Proposition de thèse

Assemblage *de-novo* de génomes complexes

Equipe : Symbiose / GenScale, Rennes
Directeur de thèse : D. Lavenier, DR CNRS

Il a fallu plus de 10 années d'efforts pour séquencer le premier génome humain. Ce travail s'est achevé au début des années 2000, a coûté environ 3×10^9 dollars et a monopolisé entre 400 et 500 laboratoires à travers le monde. Aujourd'hui, avec les progrès drastiques des biotechnologies, ces mêmes données peuvent être produites en une semaine pour un coût 100 000 fois moindre, et par une seule machine ! Les projets de séquençage fleurissent donc un peu partout et conduisent à de nouveaux défis en terme de traitement de l'information : le flot de données produit par les machines de séquençage est tel qu'il tend à surpasser les capacités d'analyse des ressources informatiques.

Un des traitements de base effectué sur ces données haut-débit est l'assemblage *de-novo*. Ce processus vise à reconstruire « in silico » le génome d'un organisme à partir des informations générées par les séquenceurs. Ces derniers transcrivent l'ADN en un texte segmenté en une multitude de fragments. L'assemblage consiste à ordonner l'ensemble de ces fragments en un seul texte. Les difficultés résident (1) dans le très grand nombre de fragments à manipuler (plusieurs centaines de millions, voire plusieurs milliards) ; (2) dans la gestion des erreurs : les séquenceurs sont des machines imparfaites qui produisent des textes erronés ; (3) dans la prise en compte des organismes séquencés et notamment dans la gestion des zones répétées au sein des génomes complexes.

D'un point de vue purement informatique, les défis sont multiples. Il faut définir des structures de données capables d'agréger l'ensemble des informations produites par les séquenceurs. Ces dernières doivent être soit suffisamment compacte pour tenir dans la mémoire centrale d'un multiprocesseur, soit posséder des propriétés distributives permettant de dispatcher l'information sur des machines à mémoire distribuée. Dans les 2 cas, il faut également développer une algorithmique adaptée pour que les traitements passent à l'échelle dans une optique de séquençage haut débit. En clair, les temps de traitements ne doivent pas être supérieurs aux temps de production des données ce qui n'est clairement pas le cas, aujourd'hui, pour des génomes complexes. Pour résumer, le défi de l'assemblage est l'élaboration d'une stratégie rapide, de faible empreinte mémoire (relativement à la masse de données manipulée) et qui restitue de manière fiable le texte des génomes à partir des séquenceurs haut-débit.

L'équipe Symbiose/GenScale, à travers la thèse de R. Chikhi, développe une méthodologie d'assemblage qui s'appuie sur les dernières avancées des biotechnologies (pair read, mate-pair) et qui dès la conception s'oriente vers l'usage du parallélisme. L'équipe participe au niveau international aux compétitions Assemblathon et collabore activement avec des

laboratoires de biologie ou de génomique sur des cas concrets d'assemblage de génomes complexes.

La thèse s'inscrit dans la poursuite de ces travaux où de nombreux verrous restent à lever, notamment lorsqu'il s'agit d'assembler des génomes fortement répétés. Il y a à la fois des aspects théoriques à explorer et des aspects plus pratiques à développer pour répondre aux problèmes critiques actuels de l'assemblage. Nous voyons ce traitement comme un processus complexe, construit à partir de briques logicielles spécifiques et qui doivent être judicieusement agencées pour « customiser » le calcul en fonction de la nature des organismes. Chaque brique est en soi un problème bioinformatique à part entière qui, s'il est imparfaitement résolu, peut conduire à une inefficacité de l'ensemble.

Le projet précis de la thèse sera élaboré en fonction des aspirations et des compétences du candidat. Les mots clef principaux sont : algorithmique, structure de données, indexation, masse de données, parallélisme, optimisation, assemblage, génomique.

Contact : D. Lavenier. Merci de faire parvenir un CV à lavenier@irisa.fr

Références

E. Pennisi
Will Computers Crash Genomics?
Science 11 February 2011: Vol. 331 no. 6018 pp. 666-668
http://www.stanford.edu/class/cs262/papers/WillComputersCrashGenomics.pdf

R. Chikhi, D. Lavenier,
Localized Genome Assembly from Reads to Scaffolds: Practical Traversal of the Paired String Graph
LNCS, Algorithms in Bioinformatics , vol 6833, 2011
http://www.springerlink.com/content/f5g305j5k73x3k14/

R. Chikhi, G. Chapuis, D. Lavenier
Parallel and memory-efficient reads indexing for genome assembly,
Workshop on Parallel Computational Biology , Torun, Poland, 2011
http://irisa.lavenier.net/PDF/Lav11cg.pdf