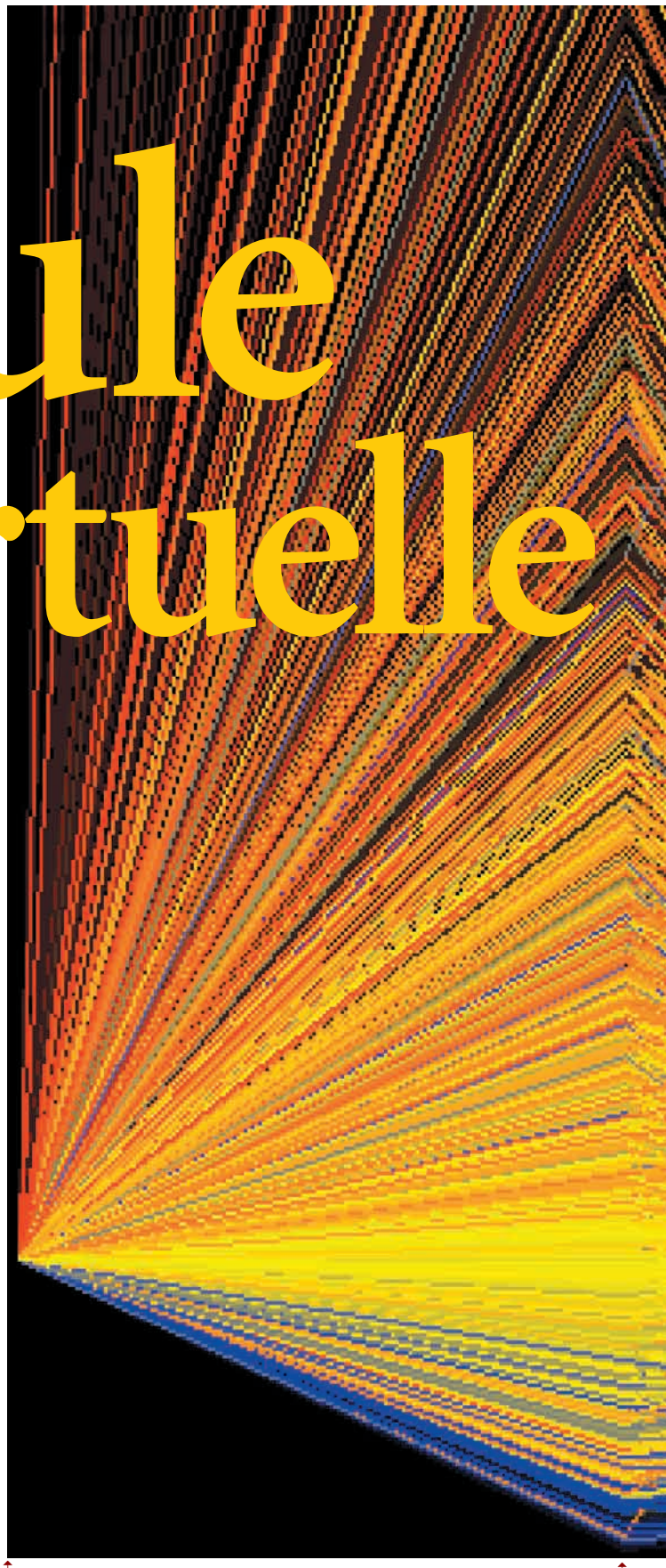
Et la cellule devint virtuelle



1 2

« *Si quelqu'un devait analyser les idées du moment et les slogans à la mode, il trouverait "systèmes" au sommet de la liste* », écrivait, en 1967, l'influent théoricien de la biologie Ludwig von Bertalanffy [1]. Il est décédé en 1972, mais son propos rencontre aujourd'hui une résonance nouvelle : en ce début de XXIe siècle, impossible d'échapper à la « *systems biology* ». Les programmes de recherche qui en appellent à cette « biologie intégrative » ou « biologie systémique » se succèdent, les revues scientifiques telles que *Nature* ou *Science* lui consacrent des dossiers entiers [2].

De quoi s'agit-il ? En termes lapidaires, de revenir à l'approche dégagée par Bertalanffy : une cellule, un organe ou un organisme sont autant de systèmes dont le fonctionnement dépend certes des entités en présence, mais aussi et surtout des interactions qui les lient. Autrement dit, le tout est plus que la somme des parties.

C'est réinventer la roue ! s'exclament certains, tant il est vrai que les biologistes – en particulier les physiologistes – ont toujours accordé de l'importance aux interactions. Pourtant, il y a bien nouveauté. Les techniques d'analyse à haut débit de la génomique et de la protéomique ont engendré une si grande quantité de données qu'il est aujourd'hui envisageable, en faisant appel aux mathématiques et à l'informatique, de réaliser la modélisation et la simulation de systèmes qui intègrent des niveaux jusque-là étudiés de façon distincte [3].

Parmi ces systèmes, figure, bien sûr, la cellule. Et le terme de « cellule virtuelle » commence, peu ou prou, à émerger sur la scène scientifique. Que veut-on dire par là ? Quels sont les objectifs visés, et les approches employées ? Est-ce une recherche purement académique, ou l'industrie s'y intéresse-t-elle aussi ? Le point, dans ce dossier de *La Recherche*.

CE DIAGRAMME PERMET DE VISUALISER L'EXPRESSION DE 22 500 GÈNES dans des cellules de prostate normales ou cancéreuses, et traitées ou non avec un produit qui ôte les groupements méthyle de l'ADN. Les données, obtenues grâce à des puces à ADN, ont été analysées informatiquement. Chaque ligne correspond à l'expression d'un gène, et chaque point d'inflexion, à une situation : cellule saine (1) ; cellule saine traitée (2) ; cellule cancéreuse (3) ; cellule cancéreuse traitée (4). La couleur rouge signale une forte expression, la bleue, une faible. Reste à comprendre le sens de ces données...

[1] L. von Bertalanffy, *General System Theory : Foundations, Development, Applications,* George Braziller Inc., New York, 1976 (édition révisée).

[2] Dossiers spéciaux « Systems biology » : *Science, 295,* 1661, 2002 ; *Nature, 420,* 205, 2002 ; *Nature Biotechnology, 22,* 1249, 2004.

[3] H. Kitano (éd.), *Foundations of Systems Biology,* MIT Press, 2001.

1-Le vivant en équations

Comment mettre à profit l'énorme quantité de données fournies par la génomique et la protéomique ? Ces informations statiques, figées dans des réseaux complexes, ne révèlent rien, en elles-mêmes, de la dynamique cellulaire. D'où l'ambition actuelle des bioinformaticiens : construire des modèles qui permettent d'accéder à la physiologie du « système cellule ».

François Rechenmann et **Hidde de Jong**
sont chercheurs au sein du projet Helix « Informatique et génomique » de l'unité de recherche Inria Rhône-Alpes.
Francois.Rechenmann@inria.fr
Hidde.de-Jong@inria.fr
www.inrialpes.fr/helix

À l'heure où le fonctionnement de systèmes aussi complexes qu'une sonde spatiale ou qu'un microprocesseur est entièrement simulé sur ordinateur avant même qu'un premier prototype soit réalisé, certains biologistes et informaticiens se prennent à rêver : pourrait-on, un jour, simuler l'action d'un nouveau médicament sur la cellule, *via* quelques « clics » de souris ? Les avancées récentes de la biologie ont dégagé des quantités phénoménales de données, et les capacités de simulation informatique se comptent désormais par milliards d'opérations élémentaires par seconde : ne suffirait-il pas d'exploiter les premières grâce aux secondes pour disposer d'une « cellule virtuelle » qui reproduise à la demande le fonctionnement d'une cellule réelle ?

C'est dans cette optique qu'ont été engagés les premiers projets de modélisation globale de cellules, il y a une dizaine d'années. Dès 1996, le projet E-cell de l'université de Kyoto [1] visait ainsi l'obtention d'un modèle d'une cellule virtuelle minimale possédant un génome inspiré de celui de la bactérie *Mycoplasma genitalium*, séquencé l'année précédente. Mais son initiateur, Masaru Tomita, reconnaît à présent qu'il avait à l'époque une perception quelque peu naïve du problème : « *Nous pensions que le goulet d'étranglement de ce grand défi serait la puissance de calcul et les ressources humaines. Bien qu'ils constituent toujours des obstacles importants, nous devons faire face à beaucoup d'autres défis scientifiques majeurs* », déclare-t-il sur le site de son projet [2]. C'est le moins que l'on puisse dire.

Prenons l'exemple de la bactérie *Escherichia coli*. Depuis 1996, la base EcoCyc tente de recenser et d'organiser les données disponibles sur cet organisme modèle, qui figure parmi les plus étudiés et les mieux connus [3]. Dans sa version de septembre 2004, EcoCyc décrit 4 497 gènes qui codent presque autant de protéines (dont 3 461 sont caractérisées), 956 opérons*, 3 629 réactions biochimiques, 182 voies métaboliques, 1 133 enzymes et 197 transporteurs. Bien que s'enrichissant régulièrement de nouvelles données extraites de la littérature spécialisée, EcoCyc est loin d'être une base exhaustive, et ces nombres donnent une appréciation encore sous-évaluée de la réalité.

Qui plus est, même si elles étaient complètes, ces énumérations ne fourniraient pas une mesure adéquate de la complexité à laquelle sont confrontés les biologistes. En effet, ce sont les interactions entre molécules qui sont à l'origine de la dynamique des processus cellulaires. Et ces interactions n'obéissent pas à un schéma séquentiel simple : le bon fonctionnement cellulaire implique en effet des boucles de rétroaction positives et négatives, aux effets respectivement amplificateurs et stabilisateurs.

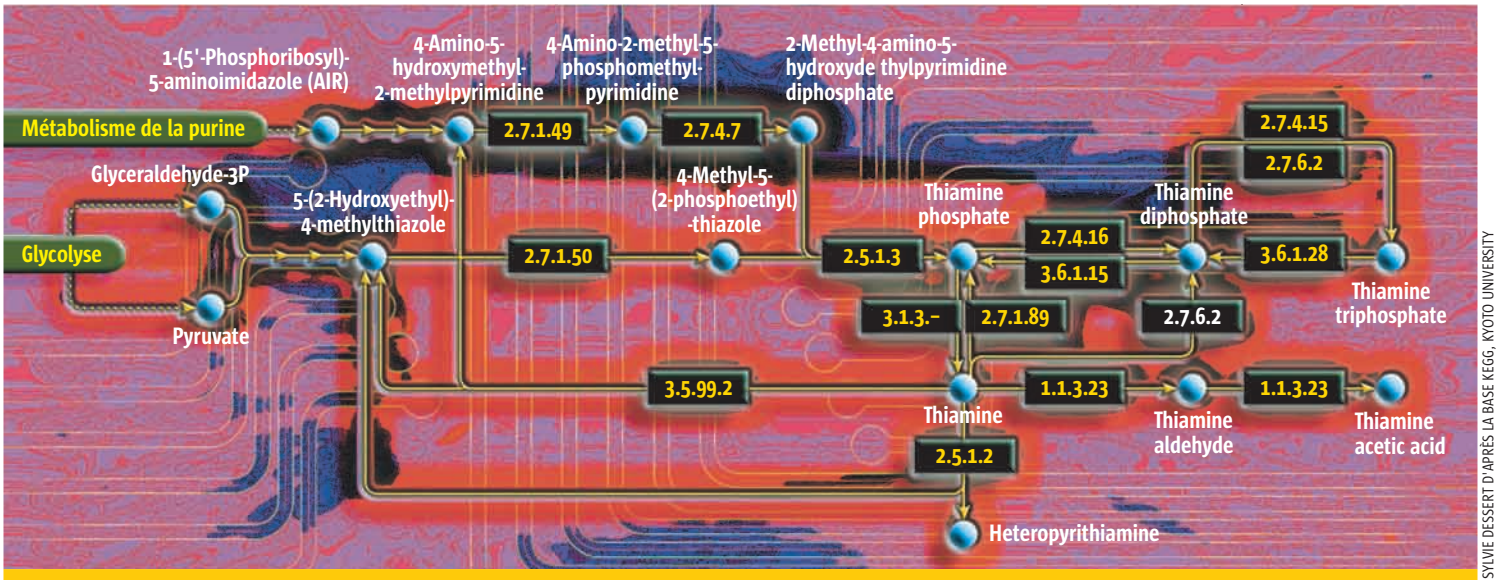
Mais même les bases de données qui décrivent des réseaux d'interactions entre entités biologiques ne peuvent prétendre éclairer vraiment le comportement de la cellule, car elles ne fournissent qu'une description statique de ces interactions. L'exemple de la base Kegg – Kyoto Encyclopaedia of Genes and Genomes – est, à cet égard, particulièrement parlant. Cette base rassemble les données disponibles sur les voies métaboliques, les processus de transcription et de traduction des gènes, les voies de signalisation intracellulaire, ou encore les processus tels que le cycle cellulaire ou le rythme circadien [4].

Des tableaux trop statiques

Il y a peu de temps encore, ces données étaient stockées uniquement sous la forme de planches, accompagnées de commentaires textuels : le tout était, par nature, parfaitement réfractaire à une exploitation systématique par des méthodes informatiques. Depuis deux ans, les promoteurs de la base tentent d'élaborer une représentation sous-jacente formelle de ces réseaux (voir illustration p. 35). Toutefois, cette représentation reste statique : il ne faut pas attendre de la base une quelconque capacité à simuler le fonctionnement des voies métaboliques, et donc à rendre compte de la dynamique cellulaire. Telle molécule y est bien décrite comme étant le précurseur de telle autre, une troisième étant impliquée dans la synthèse de la seconde à partir de la première, et une quatrième interférant avec elle, mais rien ne montre dans quelles conditions s'amorce ce processus, ni pourquoi et comment il se prolonge, ou au contraire s'arrête…

Dans ce contexte, modéliser la dynamique de la cellule et tenter de la représenter sous une forme mathématique semblent les seuls moyens d'échapper aux limites de l'esprit humain, incapable de suivre le comportement de plus de trois ou quatre éléments en interaction – et à plus forte

* **Un opéron** est un groupe de gènes adjacents dans un génome bactérien, transcrits à partir d'un même promoteur et fonctionnant de façon coordonnée.



CE GRAPHE DU MÉTABOLISME DE LA VITAMINE B1, la thiamine, est l'un des plus simples figurant dans la base de données Kegg. Les nœuds correspondent aux différents composés, et les lignes figurent les réactions qui permettent de passer des uns aux autres. Les enzymes qui catalysent ces réactions apparaissent dans les rectangles. La base Kegg fournit une description statique des interactions mais ne donne pas accès à leur dynamique.

raison de prédire ce comportement. Avec l'espoir que les simulations lancées sur ces modèles permettent d'accéder à l'évolution des propriétés du « système cellule » au cours du temps, en fonction des interactions entre les entités qui le constituent et des conditions environnementales.

Depuis leur introduction par Newton et Leibniz dans la seconde moitié du XVII^{e} siècle, les équations différentielles constituent le formalisme mathématique le plus naturel pour représenter des systèmes dynamiques (lire ci-contre « Systèmes dynamiques et équations différentielles »). Il n'est donc pas surprenant que quelques-uns des meilleurs modèles cellulaires actuels le mettent en œuvre, en représentant le système en termes de concentrations de molécules et de leur taux de variation. Les groupes de Bela Novak, à Budapest, et de John Tyson, à Blacksburg, en Virginie, codéveloppent et enrichissent depuis plus de dix ans des modèles comptant jusqu'à une cinquantaine d'équations en vue d'étudier le cycle cellulaire du crapaud *Xenopus laevis* et des levures *Saccharomyces cerevisiæ* et *Schizosaccharomyces pombe* [5].

En reprenant l'un de ces modèles, un groupe de biologistes de l'université Rockefeller a récemment effectué des allers-retours extrêmement fructueux entre formalisation et expérimentation [6]. « Fructueux » signifie-t-il que toutes les prédictions du modèle ont été confirmées par les observations ? Non, certaines ont été contredites. Et c'est bien là tout l'intérêt de la simulation que de pointer du doigt l'insuffisance de certaines hypothèses ! Les différences entre prédictions et observations signalent en effet la nécessité de considérer comme importantes des interactions jusque-là négligées. Celles-ci sont alors formalisées, puis intégrées au modèle auquel on demande ensuite de fournir de nouvelles prédictions, lesquelles sont à leur tour confrontées à l'expérimentation. ⇨

MÉTHODE Systèmes dynamiques et équations différentielles

Au niveau moléculaire, une cellule peut être vue comme un système biochimique complexe, composé d'un grand nombre de molécules de différentes espèces en interaction. Pour rendre compte de la dynamique d'un tel système, la concentration d'une espèce de molécule est modélisée par une variable, x, dont le taux de variation dans le temps, noté x', est relié aux concentrations x, y, z, etc. des espèces avec lesquelles elle interagit.

À tout instant, la variation x' de la concentration d'une espèce de molécule est égale à la différence entre son taux de production (appelons-le prodx) et son taux de consommation (consx). En raison des interactions de la molécule considérée avec d'autres molécules, x' sera donc fonction de x, y, z, etc. : x' = prodx(x, y, z,...) – consx(x, y, z,...). Une équation similaire peut être écrite pour chaque espèce de molécule : y' = prody(x, y, z,...) - consy(x, y, z,...), etc. On obtient de ce fait un ensemble d'équations différentielles dites couplées, qui font dépendre, à tout moment, l'évolution des variables de leurs propres valeurs. En pratique, une fonction telle que prodx ou consx ne fait intervenir qu'un sous-ensemble des variables, car, même dans un système dit complexe, « tout n'interagit pas avec tout ».

Dès que l'expression mathématique des fonctions s'écarte du cas simple qu'est la combinaison linéaire des variables, il est pratiquement impossible de trouver la solution analytique du système différentiel. Il faut donc recourir à la simulation qui consiste à calculer, à l'aide de méthodes numériques, les valeurs successives dans le temps des variables x, y, z,... à partir de leurs valeurs initiales données, et prédire ainsi le comportement du système. L'effet sur ce comportement de modifications de la valeur de paramètres ou de perturbations de concentrations moléculaires est facilement étudié : il suffit de relancer la simulation avec ces nouvelles conditions.

* **L'opéron lactose** est un groupe de plusieurs gènes codant les enzymes nécessaires à l'utilisation du lactose chez les bactéries. Le lactose régule sa propre dégradation en inactivant la protéine, dite « répresseur », qui inhibe l'opéron.

⇨ Toutefois, comme nous l'avons dit plus haut, les équations différentielles représentent le système en termes de concentration de molécules et de leur taux de variation. De ce fait, leur utilisation n'est pertinente que sous la double hypothèse d'un nombre élevé de ces molécules et de leur répartition homogène au sein de la cellule.

Or, le milieu cellulaire n'est pas homogène. C'est flagrant en ce qui concerne les cellules eucaryotes. Outre le noyau qui leur vaut leur nom, elles possèdent en effet différents compartiments intracellulaires comme les mitochondries, les chloroplastes (chez les plantes), le réticulum endoplasmique ou l'appareil de Golgi. Mais c'est également vrai pour les bactéries, pourtant dépourvues de ces compartiments : le cytoplasme est plus gélatineux que liquide, et les réactions biochimiques ne se déroulent pas de façon homogène dans toute la cellule.

Tenir compte de cette hétérogénéité suppose d'introduire dans les équations différentielles des variables autres que temporelles : les coordonnées spatiales. Malheureusement, le problème de la disponibilité des connaissances nécessaires à la formulation du modèle se pose de façon encore plus aiguë que dans le cas « simple » d'une cellule considérée comme un système homogène. Certes, la biologie cellulaire s'est penchée depuis longtemps sur la spécificité des réactions qui se produisent dans les compartiments intracellulaires. Et les modélisateurs ont, pour ces derniers, créé des modèles justement appelés « à compartiments ». Mais l'aspect spatial des rencontres intermoléculaires qui ont lieu dans le cytoplasme est encore peu connu. Qui plus est, la formulation et la résolution numérique des équations mises en œuvre – des équations différentielles dites « aux dérivées partielles » – sont beaucoup plus difficiles que celles des équations différentielles ordinaires.

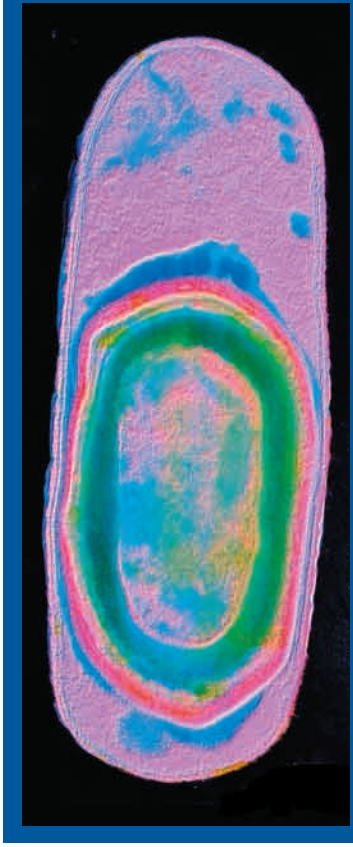
Tenir compte du hasard

Par ailleurs, la dynamique de certains processus cellulaires fait intervenir un nombre relativement faible d'entités. Ainsi, moins de dix exemplaires de la protéine répresseur de l'opéron lactose* sont présents à tout moment au sein de la bactérie *E. coli*. Dans ces conditions, le caractère aléatoire des réactions biochimiques devient plus important que lorsque les molécules en présence sont très nombreuses, ce que les équations différentielles ne permettent pas de modéliser correctement. Il est alors plus judicieux de recourir à d'autres formalismes.

C'est précisément ce qu'Adam Arkin, alors à l'université Stanford, et ses collègues ont choisi de faire, et ce dès le milieu des années quatre-vingt-dix. Ils ont appliqué des méthodes de simulation dite stochastique (lire : « Dynamique cellulaire et simulation stochastique », p. 35) dans leur modélisation de la réponse de la bactérie *E. coli* à l'infection par le bactériophage λ. Ce virus peut engager la bactérie qu'il infecte vers deux voies de développement. Dans la première, appelée « lyse », le phage se réplique en détournant la machinerie d'expression génique de son hôte, et la bactérie finit par se disloquer et libérer les particules virales dans son environnement ; dans le second, appelé « lysogénie », le génome du phage s'intègre dans le génome de la bactérie et y reste en latence. Un inventaire minutieux des réactions entre une demi-dizaine de gènes du bactériophage et les protéines régulant leur expression a conduit, en 1998, à l'élaboration d'un modèle qui explique comment, après l'infection par le phage, une population de bactéries, initialement homogène, se partitionne en des sous-populations lytiques et lysogéniques [7].

Quelles que soient leurs qualités, tous ces formalismes mathématiques exigent des informations précises sur les caractéristiques des entités en présence et sur leurs interactions. Malgré les progrès spectaculaires des méthodes d'investigation systématique des génomes, des transcriptomes* et des protéomes, les données et les connaissances biologiques disponibles sont encore très souvent insuffisantes en regard de ces demandes. Même pour des processus bien étudiés, comme le cycle cellulaire chez la levure, nos connaissances des composants du réseau et de leurs interactions sont incomplètes. En outre, les valeurs des paramètres caractérisant ces interactions sont le plus souvent très approximatives.

Sporulation chez *Bacillus subtilis*



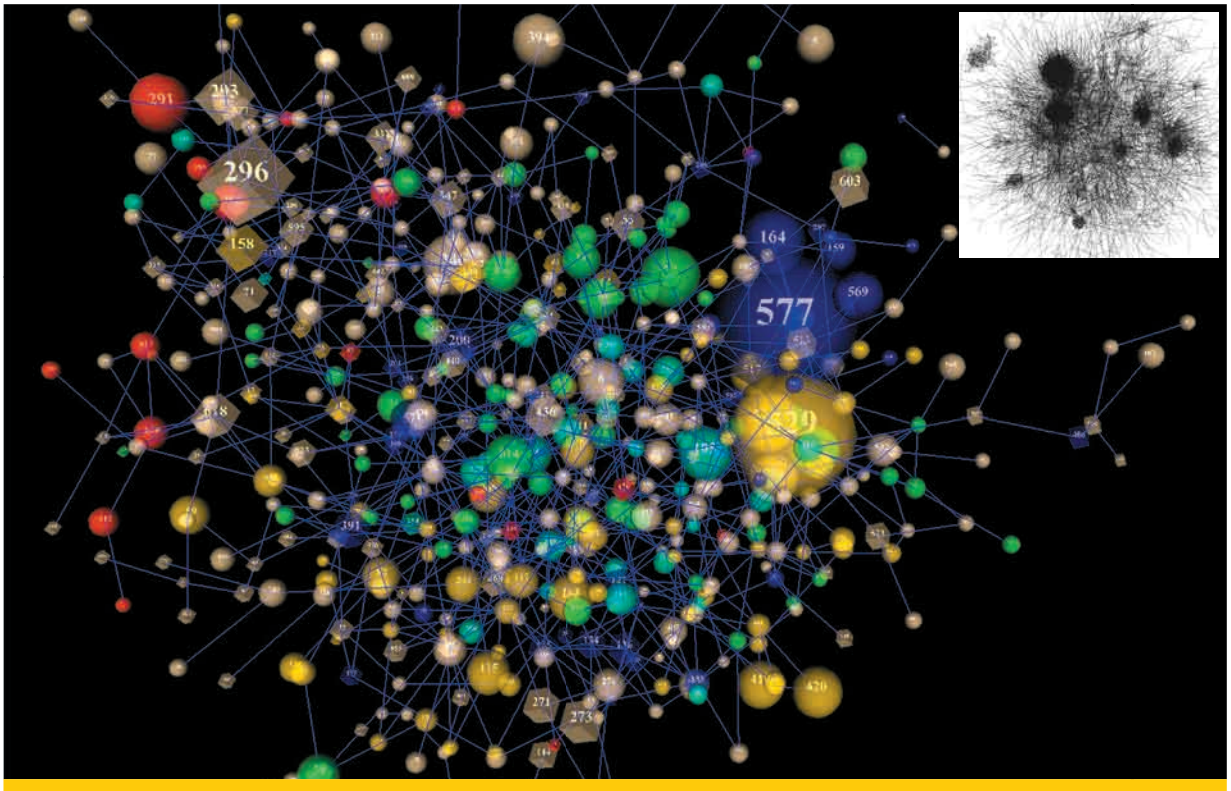
© CNRI/SPL/COSMOS

LORSQU'ELLE EST PRIVÉE DE NUTRIMENTS, la bactérie du sol *Bacillus subtilis* cesse de proliférer. Elle entame un processus qui aboutit à l'apparition d'une forme résistante – une spore – sous laquelle elle attend des jours meilleurs. Plus précisément, la division cellulaire engendre deux cellules de taille inégale. La plus petite est englobée par la plus grande, et s'y transforme (ci-contre) en une spore qui est finalement libérée dans le milieu après éclatement de la grande cellule.

L'initiation de la sporulation implique un changement radical de l'expression des gènes de la bactérie. La protéine SpoOA est au cœur de ce mécanisme. Cette protéine peut être phosphorylée, c'est-à-dire qu'un groupement phosphate y est fixé. Lorsqu'elle est sous cette forme dite SpoOA~P et que sa concentration dépasse un certain seuil, elle active plusieurs gènes nécessaires à la sporulation.

En utilisant une méthode de simulation qui permet de prédire qualitativement l'évolution de la concentration de différentes protéines de la cellule en réponse à un signal de privation de nutriments, nous avons confirmé que l'accumulation de SpoOA~P est le résultat de la compétition entre deux mécanismes de rétroaction. Dans l'un, SpoOA stimule l'expression des gènes codant les enzymes qui lui ajoutent un phosphate ; dans l'autre, elle stimule l'expression des gènes codant les enzymes qui transforment SpoOA~P en SpoOA en enlevant le phosphate.

H. de Jong *et al.*, *Bioinformatics, 19,* 336, 2003.



CES DEUX RÉSEAUX REPRÉSENTENT LES COUPLAGES FONCTIONNELS DES GÈNES DE LA LEVURE. Ils sont le résultat graphique d'une nouvelle méthode d'intégration des données de la génomique fonctionnelle. Le plus précis (en médaillon) montre que, sur les 4 681 gènes considérés, 3 285 sont regroupés en 627 modules (zones noires denses). L'image en couleur est un zoom sur 564 de ces modules, et les 950 liens les plus forts qui les unissent (traits bleus). La couleur et la forme de chaque module indiquent le type de fonction à laquelle contribuent les gènes. Par exemple, les sphères jaunes regroupent les gènes qui codent les molécules impliquées dans la transcription de l'ADN en ARN.

* **Un transcriptome** est l'ensemble des ARN messagers d'une cellule, à un moment donné et dans une situation donnée.

[1] M. Tomita, *Trends Biotechnol., 19,* 205, 2001.

[2] www.e-cell.org/about/message

[3] http://EcoCyc.org/

[4] www.genome.jp/kegg/

[5] J. J. Tyson *et al., Nat. Rev. Mol. Cell Biol., 2(12),* 908, 2001.

[6] F. R. Cross *et al., Mol. Biol. Cell, 13,* 52, 2002.

[7] A. Arkin *et al., Genetics, 149,* 1633, 1998.

[8] H. de Jong *et al., Bioinformatics, 19,* 336, 2003.

Faut-il donc abandonner l'idée de construire des modèles complexes du fait de cette insuffisance de données et de connaissances ? Une des solutions mise en œuvre consiste à tenter de combler ce manque, en lançant de vastes programmes expérimentaux ayant pour objectif la détermination ciblée des informations requises pour la formulation du modèle et la quantification de ses paramètres.

Simulation qualitative

Cependant, compte tenu du coût et de la complexité de ces tâches, plusieurs modélisateurs portent en parallèle leurs efforts sur le développement de formalismes moins exigeants en termes d'informations requises. Ainsi notre équipe a-t-elle développé une méthode de simulation dite « qualitative », qui ne requiert pas la connaissance des valeurs exactes des paramètres, mais seulement celle des inégalités qui les relient – ces données étant fournies par la littérature scientifique. Un modèle de ce type, qui met en œuvre une classe particulière d'équations différentielles (dites « linéaires par morceaux »), a permis de modéliser un processus de développement considéré comme un paradigme chez les bactéries : l'initiation de la sporulation chez *Bacillus subtilis* (lire p. 34) [8]. En l'absence de nutriments, cette bactérie peut arrêter de proliférer et se transforme alors en une spore capable ⇨

MÉTHODE Dynamique cellulaire et simulation stochastique

Les réactions biochimiques peuvent être vues comme des processus stochastiques et discrets : stochastiques, car l'intervalle de temps qui sépare deux réactions et le type de la prochaine réaction à se produire sont aléatoires ; discrets, car chaque réaction augmente ou diminue le nombre de chaque type de molécules. Si le nombre de molécules en présence est élevé, on peut, par approximation, considérer que ces réactions sont continues et se déroulent suivant un schéma déterministe. Il est alors possible de les modéliser par des équations différentielles. Mais, si le nombre de molécules est faible, cette approximation ne tient plus. Les modélisateurs doivent alors recourir à des formalismes qui prennent en compte les aspects discrets et stochastiques, en utilisant des équations appelées « équations maîtresses stochastiques ». Là encore, en raison de leur complexité, on ne sait pas résoudre analytiquement ces équations, et il faut recourir à la simulation sur ordinateur. Dans sa forme la plus simple, une simulation dans ce formalisme consiste à répéter, à partir d'un état initial, la procédure suivante : d'abord choisir aléatoirement le temps auquel la prochaine réaction se produit, ainsi que le type de cette réaction, et ensuite déterminer le nouvel état de la cellule, c'est-à-dire le nombre de molécules de chaque espèce, qui résulte de l'exécution de cette réaction. Il est nécessaire de répéter la simulation un grand nombre de fois pour obtenir une bonne approximation des probabilités des différentes successions d'états possibles de la cellule. Les approches stochastiques requièrent donc de grandes puissances de calcul. Comme pour les équations différentielles, on peut s'affranchir de l'hypothèse d'homogénéité spatiale, pour peu que soit prise en compte la diffusion des molécules et qu'une représentation de l'espace soit disponible.

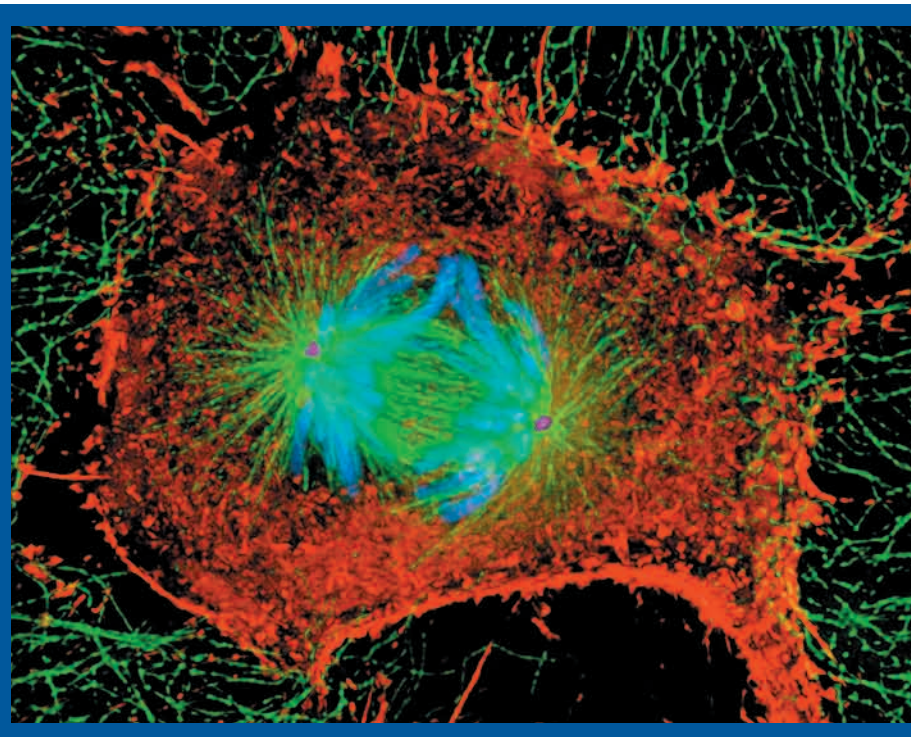
* **Un facteur de transcription** est une protéine qui déclenche la transcription d'un gène en ARN messager, en se fixant sur une portion de ce gène appelée « promoteur ».

⇨ de survivre tant que ces conditions défavorables perdurent. Le modèle a fourni des précisions sur la façon dont est régulée l'activité d'un facteur de transcription* déterminant dans la « décision » de la bactérie de sporuler ou non.
Plus radicalement, n'y aurait-il pas une troisième voie entre les descriptions statiques des bases de données, et les descriptions dynamiques obtenues grâce aux équations différentielles ou à la simulation stochastique ? Certains se le demandent, et les résultats récemment publiés par l'équipe de Bernhard Palsson, de l'université de Californie à San Diego, sont à cet égard source de remise en question.

Croissance bactérienne

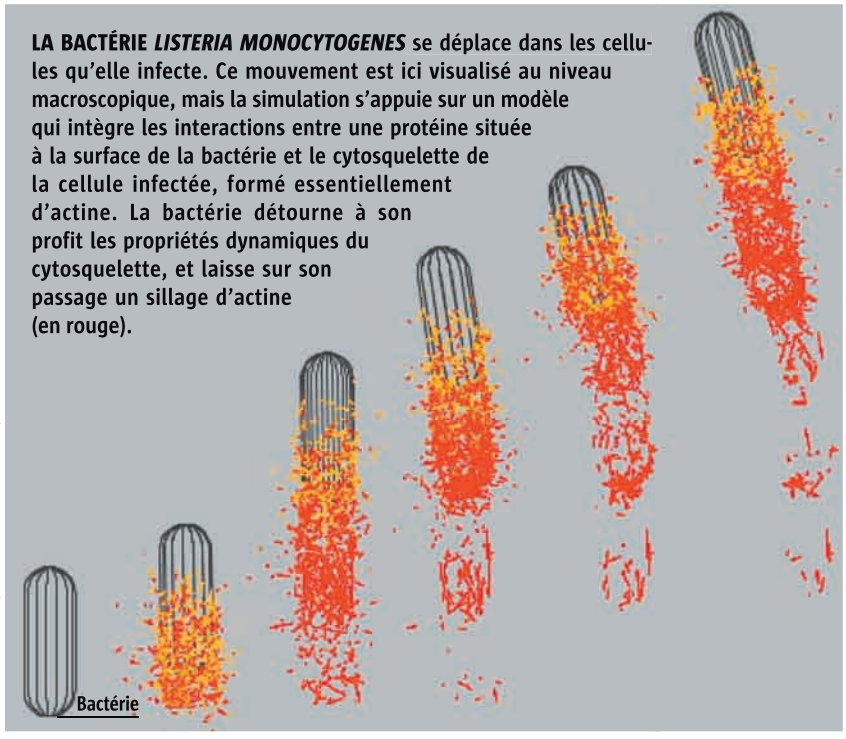
L'approche de cette équipe, qui s'intéresse au métabolisme cellulaire, consiste à calculer l'état final d'un système en fonction des conditions initiales, en négligeant la manière dont cet état est atteint. Cette démarche repose sur l'hypothèse que le système métabolique s'équilibre rapidement lorsque les conditions extérieures (par exemple, la disponibilité en certains nutriments) changent : il est du coup possible de négliger la dynamique du changement. Elle repose également sur l'hypothèse que certains processus – notamment la production d'énergie sous forme d'ATP – sont optimisés par la cellule. Parmi tous les états d'équilibre possibles, on retient donc celui pour lequel la fonction mathématique considérée – par exemple, celle qui représente la production d'ATP – est à son maximum.
Depuis une dizaine d'années, Bernhard Palsson élabore ainsi des versions de plus en plus détaillées d'un modèle qui calcule l'état d'équilibre du système constitué de réactions biochimiques au sein de la bactérie *E. coli* [9]. Ce modèle a ceci de particulier qu'il prend en compte non seulement les enzymes impliquées dans le métabolisme, mais aussi la régulation des gènes codant ces enzymes. La version la plus récente inclut quelque mille gènes (sur les plus de 4 000 connus chez cet organisme), dont 104 codent des protéines qui régulent l'expression de 479 de ces 1 000 gènes.
Le développement de ce modèle est parvenu à un stade tel qu'il a permis de prédire de façon fiable la capacité de la bactérie à croître sur tel ou tel milieu nutritif, après qu'on l'a privée de tel ou tel gène [10]. Certes, ce type de modélisations ne vaut que pour représenter des voies métaboliques, où des

Cycle cellulaire chez les mammifères



© DR ALEXEY KHODJAKOV/SPL/COSMOS

QUE CE SOIT CHEZ LA LEVURE OU CHEZ UN MAMMIFÈRE, la division cellulaire est sous le contrôle d'enzymes particulières, des kinases, qui gèrent le passage de chaque phase du cycle à la suivante. L'activité de ces kinases est régulée par d'autres protéines appelées cyclines. Alors que, chez la levure, l'association entre une cycline et une kinase particulières régule tout le cycle cellulaire, chez les eucaryotes pluricellulaires, plusieurs associations interviennent.
En s'appuyant sur leur modèle de contrôle du cycle cellulaire chez la levure *Schizosaccharomyces pombe*, John Tyson et Bela Novak ont récemment développé un modèle simplifié du contrôle du cycle cellulaire chez les cellules de mammifères. Ils ont simulé la progression de ce cycle chez des cellules traitées par une molécule qui arrête la synthèse des protéines, et donc la division de la cellule.
Les expériences *in vitro* avaient montré que, quand cette substance est appliquée juste avant que la cellule ne démarre un nouveau cycle, celui-ci est bloqué. Mais, si elle est appliquée juste après, le cycle continue comme si de rien n'était, et le blocage n'a lieu qu'à l'entrée dans le cycle suivant. Grâce à la simulation, les modélisateurs ont prédit l'évolution temporelle des concentrations de protéines consécutivement à la présence de l'inhibiteur, et ces prédictions leur ont ensuite permis d'expliquer les différences de comportement des cellules en fonction du moment où l'inhibiteur est appliqué.
B. Novak et J. J. Tyson, *J. Theor. Biol., 230,* 563, 2004.



LA BACTÉRIE *LISTERIA MONOCYTOGENES* se déplace dans les cellules qu'elle infecte. Ce mouvement est ici visualisé au niveau macroscopique, mais la simulation s'appuie sur un modèle qui intègre les interactions entre une protéine située à la surface de la bactérie et le cytosquelette de la cellule infectée, formé essentiellement d'actine. La bactérie détourne à son profit les propriétés dynamiques du cytosquelette, et laisse sur son passage un sillage d'actine (en rouge).

molécules sont transformées *via* des réactions enzymatiques. Il n'est pas utilisable pour les autres mécanismes cellulaires. Toujours est-il que, dans ce contexte particulier, la force des résultats obtenus valide *de facto* la démarche employée.

En tout état de cause, il apparaît clairement qu'une cellule et les processus qui s'y déroulent peuvent être décrits de plusieurs manières, à des niveaux de détails variables. En fait, l'analyse et la modélisation de tout système supposent de faire des choix entre ce qui est pertinent, et sera intégré dans le modèle, et ce qui ne l'est pas ou insuffisamment. Ces choix n'ont rien d'absolu ; ils sont dictés par les objectifs des modélisateurs. Le choix du formalisme, du niveau de détail, des variables, ne peut se faire qu'en fonction des questions auxquelles ils souhaitent que le modèle puisse contribuer à apporter des réponses. La vision même de ce qu'est le système à modéliser est dépendante de la problématique biologique qui est abordée. Ainsi, bien que la cellule soit physiquement délimitée par sa membrane, la frontière du « système cellule » n'est pas définie *a priori* : suivant les cas, seuls certains mécanismes cellulaires seront inclus dans le modèle, avec les entités qui y interviennent.

Le prix des modèles

Par ailleurs, il est essentiel de bien comprendre que ce n'est pas sa capacité à reproduire fidèlement le comportement observé de la cellule qui permet d'évaluer l'intérêt d'un modèle, bien qu'une telle concordance soit évidemment rassurante. L'obtention par simulation d'un comportement inattendu est beaucoup plus instructive, car elle met en évidence une insuffisance du modèle, soit que des variables et des relations aient été omises, soit qu'elles aient été incorrectement formalisées. En mettant en évidence les lacunes et les contradictions dans la connaissance qu'ont les modélisateurs, elle peut suggérer de nouvelles expériences ou de nouvelles orientations de recherches.

Dès les années soixante-dix, Herbert Simon, pionnier de l'intelligence artificielle, soulignait dans son ouvrage *La Science des systèmes, sciences de l'artificiel*, qu'on ne construit pas un modèle quand on a compris le système ; on construit un modèle pour comprendre le système [11]. C'est d'autant plus vrai en modélisation cellulaire que le modélisateur est confronté à la nécessité de trouver des équations autres que celles qui traduisent les grandes lois physiques : ces lois sont certes à l'œuvre dans la cellule, mais à elles seules, elles ne permettent pas d'accéder au niveau global de compréhension visé. Nombre des projets actuels de modélisation de cellule reconnaissent la pertinence de l'affirmation d'Herbert Simon et s'attachent désormais, à l'instar du projet pionnier qu'est E-cell, à concevoir et à expérimenter des outils informatiques aptes à faciliter la construction des modèles et surtout leur modification après la confrontation avec les données expérimentales.

Cela étant, une ultime question se pose : les modèles doivent-ils rester confinés dans ce rôle de test de cohérence des données et des connaissances ? Ou est-il légitime de les utiliser pour prédire des comportements cellulaires, par exemple en réaction à l'introduction de molécules dont on souhaite tester et évaluer l'effet thérapeutique ? La simple considération du coût de construction d'un modèle et la volonté de l'amortir poussent évidemment à répondre « oui » à cette seconde option. Mais si de tels modèles sont un jour accessibles, il conviendra de s'assurer qu'ils seront toujours utilisés en accord avec les objectifs qui auront présidé leur conception et de ne pas leur attribuer un statut de produit définitivement qualifié. Pour le généticien Richard Lewontin : « *Le prix de la métaphore est une éternelle vigilance* [12]. » C'est également le prix des modèles. ■■ **F. R. et H. de J.**

POUR EN SAVOIR PLUS

- S. Kumar et P. Bentley (éd.), *On Growth, Form and Computers,* Elsevier Academic Press, 2003.
- C. Fall, E. Marland, J. Wagner, et J.J. Tyson, *Computational Cell Biology,* Springer, 2002.
- J. Bower et H. Bolouri (éd), *Computationnal Modeling of Genetic and Biochemical Networks,* MIT Press, 2001.
- E. Fox Keller, « Génome, postgénome : quel avenir pour la biologie ? », *La Recherche,* juin 2004, p. 30.

[9] J. L. Reed et B.O. Palsson, *J. Bacteriol., 185,* 2692, 2003.

[10] M.W. Covert *et al., Nature, 429,* 92, 2004.

[11] H.A. Simon, *La Science des systèmes, sciences de l'artificiel,* Épi éditeurs, 1974.

[12] Richard C. Lewontin, *La Triple Hélice – Les gènes, l'organisme, l'environnement,* Seuil, 2003.

2-L'industrie cherche son

Si les groupes pharmaceutiques semblent, pour l'instant, plutôt indifférents à la modélisation cellulaire, plusieurs sociétés de biotechnologie ont choisi d'occuper ce créneau. Les premiers résultats sont timides, mais encourageants.

Olivier Blond
est journaliste scientifique.

Qu'il s'agisse de séquençage du génome ou d'analyse des protéines, les technologies à haut débit ont rempli les disques durs des chercheurs de quantités énormes d'informations. La question est désormais de les interpréter. « *Certains pensent que la biologie systémique* [les modèles informatiques] *détient la clé qui permettra de réassembler toutes ces informations et d'accélérer la découverte de nouveaux médicaments* », lit-on dans la revue *Nature Biotechnology* d'octobre 2004. Mais Frank Molina, du laboratoire de bioinformatique pour la biologie systémique du CNRS, à Montpellier, nous met en garde : « *Les ambitions sont énormes, et nous sommes à l'aube d'une révolution en biologie. Mais il faut savoir faire la part de la réalité et des discours marketing. Non pas que la science-fiction ait tort, mais les résultats qu'elle promet n'arriveront que dans quinze ans.* »

Relier les informations

À l'instar de nombreux acteurs de la recherche privée dans ce secteur, François Iris, président fondateur de la société Bio-Modeling Systems (BMSystems), adhère à cette prudence : « *Un fossé colossal sépare nos capacités actuelles de la réalisation d'une cellule hépatique ou neuronale virtuelle. Mais il est impossible d'attendre que ces modèles soient achevés. C'est pourquoi nous en cherchons d'autres, plus simples, qui amènent des prédictions que l'on puisse confronter à des expériences.* »

Pour l'instant, les projets privés sont encore peu nombreux, et la plupart sont l'œuvre de petites entreprises de biotechnologie. En effet, les industriels de la pharmacie hésitent encore à entrer dans la course. « *Ils sont méfiants, car trop d'argent a été dilapidé dans des start-up qui promettaient monts et merveilles* », explique Laurent Buffat, directeur scientifique de IT-Omics. Et si Eli Lilly peut se targuer d'avoir créé, à Singapour, le plus grand centre en ce domaine, il n'emploie pas plus de 60 personnes.

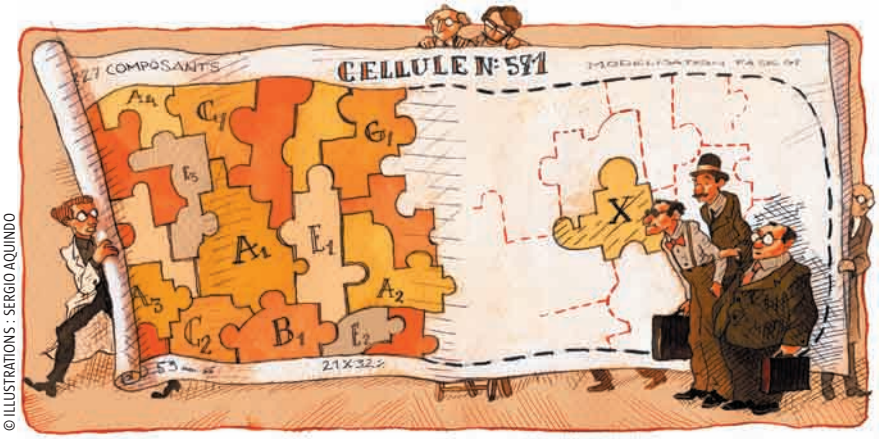
Où en est-on en France ? Parmi les acteurs, la société Atragene Bioinformatics développe des outils informatiques pour identifier de nouveaux facteurs de transcription. Ces molécules se fixent sur l'ADN en des zones particulières et contrôlent ainsi l'expression d'autres gènes. Chaque facteur de transcription en active d'autres qui, fonctionnant en cascade, jouent un rôle fondamental. Le dérèglement de la chaîne est à l'origine de nombreux cancers. Comme l'explique Alain Malpertuy, directeur scientifique d'Atragène : « *Ces facteurs de transcription sont mal connus. Chez l'humain, on estime leur nombre à plus de 3000, mais on ne connaît qu'environ 300 sites de fixation.* »

Atragene Bioinformatics commercialise deux logiciels permettant d'analyser les profils d'expression des gènes obtenus grâce à des puces à ADN*, et d'identifier d'éventuels éléments génétiques de régulation. Elle met également en place des plates-formes d'analyse bioinformatique. L'enjeu est ici d'acquérir suffisamment d'informations pour construire des modèles (relativement simples) qui expliquent la régulation des processus étudiés.

Certaines « biotechs » veulent aller plus loin : mettre en relation non pas quelques dizaines de molécules, mais des milliers. C'est l'ambition de IT-Omics. Cette société construit des outils de gestion de données appelés graphes. L'idée est de fouiller automatiquement l'ensemble de la littérature scientifique et d'en extraire des « relations » entre des objets biologiques – gènes, protéines ou autres molécules. Concrètement, cela signifie croiser la base de publications Medline (qui contient plus de 12 millions d'articles de biologie et médecine, la quasi-totalité de ceux publiés depuis 1960) avec les bases de données de gènes ou de protéines (GenBank et Swiss-Prot, par exemple), en recherchant, dans les résumés des articles, toute mention d'un lien entre deux entités A et B.

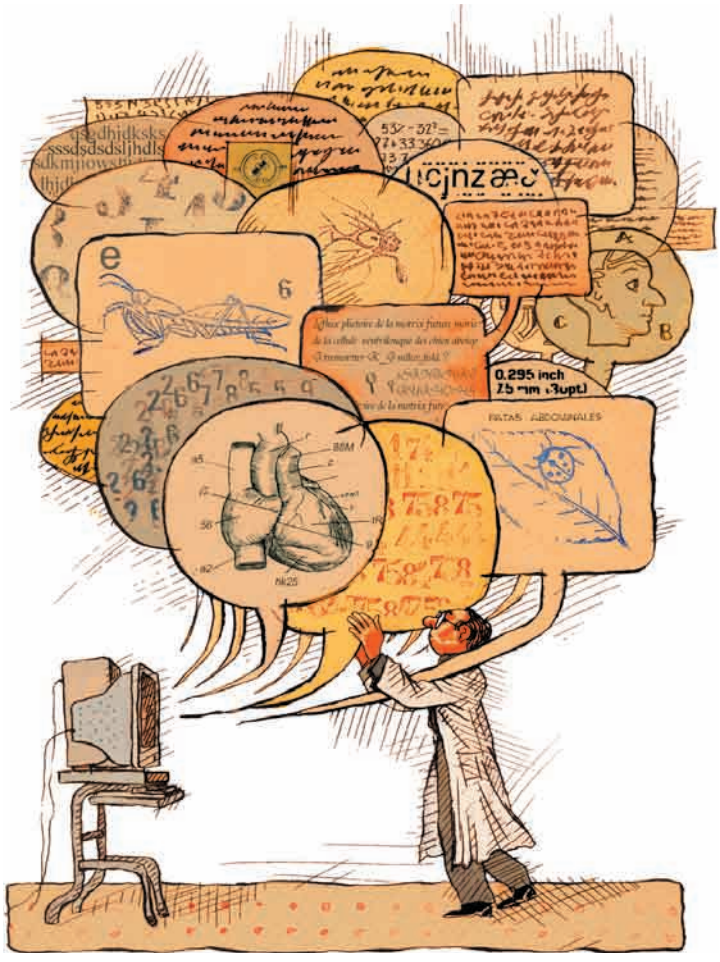
« *Le graphe de IT-Omics contient 500 000 relations entre 18 000 objets*, explique Laurent Buffat. *C'est une sorte de représentation pragmatique de la connaissance* ». Comment utiliser un tel outil ? « *Il faut poser une question précise. Sans cela, on se noie dans les données.* » Il peut s'agir d'interpréter les résultats d'une puce à ADN sur laquelle on a observé la surexpression

***Une puce à ADN** est un dispositif qui permet de visualiser l'expression d'un très grand nombre de gènes (plus de 10 000) de façon rapide et simultanée.



© ILLUSTRATIONS : SERGIO AQUINDO

modèle



simultanée de 1 500 gènes : on plonge ces données dans le graphe et on essaie d'en sortir, par exemple, un lien fonctionnel entre ces gènes. La société a déjà vendu plusieurs licences de son logiciel à des groupes pharmaceutiques comme Aventis ou UCB*. IT-Omics assure aussi des fonctions de service, en proposant des interprétations à la carte.

Toutefois, selon François Iris, de BMSystems, le système des graphes souffre de deux inconvénients majeurs : d'une part, il est difficile de savoir si l'information est fiable, car la littérature scientifique regorge d'erreurs. D'autre part, la complexité d'un graphe croît exponentiellement avec le nombre de ses éléments, jusqu'à devenir ingérable. BMSystems a donc élaboré un système de « sélection négative » : un outil semi-automatique de réfutation qui génère des hypothèses de relation entre entités cellulaires et les teste les unes après les autres en les recroisant avec les bases de données. Le but est d'en éliminer le plus possible jusqu'à n'en conserver qu'un nombre suffisamment restreint pour permettre de proposer des expériences. « *À l'inverse des graphes, plus l'information est riche et diversifiée, meilleures sont nos hypothèses. Nous avons même un problème quand il n'y a pas assez d'informations. Quant aux incohérences internes, elles nous servent car elles permettent d'éliminer les hypothèses fausses* », explique François Iris.

Débouché médical

Le « business modèle » de BMSystems diffère lui aussi de celui de ses concurrents. Manuel Géa, PDG de l'entreprise, a choisi non de vendre une technologie particulière (le logiciel reste confidentiel) ou de développer un modèle exhaustif, mais de proposer des « solutions » à des problèmes industriels. Par exemple, BMSystems a testé son approche avec une forme de cancer du sein. Ses chercheurs ont identifié plusieurs voies métaboliques, qu'ils ont rassemblées dans un modèle. Et il est apparu que le blocage simultané de trois d'entre elles devait stopper la prolifération cancéreuse. Or, en pratique, chacune de ces voies peut être bloquée par une molécule peu toxique, à l'inverse des anticancéreux classiques. Sur des cellules en culture, chaque molécule utilisée seule n'apporte qu'une légère amélioration. Mais, utilisées ensemble, elles sont efficaces dans 75 % des cas ! Ces résultats ont été obtenus indépendamment par trois hôpitaux publics français et publiés [1] – sauf pour la troisième molécule, qui reste secrète. Manuel Géa affirme être en négociation avec un industriel pour développer cette approche thérapeutique. Il affirme également que BMSystems a déjà réalisé neuf modèles prédictifs de pathologies humaines, dont deux ont déjà été validés et publiés (pour le cancer du sein), et deux autres (dont un pour la maladie de Creutzfeldt-Jakob) sont en « évaluation expérimentale » dans des laboratoires de recherche publics.

Certains des projets de la recherche privée portent donc leurs fruits. Mais comparés à ceux développés dans le monde académique, les systèmes utilisés ne sont pas des modèles itératifs évoluant avec le temps (voir l'article de F. Rechenmann et H. de Jong). Par ailleurs, reconnaît Laurent Buffat : « *Nos modèles sont des outils de connaissance et d'interprétation des données qui permettent d'identifier, dans le flux des données, ce qui n'est pas connu et ce qui est intéressant. Ce ne sont pas encore des modèles qui permettent une approche rationnelle de la recherche pharmaceutique.* » On est encore loin des méthodes qui permettraient enfin à l'industrie pharmaceutique de découvrir de nouvelles catégories de molécules. Loin aussi de la toxicogénomique : prévoir si une molécule aura des effets secondaires néfastes et dans quelle mesure. Mais cela pourrait changer. ▮▮ **O. B.**

*__UCB__ est une société pharmaceutique basée à Bruxelles et productrice, notamment, des anti-allergiques Zyrtec et Atarax.

POUR EN SAVOIR PLUS

▮ Les sites des sociétés
www.atragene.com/atragene/
www.it-omics.com/v3gb/index.html
www.bmsystems.net/

[1] F. Gadal, *Nucl. Acids Res.*, *31*, 5789, 2003 ; F. Gadal *et al.*, *J. Mol. Endoc.*, sous presse.