

Europe et la génomique en première ligne

Cette Assemblée Générale d'automne a permis de faire un point complet sur les activités et perspectives pour l'Association, notamment dans le cadre d'EuroHPC avec la mise en place de la nouvelle Commission Européenne. Elle a aussi été l'occasion de recevoir Jean-Marc Grognet, le Directeur Général de Genopole qui, après nous avoir rappelé ce qu'était la génomique et ses avancées, nous a expliqué pourquoi le HPC était indispensable dans les travaux du Genopole et pourquoi il avait souhaité se rapprocher de Teratec.



La 28^{ème} Assemblée Générale de Teratec, s'est déroulée le 10 octobre dans l'amphithéâtre du TGCC du CEA. Après avoir souhaité la bienvenue aux personnes présentes, représentant 55 des 85 sociétés membres de l'association, ainsi que le CEA, qui hébergeait cette AG, et les personnels de **Teratec**, **Daniel Verwaerde**, le nouveau Président de l'Association depuis octobre 2018, a présenté l'ordre du jour, en insistant sur l'intervention de **Jean-Marc Grognet**, le Directeur Général de **Genopole**

« Nous avons une relation de longue date avec le **Genopole**, mais nous avons décidé de renforcer les collaborations entre nos deux structures depuis un peu plus d'un an. C'est pourquoi nous accueillerons lors de cette Assemblée Générale **Jean-Marc Grognet**, qui nous présentera sa vision de notre partenariat ». L'Assemblée Générale a ensuite démarré suivant le planning habituel.

1. Toujours de nouveaux membres

Hervé Mouren, Directeur de Teratec, a présenté les candidatures de quatre nouveaux adhérents pour 2019 : **NAG**, **Tweag.io**, **Western Digital** et **Xilinx**. « Et les choses se mettent en place rapidement puisque NAG organise aujourd'hui même sur le Campus Teratec un séminaire d'une journée entière qui rassemble une trentaine de personnes ».

Ils nous ont rejoints en 2019 :

- NAG
- Tweag.io
- Western Digital
- Xilinx

Les quatre sociétés ont été accueillies dans l'association par un vote à l'unanimité. Après ces nouvelles arrivées, Teratec compte désormais 85 organismes membres. Chacun des nouveaux membres est ensuite venu présenter ses activités.

• **NAG** : des bibliothèques présentes dans toutes les applications

À commencer par **François Cassier**, Account Manager Europe de Numerical Algorithms Group Ltd (**NAG**), qui expliqué ce qu'était cette structure et comment elle fonctionne. « La collaboration est dans les gènes de NAG depuis sa création en 1970, puisque c'est à la base un projet collaboratif entre plusieurs universités britanniques dans le domaine du calcul numérique. Nous sommes devenus une société commerciale indépendante en 1976 avec la particularité d'être à but non-lucratif, vivant de ses revenus (produits et services). Nous avons réalisé un chiffre d'affaires de 8,4 m€ sur notre dernier exercice fiscal et employons 70 personnes ».



Cette structure britannique basée à Oxford, avec une entité HPC à Manchester, dispose de filiales aux USA et au Japon, ainsi que de bureaux en France et en Allemagne. NAG dispose de 3 000 sites clients dans le monde sur des marchés industriels, >>

financiers et pétroliers. NAG travaille aussi avec les fabricants de processeurs (**AMD, ARM, Intel...**) et des éditeurs de logiciels.

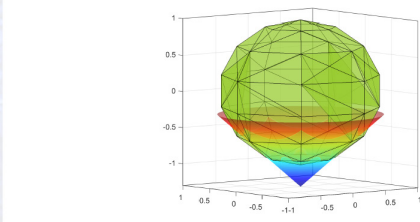
« *Le fait d'être une structure à but non-lucratif, nous met à l'abri des acquisitions, ce qui nous permet d'avoir une vision à long terme de nos missions. Elles sont depuis 50 ans de deux ordres. D'une part, proposer à nos clients des solutions numériques robustes et performantes pour qu'ils puissent résoudre leurs problèmes techniques et, d'autre part, les aider à créer la future génération des utilisateurs du calcul numérique dans leurs communautés.* »

NAG est bien connu pour sa librairie, qui comporte des centaines de composants mathématiques et statistiques, que les développeurs intègrent à leurs applications de calcul et d'analyse de données pour en améliorer les performances, dans les environnements Fortran, C, C++, Python, Matlab... « *Une bibliothèque qui est largement utilisée par le TGCC du CEA qui nous accueille pour cette AG* ».

Outre cette bibliothèque, NAG propose : des compilateurs Fortran très performant pour la recherche d'erreurs; les outils **DCO** dédiés à la différentiation algorithmique (calcul des dérivées d'un code) qui est une alternative à la différence finie; l'outil **Origami**, qui est un environnement de développement et d'exécution de Grid et Cloud Computing.

SOCP – closer look

Feasible region of an SOCP problem in 3D:



- ▶ introduces curvature (quadratic info) to otherwise linear model
- ▶ very powerful
→ SOCP could be the first choice after LP
- ▶ reformulation needed

©NAG

NAG propose aussi un certain nombre de services numériques tel le développement d'algorithmes ou plus orientés HPC comme de la modernisation et de la parallélisation de code, de l'évaluation technologique et du benchmarking, ainsi que du conseil à la mise en place de systèmes HPC et de l'assistance à l'approvisionnement.

« *Nous avons une longue histoire de collaboration avec les scientifiques et les ingénieurs à travers le monde qu'ils soient dans le domaine privé ou public, académique ou industriel. Par exemple, nous travaillons actuellement avec les Universités de Lancaster, Manchester, Oxford, Sheffield et Warwick pour enrichir la librairie NAG avec de nouvelles fonctions, ou encore avec l'Université de Stanford sur de l'optimisation mathématique. Nous avons aussi un partenariat avec l'Université d'Aix-la-Chapelle en Allemagne sur la différentiation algorithmique pour des applications industrielles et financières. Nous travaillons enfin sur des projets européens comme POP 2, qui vise à améliorer les codes, ou POEMA avec l'INT Marie Curie pour de l'optimisation polynomiale.* »

NAG est très impliqué dans la communauté numérique, puisqu'il participe à de nombreuses instances normatives (langages

Fortran et **MathML**, système de fichiers Lustre **OpenSFS**, librairies **Blas & Lapack**). Il collabore aussi avec de nombreuses universités, pour faire la promotion des carrières en sciences, technologies, ingénierie et mathématiques, ainsi qu'à l'initiative **Women in HPC**.

Suite à une question de la salle, François Cassier a précisé : « *Nous travaillons effectivement principalement autour des CPU, mais nous avons aussi une activité de conseil autour des accélérateurs GPU.* »

• **Tweag.io** : l'Open Source pour des applications rapides, correctes et maintenables

Ensuite **Benjamin Robin**, Chief Financial Officer, est venu présenter Tweag.io et ses activités. « *Nous sommes un laboratoire d'innovation logicielle qui a été créé en 2013, par son actuel président, Mathieu Boespflug, et un petit groupe de mathématiciens et d'experts en méthode formelle et systèmes distribués, qui travaillaient à l'origine notamment pour le laboratoire pharmaceutique américain Amgen. Ils ont fini par développer un module Open Source pour la communauté Haskell, un langage de programmation fonctionnelle ancien, qui refait surface avec la capacité que l'on a d'intégrer différents types de langages dans un même code.* »



©TERATEC

Ce savoir-faire de développeur a été remarqué par des industriels qui ont confié des développements autres à l'équipe, qui compte aujourd'hui une quarantaine d'ingénieurs. « *Grâce aux nouvelles technologies, nous pouvons facilement aller chercher les talents là où ils se trouvent, ce qui explique que nous avons aujourd'hui des salariés d'une vingtaine de nationalités dont la plupart travaille depuis leur pays de résidence. Fort de ces différentes expertises, on arrive à répondre aux enjeux de nos clients dans des domaines très différents (IA, Blockchain, Bio-informatique, Infrastructures Cloud...).* »

C'est la complémentarité de ces différentes expertises qui permet à Tweag.io d'avoir 80 % de clients internationaux, qu'il accompagne des premiers développements à la mise en service de leurs applications sur des thématiques généralement pointues. « *Notre engagement vis-à-vis des travaux pour nos partenaires s'articule autour de trois qualificatifs : Rapide dans l'exécution; Correct par construction; Maintenable dans le temps. C'est ainsi que nous avons eu deux contributions majeures. Notre équipe comprend les inventeurs de Nix, qui est un langage de programmation décrivant de manière uniforme et reproductible les systèmes de configuration des ressources dans le Cloud. Elle comprend aussi les principaux contributeurs du langage Haskell, dont la rigidité dans les années quatre-vingt/90 est aujourd'hui oubliée, et qui fait son trou pour tous les sujets qui demandent un niveau d'exactitude des données extrêmement fort (analyses médicales, Blockchain...).* »

>>

Tweag.io a ainsi développé en Open Source un certain nombre de bibliothèques et d'outils qui permettent de passer d'un langage à l'autre, notamment entre **R** et Haskell. Tweag.io travaille aussi sur des transcriptions entre le moteur de production Open Source **Bazel**, développé par **Google**, et Haskell, ou entre le framework Open Source de calcul **Spark** et Haskell.

« Nous faisons en sorte avec nos clients, qui globalement jouent le jeu, que la plus grande partie des projets sur lesquels nous travaillons soit Open Sourced. Ainsi nous avons travaillé par exemple avec **Pfizer** sur un processus scientifique reproductible pour séquencer et analyser les gènes CRISPR de manière automatisée pour réduire les coûts et les délais. En bio-pharma, on travaille avec l'entité lyonnaise **Novadiscovery** pour laquelle nous avons créé en Haskell un outil de simulation in-silico de la physiologie humaine, afin de réduire le coût des essais cliniques en prédisant les interactions biochimiques intervenant dans le cadre d'un traitement, de manière à supprimer un certain nombre d'essais physiques, pour ce qu'il est possible de prendre en charge ».



©TWEAG.IO

« On travaille en matière de Blockchain avec **IOHK**, le créateur de la crypto-monnaie **Cardano**, sur le langage intelligent **Plutus** qui lui sert de support. Enfin, on a travaillé avec le **CEA** sur l'industrialisation de l'analyse et du traitement des logs, des bugs et des anomalies générées par les super-calculateurs, qui sont, compte tenu du volume de données dont on parle, absolument énormes en termes de traitement humain, alors que l'on est capable en MLP de dire à la machine de retrouver certains mots-clés, de recouper les informations, d'éliminer les déchets et de ne présenter aux opérateurs que ce qui "mérite" d'être traité ».

• **Western Digital** : faire face à l'explosion du volume de données

Davide Villa, Director EMEA Business Development de **Western Digital** a ensuite décrit les activités de sa société qui, outre les disques externes WD, regroupe aussi les marques **SanDisk**, leader des clés USB et **HGST**, l'ancien **IBM Storage**, pour les stockages de masse.

« Notre histoire est synonyme d'innovation depuis les premiers disques durs apparus en 1956 chez IBM, en passant par les premières mémoires Flash avec **SunDisk** en 1988, jusqu'au stockage 96 layer QLC NAND introduit en 2018 pour sauvegarder les ZB de données générés aujourd'hui par les plus grandes applications ». Une culture de l'innovation qui se traduit par le fait que Western Digital dispose du 4e portefeuille commercial de brevets dans le monde, derrière **Amazon**, **Google** et **Apple**.



©TERATEC

Une innovation qui va des puces aux systèmes complets. « Nous sommes l'un des 5 fabricants mondiaux restants de mémoires Flash, pour lesquels nous développons en interne notre silicium, notre firmware, nos contrôleurs qui sont utilisés dans les disques statiques à semi-conducteurs SSD et dans les disques durs HDD. Mais au-delà de ces composants, nous les assemblons pour faire des systèmes intégrés demandant beaucoup de stockage et de performances (Big Data/Fast Data) ».

Dans les systèmes d'entreprise, l'offre de Western Digital va des disques de stockage HDD et SSD jusqu'aux systèmes les plus complexes allant jusqu'à 1,4 Po dans des boîtiers 4U. Et le besoin est énorme : « Aujourd'hui 90 % des données qui existent ont moins de 2 ans d'existence. Et la tendance ne va pas changer. Il faut donc proposer au marché des systèmes de stockage toujours plus efficaces pour absorber cette quantité phénoménale de données, qui est actuellement de l'ordre 2,5 EB produit par jour, soit 1018 ou un milliard de milliards d'octets ! ».

D'après **IDC**, on aurait déjà produit 35 ZB, soit 10^{21} , de données dans le monde, alors que la base installée mondiale de stockage n'est capable d'en absorber que 5 ZB. Et la situation ne va pas s'arranger, car le même organisme prévoit qu'en 2025, 175 ZB de données auront été créés, alors que l'ensemble des fournisseurs n'auront pu faire passer la capacité de stockage mondiale que de 5 à 22 ZB. « Nous avons donc une pression énorme pour développer plus rapidement des systèmes de stockage plus grands, afin d'absorber la plus grande partie de ces données, d'autant plus que Western Digital représente 43 % des capacités de stockage actuelles dans le monde ».



©WESTERN DIGITAL



• Xilinx : des FPGA pour accélérer les traitements des données

Thomas Boudrot, Sr. Director Business Development de **Xilinx**, a clos la session de présentation des nouveaux membres. « Mon rôle est aujourd'hui d'évangéliser notre savoir-faire en FPGA dans le monde du Data Center et du HPC. Notre chiffre d'affaires est de 3 B\$ pour 4 400 employés et plus de 20 000 clients. Nous avons inventé le Field Programmable Gates Array (FPGA) ou réseau de portes programmable après fabrication, voici une trentaine d'années. Nous sommes actifs sur 8 grands marchés : les Télécoms qui représentent encore plus de 36% de notre chiffre d'affaires; mais aussi Aerospace & Defense; Industrial & Medical; Test, Mesure & Emulation; Automotive & Transportations; Audio, Video & Broadcast; Consumer; Data Center & HPC ».



© TERATEC

Ce dernier segment s'est depuis 2 ans tourné vers le FPGA. La nature même du composant est une architecture massivement parallèle, ce qui permet de répondre mieux que des architectures traditionnelles à base de CPU aux besoins de traitement toujours plus grands. « C'est pourquoi, plutôt que de continuer à proposer des composants, nous avons développé une plate-forme d'accélération à base de cartes PCI Express, **Alveo**, qui se place directement dans n'importe quel serveur pour accélérer du calcul, du réseau ou même du stockage ».

Le développement de ces solutions est dû plusieurs facteurs : on passe du 'CPU-Centric' vers de l'informatique personnalisée et distribuée; l'avènement de l'Intelligence Artificielle et du Machine Learning, couplé à d'autres applications, change la donne. « Les GPU peuvent très bien se sortir du Machine Learning seul, par contre si on le couple à des applications (analyse génomique; encodage/décodage vidéo; compression/encryption...) la structure flexible du FPGA est beaucoup plus adaptée. C'est ce que l'on adresse avec nos cartes d'accélération qui sont déployées soit dans le Cloud (**Microsoft Azure**, **Amazon**, **Alibaba**, **Tencent**...), soit On Premise chez le client ».

Suivant les applications, le facteur d'accélération du FPGA face à une architecture de type CPU, peut aller de 5 à 90. Ainsi, une analyse de génome, qui demandait 23 heures au spécialiste américain **Illumina**, est réduite à 12 minutes en ajoutant au serveur de calcul une carte d'accélération à base de FPGA. On approche le temps réel !



© XILINX

Un des freins évoqués à l'utilisation du FPGA est le besoin de programmation. Le fait que l'on ne puisse jusqu'à maintenant le programmer qu'en VHDL ou en langage de bas niveau, a limité son utilisation. « Nous avons annoncé début octobre la plate-forme de développement logiciel **Vitis** dédiée à la programmation de nos FPGA dans des langages **C**, **C++**, **Python**, capable d'utiliser des bibliothèques de fonctions mathématiques Open Source, qui seront mappées sur nos cœurs d'accélération. Et si vous utilisez des frameworks pour le Machine Learning de type **TensorFlow**, **Cafe** ou même **FFmpeg** pour ce qui est de la vidéo, restez dans ces frameworks et notre outil va directement convertir votre code et le mapper sur nos plateformes, ce qui permet aux utilisateurs de cibler des applications dans le Data Analytics, l'Image Processing, le Video Processing, le Machine Learning, les Life Sciences & HPC, le Financial Computing, en utilisant des bibliothèques Open Source ou développées par des partenaires spécialisés ».

2. Le défi numérique de la médecine Génomique

Daniel Verwaerde a ensuite accueilli Jean-Marc Grognet, le Directeur Général de Genopole. « Jean-Marc Grognet est un grand scientifique. Biologiste de formation il a fait une bonne partie de sa carrière au CEA, avant de prendre en 2017 la direction du Genopole. Nous avons depuis renforcé nos liens et considéré qu'un partenariat entre Genopole et Teratec serait bénéfique pour nos deux structures, car le numérique a une place grandissante dans le décodage du génome. C'est ce qu'il va nous rappeler ».



© TERATEC

« Le Genopole est une vision qui a été portée il y a 20 ans pour créer un biocluster, c'est-à-dire un campus d'innovation qui rassemble sur un territoire le plus resserré possible (Evry, Corbeil, Courcouronnes) autour d'une thématique focalisée de grande importance (la génomique), le triangle de la connaissance (l'enseignement supérieur, des acteurs de recherche académique, des industriels). Et nous avons mis en plus quelque chose d'essentiel au milieu du triangle, le patient. La première raison est historique parce que l'**AFM-Téléthon**, qui lutte contre les maladies neuromusculaires, a eu ses premiers laboratoires sur le site d'Evry. La seconde raison est que le Centre Hospitalier Sud-Francilien (**CHSF**), le plus important en Ile-de-France hors Paris avec 1 000 lits et 3 000 soignants, est au cœur de notre territoire », explique Jean-Marc Grognet.

• 5 600 personnes autour de la Génomique

Et la réussite a été au rendez-vous, puisque 20 ans après le campus d'innovation Genopole c'est 5 600 personnes dont 2 400 emplois directs. Elles se répartissent dans 16 laboratoires



académiques sous tutelle de l'INSERM, du CNRS, du CEA, des **Universités d'Evry Paris Saclay** et de **Paris-Sud** (1 000 personnes), ainsi que dans 96 entreprises labellisées (1 500 personnes) et 29 plateformes technologiques de pointe.

Celles-ci permettent de doter un laboratoire ou une entreprise, qui ont les compétences techniques, d'un matériel de pointe qu'elles n'auraient pas forcément les moyens de financer, et de les aider à l'opérer et à l'ouvrir à l'ensemble des acteurs du site, pour faciliter la mutualisation des savoirs. Par exemple, l'**Ecole des Mines** de Paris a sur Evry un laboratoire travaillant sur les matériaux, qui a été doté d'un microscope électronique, aujourd'hui accessible aux applications de biologie des laboratoires et entreprises membres du Genopole.

©Christophe Hargoues - Genopole



Ce sont aussi de grandes écoles : École nationale supérieure d'informatique pour l'industrie et l'entreprise (**ENSIIE**) ; **Telecom Sud Paris** ; Institut Mines-Télécom Business School (**IMT BS**).

Enfin, c'est une réussite économique puisque les entreprises du Genopole lèvent en moyenne 70 à 80 millions d'Euro par an. Dernière opération en date, **Ynsect**, spécialiste des protéines d'insectes, créé il y a 5 ans au Genopole, vient de lever 125 m\$.

Aujourd'hui le Genopole c'est 5 campus, qui devraient être rapidement complétés par 2 autres, qui, outre des laboratoires, hébergeront aussi des unités de production.

• L'ADN de Genopole c'est l'ADN

« La génomique est depuis un siècle une façon de répondre à la question très simple du caractère héréditaire de certaines choses et aux exceptions constatées. Pourquoi l'enfant d'un couple aux yeux bleus n'a-t-il pas les yeux bleus ? On savait aussi qu'une partie des maladies avait une composante génétique que la génomique essayait d'expliquer. La réponse s'est trouvée dans l'ADN, c'est-à-dire cette molécule qui est dans le noyau de nos cellules ».

L'ADN est une longue chaîne qui mesure en moyenne 2 m, ce qui a un impact sur la quantité d'informations portée par cette molécule. Le noyau de chaque cellule est structuré en 23 paires de chromosomes chez l'homme. Des protéines associées à l'ADN sont présentes dans ces chromosomes. Cette molécule d'ADN est formée d'un enchaînement de 4 types de motifs moléculaires (les bases) qui représente toute l'information portée par la cellule, soit 3 milliards de paires de bases !

« L'ADN total d'un organisme représente son génome. La quasi-totalité des 70 000 milliards de cellules qui constituent le corps

humain possèdent le même ADN. Décrypter le génome, c'est donc acquérir ces 3 milliards d'informations, soit 3×10^9 ».

Il suffit que l'enchaînement de la molécule d'ADN comporte des erreurs au niveau de l'une des bases pour que les capacités des protéines produites soient modifiées, entraînant des anomalies héréditaires ou des maladies génétiques.



©Christophe Hargoues - Genopole

• Séquencer le génome

Jusque dans les années quatre-vingt-dix, on a essayé de comprendre ces mécanismes, restait à séquencer le génome. Si l'on essayait de prendre la molécule et d'en lire chaque fragment d'ADN, à raison de 1 seconde pour chaque base, cela prendrait environ 100 ans pour lire le génome entier d'un individu !

D'où le recours au calcul massivement parallèle en découpant l'ADN en une multitude de fragments monobins. Des séquenceurs à haut débit exécutant des protocoles complexes lisent la suite de bases de chaque fragment. Les fragments d'ADN lus sont ensuite réassemblés par analyse informatique. L'ordinateur reconstitue les génomes et les stocke dans de grandes bases de données.

Pour effectuer le premier séquençage du génome humain, le **Human Genome Project** a été mis en place en 1990. Une dizaine de laboratoires dans le monde se sont consacrés chacun au séquençage d'un chromosome. Pour la France, la séquence du chromosome 14 a été révélée en 2001 par **Genoscope** (laboratoire du CEA situé sur le Genopole). Le séquençage du génome humain complet fut achevé en 2003.

Reste à trouver la signification de ce message en essayant de retrouver dans le génome les gènes, des séquences spécifiques de bases, environ 22000 pour l'être humain.

• De 100 M\$ à 100 \$ en 20 ans

Ce séquençage a un coût. Si celui du premier séquençage par le Human Genome Project a pu être estimé à 100 M\$, on a ensuite suivi la Loi de Moore en gagnant un facteur 2 environ tous les 18 mois jusqu'en 2007. Un séquençage coûtait alors environ 10 M\$. *« C'est alors que la société américaine **Illumina** a trouvé une méthode fracassante pour paralléliser massivement le séquençage, ce qui a fait chuter drastiquement les coûts pour arriver aujourd'hui à 1 000 \$ par séquençage ».*

>>

« Cela veut dire qu'on est capable aujourd'hui d'avoir un séquençage de génome pour un prix similaire à une analyse biologique complexe. Ce qui devient acceptable dans une pratique médicale classique. La question maintenant n'est plus de savoir si l'on va arriver à 100 \$ un jour, mais de savoir dans combien de temps, 3 ans, 6 ans ? »

Mais outre la chute du coût, le délai a lui aussi chuté drastiquement. Les dernières machines de séquençage traitent 48 génomes humains en parallèle en 44 heures. « Avec les technologies moléculaires, la robotique, l'informatique et l'Intelligence Artificielle, la biologie vit une révolution méthodologique. Nous entrons dans l'ère du Big Data et du séquençage pour tous ». Ce qui permet d'envisager de multiples applications nouvelles.

Tout d'abord explorer la diversité du monde vivant en connaissant mieux les espèces vivantes, c'est la génomique. Mais aussi de découvrir des écosystèmes complexes formés de toute une communauté d'organismes grâce à leur ADN, c'est la métagénomique. Cela va aussi aider à comprendre les mécanismes du vivant, en déterminant la fonction des gènes repérés dans les génomes, c'est la génomique fonctionnelle. Enfin, on va mieux comprendre l'humain et sa santé, c'est la génomique médicale.

• La génomique annonce le retour des grands explorateurs

La génomique permet aussi de voir ce que l'on ne savait pas voir avant. Par exemple, l'expédition **Tara Oceans**, grâce à ses centaines de milliers de prélèvements d'eau effectués dans toutes les mers du monde et aux séquençages effectués dessus, a recensé 117 millions de gènes dont plus de la moitié étaient jusque-là inconnus. On repère ainsi des espèces que l'on n'avait jamais vues ! De même, le séquençage a totalement changé la vision du microbiote intestinal. De quelques espèces de bactéries identifiées il y a 40 ans, on est passé à plusieurs centaines aujourd'hui et il en reste encore à découvrir !



© Tara Expeditions

Connaitre le réarrangement génomique d'une tumeur peut, par analyse de ses caractéristiques génétiques, donner des informations précieuses aux médecins pour orienter les traitements. Pour cela, il va falloir faire en amont un travail d'acquisition, d'interprétation, de digestion et de présentation des données, pour ne pas les submerger sous une masse d'informations inexploitable.

Cette médecine génomique ou personnalisée se divise en deux grandes branches. L'analyse du génome va permettre de détecter précisément les prédispositions d'un individu à des pathologies en fonction de sa carte d'identité génétique et de faire de la médecine préventive en lui donnant des conseils de comportement, voire en lui prescrivant des traitements préventifs pour diminuer les risques.

L'analyse du génome va aussi permettre de détecter des pathologies rares en posant le bon diagnostic du premier coup. Dans le cas de maladie avérée, l'analyse du génome va permettre d'évaluer différents traitements et de choisir du premier coup celui qui sera le plus efficace et le plus rapide, avec le moins d'effets secondaires, c'est la médecine curative. « Depuis environ deux ans en cancérologie, il est devenu possible non plus de traiter une pathologie rénale ou digestive, mais une mutation génétique avec un médicament dédié quelle que soit la localisation de la tumeur. C'est un changement radical de la vision du traitement ».

Tout cela est au cœur du grand plan **France Médecine Génomique 2025**, né en partie au Genopole, qui vise à doter la France d'un grand nombre de plateformes dédiées au séquençage qui enverront les informations obtenues à un Collecteur Analyseur de Données (CAD) qui en assurera le traitement et l'interprétation pour chaque patient et alimentera un centre de recherche, le **Crefix**, qui validera les procédures et les appareils.

>>



© S. Bollet - Tara Expeditions

• Personnaliser l'approche médicale

Le séquençage du génome de chaque individu va révolutionner les pratiques médicales en personnalisant le diagnostic et les traitements. Mais pour que cela soit efficace, il va falloir pouvoir connaître et interpréter en temps réel les informations contenues dans le génome, d'où un besoin grandissant en puissance et performances de calcul. « L'idée est de pouvoir traiter de plus en plus rapidement des flots de données hétérogènes grâce à des algorithmes, notamment d'Intelligence Artificielle, et des outils bio-informatiques dans un contexte d'interdisciplinarité médicale croissante pour poser le bon diagnostic ».

• Génomique et Big Data

Contrairement à beaucoup d'applications de Big Data où l'on traite quelques informations sur un très grand nombre d'individus, en médecine génomique on traite un très grand nombre de descripteurs (code génétique, mutations...) sur un très faible nombre d'individus.

« On estime que l'on va avoir 10^{21} bytes d'information à traiter par individu à multiplier par plusieurs centaines de milliers de patients par an. Ce qui, outre la puissance de traitement, va aussi nécessiter de très gros moyens de stockage, environ 10 ExaBytes par an ».

Cela pose aussi des questions car si l'acquisition n'est plus un problème, les données doivent rester confidentielles, mais l'on doit être capable de les rattacher à un dossier médical pour qu'elles aient un intérêt. Comment fait-on pour garantir cette confidentialité? À qui appartiennent ces données, au patient, au médecin qui a fait la prescription, au centre de séquençage, à la collectivité donc l'état? Qui va être chargé de les conserver et avec quelle pérennité, car les informations d'aujourd'hui pourront être utiles pour traiter une pathologie dans 30 ans? Qui les interprétera et qui les réinterprétera au fil des évolutions de la science?

• Pas de génomique sans numérique

« C'est pourquoi nous nous sommes rapprochés de **Teratec**, car vous avez compris comme nous que la médecine du futur sera basée sur la génomique et que pour cela nous avons besoin d'une informatique de calcul toujours plus performante pour le traitement sécurisé de très gros volumes de données, tout en garantissant leur pérennité sur le long très terme. Des préoccupations qui sont au cœur des activités de Teratec. Nous envisageons d'ailleurs de créer le premier institut de "génomique numérique" au monde, pour répondre à ces besoins, car nous avons sur notre territoire les centres de séquençage parmi les plus puissants d'Europe et des structures, telle **Teratec**, ayant de très fortes compétences en calcul numérique haute performance (HPC) », conclut Jean-Marc Grognet.

Suite à une question de **Christian Saguez**, Jean-Marc Grognet a précisé que l'étude du génome ne ferait pas tout en médecine personnalisée. « La génomique toute seule n'apportera pas la réponse à tout. Au cours des deux dernières décennies, on a découvert que tout n'était pas dans le génome. Contrairement à ce que l'on pensait notre génome n'est pas immuable et que nous pouvons par notre expérience de vie et notre exposition des facteurs extérieurs (radiations ionisantes, produits chimiques...) connaître des modifications de notre génome, qui seraient transmissibles à d'autres générations ».

Cela va à l'encontre du dogme Lamarck/Darwin où Lamarck défendait l'hérédité des caractères acquis tandis que Darwin disait c'est immuable. « On pensait que Lamarck avait tort, mais on ne se rend compte qu'aujourd'hui que la réponse est beaucoup moins ferme. On montre, par exemple, qu'une souris pouvait transmettre à sa descendance et à la descendance de celle-ci, la peur issue d'un traumatisme. Il y a donc une transmission des caractères acquis. On a aussi découvert que des virus étaient capables de s'intégrer dans le génome et d'être ainsi transmis à une génération suivante. Outre la génomique la médecine de demain intégrera donc d'autres paramètres comme l'expérience de vie et l'exposition des facteurs extérieurs. Ce qui suppose le traitement de multiples données à la fois personnelles et collectives ».

Suite à une question de **Karim Azoum** sur l'utilisation du Machine Learning et de l'IA dans les travaux de Genopole, Jean-Marc Grognet a précisé : « Je ne sais pas si on peut parler d'Intelligence Artificielle, mais nous avons aussi toute une activité qui consiste à connaître les réseaux d'interaction entre les gènes. Cette partie est extrêmement importante car l'interaction des réseaux de gènes a une responsabilité dans les pathologies. D'autres équipent travaillent aussi sur la prédiction de pathologies ».

3. Grandes initiatives européennes

Daniel Verwaerde a ensuite repris la parole pour présenter le rapport d'activité de l'association.



©TERATEC

« Concernant les grandes initiatives européennes, le point marquant des dernières semaines a été, suite aux élections européennes, le changement de la Commission. Nous sommes passés de la **Commission Junker** (2014-2019) à la **Commission Van Der Leyen** (2019-2024). Cela se traduit par des changements assez importants pour ce qui concerne le rattachement de l'activité des membres de notre association, qu'il s'agisse de HPC et plus généralement des technologies de l'information. En effet, dans la Commission précédente le HPC et la **DG Connect**, qui en a la responsabilité, étaient rattachés à la Commissaire à la Société et à l'Economie Numérique, **Mariya Gabriel**, qui dépendait du Vice-président **Marché Unique Numérique**. Cela s'était traduit par la création de la **Joint Undertaking EuroHPC**, indépendante de la Commission et pilotée par DG Connect. Ce qui est entrain de changer, c'est que cette activité du HPC et, d'une manière plus générale, de la digitalisation de l'économie et de la société, est jugée encore bien plus importante par cette nouvelle Commission et du coup le pilotage quasi-opérationnel ne sera plus effectué par un simple Commissaire, mais par une des deux Vice-présidentes, celle qui est chargée d'une Europe préparée à l'Ere numérique, **Margrethe Vestager**. Du coup, on aura en dessous d'elle en rôle d'exécutant, un HPC et une DG Connect qui seront rattachés au Commissaire au marché intérieur Industrie, Défense et Espace ».

• Préparer le futur de EuroHPC

Rappelons que EuroHPC rassemble la Commission et 29 états membres avec un budget de l'ordre de 1 milliard d'Euro sur la période 2019-2020, financé pour moitié par la Commission, puis de 7 milliards pour la période 2021-2027. La mission est double : acquérir en Europe de très gros ordinateurs, entrant dans le Top 5 mondial, pour les mettre à la disposition des chercheurs et de l'industrie; développer en Europe une capacité industrielle complète pour produire, vendre, installer, maintenir ces super-calculateurs. « **Mariya Gabriel** avait coutume de dire que 30 % des super-calculateurs mondiaux étaient utilisés en Europe et que son ambition était que 30 % soient produits en Europe ».

>>

Le programme EuroHPC avance rapidement. Pour les machines de 150 à 200 PFlops, 3 sites ont été retenus (Kajaani en Finlande, Barcelone en Espagne, Bologne en Italie), les appels d'offres sont lancés et les installations devraient avoir lieu fin 2020. Le budget total sur 5 ans est de l'ordre de 650 m€. « *La France et l'Allemagne ne font pas partie de ces trois consortiums, préférant se réserver pour la phase suivante qui concernera des machines exaflopiques 5 à 10 fois plus performantes* ». 5 autres machines de 5 à 30 PFlops vont aussi être acquises pour un budget total sur 5 ans de 180 m€. Elles seront installées au Luxembourg, au Portugal, en République Tchèque, en Slovaquie et en Bulgarie, à partir de la mi-2020. « *Ces sites n'ayant pas forcément une forte antériorité dans l'exploitation de ce genre de machines, les membres de Teratec ayant cette expertise pourraient trouver là de nouveaux débouchés* ».

« *Nous préparons actuellement à Bruxelles l'évolution d'EuroHPC pour la décennie à venir. Il est prévu de continuer à acheter des machines de manière à ce qu'il y ait toujours en Europe l'une des 5 machines les plus performantes du monde. Nous avons aussi la volonté d'avoir une chaîne technologique au niveau permettant d'y arriver et un écosystème de développement d'applications adapté. Enfin, la Commission souhaite que toutes ces machines disposent d'une interconnexion extrêmement rapide pour en faciliter l'accès et l'usage* ».

Le rôle de Teratec dans EuroHPC est triple. « *Tout d'abord et je viens de l'évoquer, nous participons activement à 2 'Advisory Groups' (Infrastructure Advisory Group, **INFRAG**; Research and Innovation Advisory Group, **RIAG**), dont la mission est de rédiger des documents stratégiques pour la sélection des plateformes et la définition de la feuille de route technologique, afin de continuer à faire jeu égal avec les autres pays leaders mondiaux du domaine. Nous travaillons aussi sur la rédaction du « Vision Paper 2021/2030 » pour la feuille de route EuroHPC* ».



©EUROHPC

• Exprimer les besoins des utilisateurs

Deuxièmement, à la demande de la Commission et de la DG Connect, Teratec a été chargé de produire un document d'expression du besoin industriel en matière d'utilisation des super-calculateurs d'EuroHPC. L'objectif pour la Commission étant de s'en servir pour définir sa politique d'ouverture de ces super-calculateurs aux industriels, des PME aux Grands Groupes.

Troisièmement, « *Nous participons également à 3 grands Appels à Projets lancés en juillet dernier : EuroHPC 03 'Industrial software codes for extreme scale computing environments and application' ; EuroHPC 04 'Competence centers' ; EuroHPC 05 'Stimulating the innovation potential of SME's' qui est la suite du projet européen de **FORTISSMO**, cousin du français **SIMSEO*** ».

« *Bien évidemment dans ce contexte, Teratec est prêt à être le facilitateur de montage de projets par ses membres dans le cadre des différentes actions de EuroHPC* ».

• Devenir le Centre de Compétence HPC français

L'action EuroHPC 04 vise à créer un unique Centre de Compétence HPC dans chaque pays européen. Il servira d'interface avec la Commission et les autres états membres, pour décliner la politique européenne en la matière, et apportera les services suivants à l'ensemble des acteurs économiques du pays :

- Diffuser les usages des technologies HPC/HPDA;
- Proposer un ensemble complet de services (conseil, études spécifiques);
- Faciliter l'accès des industriels aux capacités matérielles et logicielles HPC/HPDA;
- Apporter un support pour le déploiement des outils;
- Assurer la formation et l'information;
- Rassembler les informations permettant de mesurer l'apport du Centre de Compétence.

« *Le Gouvernement français a désigné Teratec auprès des autorités de Bruxelles comme étant son porte-bannière pour le Centre de Compétence français. Nous avons donc fait avec le **CERFACS**, dont le travail depuis des années auprès des industriels le rend plus que légitime dans cette structure, une proposition d'organisation aux pouvoirs publics français. Le comité d'organisation comportera 5 membres : **Teratec** (Présidence); **CERFACS** (Vice-Présidence); **France Industries**, qui régit les 17 branches industrielles françaises; **GENCI**, qui est le maître d'ouvrage pour les grands équipements de calcul destinés à la recherche; la **Maison de la Simulation**, une organisation qui regroupe le **CNRS**, le **CEA** et les **Universités de Paris-Saclay**, de **Paris-Sud (UPS)** et **Versailles Saint-Quentin-en-Yvelines (UVSQ)** ».*

L'objectif du Centre de Compétence est d'avoir une liaison permanente avec le 'Groupe Miroir' EuroHPC, ainsi qu'avec les grands organismes de recherche pouvant apporter des technologies et des savoirs (**CEA**, **CNRS**, **INRIA**, **AFneT**, **AMIES**, Centres techniques, Pôles de compétitivité, IRT...). De même, le Centre de Compétence associera le monde industriel et ceux de l'enseignement et de la recherche à travers des réunions régulières. Le budget (EuroHPC + Etat) sera de 1 m€ par an, ce qui permettra de financer 8 personnes et 200 K€ d'investissement.

« *L'Union Européenne et EuroHPC sont importants pour nous car on sent bien que les canaux franco-français vont devenir européens et que si l'on veut poursuivre notre action au profit du tissu économique français, il est extrêmement important que nous jouions totalement le jeu de EuroHPC* », a conclu Daniel Verwaerde sur ce chapitre.

4. Point sur les activités

Hervé Mouren, directeur de Teratec, a ensuite repris la main pour faire le point sur l'ensemble des activités de l'Association en cours.

• Les programmes européens

« *Teratec participe à des programmes européens depuis 4 ans tel **POP**, **Excellerat** ou **Focus CoE**. **POP** est l'un des Centres d'Excellence (CoE) du Programme-cadre de recherche **H2020**. Trans-sectoriel, il est dédié à la performance des applications. C'est une mécanique remarquable* ». **POP** permet à n'importe quelle organisation

>>

de rencontrer les organismes compétents sur l'analyse des performances et l'optimisation de leurs logiciels. Ceux-ci analysent gratuitement la performance d'un code et indiquent à son auteur ce qu'il faut faire pour l'améliorer très significativement. Ce service est payé par la Commission Européenne, qui y a attribué un budget de 7 m€. « Je vous engage fortement à utiliser ce service pour améliorer les performances de vos codes ».

Autre programme européen, **Excellerat**, qui est un Centre d'Excellence assurant la promotion de la simulation numérique dans les applications d'ingénierie à l'aide de 6 codes de calcul et 11 cas d'usage. Enfin, **Focus CoE** est une sorte de tour de contrôle qui regarde comment se développent les résultats de tous les CoE sectoriels européens.

« Entre ces opérations, qui sont de qualité, plus le lancement des opérations EuroHPC et le rôle que Teratec va avoir en tant que Centre de Compétence, cela va transformer notre image. Teratec va être vu comme le Centre de Compétence français sur ces sujets, et comme en plus à la demande de la Commission on va être embarqués dans le mécanisme de coordination des différents Centres de Compétence en Europe, notre image européenne va se renforcer et nous allons être dotés de moyens tout à fait significatifs ».

Mais ces visées européennes ne doivent pas faire oublier les activités françaises. « Le programme **SiMSEO**, qui s'est terminé fin septembre, a été un franc succès. On nous demandait de mettre le pied à l'étrier de la simulation à 100 PME dans les mondes de la mécanique et du BTP. On en a fait basculer 360! Nous nous sommes pour cela appuyé à la fois sur les organisations professionnelles et sur des éditeurs spécialisés, pour proposer des catalogues d'offres logicielles et de services validées. Les PME ont ainsi investi plus de 2,7 m€, subventionnés à hauteur de 50 % par **BPI France**, dans le cadre des **Investissements d'Avenir** ».

• Les initiatives industrielles

Il y a eu un léger ralentissement sur les lancements des initiatives industrielles cette année, car il faut maintenant avancer celles lancées.

Celle sur le calcul quantique Teratec Quantum Computing Initiative (**TQCI**), qui rassemble de grandes entreprises (**Atos, CEA, EDF, IFP, Total...**), devrait être confortée par une initiative publique, visible en fin d'année, qui devrait nous fournir la possibilité de construire un véritable projet industriel.



©TERATEC

Sur la Fabrication Additive, un premier travail a été fait montrant l'intérêt du sujet et les points d'application que l'on pourrait imaginer dans un projet industriel.

Quant aux autres (Systèmes Autonomes, Ressources Naturelles, Agroalimentaire), ils sont dans des phases de démarrage ou de conception.

• Le Campus Teratec

Hervé Mouren a ensuite mentionné le bon fonctionnement du **Campus Teratec** avec 22 structures installées, dont 4 laboratoires de recherche, regroupant 250 personnes.

Il a insisté sur une initiative qui aura lieu sur le Campus le 28 novembre : « Nous allons recevoir 200 étudiants pour répondre aux besoins de recrutement des entreprises, qu'il s'agisse de petites ou de grandes structures. Les Universités de Versailles Saint-Quentin-en-Yvelines (**UVSQ**) et **Léonard de Vinci** vont faire venir sur le Campus, dans le cadre de leur cursus avec leurs professeurs, leurs élèves de dernière année de Master. Ils rencontreront les entreprises membres de Teratec qui leur présenteront leurs savoir-faire ».

Le Campus Teratec

Entreprises Technologiques

• Atos • CEA • ESI • Intel • Egis •

Pépinière d'entreprises

• AS+ • Assystem Technologies • Avise Conseil • CesiMat •
• Distène • ETP4HPC • Fullscale • Ingérop • Mantenna •
• MicroTrans • NumTech • ParaTools • RTSolution •
• Sema Software • Synomen • Teratec • WizYoo •

Laboratoires de recherche industrielle

• Extreme Computing • Exascale Computing Research •
• Intel Big Data Analytics • CesiMat •

• Communication et Forum Teratec

La communication de Teratec suit son cours normal avec : la publication d'un 7^e supplément de *L'Usine Nouvelle* et *Industrie & Technologies*, consacré à *Une année de simulation*, qui est diffusé à 55000 exemplaires; la mise à jour de la plaquette Teratec; la diffusion de 6 Newsletters entre janvier et juin 2019 diffusées à 60000 contacts.

« Je voudrais mettre en avant la très bonne tenue du **Forum Teratec 2019** avec quelques originalités dont les tables rondes, qui ont bénéficié d'un retour très positif et qui devraient être reconduites l'année prochaine. Concernant les ateliers et l'exposition, nous sommes à la limite du système dans les locaux actuels. Nous allons donc être amenés à réfléchir à la façon dont nous allons le faire évoluer ».

Il a ensuite annoncé que le **Forum Teratec 2020** aurait lieu les 16 et 17 juin 2020 à l'Ecole Polytechnique. Les appels à communications, à animation d'ateliers, ainsi qu'à devenir sponsors et exposants de cette édition sont lancés.

Enfin **Hélène Bringer Garlain**, la trésorière de l'association, est venue confirmer que le budget 2019 s'exécutait conformément à ce qui était prévu.

Daniel Verwaerde a ensuite clos cette Assemblée Générale en rappelant que l'Association ne vivait que grâce à ses membres et à leur participation active, et en rappelant qu'il était à l'écoute de tous.

Jean-François Prevéraud