MI
Mission pour l'Interdisciplinarité

DEFI INTERDISCIPLINAIRE : « *MASTODONS* »

Appel A Projets 2017 - Formulaire de candidature

Ce formulaire de candidature (libellé : AAP2017-MASTODONS–nomcandidat) doit être obligatoirement déposé par le porteur du projet (dûment visé par son directeur d'unité) sur le site à l'adresse https://sigap.cnrs.fr/sigap/web/connexion.php (sous format word et ou pdf)
Dans le dossier « Défi MASTODONS - AAP 2017 »

DATE LIMITE de CANDIDATURE : 13 DECEMBRE 2016 à midi

Civilité –Nom/Prénom du porteur du projet	M. Dominique LAVENIER
Titre long (max 150 caractères)	Correction des données de séquençage de 3ème génération
Acronyme du projet	C3G

Mots clés : Génomique, Séquençage, Correction, Bio-informatique, Algorithmique

Résumé scientifique du projet

Les applications du séquençage haut débit couvrent désormais toutes les sciences de la vie : de la médecine à l'agronomie. Le séquençage de 3ème génération produit de très longues lectures, mais elles sont extrêmement bruitées ce qui impacte fortement la qualité des analyses bio-informatiques. Le défi du projet **C3G** est d'amener ce type de données à un **haut niveau de qualité** à travers le développement de nouvelles stratégies de correction.

Demande budgétaire première année	
Déplacements : Missions/accueils (conférences,….	16 000 €
Organisation manifestations : Colloques/ateliers/journées techniques	6 000 €
Petits Equipements : (montant unitaire inférieur à 800 € HT)	€
Equipement amortissable : (montant unitaire supérieur à 800 € HT)	€
Prestations de service étroitement lié à la mise en œuvre du projet :	€
Rémunérations de personnels avec approbation préalable de la MI :	€

Description du projet

1. Introduction

Le séquençage haut débit (SHD) est maintenant une technologie omniprésente en Sciences de la Vie dont les applications couvrent tous les domaines de la médecine à la biologie moléculaire en passant par l'agronomie ou l'écologie. Elle sert aussi bien à estimer la biodiversité qu'à comprendre la formation des tumeurs cancéreuses.

Les SHD de 2ème génération produisent des lectures courtes (< 250 nucléotides) de très haute qualité (<1 % d'erreur). Les SHD de 3ème génération (SHD3) produisent des lectures longues (3K à 20K nucléotides) mais de faible qualité (15 à 18% d'erreur). Deux firmes, Pacific Biosciences (PacBio) [1] et Oxford Nanopore Technology (ONT) [2] se partagent le marché. Le séquenceur MinIon d'ONT sous forme de clé USB attise les espoirs d'utilisations en temps réel sur le terrain ou en environnement complexe. Cependant, la piètre qualité des lectures longues, due au fort taux d'erreur, représente un frein considérable.

Le consortium des plateformes de SHD France Génomique en fait un défi majeur comme en atteste l'affluence aux ateliers "*Long Reads*" qu'il organise. En effet, la qualité met en défaut la plupart des algorithmes et empêchent les pipelines d'analyse bio-informatique de fonctionner correctement. Même des tâches simples (comparer ou classifier des séquences) deviennent difficiles [3]. Il est crucial d'évaluer l'impact de ces taux d'erreur sur ces analyses et de concevoir des solutions appropriées. Ce projet vise à proposer des outils de correction de séquence adaptés aux SHD3 pour relever ce défi. Ces outils seront mis à l'épreuve des données réelles en collaborations avec les partenaires biologistes.

Actuellement, grâce aux SHD, de nombreux projets de séquençage de génomes fleurissent. Mais ils buttent sur la construction de séquences complètes : le génome obtenu reste fragmenté à cause des régions répétées. Par exemple, l'assemblage du génome du cacao (une plante diploïde) comprend 4792 séquences (*scaffolds*) au lieu de 10 chromosomes [4]. Les lectures longues PacBio ou MinIon, parce qu'elles sont plus longues que de nombreuses répétitions génomiques, aident à résoudre ces incertitudes. Cependant les premiers essais restent infructueux à cause des taux d'erreur trop élevés de ces lectures. Un problème important vient des types d'erreurs qui sont en majorité des insertions et délétions, plutôt que des substitutions (<20% de substitutions).

Des outils de correction de type "hybride" ont été proposé pour effectuer des corrections sur de longues lectures bruitées. Ils utilisent des lectures courtes très fiables [5][1]. Par exemple, le LIRMM avec l'Université d'Helsinki a conçu le logiciel LoRDEC [6], extrêmement efficace en temps et en mémoire. Cependant, la correction hybride reste un second choix qui oblige à effectuer deux types de séquençage. La correction non hybride, i.e. basée uniquement sur des lectures longues reste un enjeu majeur. C'est un de nos objectifs.

L'utilisation de structures de données d'indexation de séquences qui allient compression et efficacité en temps [7][8] sont une des clefs des algorithmes pour analyser des données bruitées. Elles permettent d'augmenter les requêtes pour analyser ces séquences en temps raisonnable. L'expérience du consortium dans ce domaine est un atout, en particulier pour le critère du passage à l'échelle.

Le projet **C3G** est structuré en trois parties :

1. Correction hybride
2. Correction non-hybride
3. Validation

L'objectif des deux premières parties est la mise au point de stratégies pour corriger les gros volumes de données produit par le séquençage de 3ème génération et les amener à un niveau de qualité compatible avec les pipeline d'analyse bio-informatique. Le défi est double : (1) proposer des méthodes efficaces et (2) proposer des méthodes qui passent à l'échelle. La troisième partie a pour but de valider les

développements algorithmiques sur des données réelles et de les mettre à l'épreuve sur un projet de séquençage de plusieurs génomes complexes.

2. Correction hybride

La plupart des projets de séquençage actuels combinent maintenant des données de 2ème et 3ème générations, et probablement pour quelques années encore. Il est donc important de proposer des méthodes de correction qui utilisent pleinement ces deux technologies. Un des défis est que ces méthodes passent à l'échelle au regard des projets de séquençage qui se profilent.

L'objectif d'une correction hybride est de tirer parti du meilleur des technologies de 2ème et 3ème génération. La première produit une très grande quantité de petites séquences de très bonne qualité. La seconde produit un nombre raisonnable de grandes séquences de très mauvaise qualité. L'idée est donc de corriger les grandes séquences à l'aide des petites. Le logiciel LoRDEC, publié en 2014, réalise cette opération. La correction s'effectue par programmation dynamique et donne d'excellents résultats.

Le but de cette partie est d'aller encore plus loin à l'aide d'un graphe de de-Bruijn construit à partir des courtes séquences qui, rappelons-le, sont de bonne qualité. Ainsi, le graphe contient un ensemble de chemins dont certains représentent des parties du génome et dont la qualité reflète celles des données. La difficulté est que le parcours d'un graphe de de-Bruijn engendre une multitude de chemins possibles à cause des répétitions qui jalonnent les génomes. Les longues séquences de 3ème génération, même si elles sont bruitées, permettent cependant d'aiguiller les parcours en levant les ambiguïtés devant plusieurs choix. Par rapport aux solutions actuelles, on peut relever deux principaux avantages : (1) obtenir des chemins qui peuvent dépasser la longueur des longues séquences d'entrée (une séquence peut connecter deux longs chemins non ambigu) ; (2) réduire significativement les temps de calculs par rapport à une approche basée sur la programmation dynamique.

A terme, on peut espérer que cette stratégie limite l'impact de la qualité des données de 3ème génération sur les résultats. Elles ne servent qu'à guider le cheminement dans le graphe. Le résultat est un chemin basé sur la qualité des séquences de 2ème génération qui, aujourd'hui possède un taux d'erreur largement inférieur à 1%.

3. Correction non-hybride

Un nombre grandissant de projets de séquençage visent l'exploitation de données SHD3 uniquement. Le défi, ici, est d'augmenter la qualité de ces données par un traitement qui ne repose que sur ces données bruitées. Les technologies évoluent rapidement, mais il faudra encore compter pour les années qui viennent sur des taux d'erreur compris entre 5 et 15 %.

Les méthodes de correction s'appuient principalement sur la redondance de l'information. En effet, le séquençage d'un génome consiste à produire un nombre important de lectures qui se chevauchent plus ou moins fortement le long du génome. On parle ici de la couverture de séquençage. Plus elle est importante, plus la redondance d'information est forte. Dans les projets de séquençage actuels, la couverture varie de 30 à 100. Il y a donc un fort potentiel à exploiter cette redondance pour développer des stratégies de correction.

C3G explorera au moins trois pistes sur lesquelles les partenaires ont entamé des réflexions et des travaux de recherche préliminaires.

La première piste propose de réutiliser des techniques issues de la compression de données. La redondance d'information permet en théorie de construire un consensus et de corriger les lectures qui diffèrent de ce consensus. Cela nécessite d'une manière ou d'une autre de regrouper les informations similaires, par exemple en comparant toutes les lectures deux à deux, ce qui est évidemment bien trop coûteux. Il se trouve que les programmes de compression de données ont à traiter un problème similaire : extraire les informations redondantes afin de ne pas les stocker plusieurs fois. Ainsi, le programme Orcom [9] comporte une première étape qui regroupe les séquences similaires par une technique basée sur les minimiseurs (des k-mers avec certaines propriétés), qui est extrêmement rapide et passe bien à l'échelle. Nous proposons de réutiliser cette première étape comme pré-traitement, qu'il faudra probablement adapter aux

contraintes des longues lectures. Il sera ensuite possible dans chaque petit groupe d'appliquer des méthodes plus coûteuses algorithmiquement pour corriger les lectures.

La seconde piste se base sur la détection de kmers solides. Mêmes si les lectures longues sont très bruitées, elles contiennent de courtes séquences (kmers) dont on fait l'hypothèse que leurs occurrences sont supérieures aux kmers qui contiennent des erreurs par rapport à la séquence génomique. L'idée est d'extraire d'abord ces kmers de l'ensemble des lectures longues pour qu'ils puissent être utilisés comme ancres afin de regrouper des lectures qui se chevauchent. Cette approche brise la complexité de recherche de similarité entre toutes les longues lectures, ce qui constitue actuellement un sérieux verrou dès lors que le nombre de séquences dépasse le million. Une fois les chevauchements détectés, tous les kmers solides qui appartiennent aux séquences chevauchantes sont ordonnés et une recherche de consensus sur la base d'un algorithme d'assemblage est lancé pour construire une séquence consensus.

La troisième piste exploite les mêmes informations que la seconde. Un des écueils de la piste par kmers est qu'une seule taille de kmers n'est pas suffisante pour couvrir toutes les régions du génome. Nous proposons une approche basée sur plusieurs graphes de de-Bruijn ayant des tailles de kmers différentes (approche qui réussit dans l'assemblage des génomes). La comparaison des lectures sur les chemins du graphe permet de déterminer des groupes de lectures similaires. Le groupe est raffiné en itérant cette procédure avec plusieurs valeurs de k. Ensuite, le consensus peut être calculé par des méthodes d'alignements multiples. Un premier prototype pour explorer cette piste est d'utiliser une méthode existante capable de détecter des kmers partagés par comparaison des lectures, ce que fait l'outil de correction hybride LoRDEC [6]. Nous proposons d'itérer LoRDEC en mode non hybride pour identifier les groupes de lectures similaires tout en faisant varier k, puis d'appliquer l'alignement multiple aux groupes identifiés pour les corriger.

4. Validation

Les méthodes de correction développées au cours de ce projet seront évaluées sur des jeux de données des technologies PacBio ou Oxford Nanoport Technology et sur des génomes dont on connait la référence. De tels jeux sont maintenant disponibles pour les principaux organismes modèle. Avoir le génome de référence permet de mesurer précisément la qualité de la correction.

Un autre volet de la validation sera de confronter nos méthodes à la réalité du terrain au sein d'un projet de séquençage d'envergure, le projet ALPAGA dans lequel plusieurs partenaires sont impliqués. Ce projet a pour objectif de caractériser l'évolution de génomes d'animaux en l'absence de reproduction sexuée. 20 génomes d'animaux d'espèces divergentes qui ont perdu la reproduction sexuée seront séquencés via le PIA France Génomique. Ils seront prioritairement séquencés en technologie MinIon, puis en technologie Illumina. La partie bio-informatique du projet vise, entre autres, à réaliser un assemblage de haute qualité sur ces génomes pour une analyse structurelle ultérieure.

Ce projet, qui va produire de gros volumes de données de séquençage de 2ème et 3ème génération est un cadre idéal pour l'évaluation des méthodes hybrides et non-hybrides qui seront développées dans le projet **C3G**. Le passage à l'échelle des stratégies sera notamment crucial pour produire des données de haute qualité qui alimenteront les assembleurs. Nous pourrons donc évaluer à grande échelle l'impact de la qualité des données sur l'assemblage final. Le projet ALPAGA a démarré courant 2016 et les premières données seront disponibles en 2017.

5. Participants et disciplines

- **Bioinformatique / Informatique**
  - LIRMM, CNRS, Montpellier : E. Rivals, A. Makrini, B. Cazaux, D. Paulet
  - IRISA, CNRS, Rennes : P. Peterlongo, C. Lemaitre, C. Marchet, D. Lavenier, G. Rizk, A. Limasset
  - …
- **Biologie**
  - IGEPP, INRA, Rennes : D. Tagu, F. Legeai, JC. Simon
  - ISA 1355, INRA, Sophia Antipolis : E. Danchin
  - Université Libre de Bruxelle : J.-F. Flot

Bibliographie

1. Koren,S. and Phillippy,A.M. (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.*, **23**, 110–120.
2. Mikheyev,A.S. and Tin,M.M.Y. (2014) A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.*, **14**, 1097–1102.
3. Myers,G. (2014) Efficient Local Alignment Discovery amongst Noisy Long Reads. In Brown,D., Morgenstern,B. (eds), *Algorithms in Bioinformatics*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 52–67.
4. Argout,X., Salse,J., Aury,J.-M., Guiltinan,M.J., Droc,G., Gouzy,J., Allegre,M., Chaparro,C., Legavre,T., Maximova,S.N., *et al.* (2011) The genome of Theobroma cacao. *Nat. Genet.*, **43**, 101–108.
5. Koren,S., Schatz,M.C., Walenz,B.P., Martin,J., Howard,J.T., Ganapathy,G., Wang,Z., Rasko,D.A., McCombie,W.R., Jarvis,E.D., *et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.*, **30**, 693–700.
6. Salmela,L. and Rivals,E. (2014) LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, **30**, 3506–3514
7. Drezen,E., Rizk,G., Chikhi,R., Deltel,C., Lemaitre,C., Peterlongo,P. and Lavenier,D. (2014) GATB: Genome Assembly & Analysis Tool Box. *Bioinformatics*, **30**, 2959–2961.
8. Välimäki,N. and Rivals,E. (2013) Scalable and Versatile k-mer Indexing for High-Throughput Sequencing Data. In Cai,Z., Eulenstein,O., Janies,D., Schwartz,D. (eds), *Bioinformatics Research and Applications*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 237–248
9. Grabowski, S., Deorowicz, S., & Roguski, Ł. (2015). Disk-based compression of data from genome sequencing. *Bioinformatics*, *31*(9), 1389-1395

Justification de la demande financière

La demande financière se décompose de la manière suivante :

- 3 réunions de travail entre partenaires. Il s'agit d'échanger sur les différentes stratégies étudiées et de confronter les résultats des partenaires [8000 Euros]
- 1 colloque à la fin de la première année partagé entre conférenciers invités et restitution des travaux menés dans **C3G** [6000 Euros]
- Participation à des congrès nationaux et internationaux [8000 Euros]

Visa et argumentaire du directeur d'unité du porteur :