

MODÉLISATION DES MOLÉCULES BIOLOGIQUES

HEWLETT-PACKARD ET L'EPFL SIGNENT UN ACCORD DE COLLABORATION



MARIE-CHRISTINE.SAWLEY@epfl.ch, EPFL-VPR-DAV

PRÉAMBULE

Dans un précédent article paru en janvier 2002 (<http://sic.epfl.ch/publications/FI02/fi-1-2/1-2-page14.html>) dans ce même journal, l'auteur dressait le constat de la spécificité de l'interface entre les sciences de la vie et sur la dynamique de l'évolution en cours. Fin octobre, l'EPFL et Hewlett-Packard concluaient un partenariat, première brique d'un édifice qui pourrait évoluer, appuyé par le plan d'urbanisme adapté, vers un développement important pour les institutions académiques de l'Arc Lémanique.

LE CONTENU DE L'ACCORD

L'EPFL et Hewlett Packard ont signé le 30 octobre dernier un accord de collaboration pour le soutien d'un programme en bio-simulation. L'accord conclu pour 3 ans comprend trois volets:

- i) le soutien des efforts de recherche en bio-simulation et bio-informatique par l'installation d'un serveur doté de 25 noeuds SC45, basés sur des stations quadri-processeurs Alpha EV68 à 1.25 GHz, avec 4 GB de mémoire, fonctionnant sous Tru64 Unix V5.1A (version identique à celle active sur Swiss-T1) et le gestionnaire de travaux batch LSF, le réseau rapide d'interconnexion utilisant le système Quadrics. Seize noeuds (64 CPU) sont réservés aux projets bio-simulation et bio-informatique (dont ceux des groupes du SIB nécessitant la puissance de l'architecture du SC 45), 8 noeuds (32 CPU) prévus pour reprendre la charge de Swiss-T1¹, un noeud servant de frontal et permettant l'accès depuis le réseau;
- ii) le soutien direct du programme visiteurs du Centre Bernoulli du premier semestre de l'année 2003;
- iii) la collaboration avec les experts en biotechnologie et bio-simulation de HP. De plus, HP souhaite également associer les laboratoires de recherche plus connus sous le nom de HP Labs à cette initiative pour un travail sur des projets de recherche technologique à plus long terme.

THÈME SCIENTIFIQUE: MODÉLISATION DES NUCLÉOTIDES SUR UNE VARIÉTÉ D'ÉCHELLES SPATIALES ET TEMPORELLES

Les acides nucléiques, maillons de base des chaînes d'ARN et d'ADN sont des molécules complexes dont il existe 4 sortes dénotées symboliquement par A, C, G et T (lettres

des séquences de l'ADN). Chaque cellule de notre corps comporte 3 milliards de nucléotides.

L'ADN contient toute l'information des gènes de l'organisme considéré (environ 80'000 gènes pour l'organisme humain) et la transmet à sa cellule-fille lors de la phase de reproduction. L'ARN se charge quant à lui de produire les protéines qui vont transmettre l'information qui va déclencher la réaction cellulaire, au moment et endroit choisis, réaction à une stimulation qui peut être bactérienne, virale ou pharmaco-chimique par exemple. Le génome étant décodé avec de plus en plus de précision, il devient important de commencer à comprendre, donc à prédire et à contrôler l'interaction de l'ARN et de l'ADN avec les protéines (protéomique: études de l'ensemble des PROTEINES produites par un génOME ou tissu biologique particulier).

Un premier pas vers la compréhension de cette formidable complexité consiste à modéliser les séquences spécifiques d'ADN et d'ARN isolées afin de comprendre leurs propriétés physiques. D'un point de vue du calcul, cette tâche est encore aujourd'hui un défi extrême. L'approche *standard* consiste à utiliser des codes de Dynamique Moléculaire (MD) tels que Charmm et Amber bien connus des spécialistes de biologie structurale: ces codes visent à simuler une molécule biologique en considérant chaque atome comme une boule rigide et en modélisant son interaction avec les autres atomes selon une loi d'interaction donnée par une fonction potentiel classique. Les limites de cette technique sont de deux sortes: déterminer les plages de validité de la fonction potentiel *simplifiée* reste difficile; d'autre part en prenant une fonction potentiel plus précise, il est encore impossible d'arriver à simuler le nombre d'atomes nécessaires aux molécules biologiques réelles.

Les groupes des professeurs Maddocks et Roethlisberger ont dans ce domaine des compétences complémentaires. Le premier a construit une expérience importante dans le domaine de la MD traditionnelle, modélisant les fragments d'ADN selon des modèles *continus* ou moyennés inspirés par la structure de l'ADN, selon des échelles de temps et d'espace relativement élevées (typiquement de l'ordre du micron). En revanche, le second de par son expérience de la méthode Car-Parrinello, est capable d'aller à des niveaux de précisions de la fonction potentiel beaucoup plus élevés et donc de raffiner le modèle MD en étant capable de simuler les phénomènes à des échelles de temps et d'espace (Angstroms) beaucoup plus petites.

Ces deux méthodes sont très gourmandes en ressources de calcul et en analyse complexe des résultats (visualisation).

¹ voir article en page 22 de ce numéro

LES PHASES SUIVANTES

Au-delà de ce premier accord important, le travail a déjà commencé sur la réalisation de deux phases subséquentes indépendantes:

- le redéploiement du domaine HPC à l'EPFL dans lequel Janus ne peut avoir l'exclusivité -il prend le rôle de ressource de pointe pour les applications *high end* ou nécessitant un système de communication très performant -demande un effort concerté sur des domaines tels que, par exemple:
 - calcul parallèle et distribué
 - clusters de Pc sous Linux
 - architecture du Grid
 - visualisation scientifique
 - accès aux données et analyse des données complexes

Cet effort permettra de faciliter la recherche de la meilleure infrastructure pour une application concernée à un coût *sustainable*, l'échange de savoir et savoir-faire et la mise sur pied d'une planification à long terme propre à assurer les ressources nécessaires à l'évolution planifiée entre les systèmes, selon le modèle nommé de manière imagée *prune and grow* (littéralement taille et repousse). Voir l'appel sur ce sujet dans ce même numéro.

- le montage d'une plate-forme collaborative avec le SIB (développement de VITAL IT) sous la forme d'une grille de calcul comportant des ressources de la partition bio-simulation du SC45 Janus, et dont les nœuds complémentaires, les services et autres boîtes à outils sont en cours de dimensionnement par une *task force* ad hoc.

LE DÉVELOPPEMENT POTENTIEL AVEC LE SIB ET LES RAISONS D'UNE COLLABORATION INTER-INSTITUTION

La biologie, à l'origine une science d'observation, est devenue une science *prédictive et systémique*. Dans la période que l'on appelle la *post-génomique*, l'intérêt va s'étendre de la bio-informatique traditionnelle, basée sur l'exploitation des bases de données biologiques domaine dans lequel le SIB a bâti une excellente réputation au niveau mondial, aux développements de modèles et de prédictions quantitatives des fonctions biologiques. Les fonctions du génome, les mécanismes cellulaires, la biologie structurale et moléculaire reposent sur une grande richesse de données expérimentales mais explorent actuellement des voies tracées par les modèles mathématiques et de la simulation numérique fondées essentiellement sur la chimie et la physique computationnelles, les mathématiques et l'informatique.

Le SIB, fondation d'intérêt public créée en 1998, réunit 7 groupes de recherche issus de l'Université et de l'Hôpital de Genève, de l'ISREC, de l'Institut Ludwig, de l'Université de Lausanne et de l'Université de Bâle. Il a pour but de promouvoir l'enseignement, la recherche (dont le développement des banques de données et d'outils informatiques) et les activités de service dans le domaine de la bio-informatique. Les groupes du SIB ont acquis dans ce domaine une expérience remarquable, et l'attrait de notre région dans le

² European Biological Institute

³ American National Institute of Health

domaine de la bio-informatique est confirmé par l'annonce récente d'un investissement de 15 millions de dollars à l'EBI² et au SIB par le NIH³.

De son côté, l'EPFL vient d'être classée parmi les 50 meilleures universités scientifiques au niveau mondial, et ce résultat n'a pu être obtenu que grâce à la vision et aux efforts des directions qui se sont succédées depuis la fédéralisation de l'EPFL il y a 33 ans et aux résultats des équipes d'enseignement et de recherches qu'elles ont su accueillir. Les orientations stratégiques de ces deux dernières années visant notamment à créer un pôle d'excellence pour la recherche et de formation dans divers secteurs des sciences de la vie, et parmi celles-ci, la bio-simulation, la bio-informatique et les biomathématiques. Vu l'excellence du terreau en sciences de base et en informatique déjà présent sur le campus, la meilleure façon de commencer était de bâtir sur les forces actuelles et de monter en puissance en s'appuyant sur les meilleures collaborations inter-institutionnelles possibles.

Plusieurs pistes non exclusives se dessinent dès 2001 pour une collaboration entre l'EPFL et le SIB: la constitution d'une plate-forme collaborative de support et développement pour les applications de calcul intensif de bio-informatique moléculaire en est une, et reçoit alors le nom de code de *VITAL IT*. Compaq d'abord, puis *New HP* ayant décidé de faire du marché de la Science du vivant une forte priorité, souhaitait collaborer avec des institutions de pointe dans ce domaine. HP avait annoncé son intention d'entrer dans un accord de partenariat sur les thèmes précités avec un apport substantiel et rejoignait le groupe de réflexion *VITAL IT*. L'EPFL est depuis entrée au Conseil de Fondation de l'Institut Suisse de bio-informatique en juillet 2002.

Une analogie qui peut être faite pour expliquer la nature des avancées que nous observerons ces prochaines années est la météorologie: il y a 25 ans, les offices météo collectaient des masses de données (observation au sol, ballons sondes, stations météo en altitude, etc.) et les interprétaient de manière à faire des prédictions sur la région en question selon des méthodes largement empiriques et sans capacité d'analyse et de traitement suffisante. D'autre part, les mathématiciens et les physiciens développaient des modèles numériques théoriques pour simuler le comportement de fluides compressibles comme l'air et l'atmosphère. Petit à petit les deux mondes ont intensifié leur collaboration de manière à intégrer dans l'analyse des données de plus en plus d'outils et méthodes issus des sciences fondamentales, à apprendre à sélectionner les modèles adaptés suivant les échelles de temps et d'espace considérées; aujourd'hui la transition est faite, la production de prévisions peut se faire selon des modèles allant de quelques heures à quelques jours et couvrant des zones plus ou moins étendues.

Toute analogie a ses limites, et celle-ci ne fait pas exception. Nous pouvons toutefois simplifier en voyant le bio-informaticien comme le scientifique spécialiste de l'analyse de l'information issue des échantillons biologiques, alors que le spécialiste en bio-simulation est celui qui tente de comprendre la formation et le fonctionnement des systèmes biologiques par des méthodes théoriques inspirées des sciences fondamentales et de l'informatique: les avancées les plus spectaculaires se feront à l'interface, les groupes capables d'intégrer et de collaborer seront les premiers à faire ces découvertes. ■