Défis computationnels des séquençage et phénotypage haut-débit en science de la vie

Eric Rivals (rivals@lirmm.fr), Esther Pacitti (pacitti@lirmm.fr)

1 Introduction, contexte scientifique et vision du domaine

La biologie et ses applications, de la médecine à l'agronomie ou l'écologie, deviennent des sciences productrices des données massives et par là exigentdes approches computationnelles pour analyser ses données. Les nouvelles technologies de Séquençage à Haut Débit (SHD) apparues en 2005, et aussidites Séquençage de Nouvelle Génération (NGS), révolutionnent la manière dont sont posées et résolues les questions de recherches en science du vivant. Par exemple, pour évaluer la biodiversité d'un espace, au lieu de déterminer patiemment les espèces après prélèvement, on peut aujourd'hui séquencer l'ADN des espèces présentes ou ayant laissé des traces dans un échantillon « environnemental » (sol, eau, air, instestin, etc). Une seule expérience de séquençage (ici de type métagénomique) produit jusqu'à plusieurs centaines de millions de courtes séquences (entre 30 et 120 nucléotides), nommées « reads ». Ces reads sont ensuite groupés en catégories représentant les espèces, et ainsi leur nombre et abondance relative permettent d'estimer la biodiversité. La question devient alors computationnelle.

Dans le cadre médical, pour diagnostiquer ou suivre l'évolution d'un cancer et du patient, on veut surveiller les dysfonctionnements biologiques au niveau moléculaire. On peut grâce au SHD capturer l'ensemble des gènes activés dans les cellules cancéreuses en séquençant les ARNs qu'ils produisent dans un échantillon. Les reads sont comparés au génome de l'espèce (et bientôt au génome de l'individu) pour déterminer quels gènes sont activés, sous quelle forme, et à quelle abondance moléculaire. Un catalogue précis des gènes actifs peut permettre d'identifier la forme du cancer, de suivre l'évolution du patient durant le traitement. Aujourd'hui, pour les applications dites de « médecine personnalisée », le point de blocage se situe clairement au niveau de l'analyse bioinformatique des reads. Aujourd'hui, les méthodes ne délivrent que des catalogues incomplets, peu précis, sous-estiment la diversité des formes d'ARN, etc. En outre, les algorithmes atteignent leur limite d'efficacité avec les débits actuels des séquençeurs, alors qu'étant donné l'ampleur du marché, les technologies évoluent à un rythme soutenue et promettent régulièrement des accroissement des débits et des longueurs de reads. Un mot clé est la passage à l'échelle des algorithmes de traitement des séquences primaires.

L'emploi du SHD qu'il soit de type génomique, transcriptomique, épigénomique, ou métagénomique, touche bien d'autres domaines des sciences de la vie : fabrication alimentaire, défense, écologie, amélioration des plantes, etc.

La mesure automatisée des phénotypes (caractères observables d'un organisme) permet d'identifier les conséquences de variations alléliques en termes de morphologie, de croissance ou de métabolisme dans un environnement donné. Dans le domaine végétal, les méthodes de génétique quantitative permettent d'identifier les gènes impliqués dans des variations phénotypiques en réponse aux conditions environnementales, afin d'identifier des gènes de tolérances aux stress associés aux changements climatiques. Les plateformes haut débit de phénotypage (PHD) des plantes, par ex. Phenoarch et Phenodyn au LEPSE (INRA - SupAgro Montpellier), s'appuient sur de nouvelles technologies de traitement d'images et de capteurs. Elles produisent de grandes quantités de données (par ex 105 données par jour) à différents intervalles de temps (de minutes à des jours) et à différentes échelles depuis des échantillons de petits tissus jusqu'à la plante entière. Sept plates formes de ce type sont en cours de d'installation au champ ou en condition contrôlées, et seront mises en réseau (projet Infrastructures d'Avenir Phénome).

Pour chaque plate forme, les données brutes doivent être transformées en datasets de différente nature (variables environnementale, mesures physiologique, caractérisations 3D de l'architecture de plante, etc.). L'objectif final, l'identifier les relations entre phénotype et génotype, nécessite des méthodes statistiques, algorithmiques, de fouille de données complexes capables en outre de changer d'échelle. Pour cela, il faut aussi pouvoir

partager les données de phénotypage distribuées à grande échelle entre communautés scientifiques. Ceci nécéssite de concevoir des modèles novateurs de gestion distribuée des données capables d'utiliser une infrastructure hétérogène de calcul à grande échelle avec de multiples clouds.

Le domaine du traitement des données de SHD et PHD est diversifié de par ses applications, les questions d'abordées, mais surtout il aiguise, suscite, engendre de nouvelles approches en science de la vie à un rythme effréné. Cette révolution des « Big Data » permet de poser aujourd'hui des questions que l'on aurait à peine évoqué il y a une décennie. En revanche, les avancées sur le plan computationnel au sens large, c'est à dire de l'informatique, de la bioinformatique, du calcul parallèle, de la fouille, la gestion de ces types de données en sont à leur balbutiements. Le volume de recherche consacré à ces questions est insuffisant face à l'avalanche de données, aux progrès technologiques, et surtout à la complexité et diversité des questions biologiques posées (ERCIM news).

2 Verrous

Les verrous identifiés et visés comprennent :
- l'algorithmique du texte et des séquences (indexation, comparaison, compression) et son passage à l'échelle
- l'exploitation des architectures parallèles (multi-coeurs, grille, cloud) pour l'analyse des données
- l'invention de nouvelles approches et algorithmes pour identifier variations génomiques, épigénomiques, transcriptomiques ou classifier les données du méta-génome
- le partage et la fouille de données à grande échelle
- l'intégration de données sur les versants technique et biologique (lien génotype-phénotype).

Étant donnée l'ampleur de la tâche, ce projet mentionne plusieurs pistes à court et moyen termes, et reste ouvert à l'accueil des partenaires qui amèneront des expertises complémentaires sur ces questions et domaines.

3 Acquis du domaine

A l'heure actuelle, la recherche en informatique et bioinformatique sur ce thème est naissante en France, un peu plus avancée en Europe (Allemagne, Grande-Bretagne, pays nordiques), et plus développée aux USA et Canada. Les acquis méthodologiques originaux concernent généralement pour une grande majorité des méthodes :

1. de comparaison de reads par rapport aux génomes, dites de « mapping »
2. d'assemblages des génomes
3. de classification par espèces des reads de métagénomiques
4. de gestion de données à grande échelle (P2P, cluster, cloud)
5. d'analyse génétique des plantes.

Un effort sensible sur l'exploitation de ces méthodes aux processeurs graphiques ou en cloud computing a donné des avancées pragmatiques. Quelques articles épars traitent de compression, d'indexation, et de fouille de données.

4 Participants seniors

- Eric Rivals, directeur de recherche CNRS (LIRMM, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR CNRS 5506 Université Montpellier II) et directeur du GDR Bio-Informatique Moléculaire
- Ester Pacitti, professeur UM2, responsable adjointe de l'équipe Zenith (LIRMM), directrice adjointe du département informatique du LIRMM
- Hélène Touzet, directeur de recherche CNRS (LIFL, Laboratoire d'Informatique Fondamentale de Lille, UMR CNRS 8022 Université Lille 1)
- Dominique Lavenier, directeur de recherche CNRS (IRISA, Institut de Recherche en Informatique et Systèmes Aléatoires, UMR CNRS 6074 Université Rennes I)

– Pascal Barbry, directeur de recherche CNRS (Institut de Pharmacologie Moléculaire et Cellulaire, UMR CNRS 7275 Université de Nice Sophia Antipolis)
– Nicolas Galtier, directeur de recherche CNRS (Institut des Sciences de l'Evolution de Montpellier, CNRS 5554 Université Montpellier II)
– François Tardieu, directeur de recherche INRA, responsable de l'équipe Mage au LEPSE (INRA–SupAgro)
– Patrick Valduriez, directeur de recherche INRIA, responsable de l'équipe Zenith au LIRMM

Eric Rivals, Hélène Touzet et Dominique Lavenier représentent des équipes d'informatique et de bio-informatique avec une activité reconnue en algorithmique des séquences haut-débit et en calcul intensif. Pascal Barbry apparaît comme partenaire pour l'infrastructure nationale de plateformes de séquençage France Génomique. Nicolas Galtier est responsable du pro jet ERC PopPhyl qui explore les transcriptomes de 30 plantes et animaux non modèles. F. Tardieu est expert en analyse génétique et modélisation du développement des plantes. Il coordonne le projet national Phenome. E. Pacitti et P. Valduriez sont experts en gestion de données distribuées et parallèles (P2P, grid, cloud).

5 Axes de recherches

Le problème du développement d'outils informatiques pour l'analyse de données de séquençage a pris son essor avec les premiers séquençages à grande échelle. C'est un domaine bien identifié de la bio-informatique. Mais comme nous l'avons décrit, la quantité inédite des données générées, leur diversité, les impératifs de production et d'analyse, obligent à repenser toute la chaîne de traitement. D'autre part, le données de phénotypage accessibles sous forme transformée exigent une organisation distribuée qui n'existe pas à cette échelle. Nous proposons trois grands axes de recherche.

5.1 Traitement primaire des séquences : algorithmique et parallélisme.

Les données brutes de séquençage se présentent sous la forme de millions de petites séquences courtes, non localisées, qui doivent avant toute analyse ultérieure être identifiées : assemblage ou alignement sur des génomes connus, mais aussi correction. Il est communément admis que la croissance du volume des données de séquençage excéde maintenant la croissance décrite par la loi de Moore pour la puissance des processeurs des ordinateurs. Les gains en temps de calcul doivent donc venir de progrès algorithmiques combinés à une utilisation optimale des architectures parallèles disponibles : grilles mais aussi machines multi-coeurs et GPU (Graphic processing units). C'est un nouveau paradigme, qui posent de nouvelles questions propres à l'analyse de séquences. Par exemple, les algorithmes doivent établir un bon compromis entre l'utilisation de la mémoire globale lente et les petites mémoires locales rapides, entre le calcul et les coûts de transfert des données. Ces questions se posent au niveau de toutes les techniques habituelles mises en oeuvre en algorithmique du texte : filtrage, indexation, programmation dynamique, ... Peu de travaux de travaux de recherche existent dans ce sens. Au niveau international, cette communauté est émergente (cf. rubrique Bioinformatics and Computational Biology de gpucomputing.net), mais plusieurs indicateurs montrent que l'attente est très forte. Le Beijing Genomics Institute, plus grand centre international de séquençage, vient ainsi récemment de mettre en place une ferme de GPU dédiée à l'analyse de séquences. Un aspect algorithmique important, mais pas nécessairement parallèle, concerne l'indexation, c'est à dire la structuration en mémoire des informations concernant des sous-parties de séquences et leur positions dans l'ensemble de séquences ou le génome. Cette piste très ouverte engendrera des progrès qui permettront le passage à l'échelle des méthodes de bioinformatiques.

5.2 Prediction d'événements biologiques

Les reads, même lorsqu'ils ont subi le traitement primaire, ne livrent pas directement d'informations biologiques. L'ensemble des informations contenues dans les reads et dans leur positionnement par rapport à un génome de référence permet ensuite de prédire des candidats de ce que nous appelons génériquement « événements biologiques » : mutations ponctuelles, réarrangements génomiques, variations en nombre de copies, production d'un ARN isoforme à partir d'un gène, production d'ARN chimériques, expression d'ARN non codants, modifications post-transcriptionnelles des d'ARN (édition), identification de sites de liaison à l'ADN ou l'ARN, altération épigénomique de la chromatine, etc. Chaque type d'événements demande des algorithmes de prédiction prenant en compte des critères spécifiques. Cependant aucun de ces types d'événements ne sont

connus avec précision : nous ne disposons pas de définition formelle directement traduisibles en informatique. Les modèles sont plus ou moins précis et de meilleurs modèles seront développés grâce aux résultats des analyses à grande échelle. L'interaction avec des biologistes est indispensable pour contrôler, préciser, paramétrer, et valider les méthodes. En outre ces questions se déclinent suivant le contexte de l'étude : prédire des mutations fréquentes dans une population d'individus (biologie évolutive) diffère de la prédiction de mutations récurrentes au sein de tumeurs d'un même type (biologie médicale). Le consortium propose de lancer des efforts de recherche sur ces thèmes pour développer des méthodes sophistiquées utilisant des techniques de recherche de motif, d'apprentissage automatique, d'algorithmique combinatoire, etc. La finesse de ces analyses secondaires et la qualité des informations biologiques en dépendent.

5.3 Partage et analyse de données

Nous proposons d'exploiter deux approches complémentaires pour le partage de données à grande échelle : le pair-à-pair (P2P) et le cloud. Le P2P est bien adapté aux applications collaboratives nécessitant la décentralisation du stockage de données. Un pair peut représenter une organisation ou un utilisateur qui souhaite partager ses données tout en conservant ses données en local. L'approche cloud permet d'exploiter les techniques de gestion efficace de Big Data (MapReduce, NoSql, etc) en offrant un stockage efficace et très fiable. Dans le domaine du phénotypage, les données partagées sont les données transformées qu'on retrouve à partir de documents scientifiques (e.g. papiers de conférences, rapports techniques, etc) qui les décrivent. Nous considérons que les données transformées sont stockées dans des fichiers de façon ad-hoc. Pour faciliter le partage de ces données, nous comptons proposer des solutions pour la recommandation décentralisée en nous appuyant sur les approches cloud et P2P ainsi qu'en exploitant des méthodes issues des réseaux sociaux. L'approche développée pourra également être exploitée pour le partage données de génotypage. Concernant l'analyse de ces données transformées, le problème concerne l'extraction de différents types de connaissances (des corrélations fréquentes, des séquences fréquentes, des groupes d'enregistrements partageant des valeurs similaires, etc.). Dans ce projet, nous proposons de nous concentrer sur la découverte de corrélations fréquentes. Le problème principal sera d'identifier et de contourner une grande quantité de corrélations dites « parasites » qui sont fréquentes mais peu informatives, afin de découvrir des corrélations plus pertinentes. Nous envisageons deux approches principales pour contourner le problème des corrélations parasites et répondre au défi de la découverte de corrélations informatives. La première piste consiste à identifier les corrélations parasites dans un premier temps et de reprendre le processus d'extraction par étapes successives (creuser un peu plus profond dans les données à chaque étape). La deuxième piste consiste à considérer les données d'origine comme des données biaisées par les corrélations parasites. L'objectif est alors de travailler directement sur ces données en les « dé-biaisant » pendant le processus d'extraction.

6 Disciplines

- Biologie moléculaire
- Biologie évolutive, agronomie, écologie
- Médecine
- Informatique
- Bioinformatique