DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Benjamin CHURCHEWARD

Models and methods for genome-resolved metagenomics

Thèse présentée et soutenue à Nantes, le 13 Décembre 2022
Unité de recherche : Nantes Université

Rapporteurs avant soutenance :

Lucie BITTNER — Maître de conférence Université Paris Sorbonne
Eric PELLETIER — Directeur de Recherche CEA

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Dominique LAVENIER	Directeur de Recherche Université de Rennes
Examinateurs :	Silvia ACINAS	Associate Professor ICM-CSIC Barcelona
	Mathieu ALMEIDA	Chargé de Recherche INRAE
Dir. de thèse :	Guillaume FERTIN	Professeur Nantes Université, LS2N
Encadrant. de thèse :	Samuel CHAFFRON	Chargé de Recherche CNRS, LS2N

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

Titre : Modèles et méthodes pour la métagénomique à résolution génomique

Mot clés : Binning, assemblage, métagénomique, ASP

Résumé : La reconstruction de génomes à partie de données métagénomiques, aussi appelés MAG) représente une étape majeure dans l'étude des communautés microbiennes. La reconstruction de MAGs souffre néanmoins de limitations, telles que la nature fragmentée de ces MAGs, les difficultés inhérentes à la reconstruction du pangenome, ou la capture des variations entre souches d'une même espèce. Dans cette thèse, le problème du binning a été appréhendé à travers un modèle de clustering suivant le paradigme de la logique déclarative, l'objectif étant de maximiser l'information sur les génomes présents grâce à l'exploration de l'ensemble des solutions de binning possible. Ce modèle de binning incluant métrique compositionnelle, mesure d'abondance et occurrence de gènes marqueurs a été implémenté en langage ASP. Nous nous sommes ensuite concentrés sur l'optimisation du processus d'assemblage, étape préliminaire clé de la classification de contigs, avec pour objectif d'encore améliorer la reconstruction de MAGs. Nous avons développé une approche automatique pour guider le processus de co-assemblage, couplant des distances métagénomiques avec une méthode d'optimisation du clustering. Cette approche a été intégrée dans un nouveau workflow de reconstruction de MAGs, MAGNETO, qui intègre également des stratégies assemblage-binning complémentaires.

Title: Models and methods for genome-resolved metagenomics

Keywords: Binning, co-assembly, metagenomics, ASP

Abstract: The reconstruction of individual genomes from metagenomic data, also called MAGs have constituted a major milestone in the study of microbial communities. However, the recovery of MAGs still suffers several limitations, including the mosaic and population nature of these MAGs, the inherent difficulties to assemble pangenomes, and the recovery of strain-level variations for a given species. In this thesis, a declarative programming framework was designed and used to resolve the genome binning problem through a constrained clustering approach, with the goal to explore several optimal binning solutions, informing us about the organization dynamics of naturally occurring genomes. A novel genome binning model integrating compositional and abundance information as well as constraints on single-copy core genes was designed and implemented using the ASP language. With the goal to further enhance the recovery of MAGs, we focused on optimizing the assembly process, a key genome binning preprocessing step. We developed an automated approach to guide the co-assembly process, combining metagenomic compositional distances with an optimal clustering method. These developments were implemented into a novel genome-resolved metagenomics workflow called MAGNETO, integrating complementary assembly-binning strategies.