Peter Kunszt connaît les défis qu'engendre la recherche avec les big data.

Peter Kunszt est responsable de SyBIT, le projet de bioinformatique de SystemsX.ch

«SyBIT prépare les scientifiques aux big data»

Dans le domaine de la biologie des systèmes, les big data ont depuis longtemps la réputation d'être une source précieuse d'informations. Pourtant, les connaissances sur les processus dans les systèmes biologiques ne peuvent être accumulées que si on parvient à extraire les informations pertinentes du volume gigantesque de données. Peter Kunszt et son équipe viennent en aide aux chercheurs pour automatiser de tels processus. Grâce à SyBIT, le projet de soutien TI, ils contribuent à ce que les scientifiques suisses fassent partie, à long terme, de l'élite mondiale.

D'où vient ce volume énorme de données dans la biologie des systèmes?

Ce sont les importants développements au niveau des technologies permettant d'observer les systèmes biologiques qui ont rendu possible la recherche dans le domaine de la biologie des systèmes. Des progrès considérables ont par exemple été faits au niveau de l'appareillage servant à séquencer l'ADN. Mais les progrès ont également été impressionnants dans les domaines de la spectrométrie de masse, et tout récemment, de la microscopie. Tous ces appareils génèrent des quantités croissantes de données, ce qui peut être comparé aux caméras digitales de plus en plus performantes: chaque année arrivent sur le marché de nouveaux modèles, offrant un nombre croissant de mégapixels et requérant de nouveaux supports de stockage d'information à capacité de plus en plus grande. La même tendance s'observe dans la recherche, bien que dans une toute autre dimension. C'est la raison pour laquelle nous parlons de big data.

Que signifie ce terme pour vous?

Mis à part la quantité, les big data sont caractérisés par la complexité des données, en particulier dans le domaine de la biologie des systèmes. On se trouve souvent confronté à différentes catégories de valeurs dont il faut clarifier les liens. De telles données sont donc difficiles à interpréter. La vitesse est un autre aspect à considérer; lorsqu'on possède d'importantes quantités de données complexes, il devient de plus en plus difficile de les exploiter dans un délai raisonnable; sans compter l'incertitude des données, donc la qualité et la fiabilité des informations, car les valeurs peuvent être sujettes à des erreurs de mesure.

Lequel de ces aspects représente le défi le plus important?

Il y a toujours une relation étroite entre quantité, complexité et vitesse. Tout en tenant compte de ces trois facteurs, notre mission est de trouver, en collaboration avec les chercheurs, les meilleures solutions permettant de répondre rapidement et avec les techniques actuellement disponibles à leurs questions scientifiques.

Pouvez-vous préciser ce point? Comment le projet SyBIT soutient-il les chercheurs de SystemsX.ch?

Cela dépend du projet. Selon les besoins des chercheurs, nous les aidons à assembler l'équipement informatique et les logiciels ou nous organisons l'accès à des ordinateurs centraux. Mais nous soutenons également les scientifiques au niveau des analyses et des évaluations ainsi que de la gestion et du stockage de leurs données.

Christa Smith

Dans le cadre du projet SyBIT, nous mettons à disposition tout le savoir dans le domaine de la gestion des données et aidons à automatiser et à rendre efficace les différentes étapes. Nous préparons pour ainsi dire les scientifiques aux big data, afin qu'ils puissent profiter pleinement du potentiel de ces nouvelles technologies.

Les besoins en support informatique augmentent-ils?

Oui, très clairement. Le volume et la complexité des données vont en croissant, ce qui signifie pour les scientifiques que la gestion de leurs données devient toujours plus difficile et prend de plus en plus de temps. On s'imagine qu'il s'agit là d'une tâche facile, mais avec le grand volume d'informations tel qu'il se rencontre par exemple au niveau de la spectroscopie de masse, déjà la saisie des données est une étape compliquée. Il convient d'annoter et de classer correctement les données d'expériences de toutes sortes si l'on tient plus tard à être en mesure de les réattribuer au bon projet et, au besoin, de les reproduire. Pour l'analyse et l'évaluation des volumes énormes de données, il est souvent nécessaire de faire appel à des algorithmes capables d'identifier automatiquement les caractéristiques et les motifs intéressants.

Pouvez-vous nous citer un exemple concret?

Nous soutenons actuellement le projet MorphogenetiX, dans le cadre duquel les chercheurs se servent de la microscopie en trois dimensions pour étudier la spécialisation de cellules. Grâce à cette nouvelle technologie, la fixation des échantillons avec du formol devient superflue, et il est ainsi possible de les filmer vivants. Le microscope 3D prend jusqu'à 700 images par seconde, et les scientifiques sont ainsi à même d'observer la division cellulaire et de démontrer comment se forme par exemple une cellule spécialisée du cerveau.

Le volume des informations ainsi générées est énorme. SyBIT vient donc en aide aux chercheurs du projet MorphogenetiX pour l'évaluation des données assistée par ordinateur. Un de mes collaborateurs travaille plusieurs mois sur place et, avec les spécialistes du projet, teste les algorithmes développés, dans le but d'automatiser l'exploitation de cette masse de données.

Les grands volumes de données nécessitent une grande capacité de stockage. Que garder? Qu'ignorer?

Afin de pouvoir trancher sur l'importance de nos données, nous devons parfaitement les connaître. Dans le cadre de cette recherche qui, grâce à des technologies innovantes, produit des données en grand nombre, il convient en premier lieu de comprendre ces informations et d'identifier les motifs et les connexités. Particulièrement dans la recherche fondamentale, on ne comprend souvent pas tout de suite ce qui se cache derrière les résultats, ce qui explique pourquoi les chercheurs souhaitent généralement stocker toutes les données. Souvent, ce n'est que vers la fin d'un projet que devient clair quelles informations sont pertinentes au projet et lesquelles peuvent être éliminées du fait qu'il sera plus tard possible de les reproduire sans difficultés et même plus précisément.

Et comment assurez-vous qu'à l'avenir les données seront toujours accessibles?

Malheureusement, la sauvegarde à long terme des données est encore un problème irrésolu. Dans les domaines de la génomique et de la protéomique, il existe déjà des banques de données centralisées internationales, mais à longue échéance, leur financement n'est pas encore garanti. Pour l'archivage de données issues de procédés d'imagerie, par exemple, une solution n'est pas encore à portée de main. Après l'expiration de SyBIT, aucune institution ne gérera de telles informations.

Qui, à votre avis, en serait responsable?

À mon avis, l'archivage des données incomberait aux bibliothèques. Il n'est pas acceptable que les scientifiques doivent payer pour faire conserver leurs informations. L'Etat doit élaborer des solutions. Heureusement que le problème a déjà été reconnu. A l'heure actuelle, plusieurs options sont en voie d'être examinées et aussi discutées sur le plan politique.

SyBIT touche à sa fin en 2018, en même temps que SystemsX.ch. Comment le soutien informatique sera-t-il assuré par après?

Par des groupes de soutien locaux. L'idée est née dans l'Arc lémanique où, en 2004 déjà, le groupe «Vital-IT» a été mis sur pied. Celui-ci offre puissance informatique, mémoire et soutien dans le domaine bioinformatique. A l'exemple de ce groupe, nous avons également réussi à établir des partenaires locaux de SyBIT dans les Universités de Zurich et de Bâle ainsi qu'à l'EPF Zurich. Le soutien des scientifiques sera donc assuré après l'expiration de SyBIT et de SystemsX.ch.

Et que faut-il entreprendre pour ancrer ce savoir-faire dans la communauté scientifique?

Fort heureusement, nous pouvons aujourd'hui déjà observer que les chercheurs de SystemsX.ch appliquent les connaissances acquises dans de nouveaux projets. Les spécialistes des équipes de soutien apportent également leur aide à des projets ne faisant pas partie de SystemsX.ch, et ainsi les compétences acquises sont également transmises à d'autres groupes de recherche.

Les hautes écoles sont-elles prêtes à affronter la recherche à fort volume de données après la fin de SyBIT ?

Oui, en principe. Les équipes de soutien locales sont ancrées dans les hautes écoles sous forme de services en bioinformatique, et le SIB Institut Suisse de Bioinformatique en assumera la coordination après l'expiration de SyBIT. Ces dernières années, nous avons en outre aidé les institutions partenaires de SystemsX.ch à mettre en place l'infrastructure informatique nécessaire. Il convient maintenant de mettre en réseau les différentes ressources locales, afin que les hautes écoles puissent profiter de tous les services et de toutes les infrastructures spécialisées disponibles. Cela contribue également à ce qu'à long terme les chercheurs suisses continuent à faire partie de l'élite mondiale dans le domaine de la biologie des systèmes.