

Proposition de thèse

Mémoire d'Indexation Active pour la Génomique

Directeur de thèse : D. Lavenier, Prof. ENS Cachan / Bretagne

Lieu : Equipe Symbiose Irisa/Inria Rennes

Contact : lavenier@irisa.fr – 02 99 84 72 17

<http://irisa.lavenier.net>

Contexte

La recherche dans les banques de données génomiques est un traitement de base de la bioinformatique. Etant donnée une requête, par exemple une séquence d'ADN représentant un gène, l'objectif est de récupérer tout ou partie des séquences de la banque qui *ressemble* à cette séquence. Avec les progrès continus des biotechnologies, les tailles des banques deviennent gigantesques et augmentent d'autant le temps de réponses des interrogations.

Pour limiter ce temps, une technique consiste à indexer la banque de manière à ne considérer, lors d'une interrogation, qu'une petite fraction des données. Cette approche, beaucoup plus rapide, demande cependant des espaces de stockage encore plus volumineux que les données brutes et une mémoire à accès rapide que ne peut procurer un disque magnétique classique.

L'architecture ReMIX, développée au sein de l'équipe Symbiose, a été une première tentative, à travers l'usage de mémoire FLASH, pour proposer une architecture nouvelle de mémoire d'indexation. Sur cette architecture, les interrogations des banques sont effectivement beaucoup plus rapides. Par contre, les expérimentations ont clairement mis en évidence que la phase d'indexation, c'est-à-dire la phase de mise en forme des banques, constituait un goulot d'étranglement majeur, parce que réalisée à l'extérieur sur des ordinateurs possédant une mémoire limitée.

Proposition de thèse

L'objectif de la thèse est de pousser plus loin le concept élaboré dans l'architecture ReMIX en introduisant la faculté d'auto indexation. L'idée est que l'étape d'indexation, auparavant explicite et à la charge des usagers, devienne transparente et rapide, si possible au rythme où sont stockées les données. Pour intégrer de telles fonctionnalités, l'architecture doit être complètement reconsidérée et probablement adossée à un modèle de programmation parallèle (type MAP-REDUCE, par exemple) permettant de spécifier à minima les structures d'index.

Dans un contexte où l'accroissement rapide des volumes des données génomiques soulève de véritables défis dans le domaine de la bioinformatique, le travail de cette thèse consistera à apporter des solutions originales en mixant des approches architecturales, algorithmiques et technologiques.

Bibliographie

Site WEB ReMIX : <http://irisa.lavenier.net/research/remix.html>