaires constructifs. Ce travail a été financé par l'Agence Nationale de la Recherche [ANR-PDR-080124-03-01].

References

[1] M. Huynen, B. Snel, W. Lathe, and P. Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*, 10, 1204-1210, 2000.

[2] X.H. Zheng, F. Lu, Z.Y. Wang, F. Zhong, J. Hoover, and R. Mural, Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics*, 21, 703-710, 2005.

[3] T. Davidsen, E. Beck, A. Ganapathy, R. Montgomery, N. Zafar, Q. Yang, R. Madupu, P. Goetz, K. Galinsky, O. White, *et al.,* The comprehensive microbial resource. *Nucleic Acids Res*, 38, D340-345, 2010.

[4] J. Blom, S.P. Albaum, D. Doppmeier, A. Puhler, F.J. Vorholter, M. Zakrzewski, and A. Goesmann, EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics*, 10, 154, 2009.

[5] E. Courcelle, Y. Beausse, S. Letort, O. Stahl, R. Fremez, C. Ngom-Bru, J. Gouzy, and T. Faraut, Narcisse: a mirror view of conserved syntenies. *Nucleic Acids Res*, 36, D485-490, 2008.

[6] C. Soderlund, M. Bomhoff, W.M. and Nelson, SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res*, 39, e68, 2011.

[7] A.U. Sinha, J. and Meller, Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 8, 82, 2007.

[8] D.R. Riley, S.V. Angiuoli, J. Crabtree, J.C. Dunning Hotopp, and H. Tettelin, Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics*, 28, 160-166, 2012.

[9] S.J. McKay, I.A. Vergara, J.E. and Stajich, Using the Generic Synteny Browser (GBrowse_syn). *Curr Protoc Bioinformatics*, Chapter 9, Unit 9 12, 2010.

[10] D. Vallenet, L. Labarre, Z. Rouy, V. Barbe, S. Bocs, S., Cruveiller, A. Lajus, G. Pascal, C. Scarpelli, and C. Medigue, MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res*, 34, 53-65, 2006.

[11] Y.Wang, H.Tang, J.D. Debarry, X. Tan, J. Li, X. Wang, T.H. Lee, H. Jin, B. Marler, H. Guo, *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*, 40, e49, 2012.

[12] T. Derrien, C. Andre, F. Galibert, C. and Hitte, AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. *Bioinformatics*, 23, 498-499, 2007.

[13] M. Muffato, A. Louis, C.E. Poisnel, and H. Roest Crollius, Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 26, 1119-1121, 2010.

[14] F. Lemoine, B. Labedan, and O. Lespinet, SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes. *BMC Bioinformatics*, 9, 536, 2008.

[15] K.E. Nelson, D.E. Fouts, E.F. Mongodin, J. Ravel, R.T. DeBoy, J.F. Kolonay, D.A. Rasko, S.V. Angiuoli, S.R. Gill, I.T. Paulsen, *et al.* Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen Listeria monocytogenes reveal new insights into the core genome components of this species. *Nucleic Acids Res*, 32, 2386-2395, 2004.

[16] T.D. Read, G.S. Myers, R.C. Brunham, W.C. Nelson, I.T. Paulsen, J. Heidelberg, E. Holtzapple, H. Khouri, N.B. Federova, H.A. Carty, *et al.* Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. *Nucleic Acids Res*, 31, 2134-2147, 2003.