Rapport d'activité
Qualification CNU 2024

Victor EPAIN

[redacted]

https://vepain.gitlab.io/

1 Parcours Universitaire

2020 — 2023 — **THÈSE EN INFORMATIQUE**
3 ANS — Centre Inria de l'Université de Rennes, sur bourse nationale Inria — INRAe
OBTENUE LE — École doctorale Mathématiques, informatique, signal et électronique et télécommunications
27 NOV. 2023 — Assemblage de fragments ADN : structures de graphes et échafaudage de génomes de chloroplastes – analyses comparatives, formulations et implémentations

Directeur de thèse :
Rumen ANDONOV, Professeur des Universités, IRISA, Université de Rennes

Co-directeur de thèse :
Jean-Francois GIBRAT, Directeur de recherche, INRAe, Université Paris-Saclay

Encadrant :
Dominique LAVENIER, Directeur de recherche, CNRS, IRISA, Rennes

Rapporteurs :
Éric ANGEL, Professeur des Universités, IBISC, Université Paris-Saclay
Annie CHÂTEAU, Maitresse de conférence — HDR, LIRMM, Université de Montpellier

Membres du jury :
En supplément de D. LAVENIER, É. ANGEL et A. CHÂTEAU
Élisa FROMONT, Professeur des Universités, IRISA, Université de Rennes
Camille MARCHET, Chargée de recherche CNRS, CRIStAL, Lille
Mathias WELLER, Professor, Institu für Softwaretechnik und Theoretische Informatik, Berlin

2018 — 2020 — **MASTER EN BIOINFORMATIQUE**
2 ANS — Université de Rennes
Master mention Bio-Informatique — mention très bien
Mémoire : Production d'une séquence consensus à partir du positionnement de longues lectures ADN

Encadrants :
Rumen ANDONOV, Professeur des Universités, IRISA, Université de Rennes
Dominique LAVENIER, Directeur de recherche, CNRS, IRISA, Rennes

2017 — 2018 — **3ÈME ANNÉE DE LICENCE EN INFORMATIQUE**
1 AN — Université de Rennes — ISTIC

Licence mention Informatique — mention assez bien
Rapport de stage : Extraction de gènes de chloroplastes et recherche de la profondeur génomique pour trois espèces de plantes du genre *Brassica*

Encadrants :

Rumen ANDONOV, Professeur des Universités, IRISA, Université de Rennes
Dominique LAVENIER, Directeur de recherche, CNRS, IRISA, Rennes

2015 — 2016 (2 ans) **CLASSE PRÉPARATOIRE MPSI-MP, SPÉCIALITÉ MATHS-PHYSIQUE**
Lycée Dupuy de Lôme, Lorient

2 Expériences professionnelles

DE DÉC. 2024 À DÉC. 2025 (1 an) **CHERCHEUR POSTDOCTORANT**
Université de Milan-Bicocca, Italy, sur bourse européenne programme H2020-MSCA-RISE-2019, action H2020/Marie Curie, code 24A196
CDD
Construction et recherche de génomes de plasmides dans des graphes de pangénomes à partir d'assemblages métagénomiques

DE JUI. 2024 À AUJOURD'HUI (2 ans) **MEMBRE DU BUREAU DE L'ASSOCIATION POUR LA LIBERTÉ ACADÉMIQUE (ALIA)**
Association pour la liberté académique (ALIA), France
Bénévole
Association loi 1901 d'intérêt général dont l'action se situe à l'interface entre la science et la société. Rédaction de communiqués de presse, mise en place d'un site Internet, organisation de réunion, communication

DE DÉC. 2023 À DÉC. 2024 (1 an) **CHERCHEUR POSTDOCTORANT**
Sans affiliation, Lorient
Sur les ressources de l'assurance chômage
Continuation des activités de recherche, apprentissage de Rust, candidature à des concours postdoc sur financements publics

2022 — 2024 (2 ans) **TRÉSORIER DE L'ASSOCIATION NICOMAQUE**
Association Nicomaque, Rennes
Bénévole
Association loi 1901 de doctorants et docteurs bretons pour la vulgarisation scientifique

2020 — 2023 (3 ans) **DOCTORANT CONTRACTUEL**
Centre Inria de l'Université de Rennes, sur bourse nationale Inria — INRAe
CDD
Assemblage de fragments ADN : structures de graphes et échafaudage de génomes de chloroplastes – analyses comparatives, formulations et implémentations

Directeur de thèse :

Rumen ANDONOV, Professeur des Universités, IRISA, Université de Rennes

Co-directeur de thèse :

Jean-Francois GIBRAT, Directeur de recherche, INRAe, Université Paris-Saclay

Encadrant :

Dominique LAVENIER, Directeur de recherche, CNRS, IRISA, Rennes

2021 — 2022 (1 semestre) **ENSEIGNANT VACATAIRE EN INFORMATIQUE ET EN BIOINFORMATIQUE**

Université de Rennes
Enseignant vacataire
Enseignement des méthodes algorithmique sur les graphes en L3 Informatique et L3 Méthodes Informatiques Appliquées à la Gestion des Entreprises
Enseignement en recherche opérationnelle en M1 Méthodes Informatiques Appliquées à la Gestion des Entreprises
Enseignement en algorithmique en M2 Bioinformatique

2020 — 2021 **Enseignant vacataire en informatique**
1 semestre Université de Rennes
Enseignant vacataire
Enseignement des méthodes algorithmique sur les graphes en L3 Informatique et L3 Méthodes Informatiques Appliquées à la Gestion des Entreprises
Enseignement en recherche opérationnelle en M1 Méthodes Informatiques Appliquées à la Gestion des Entreprises

2020 **Stage de 2ème année de Master en Bioinformatique**
6 mois Centre Inria de l'Université de Rennes
Stagiaire
Production d'une séquence consensus à partir du positionnement de longues lectures ADN
Encadrants :
Rumen Andonov, Professeur des Universités, IRISA, Université de Rennes
Dominique Lavenier, Directeur de recherche, CNRS, IRISA, Rennes

2019 **Stage de 1ère année de Master en Bioinformatique**
3 mois Centre Inria de l'Université de Rennes
Stagiaire
Programmation linaire en nombres entiers pour l'assemblage *de novo* de longues lectures ADN (*De novo* long reads assembly using integer linear programming)
Encadrants :
Rumen Andonov, Professeur des Universités, IRISA, Université de Rennes
Dominique Lavenier, Directeur de recherche, CNRS, IRISA, Rennes

2018 — 2019 **Secrétaire Général de l'association étudiante e-BIGO**
1 an Rennes
Bénévole
Association loi 1901 des étudiants du Master en Bioinformatique de l'Université de Rennes
Rédaction de comptes rendus de réunion et d'Assemblées Générales

2018 **Stage de 3ème année de Licence en Informatique**
2 mois Centre Inria de l'Université de Rennes
Stagiaire
Extraction de gènes de chloroplastes et recherche de la profondeur génomique pour trois espèces de plantes du genre *Brassica*
Encadrants :
Rumen Andonov, Professeur des Universités, IRISA, Université de Rennes
Dominique Lavenier, Directeur de recherche, CNRS, IRISA, Rennes

3 Activités d'enseignement

Statut	Année d'exercice	Établissement d'exercice	Public	Niveau	Nom de la matière	Volume horaire total en heures équivalent TD	Effectifs	Nature	Responsabilités	Supports d'enseignement éventuels
Vacataire	2021 — 2022	Université de Rennes	Informatique	L3	Modèles et Algorithmes des Graphes	48	40	TD	Actualisation des sujets de TDs Création et correction d'examens	
Vacataire	2021 — 2022	Université de Rennes	Informatique	M1	Recherche Opérationnelle	12	20	TP	Actualisation des sujets de TPs Création et correction d'examens	
Vacataire	2021 — 2022	Université de Rennes	Bioinformatique	M2	Algorithmique en Bioinformatique	11	30	CM (2h) TD (4h) TP (6h)	Actualisation du CM Actualisation des sujets de TD	
Vacataire	2020 — 2021	Université de Rennes	Informatique	L3	Modèles et Algorithmes des Graphes	48	40	TD	Actualisation des sujets de TDs Création et correction d'examens	
Vacataire	2020 — 2021	Université de Rennes	Informatique	M1	Recherche Opérationnelle	12	20	TP	Transformation des sujets de TPs pour utiliser Python Création et correction d'examens	

4 Activités de recherche

2024 **Formalisation du problème de l'assemblage d'haplotypes par un problème de pavage**

1 an Lorient, télétravail indépendant

Formalisation — modélisation — programmation

Contexte L'assemblage d'haplotypes à partir de fragments ADN est une tâche délicate, car, par définition, deux haplotypes ont des séquences ADN très similaires pouvant varier sur quelques positions génomiques. La plupart des assembleurs fusionnent des fragments ADN de différents haplotypes pour former une même séquence, ce qui ne permet pas de rendre compte de la population génomique d'un échantillon.

Entrée Étant donné une fenêtre d'alignement de lectures ADN sur un contig (séquence ADN résultat d'un premier assemblage des lectures ADN), on obtient une matrice binaire où chaque colonne représente une variation nucléotidique (*SNP*) identifiée sur les lectures alignées (les lignes). À la lecture (ligne) i il y a un 1 à la position (colonne) j si le SNP (i, j) est le SNP majoritaire par rapport aux autres SNP à la même position, un 0 sinon.

Problème Trouver un partitionnement des lectures (lignes) qui rende compte des similarités de SNP, sans connaitre le nombre de parties.

Méthode Nous proposons une méthodologie en trois étapes : **(i)** les colonnes sont permutées pour former des groupes de colonnes (*bandes*, au nombre inconnu de m), représentant chacune un bipartitionnement des lectures en quasi-bicliques denses et creuses ; **(ii)** création d'un vecteur binaire (*vecteur caractéristique* v_i) pour chaque lecture i, où $\forall 1 \leq k \leq m$, $v_i[k]$ vaut 0 si la lecture fait partie d'une quasi-biclique creuse à la $k^{\text{ème}}$ bande, 1 si elle fait partie d'une quasi-biclique dense ; **(iii)** deux lectures participent au même haplotype si et seulement si leurs vecteurs caractéristiques sont identiques. En général, nous souhaitons minimiser le nombre de bandes. Chaque quasi-biclique est obtenue via la résolution d'un problème linéaire mixte en nombres entiers. Une bande est obtenue par recherche itérative de la quasi-bicliques maximisant le nombre de 1, jusqu'à avoir assigné toutes les lectures.

Résultats Nous avons testé notre méthode sur des lectures ADN provenant de souches bactériennes connues pour être difficilement séparables. Après partitionnement des lectures, nous assemblons les parties indépendamment et obtenons de nouveaux contigs ne confondant pas les haplotypes lorsqu'ils diffèrent. Les nouveaux contigs passent à travers une suite de programme communément utilisés pour obtenir les séquences finales de chaque haplotype. Nous comparons ensuite les séquences finales avec les vraies séquences des bactéries initiales pour valider la qualité de l'assemblage.

Production scientifique Un papier est en cours de rédaction pour soumission à une revue internationale à comité de lecture.

Collègues Roland Faure ; Tam Khac Minh Truong ; Riccardo Vicedomini ; Rumen Andonov

2024 **Partitionnement de colonnes de matrices binaires creuses très larges**

1 an Lorient, télétravail indépendant

Modélisation — programmation

Problème Étant donnés une matrice binaire creuse de dimensions de l'ordre de grandeur de 10^4 à 10^9 lignes et colonnes, un réel $\epsilon \in (0, 1]$,deux entiers $minlignes$ et $mincolonnes$, trouver un nombre inconnu de m sous-matrices denses avec de densité supérieure à ϵ, de taille supérieure à $minlignes \times mincolonnes$ et de périmètre (nombre de lignes + nombre de colonnes) maximal.

Méthode Nous proposons une heuristique de recherche de quasi-bicliques denses en nous inspirant de l'heuristique relaxée du problème du sac-à-dos en nombre entiers.

Résultats Deux résultats : **(i)** implémentation d'une structure de matrices binaires stockant les 1 et passant à l'échelle ; **(ii)** recherches itératives de la quasi-biclique de périmètre maximal respectant les contraintes du problème.

Production logicielle La librairie `csvbinmatrix` Rust répond au point (i) et s'appuie sur une publication extérieure qui n'avait pas bénéficié d'une implémentation. Le programme (ii) est pour le moment en source fermée.

Collègue Rumen ANDONOV

2020 À 2024 **PROPOSITION ET COMPARAISON D'IMPLÉMENTATIONS DE STRUCTURES DE GRAPHES REPRÉSENTANT DES LIENS ENTRE DES FRAGMENTS ADN**

4 ANS Centre Inria de l'Université de Rennes

Description de structures de données de graphes — comparaison théorique en mémoire et coûts algorithmiques — programmation — comparaison en pratique

Contexte Les graphes sont des structures de données adaptées pour représenter des liens entre des fragments ADN. Trois structures sont présentes dans la littérature : chronologiquement, la première s'appuie sur un graphe dirigé, la deuxième un graphe bidirigé, la troisième enfin sur un graphe non dirigé. Toutes ces structures représentent les fragments ADN, ainsi que leur miroir (version inverse-complémentaire). On parle alors d'orientation de fragment. Ainsi, chaque fragment est représenté par ces deux versions dans les graphes, soit directement, soit en fonction d'une marche. Cela implique qu'un lien entre deux fragments est également doublement représenté (si le fragment orienté u va vers le fragment orienté v, alors on retrouve le lien $\bar{v} \to \bar{u}$ dans le graphe, où $\bar{v}$ et $\bar{u}$ sont respectivement les miroirs de v et u). Bien que ces structures de graphes soient omniprésentes tout le long des processus d'assemblage de fragments ADN, aucune comparaison de ces structures n'a été décrite dans la littérature.

Résultat Pour chacune de ces structures, est proposée au moins une implémentation basée sur des listes d'adjacences. Des algorithmes pour des opérations basiques, telles que l'accès aux voisins d'un fragment orienté, ou par exemple l'ajout ou la suppression d'un fragment ou d'un lien ont été proposés. Pour chaque implémentation, le coût mémoire a été analysé et comparé théoriquement et en pratique. De même, le coût de chaque algorithme a été comparé théoriquement et en pratique.

Production logicielle Dans un souci de reproductibilité, les implémentations et les programmes de comparaison en Rust sont disponibles sur GitLab : `rustrevsymg` (branche `develop`).

Publication Les comparaisons théoriques sont disponibles au quatrième chapitre de thèse Assemblage de fragments ADN : structures de graphes et échafaudage de génomes de chloroplastes – analyses comparatives, formulations et implémentations. Un papier est en cours de rédaction pour soumission à une revue internationale à comité de lecture.

Encadrants de thèse Rumen ANDONOV, Dominique LAVENIER, Jean-Francois GIBRAT

2021 À 2023 **ÉCHAFAUDAGE DE GÉNOMES DE CHLOROPLASTES EN RENDANT COMPTE DE LEURS HAPLOTYPES STRUCTURAUX**

2 ANS Centre Inria de l'Université de Rennes

Formalisation à partir de connaissance — modélisation — programmation — évaluation sur données synthétiques et semi-synthétiques

Contexte Les chloroplastes sont des organites peuplant les cellules de plantes vertes et interviennent dans le processus de photosynthèse. Les génomes de chloroplastes d'intérêts sont circulaires et possèdent pour la plupart une paire de régions répétées inverses-complémentaires. Lors de la réplication ADN, la région génomique entre ces répétitions peut se retrouver inversée et complémentarisée, formant ainsi un autre génome. Cette inversion *flip-flop* est réversible,

et implique la coexistence de plusieurs formes de génomes de chloroplastes dans une même cellule de plante. Puisque les génomes diffèrent en forme par la transformation d'une région, on parle alors d'*haplotypes structuraux*. Aucun assembleur ne se focalise sur la reconstruction de ces haplotypes structuraux avec la connaissance a priori sur les structures de génomes de chloroplastes. L'échafaudage est une partie du processus de l'assemblage qui se concentre sur la résolution des régions répétées.

Résultat En 3 étapes : **(i)** Formalisation et définition du problème de l'échafaudage de génomes de chloroplastes avec connaissances structurales a priori ; **(ii)** modélisation par programmation linéaire en nombres entiers de recherche de chemins dans un graphe de fragments ADN sous contraintes structurelles ; **(iii)** implémentation et évaluation sur données synthétiques et semi-synthétiques.

Production logicielle L'échafaudeur est disponible et installable via PyPI : `khloraa_scaffolding`, les évaluations sont reproductibles via la documentation accolée au répertoire `khloraascaf_results`

Visibilité

2021 — 2022 (2 MOIS) Visite scientifique à l'Université Heinrich Heine Düsseldorf, Allemagne, équipe ALBI, responsable : Gunnar KLAU

Encadrement

2021 (3 MOIS) Co-supervision de Pauline HAMON-GIRAUD, en stage de fin de première année de Master Bioinformatique, Université de Rennes

Encadrants de thèse Rumen ANDONOV, Dominique LAVENIER, Jean-Francois GIBRAT

2019 (3 MOIS) **PARTITIONNEMENT DE GRAPHES DE CHEVAUCHEMENTS POUR L'ASSEMBLAGE LONGUES LECTURES**
Centre Inria de l'Université de Rennes
Laboratoire national de Los Alamos

Contexte À partir des chevauchements entre des longs fragments ADN (obtenus par un séquençage de troisième technologie), on peut construire un graphe dirigé où les sommets représentent les fragments orientés (la séquence dans sa version séquencée originale, ou son « miroir » i.e. sa séquence inverse-complémentaire). Deux sommets sont reliés par un arc si, et seulement s'il existe un chevauchement entre les deux fragments orientés associés à ces sommets. On dénote alors ce graphe par *graphe de chevauchements*. À cause des répétitions dans le génome, tous les chemins de ce graphe ne représentent pas forcément de vraies sous-séquences du génome.

Méthode **(i)** partitionner les sommets du graphe de chevauchements avec METIS ; **(ii)** construire un graphe de parties ; **(iii)** trouver le chemin le plus long dans le graphe de parties ; **(iv)** revenir à l'échelle du graphe de chevauchements et trouver le chemin le plus long dans chaque partie, dans l'ordre des parties déterminé à l'étape (iii).

Visibilité

2019 (1 MOIS) Visite scientifique d'un mois financée par la bourse HipcoGen (High-Performance Combinatorial Optimization for Computational Genomics), dans le contexte de l'équipe associée Inria-LANL

Liste des publications

Revue internationale avec comité de lecture

V. EPAIN et R. ANDONOV, « Global Exact Optimisations for Chloroplast Structural Haplotype Scaffolding », *Algorithms for Molecular Biology*, t. 19, n° 1, p. 5, 6 fév. 2024, ISSN : 1748-7188. DOI : 10.1186/s13015-023-

00243-1

Type d'article Article long

Nombre de pages 34

Contributions Conception de la méthode, des modèles linéaires en nombres entiers, des algorithmes ; prouver la NP-complétude ; prouver que la réduction de la taille des ensembles de définitions permettait toujours de trouver toutes les solutions distinctes ; implémentation en Python ; tester les méthodes sur les données ; rédaction du papier.

Conférence nationale avec comité de lecture

V. EPAIN et al., « Optimal Scaffolding for Chloroplasts' Inverted Repeats », présenté à JOBIM2022, 5 juill. 2022

Type d'article Article de conférence

Nombre de pages 10

Contributions Conception de la méthode, du modèle linéaire en nombres entiers, des algorithmes ; implémentation en Python ; tester les méthodes sur les données ; rédaction du papier.

Liste des actes JOBIM2022 https://jobim2022.sciencesconf.org/data/pages/JOBIM2022_proceedings_oral.pdf

5 Activités et responsabilités collectives et administratives

2023 — 6 MOIS — 1H / SEMAINE

ORGANISATION DE L'ATELIER ÉTHIQUE INRIA RENNES DU 4 JUILLET 2023

Centre Inria de l'Université de Rennes

Médiation scientifique — organisation évènementielle et communication

Contexte Le 4 juillet 2023, Enka BLANCHARD et Bernard FRIOT étaient invités pour présenter leurs travaux de recherche respectivement en sciences humaines et sociales, et en sociologie du travail et économie. Les présentations et les échanges articulèrent critique du productivisme scientifique et capacités pour produire de la recherche éthique.

Valorisation Le Collège Doctoral de Bretagne a reconnu cet atelier comme formation à l'éthique, code AMETHIS : ETH-CDB-07

— Le site internet de l'évènement

— La retransmission vidéo de la conférence (en deux parties)

2022 — 2023 — 2 ANS — 1H / SEMAINE

ORGANISATION DE SÉMINAIRES ET D'ÉVÈNEMENTS DU DÉPARTEMENT DATA KNOWLEDGE MANAGEMENT (DKM)

IRISA

Médiation scientifique — organisation événementielle et communication

Chaque année, le département invite tous les mois un ou une scientifique à présenter ses travaux au département. Le département organise également ses journées dédiées (sur un jour et demi).

— Le programme des journées DKM de l'année 2023

— Le programme des journées DKM de l'année 2022

2022 — 2023 — 2 ANS — TOTAL 20H

COANIMATEUR DE LA SESSION RECHERCHE OPÉRATIONNELLE EN BIO-INFORMATIQUE AU CONGRÈS NATIONAL DE RECHERCHE OPÉRATIONNELLE ET AIDE À LA DÉCISION EN FRANCE (ROADEF)

Congrès ROADEF Rennes 20–23 Fév. 2023

Congrès ROADEF Lyon 23–22 Fév. 2022

Organisation et animation de la session (attribution des relecteurs des résumés soumis à la session, introduction de l'orateurice, prise des questions)

— Liste des sessions de la ROADEF2023

— Liste des sessions de la ROADEF2022

2021 — 2023
3 ANS
TOTAL 24H

ORGANISATION DES ENTRAINEMENTS AUX ORAUX DE STAGES DES PREMIÈRES ET SECONDES ANNÉES DE MASTER DEVANT LES ÉQUIPES DU DÉPARTEMENT DKM

IRISA

Organisation de l'emploi du temps de passages des entrainements, gestion du temps de parole et des questions pour mise en situations réelles d'examen