Actions de Recherche Coopérative INRIA 2006-2007

Optimisation de graines et indexation des banques génomiques sur mémoire FLASH reconfigurable
Application à l'identification de nouvelles protéines mitochondriales

D. Lavenier
Equipe Symbiose IRISA - Rennes

Résumé

La recherche - *par le contenu* - dans les banques d'ADN fait appel à des heuristiques basées essentiellement sur l'identification de courtes zones fortement similaires. Celles-ci reposent sur le concept de *graines* qui déterminent fortement la qualité des résultats. Pour accélérer la recherche, on peut indexer les banques à partir des ces graines. Ainsi, au lieu de parcourir systématiquement les banques pour rechercher des similitudes, on pointe directement sur l'information pertinente. Le corollaire est de disposer d'une mémoire suffisamment grande et rapide pour stocker l'index. Le projet Symbiose a développé un prototype de mémoire reconfigurable de grande capacité permettant de stocker des index de plusieurs centaines de giga octets, basée sur la technologie FLASH. Ce prototype (ReMIX) est opérationnel depuis l'automne 2005.

L'objectif de cette ARC est d'exploiter les propriétés du prototype ReMIX pour la recherche par le contenu dans les banques d'ADN en combinant la reconfigurabilité du système avec des méthodes d'indexation issues de graines non conventionnelles. Il s'agit, en fonction de contraintes de sensibilité et/ou de sélectivité définies par l'usager, de générer automatiquement une structure d'index ad hoc, ainsi que la structure matérielle correspondante pour accéder et traiter cet index.

Quatre équipes de recherches sont impliquées :

- le projet INRIA Symbiose qui a développé le prototype ReMIX ;
- l'équipe de bio-informatique du LIFL qui, en coopération avec le projet INRIA ADAGE, travaille sur des techniques de graines optimisées ;
- l'équipe « IP Design » du LESTER dont une des activités est la synthèse automatique d'architecture ;
- l'unité INSERM U694 du CHU d'Angers qui, via une application de recherche intensive dans les génomes, apporte ses compétences en génomique.

1. Description des activités scientifiques envisagées

1.1 Contexte

Cette ARC se propose de réunir quatre équipes ayant des compétences différentes sur un projet fédérateur centré sur la recherche par le contenu dans les banques de séquences d'ADN. Elle s'appuie sur un projet démarré en 2004 - et soutenue par l'ACI Masse de Données (projet ReMIX : http://www.irisa.fr/remix/) – qui consiste à concevoir un système ayant un accès rapide à un espace mémoire important. Dans ce contexte, l'équipe Symbiose a développé un prototype possédant une mémoire

FLASH de 512 giga octets pilotée par des composants reconfigurables qui, suivant l'application, reconfigurent les accès mémoire et le traitement associés aux données. En dehors de l'équipe Symbiose, les trois autres équipes ne sont pas partenaires de l'ACI ReMIX.

Concrètement, le prototype se présente sous un petit cluster de 4 PCs équipés chacun de 2 cartes PCI sur lesquelles on trouve 64 Go de mémoire FLASH connectée à un composant reconfigurable Xilinx Virtex-2 Pro. Un environnement de programmation permet d'avoir une vision unifiée de l'espace mémoire disponible.

L'originalité de ReMIX réside dans le fait de disposer de capacité de calcul au plus près des données. Pour des applications ciblées sur la recherche dans les masses de données, cela permet d'éliminer extrêmement rapidement (au vol) une grande quantité de données et, ainsi, consacrer les ressources en calcul du processeur hôte uniquement aux informations pertinentes.

Grâce à sa capacité mémoire importante et à son accès rapide, ReMIX peut stocker des structures de données sophistiquées qui évitent de balayer l'ensemble des données pour rechercher de l'information. Dans le cas des banques d'ADN, plutôt que de stocker linéairement les séquences on peut imaginer un schéma d'indexation qui permet de ne pointer que vers des sous ensembles de séquences – ou portion de séquences. Il faut alors définir des critères permettant d'indexer ces sous-ensembles. Ce type de schéma d'indexation est employé dans le logiciel BLAST (le logiciel le plus utilisé en bioinformatique) : la confrontation d'une séquence d'ADN contre une banque consiste d'abord à indexer la requête sur la base de mots (graines) de 10 à 12 caractères. La taille des mots définie la sensibilité et la rapidité de la recherche.

Dans notre cas, sur ReMIX, c'est la banque qui est indexée de manière permanente dans la mémoire FLASH. Le choix des graines est donc extrêmement important pour garantir une certaine sensibilité. De plus, à sensibilité fixée, il peut exister plusieurs graines qui peuvent influer sur la taille de l'index. En général, les études portent sur l'optimisation des graines pour favoriser une sensibilité maximale. Ici, la composante « taille de l'index » doit être largement considérée pour minimiser l'espace de stockage. **La premier axe de l'ARC sera donc de considérer le problème de l'optimisation de graines sous un angle différent : pour une sensibilité donnée, quelles sont les graines qui procurent une indexation minimale.**

Nous proposons ensuite d'expérimenter les divers modèles d'indexation sur une étude menée actuellement au sein de l'unité INSERM U694 du CHU d'Angers et qui porte sur l'identification de nouvelles protéines mitochondriales. Les mitochondries sont des organites qui jouent un rôle déterminant dans la production énergétique et dans la mort cellulaire programmée des cellules eucaryotes. Un grand nombre de processus pathologiques impliquent les mitochondries (cancer, diabète, obésité, maladies génétiques et neurodégénératives). Les mitochondries comportent approximativement 1500 protéines dont seules 500 environ sont connues et caractérisées chez l'homme. L'un des défis pour les prochaines années consistera à identifier l'ensemble de ces protéines. Etant donné que 60% des protéines mitochondriales sont issues de d'organismes procaryotes, la comparaison des génomes eucaryotes et procaryotes permettra un criblage à haut débit de nouvelles protéines mitochondriales et ainsi d'améliorer la compréhension de la physiopathologie mitochondriale.

En terme informatique, cela revient à comparer plusieurs centaines de milliers de protéines d'origine procaryote contre quelques génomes eucaryotes (comme l'homme, par exemple). Plusieurs essais ont déjà eu lieu avec le programme BLAST et ont permis de valider cette méthode de travail sur des machines de l'IDRIS (IBM-SP4 - 12 nœuds SMP P690+). Il existe donc une base réelle sur laquelle on peut s'appuyer pour comparer nos modèles d'indexation. **Le second axe de l'ARC sera ainsi de comparer de nouveaux modèles d'indexation par rapport à une base de référence, tant en terme de sensibilité que de temps de calcul.**

Enfin, l'expérimentation de plusieurs modèles n'a de sens que si l'on est capable de porter rapidement ces modèles sur le prototype ReMIX. Ici, cela revient à implémenter dans les composants reconfigurables des filtres adaptés aux différentes graines. Une approche manuelle pour spécifier en VHDL chaque modèle n'est pas vraiment réaliste car trop coûteuse en temps. Nous proposons d'automatiser cette étape à partir d'une description algorithmique (en C) en adaptant à ReMIX l'outil de synthèse GAUT développé par l'équipe « IP Design » du Lester. L'architecture de ReMIX est en fait très proche du modèle d'exécution proposé par GAUT : le traitement effectué sur les données en provenance de la mémoire alimente un pipeline d'opérateurs que GAUT optimise en fonction de contraintes telles que la latence, le nombre de ressources, etc. **Le troisième axe de l'ARC sera d'automatiser la « programmation » de la mémoire reconfigurable en intégrant l'outil de synthèse GAUT à l'environnement logiciel de ReMIX**.

Les trois sections suivantes présentent plus en détail les trois axes de recherches qui seront développés dans cette Action de Recherche Coopérative.

1.2 Conception de graines

En génomique, les méthodes de recherches les plus fréquemment utilisées sont majoritairement basées sur des heuristiques. La qualité est mesurée par le ratio entre la sensibilité (taux de détection de similarités jugées intéressantes) et la sélectivité (lié à la rapidité de l'algorithme).

Dans un contexte où la croissance des bases augmente rapidement, et où les recherches biologiques actuelles ne se focalisent plus uniquement sur des régions fortement similaires, mais de plus en plus altérées par des mutations, un gain à la fois en sensibilité et en sélectivité est nécessaire.

Ce gain peut être obtenu par amélioration des heuristiques existantes et des modèles sous-jacents : les heuristiques les plus couramment utilisées sont rassemblées sont le concept de « graine » : une graine (représentée selon différents modèles comme les *graines contiguës*, *graines espacées*, *graine à transitions)* est une expression régulière d'un type particulier qui permet de générer un index.

Le principe de graine généralement adopté est celui de graine dite « optimisée » : le choix de l'expression régulière associée à la graine est réalisé de manière à maximiser, pour une sélectivité donnée, le taux de détection de similarités. L'utilisation de graines espacées ou des graines à transitions permet d'obtenir un gain en sensibilité, gain qui peut encore être amélioré par utilisation de graines multiples.

Le projet INRIA ADAGE (G. Koucherov et L. Noé) a acquit une expérience pratique et théorique concernant la conception et l'utilisation de graines, en usage au travers des logiciels YASS et HEDERA. Leur conception demande de calculer la sensibilité selon un modèle d'alignement donné qui utilise des automates classiques déterministes mêlés à des automates probabilistes.

En pratique, la mise en place de plusieurs graines dans un index, même si elle est réalisée de manière séquentielle sur des architectures classiques, améliore grandement le ratio sensibilité/sélectivité. Cependant, dans YASS, le coût mémoire impliqué par un index associant plusieurs graines limite l'usage à deux graines.

Dans le cadre de ce projet, il s'agit d'associer notre connaissance sur la conception de graines et dans une moindre mesure sur la génération d'index afin de l'adapter à une architecture parallèle qui dispose d'une grande quantité de mémoire distribuée. Il est clair que cette architecture permet non seulement de répondre aux exigences issues des graines multiples, mais également de concevoir un index distribué utilisable à l'aide d'un parallélisme efficace.

Les contraintes qui sont posées proviennent tout d'abord de la génération des graines, car cette génération et la sélection qui en découle doivent être adaptées au type d'architecture et à la quantité de mémoire. Elle doit également prendre en compte le parallélisme choisi.

1.3 Application génomique : analyse à haut débit de nouvelles protéines mitochondriales par génomique comparative

La disponibilité des séquences génomiques d'un nombre croissant d'organismes a contribuée à relancer la recherche sur l'origine et l'évolution de la mitochondrie moderne. Selon la théorie endosymbiotique, les mitochondries possèderaient une origine monophylétique unique. Au cours de l'évolution, la majorité des gènes de l'endosymbionte originel auraient été perdus ou bien transférés vers le noyau de la cellule eucaryote hôte. Les nombreux pseudo gènes mitochondriaux présents dans le génome attestent, en effet, d'un processus de transfert tout au long de l'évolution.

La taille du protéome mitochondrial humain est estimée à plus d'un millier de protéines. Seules 13 protéines sont codées par l'ADN mitochondrial, vestige du génome de l'endosymbionte. Toutes les autres protéines sont codées par le génome nucléaire. Chez la levure 50-60% des protéines mitochondriales ont des homologues chez les procaryotes alors que 40-50% n'en ont pas. Les protéines mitochondriales possédant un homologue procaryote résultent probablement du transfert des gènes de l'endosymbionte vers le noyau tandis que les protéines non homologues à des protéines procaryotes résulteraient d'un phénomène « d'enrichissement » du protéome mitochondrial par de nouvelles protéines et donc de nouvelles fonctions.

Récemment, l'équipe INSERM U694 (CHU Angers) a comparé les séquences des 393 protéines mitochondriales humaines avec celles de 256 953 protéines procaryotes. Seules 64% des protéines mitochondriales humaines sont homologues à des protéines procaryotes, ce qui témoigne de leur double origine évolutive procaryote et eucaryote. De plus, il a été montré que 88% des protéines mitochondriales humaines possédant un homologue procaryote sont significativement plus grandes que leur homologue procaryote. Cette différence s'explique par la présence d'une extension N-terminale. Ces séquences additionnelles ne présentent pas d'homologie avec des

protéines procaryotes ; elles seraient probablement d'origine eucaryote. La majorité des protéines homologues à des protéines procaryotes seraient donc des protéines chimères eucaryotes-procaryotes. Etant donné que seule la moitié des protéines mitochondriales est actuellement connue, plus de 500 protéines mitochondriales restent à identifier ce qui constitue un enjeu majeur pour les prochaines années.

Le fait que les protéines mitochondriales d'origine procaryote possèdent dans leur grande majorité une séquence additionnelle N-terminale (par rapport à leur ancêtre procaryote) procure un mode de criblage potentiel de nouvelles protéines mitochondriales non identifiées à ce jour.

Pour confirmer cette hypothèse, l'idée est de comparer l'ensemble des protéines procaryotes connues contre le génome humain et sélectionner celles qui possèdent cette extension N-terminale particulière aux protéines mitochondriales. Un test réalisé sur le chromosome 19 a d'ores et déjà révélé la capacité de cette démarche à identifier les protéines mitochondriales déjà connues.

L'objectif est donc de comparer les 270 génomes d'archaea et d'eubactéries contre des génomes eucaryotes (comme *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*). Avec le logiciel BLAST, combiné à un outil de filtrage maison, le temps de calcul est estimé à plus de 10 jours sur le supercalculateur scalaire IBM-SP4 de l'IDRIS (12 nœuds SMP P690+ avec 32 processeurs / RAM 128 Go).

L'usage d'un système comme ReMIX est intéressant à double titre : (1) l'usage du logiciel BLAST associé à une batterie de filtres maison n'est pas forcément le moyen le plus adapté à nos besoins. Programmer ReMIX en adéquation directe avec notre problématique sera beaucoup plus efficace ; (2) l'usage de gros moyens de calcul (tels que ceux disponibles à l'IDRIS) est une procédure relativement lourde et qui supporte mal les phases de mise au point pour des codes non disponibles en standard. La mise à disposition de ReMIX, aussi bien pour calibrer l'application, que pour exécuter un long calcul, sera immédiate.

1.4 Synthèse d'architecture

Avec la complexité actuelle des applications à intégrer, le critère "délais de mise sur le marché" (*time to market*) devient d'importance comparable, voire supérieure aux facteurs coûts en surface et en vitesse des circuits. L'automatisation la plus complète possible du processus de conception répond à cette problématique : non seulement l'automatisation raccourcit le cycle de conception, mais elle permet aussi d'explorer différentes solutions puisque plusieurs solutions de conception peuvent être générées et évaluées rapidement.

De manière similaire à la compilation d'un programme C en un programme en assembleur dans le domaine du logiciel, la synthèse de haut niveau est un processus de transformation d'une description algorithmique, écrite dans un langage de haut niveau (C, C++, …), en une description structurelle de niveau RTL dans le domaine du matériel. A partir d'une description comportementale de l'application, qui spécifie les traitements (opérations, structures de contrôle, assignations) à opérer de manière comportementale, c'est à dire sans directive d'implémentation architecturale, un processus de synthèse d'architecture permet donc de générer, sur la base d'un modèle d'architecture cible, une architecture dédiée qui satisfait les contraintes du concepteur (temps, surface, consommation, …).

La première étape du processus de synthèse consiste en une compilation de la description comportementale : analyse de code, réduction, transformations et génération d'une représentation interne de type graphe flot de données et/ou de contrôle qui permet de tirer profit du parallélisme potentiel de l'application. Les phases d'allocation, d'ordonnancement et d'assignation permettent ensuite de transformer cette représentation comportementale en une structure matérielle. L'allocation permet de déterminer dans la bibliothèque technologique établie au préalable par le concepteur les opérateurs arithmétiques nécessaires et les éléments de mémorisation. Le processus d'ordonnancement assigne quant à lui les opérations du graphe flot à des intervalles de temps. Enfin, l'étape d'assignation assigne les opérations aux différents opérateurs alloués, les variables aux éléments de mémorisation, les transferts de données aux bus, et ce en accord avec l'ordonnancement prévu.

Le LESTER développe un outil de synthèse de haut niveau : GAUT (web.univ-ubs/lester/www-gaut). GAUT prend en entrée une description algorithmique (C ou VHDL) et génère une architecture matérielle composée d'une unité de traitement, d'une unité mémoire et d'une unité de communication. Cette architecture est décrite en VHDL et est destinée à être synthétisée par les outil de synthèse logique du commerce (ISE/Foundation de Xilinx, Quartus de Altera, …).

Dans le cadre de cette ARC il s'agit d'adapter l'outil de synthèse GAUT à l'architecture ReMIX. La cible matérielle du prototype ReMIX (Xilinx Virtex-2 Pro) ne présente pas a priori de problème particulier. Par contre, l'outil GAUT a été conçu pour synthétiser des applications dominées par les calculs. Il faudra donc étendre la sémantique de spécification acceptée par l'outil afin de pouvoir prendre en compte les expressions algorithmiques propres à l'application et opérer leur matérialisation sur l'architecture cible.