Les applications industrielles de la bioinformatique

À l'heure où les technologies en génomique et en protéomique à haut débit envahissent les laboratoires de biologie, les sciences de la vie font face à un déluge de données extraordinairement volumineuses et complexes. Manipuler ces données et en extraire un sens biologique requièrent de nouvelles approches basées sur la modélisation et l'informatique. La bioinformatique est tout à la fois une science à l'interface entre l'informatique et la biologie et une industrie vitale pour stocker, diffuser, analyser et interpréter les données biologiques en vue de leur exploitation dans l'industrie de la santé, dans l'agroalimentaire ou encore en matière d'énergie. Cet article propose un rapide tour d'horizon du domaine, de ses acteurs et de ses défis.

Par **Jean-Philippe VERT***

Les progrès technologiques fulgurants des vingt dernières années ont eu un impact radical sur les sciences du vivant. Les techniques de puces à ADN, de protéomique, d'imagerie, ou, aujourd'hui, de séquençage génomique se sont invitées au cœur des laboratoires, offrant aux scientifiques de nouveaux outils pour scruter et quantifier le vivant jusque dans ses moindres détails.

Il est désormais possible de mesurer au niveau moléculaire l'ensemble des modifications génétiques portées par les cellules cancéreuses d'un patient pour tenter d'y répondre par un traitement complètement personnalisé [1], de quantifier les millions de microorganismes évoluant dans un milieu naturel particulier afin d'identifier de nouvelles manières de transformer la matière [2], ou d'observer les changements morphologiques induits par la suppression de chacun des milliers de gènes d'un organisme donné par vidéomicroscopie, pour identifier de nouvelles cibles thérapeutiques [3].

Les possibilités offertes par l'utilisation de ces nouvelles technologies pour étudier et comprendre le vivant ne semblent avoir d'autre limite que celle de notre imagination. De par leur caractère systématique et quantitatif, elles ouvrent notamment la voie à de nouvelles approches quantitatives pour modéliser le

* Directeur du Centre de Bioinformatique de Mines ParisTech.

vivant. Elles posent cependant de nouveaux défis qui dépassent largement le cadre traditionnel des sciences du vivant de par les vastes quantités de données complexes qu'elles génèrent. Par exemple, le projet américain TCGA (*The Cancer Genome Atlas*), qui vise à cataloguer les variations génomiques à l'origine de 10 000 cas de cancer, génère 10 téraoctets de données chaque mois et devrait en produire un total de 10 pétaoctets (1). Pour transmettre, stocker, analyser et interpréter cette masse de données afin d'aboutir à un résultat biologique, la biologie s'appuie naturellement de plus en plus sur les mathématiques et les technologies de l'information et de la communication. Une discipline nouvelle est d'ailleurs née pour répondre à ces défis depuis une quinzaine d'année : la bioinformatique, qui s'épanouit à la fois comme un domaine scientifique à part entière et comme un ensemble de technologies devenues indispensables à la recherche académique et industrielle. Dans les quelques pages qui vont suivre, nous nous efforcerons de mieux définir cette discipline et d'en illustrer les principaux domaines d'application en nous concentrant essentiellement sur les applications industrielles. Nous brosserons ensuite un rapide portrait du marché et des acteurs de ce secteur, avant de conclure en ouvrant quelques pistes de réflexion sur les défis qui nous sont lancés et qui représentent autant d'opportunités scientifiques et industrielles.

QU'EST-CE QUE LA BIOINFORMATIQUE ?

La bioinformatique est une discipline à l'interface entre la biologie et l'informatique, qui recouvre l'ensemble des technologies et des méthodes permettant de collecter, de stocker, d'analyser et d'interpréter les données biologiques. Elle désigne donc tout à la fois a) le développement *d'infrastructures* et *d'outils* (tels des systèmes de stockage de données, des logiciels de base de données et de visualisation) et b) le développement et la mise en œuvre de *méthodes* mathématiques et informatiques s'appuyant sur ces outils pour analyser des données et les interpréter biologiquement. Ces deux aspects de la discipline sont d'ailleurs souvent désignés en anglais sous deux noms différents, la *bioinformatics*, terme recouvrant le développement d'infrastructures et d'outils, et la *computational biology*, qui désigne, quant à elle, la mise au point de méthodes d'analyse spécifiques et leur utilisation pour traiter un problème biologique particulier. En schématisant à l'extrême, la *bioinformatics* se rapproche plus d'un travail d'ingénierie en infrastructures et en logiciels (pas toujours spécifique, d'ailleurs, à son domaine d'application que sont les sciences du vivant), tandis que la *computational biology* s'apparente à une discipline scientifique à part entière consistant à modéliser et à analyser des systèmes biologiques au moyen de modèles informatiques, en empruntant d'ailleurs de nombreuses approches aux mathématiques et à la physique.

LES APPLICATIONS INDUSTRIELLES DE LA BIOINFORMATIQUE

La bioinformatique, dans son ensemble, est donc une activité transverse qui peut être appliquée à de nombreux secteurs des sciences de la vie et des biotechnologies confrontés à l'étude et à l'utilisation du vivant. Elle joue de ce fait un rôle important et croissant dans de nombreuses industries, allant de la recherche biomédicale jusqu'à l'agroalimentaire, en passant par l'énergie et l'environnement.

En volume, les entreprises pharmaceutiques et biotechnologiques sont certainement les premières utilisatrices de la bioinformatique. En effet, elles utilisent de plus en plus de technologies à haut débit, comme la protéomique ou le séquençage, pour étudier les systèmes biologiques qui les intéressent. Elles s'appuient naturellement sur les outils et les méthodes de la bioinformatique pour décoder l'information biologique cachée dans les multiples données qu'elles génèrent et ainsi faciliter la traduction de ces données en avancées médicales. Un domaine particulier, au cœur de la révolution en cours, est la pharmaco-génomique, une discipline qui vise à prédire la probabilité qu'un individu réponde à un traitement en fonction de son patrimoine génétique ou de marqueurs moléculaires, ouvrant ainsi la voie à la médecine personnalisée. Cette discipline s'appuie sur des traitements informatiques et mathématiques précis nécessitant des statistiques en grande dimension et l'exploitation de nombreuses données pour identifier les combinaisons de marqueurs permettant de diagnostiquer avec précision une pathologie et de prédire l'efficacité-toxicité d'un traitement sur un individu donné. Les enjeux sociétaux et économiques de la médecine personnalisée sont considérables, puisqu'il s'agit d'améliorer la sûreté et l'efficacité des traitements en prenant en compte les spécificités biomoléculaires de chaque individu.

La bioinformatique joue également un rôle important dans l'identification de nouvelles cibles thérapeutiques correspondant à des molécules (typiquement des protéines) dont l'inhibition par un traitement adapté permettrait de traiter la pathologie. Elle fournit, d'une part, des outils pour interpréter systématiquement les résultats d'études visant à caractériser les variations moléculaires entre individus (anomalies génomiques, différences au niveau de l'expression des

(1) 1 pétaoctet = 10^{15} octets = 1 000 téraoctets = 1 000 000 gigaoctets = 1 000 000 000 mégaoctets.

gènes ou des marqueurs épigénétiques, etc.) et à les corréler avec l'apparition, puis le développement de certaines maladies. Ce travail peut permettre d'identifier au niveau moléculaire des anomalies qui sont responsables du développement de la pathologie et d'en déduire des stratégies thérapeutiques, comme l'inhibition d'une protéine mutée conférant une propriété particulière à une cellule cancéreuse. Une autre approche, complémentaire, pour identifier de nouvelles cibles thérapeutiques, consiste à modéliser mathématiquement le fonctionnement d'une cellule dans un environnement donné puis, par simulation et analyse du modèle obtenu, d'en déduire des interventions thérapeutiques possibles. Ce genre de modèle (qui est au cœur de ce que l'on appelle la *biologie des systèmes*) apporte d'ailleurs bien plus que l'identification de cibles candidates : il fournit également un cadre conceptuel et computationnel permettant d'intégrer des connaissances d'experts et des données mesurées sur des échantillons biologiques, ouvrant ainsi la voie à une compréhension holistique de mécanismes parfois très complexes [4]. La modélisation de voies de signalisation cellulaires ou de réseaux de régulation décrivant, au niveau moléculaire, comment une cellule réagit à son environnement et met en place des programmes d'activité particuliers, peut ainsi aider à comprendre et à prédire de quelle manière une cellule réagirait à une ou à plusieurs perturbations spécifiques, permettant ainsi, d'une part, d'identifier les meilleures interventions possibles et, d'autre part, d'en prédire les effets secondaires (voir la figure 1).

La bioinformatique intervient également de plus en plus et de manière multiforme, avec sa cousine la chémoinformatique, dans le processus de recherche de médicaments qui est au cœur de l'activité des entreprises pharmaceutiques. La recherche de nouvelles molécules inhibant ou, au contraire, promouvant l'activité d'une cible identifiée et susceptible de déboucher sur l'élaboration d'un nouveau médicament, est en effet un processus long et coûteux qui souffre d'une chute de productivité depuis de nombreuses années, notamment parce que de trop nombreuses molécules se révèlent inefficaces ou trop toxiques lors de la phase finale des essais cliniques. Qu'il s'agisse de modéliser les interactions moléculaires en 3D entre différentes molécules afin d'identifier la meilleure molécule à synthétiser pour inhiber une cible donnée, ou, au contraire, afin qu'elle n'interagisse pas avec une autre protéine pour garantir sa spécificité (voir la figure 2), ou encore de développer des modèles *in silico* (c'est-à-dire des modèles informatisés, ndlr) permettant de prédire la toxicité et les effets secondaires d'une molécule avant de la synthétiser et de la tester sur des patients lors d'essais cliniques, les modèles

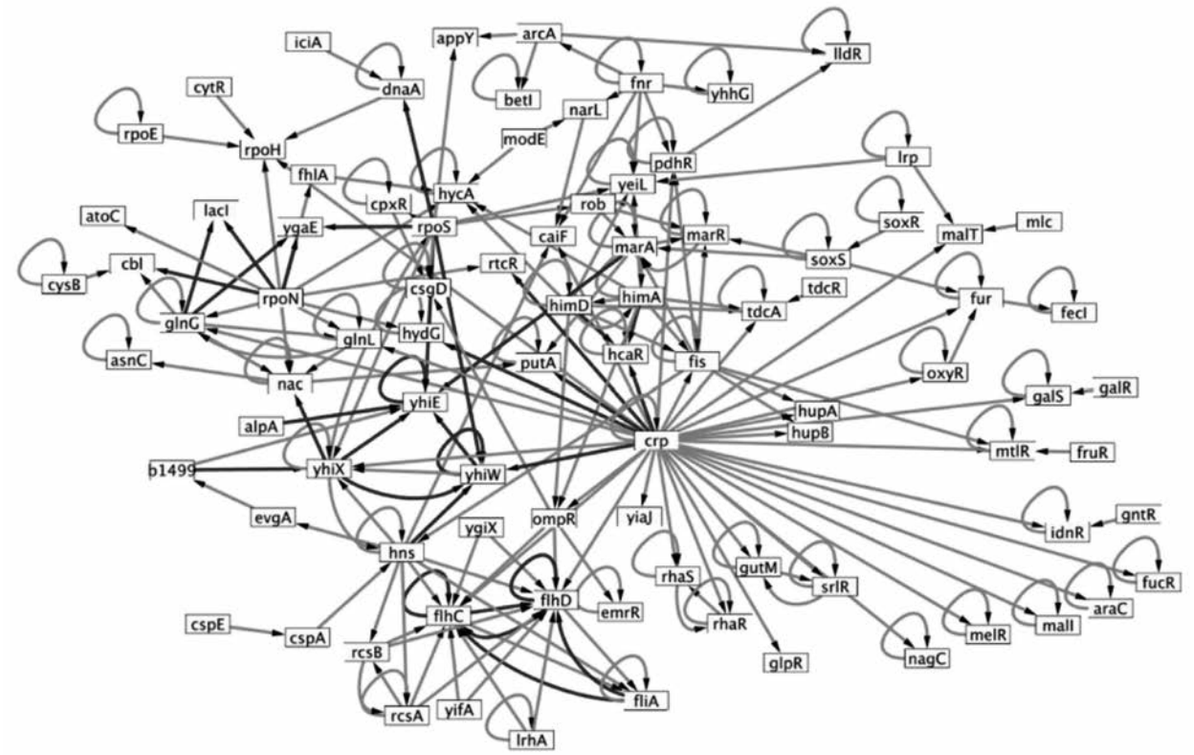

Figure 1 : La biologie des systèmes cherche à comprendre les propriétés biologiques d'un échantillon en appréhendant la complexité des interactions moléculaires.
Cette image est une représentation schématique d'un réseau de régulations décrivant la manière dont certains gènes contrôlent l'activation (ou au contraire l'inhibition) d'autres gènes.

mathématiques et les outils informatiques abondent dans l'ensemble du processus de recherche de nouveaux médicaments mis en œuvre par les entreprises pharmaceutiques.

À côté de ses applications dans les industries de la santé, la bioinformatique joue un rôle important et croissant dans l'ensemble des industries manipulant des organismes vivants et désireuses de les étudier, de les optimiser ou de les contrôler. L'industrie agroalimentaire s'appuie, par exemple, de plus en plus sur des technologies à haut débit pour disséquer et optimiser les organismes (bactéries ou levures) qu'elle utilise pour produire des aliments dans le but d'améliorer leur goût, leur odeur, leur texture ou leur valeur nutritionnelle. Une tendance similaire est observée dans l'optimisation des propriétés des aliments eux-mêmes (végétaux ou animaux) par modification génétique. Dans les deux cas, l'utilisation croissante de techniques à haut débit s'accompagne naturellement d'une utilisation croissante de la bioinformatique pour analyser et exploiter les données générées, ainsi que pour remplacer des expérimentations réelles par des simulations numériques. D'autres domaines d'application, comme les énergies renouvelables, la métagénomique ou la biologie de synthèse, suivent évidemment la même tendance.

LE MARCHÉ DE LA BIOINFORMATIQUE ET SES ACTEURS

Compte tenu de l'importance et de la variété de ses applications, la bioinformatique s'est non seulement développée en tant que discipline scientifique, mais également comme un secteur industriel à part entière et en forte croissance depuis une quinzaine d'années.

Aujourd'hui, schématiquement, on peut segmenter le marché de la bioinformatique en trois sous-marchés principaux :

– a) les logiciels d'analyse et les services associés,
– b) les contenus,

Figure 2 : Dans l'industrie pharmaceutique, les simulations en trois dimensions permettent d'optimiser virtuellement les structures des molécules permettant d'inhiber une cible donnée (ici, une protéine de la famille des tyrosines kinases, qui permet à certains cancers de se développer).

– c) et les infrastructures.
a) Le marché des logiciels d'analyse et des services associés comprend la fourniture de solutions pour analyser et exploiter les données générées par les utilisateurs, comme par exemple les données de séquençage ou de protéomique étudiées par la recherche pharmaceutique.
b) Le marché des contenus recouvre de nombreuses bases de données plus ou moins spécialisées permettant à leurs utilisateurs d'avoir accès à des connaissances, comme par exemple les cartes des réseaux biologiques ou les informations sur les gènes et les variations génomiques les plus fréquentes.
c) Enfin, le marché des infrastructures se concentre sur l'élaboration de solutions permettant à un laboratoire ou à une entreprise de stocker ses données, d'en donner l'accès aux utilisateurs et de permettre leur analyse en termes de stockage, de puissance de calcul et de réseaux.
Le marché global total de la bioinformatique est passé d'environ 840 millions de dollars en 2002 à près de 3 milliards de dollars en 2010, enregistrant une croissance régulière d'environ 25 % par an [5]. Plusieurs analystes considèrent que ce taux de croissance sera maintenu au moins dans les cinq prochaines années du fait des investissements importants réalisés par les industriels concernés dans les technologies à haut débit. En termes de volume, le marché de la bioinformatique est actuellement dominé par le secteur des contenus ; viennent ensuite les logiciels, les services d'analyse et, enfin, les infrastructures. La croissance du marché des logiciels et des services d'analyse est cependant la plus forte pour les trois segments précités.
Les acteurs de ce marché sont nombreux et variés, car les coûts d'entrée sont relativement faibles et les besoins multiples. On y retrouve aussi bien des géants des technologies de l'information comme IBM, qui a créé sa division Life Sciences Solutions dès 2000, des grandes entreprises plus spécialisées en bioinformatique comme Accelrys (600 employés, 150 millions de dollars de chiffre d'affaires), qu'une myriade de petites et moyennes entreprises *(comme SoftGenetics, DNAStar, DNAnexus ou NextBio)*, proposant souvent des solutions ou des services spécifiques pour des marchés très porteurs. D'autres acteurs importants sont les grandes entreprises pharmaceutiques ou de biotechnologie, qui ont souvent investi dans des équipes de bioinformatique en propre plus ou moins développées, et les fournisseurs de technologies pour la biologie (comme Affymetrix ou Applied Biosystems), qui vendent de plus en plus de solutions bioinformatiques permettant de récupérer, de stocker et d'analyser les données produites par leurs technologies. Enfin, de nombreuses sociétés de biotechnologie intègrent des offres de bioinformatique en complément des produits qu'elles vendent pour la biologie, comme Kinexus Bioinformatics (Canada) qui vend des produits et des prestations de protéomique et de bioinformatique.

Du point de vue de la géographie, les Etats-Unis dominent largement le marché mondial, devant l'Europe. Certains pays d'Asie enregistrent une très forte croissance, comme l'Inde, qui bénéficie d'un très grand vivier d'informaticiens de qualité. Malgré ses efforts manifestes, depuis une dizaine d'année, pour développer la bioinformatique dans son secteur académique, la France reste malheureusement en retrait par rapport à ses concurrents directs dans ce secteur industriel porteur.

LES DÉFIS À RELEVER

La bioinformatique est devenue une discipline omniprésente dans la recherche biomédicale et un secteur industriel en forte croissance. Cette croissance s'explique notamment par l'augmentation des volumes de données disponibles pour étudier les systèmes biologiques et par le basculement progressif de la biologie d'une science qualitative vers une science plus quantitative. Elle répond également à la nécessité d'améliorer la productivité des industries qui utilisent la biologie grâce à leur meilleure exploitation de ces données. De nombreux défis scientifiques et technologiques restent cependant à relever, qui représentent autant d'opportunités pour l'avenir.
Tout d'abord, les volumes de données générés en biologie croissent actuellement beaucoup plus vite que les capacités de stockage et la puissance de calcul des ordinateurs disponibles. Alors que celles-ci augmentent du double environ tous les dix-huit mois (selon la « loi » de Moore), le coût du séquençage a été divisé par 1 000 entre 2008 et 2012 (voir la figure 3 de la page suivante) et les infrastructures se sont rapidement développées. Fin 2011, le plus grand centre mondial de séquençage, basé en Chine, disposait de 167 séquenceurs capables de séquencer l'équivalent de 4 000 génomes humains par jour (voir la figure 4 de la page suivante) ! Nous sommes entrés brutalement dans une ère où le coût et les capacités de production des données s'effacent devant le coût de leur stockage, de leur transmission et (surtout) de leur analyse. Alors que l'on s'achemine rapidement vers la possibilité de séquencer un génome humain pour un coût de 1 000 $ (2), de nombreux experts ont mis en avant récemment les coûts réels, beaucoup plus élevés si l'on prend en compte toute la chaîne des traitements des données nécessaires à leur exploitation. C'est ce à quoi fait allusion Bruce Korf, ancien président de l'*American College of Medical Genetics*, lorsqu'il dit "*the $1000 genome, the $1 million interpretation*" [6]. Des géants de l'Internet ont commencé à s'intéresser aux problèmes du stockage et de la dissémination de

(2) Le séquençage du premier génome humain (en 2001) a coûté 2,7 milliards de dollars.

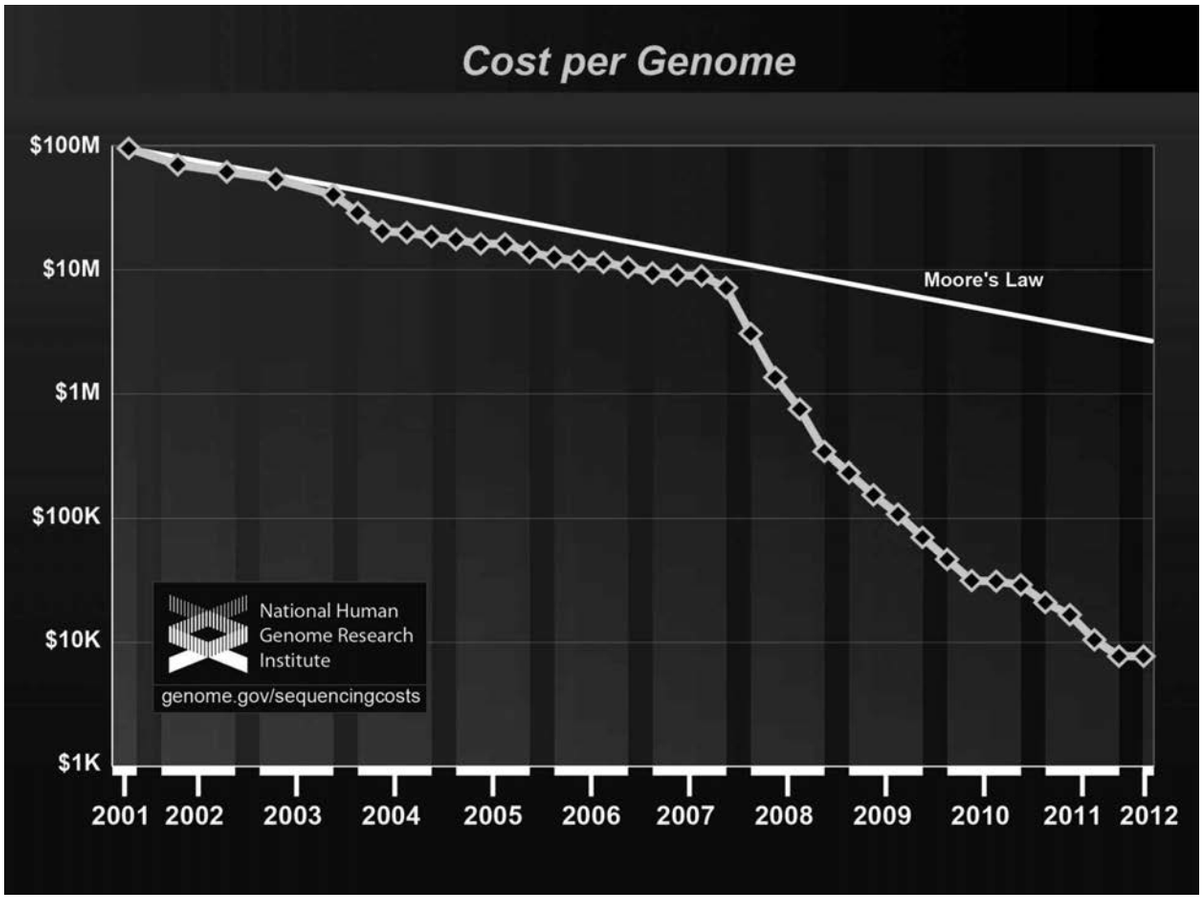

Figure 3 : Le coût du séquençage d'un génome humain est passé de presque 100 millions de dollars en 2001 à moins de 10 000 dollars en 2010. La brusque décroissance de coût à partir de 2007 correspond à l'arrivée sur le marché de techniques de séquençage massivement parallèles. En comparaison, la loi de Moore décrivant l'amélioration des performances des ordinateurs fait piètre figure (Source : KA Wetterstrand, NHGRI).

Figure 4 : L'institut BGI, basé en Chine, est le plus grand centre de séquençage génomique du monde. Il possède des centaines de séquenceurs de dernière génération qui sont capables de générer plus de 5 téraoctets de données par jour, équivalent aux données de séquençage des génomes de 4 000 êtres humains… (Source : BGI).

l'information biologique, comme Amazon qui met à la disposition de la communauté scientifique les données du projet *1000 Genomes* (correspondant aux génomes séquencés de 1 000 individus) sur son service de *cloud* (3). Il reste cependant évident que les questions de stockage, de transmission et d'algorithmes capables de s'adapter à la croissance vertigineuse des volumes de données que nous observons actuellement vont constituer très bientôt un fantastique défi à relever.

Par-delà les questions techniques liées à l'accroissement du volume des données, de nombreuses questions plus scientifiques sur la manière d'analyser les données restent ouvertes. Que faire, une fois que l'on aura cartographié systématiquement toutes les différences, au niveau moléculaire, entre des milliers d'individus, ou entre des milliers d'échantillons biologiques ? La dernière décennie nous a montré à maintes reprises que la biologie est une science d'autant plus complexe que l'on observe la réalité à une échelle plus fine, et les modèles mathématiques permettant de modéliser cette complexité restent en grande partie à inventer.

(3) http://aws.amazon.com/1000genomes

BIBLIOGRAPHIE

[1] STRATTON (M. R.) & *al.*, "The cancer genome", *Nature*, 458: 719-724, 2009.

[2] TRINGE (S. G.) & *al.*, "Comparative Metagenomics of Microbial Communities", *Science*, 308(5721): 554-557, 2005.

[3] NEUMANN (B.) & *al.*, "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes", *Nature,* 464: 721-727, 2010.

[4] BARILLOT (E.) & *al.*, *Computational Systems Biology of Cancer*, CRC Press, 2012.

[5] RNCOS E-Service Pvt. Ltd. Bioinformatics market outlook to 2015, mars 2012.

[6] DAVIES (K.), "The $ 1.000.000 Genome Interpretation", *Bio-IT World*, octobre 2010.