

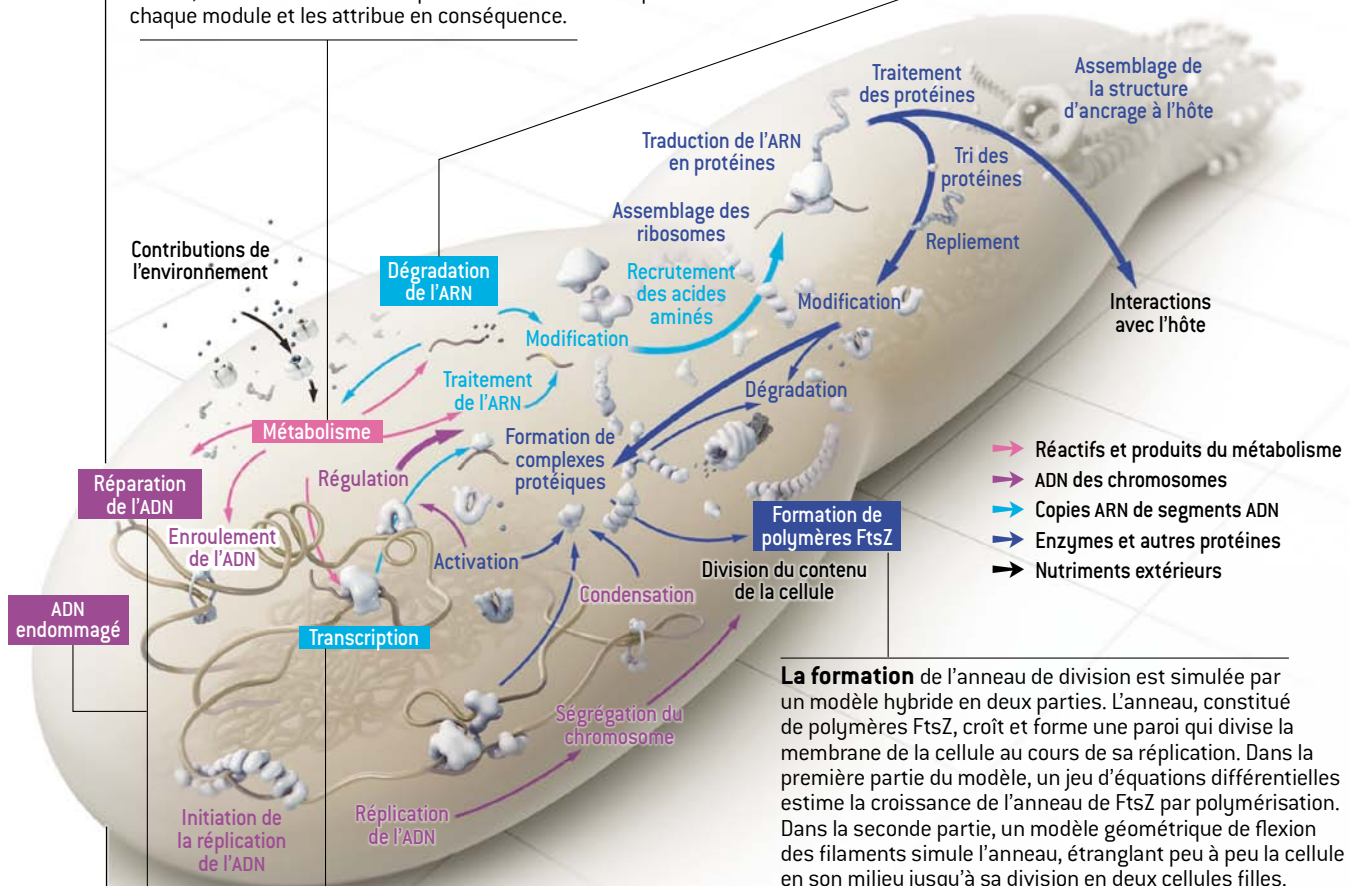
LE PRINCIPE DU SIMULATEUR

L'auteur a mis au point un modèle numérique de la bactérie infectieuse *Mycoplasma genitalium* qui représente presque tous les aspects de la vie, de la croissance et de la reproduction de ce microbe. Aucune approche mathématique ne permet de simuler toutes les fonctions biologiques de la cellule, si bien que ces fonctions ont été divisées en 28 modules distincts (*indiqués dans la cellule ci-dessous*) qui interviennent dans le traitement de l'ADN (*en violet*), de l'ARN (*en bleu clair*), des protéines (*en bleu foncé*), et de l'énergie, des nutriments et des déchets (*en rose*). Pour chaque module, les chercheurs ont sélectionné la méthode mathématique la plus appropriée (*voir ci-dessous quelques exemples de ces méthodes*).

Au début du programme, tous les modules tournent selon une séquence aléatoire, l'ensemble simulant une seconde de temps réel. Beaucoup des valeurs d'entrée sont tirées d'un grand tableau de variables représentant leurs états initiaux, d'autres sont sélectionnées à partir de fourchettes ou de calculs de probabilités. Les chercheurs peuvent simuler différents scénarios en modifiant la configuration de départ. Après ce premier pas, le programme met à jour le tableau de variables afin de refléter les sorties des modules. La séquence tourne alors pour un nouveau pas d'une seconde, met à jour le tableau, etc. La boucle continue jusqu'à ce que la cellule se divise, meure ou devienne trop vieille pour avoir un sens.

Le métabolisme des nutriments et des déchets est modélisé en utilisant l'analyse du bilan des flux. Cette méthode exploite des techniques de programmation linéaire pour calculer les vitesses de réaction qui optimisent la croissance, la production d'énergie ou toute autre caractéristique choisie par le modélisateur. Elle suppose que les réactions ont lieu assez vite pour obtenir un état stationnaire pendant la seconde correspondant à un pas de simulation. Afin d'empêcher les premiers modules d'utiliser des substances dont d'autres modules ont besoin, le simulateur estime la part de ces ressources qui revient à chaque module et les attribue en conséquence.

La dégradation et le recyclage de l'ARN et des protéines sont modélisés en utilisant des processus de comptage dits de Poisson. Ces processus utilisent un générateur de nombres aléatoires et des fonctions de probabilité pour décider si un fragment donné d'ARN ou de protéine se dégrade ou persiste jusqu'au pas suivant.



La formation de l'anneau de division est simulée par un modèle hybride en deux parties. L'anneau, constitué de polymères FtsZ, croît et forme une paroi qui divise la membrane de la cellule au cours de sa réplication. Dans la première partie du modèle, un jeu d'équations différentielles estime la croissance de l'anneau de FtsZ par polymérisation. Dans la seconde partie, un modèle géométrique de flexion des filaments simule l'anneau, étranglant peu à peu la cellule en son milieu jusqu'à sa division en deux cellules filles.

La transcription et la traduction des gènes (étapes au cours desquelles sont fabriquées de nombreuses protéines nécessaires à la croissance et la réplication des cellules) sont simulées par des algorithmes en plusieurs parties incluant des modèles de Markov (ces modèles suivent dans le temps l'état des enzymes qui copient les gènes en ARN), un algorithme probabiliste décrivant la fixation de ces enzymes sur l'ADN et une programmation linéaire pour distribuer l'énergie et autres ressources.

Les dommages subis par l'ADN et les réparations sont modélisés selon cette même méthode non déterministe.

des hypothèses sur les paramètres a et b . Cette exigence de précision dicte les expériences de laboratoire à mener pour combler les lacunes sur les vitesses de réaction et autres quantités caractéristiques du processus simulé.

Mais il paraissait clair aussi qu'une fois les modèles vérifiés, ils se substituaient à certaines expériences, permettant de réserver les travaux coûteux aux questions insolubles par les simulations seules. Les expériences simulées aideraient les biologistes à définir leurs priorités, ce qui accélérerait le rythme des découvertes. Les modèles laissaient entrevoir des outils si séduisants que certains voyaient dans la simulation de cellules entières le grand défi du XXI^e siècle.

Étudiant en thèse, j'ai été impressionné par les premiers résultats de ces pionniers de la modélisation cellulaire (voir l'encadré page 22), et ce défi est devenu mon obsession. Plus tard, quand j'ai monté mon laboratoire et développé des techniques d'imagerie de cellules isolées, ce défi est resté dans mes pensées. Et en ce soir de février, j'ai enfin vu un moyen de le relever.

Comment contraindre et vérifier ses hypothèses ?

Afin de simuler le cycle de vie d'une espèce microbienne avec une précision suffisante pour imiter ses comportements complexes, il fallait résoudre trois problèmes. Tout d'abord, il s'agissait d'encoder sous forme de formules mathématiques et d'algorithmes toutes les fonctions importantes : le flux d'énergie, de nutriments et de produits de réactions dans la cellule (en d'autres termes, le métabolisme), la synthèse et la destruction de l'ADN, de l'ARN et des protéines, et l'activité d'une multitude d'enzymes. Ensuite, il fallait trouver un cadre global pour intégrer toutes ces fonctions. Le dernier problème, enfin, était le plus compliqué à bien des égards : fixer judicieusement les limites haute et basse de chacun des quelque 1700 paramètres du modèle.

J'ai vite compris que nous aurions beau éplucher la littérature sur *M. genitalium* et ses proches parents en quête de ces paramètres (nous avons passé deux ans à extraire les

données d'environ 900 articles), il faudrait, dans certains cas, se contenter de faire des hypothèses ou d'utiliser des résultats d'expériences portant sur d'autres types de bactéries, telle *Escherichia coli*, afin d'obtenir certaines valeurs (par exemple, le temps de séjour moyen des molécules d'ARN dans la cellule avant qu'elles ne soient détruites et recyclées). Or sans un moyen de contraindre et de vérifier ces hypothèses, l'entreprise était vouée à l'échec.

C'est précisément un tel moyen qui m'est apparu ce soir d'hiver : j'ai réalisé que modéliser une cellule unique (et non un groupe de cellules, comme l'avaient

LES MODÈLES LAISSAIENT ENTREVOIR des outils si séduisants que certains voyaient dans la simulation de cellules entières le grand défi du XXI^e siècle.

fait presque toutes les études antérieures) offrait peut-être la situation recherchée. Considérons la croissance et la multiplication des cellules. Une population de cellules croît graduellement ; la naissance ou la mort d'une cellule perturbe peu cette croissance. En revanche, pour une cellule donnée, la division est un événement d'une importance capitale. Avant de se scinder en deux, l'organisme double sa masse, plus précisément les quantités d'ADN, de membrane et de toutes les protéines nécessaires à sa survie. Ainsi, si le modèle se limite à une cellule unique, l'ordinateur pourra compter et suivre chaque molécule sur tout le cycle de vie. Et il pourra vérifier les quantités lors du passage de une à deux cellules.

De plus, une cellule se reproduit à un rythme donné. *M. genitalium*, par exemple, se divise toutes les neuf à dix heures dans un environnement normal de laboratoire. Elle met rarement moins de six heures ou plus de 15. Le fait que le contenu entier de la cellule doive se dupliquer dans ce délai permet de choisir des fourchettes raisonnables pour de nombreuses variables qui, sinon, auraient été indéterminées, telles celles qui contrôlent le moment où commence la réplication de l'ADN.

J'ai réuni une équipe de physiciens, de biologistes, de modélisateurs et même un ancien ingénieur informaticien de Google, et nous avons discuté des approches mathématiques à adopter. Michael Schuler, ingénieur biomédical de l'Université Cornell,

était un pionnier des simulations cellulaires fondées sur les équations différentielles. Bernhard Palsson, qui m'avait encadré lors de mes études à San Diego, avait développé une technique qui modélisait bien le métabolisme. D'autres avaient montré que le hasard est un élément important dans la transcription des gènes, ou que la division des cellules implique un changement géométrique de la membrane cellulaire.

Aucune des méthodes développées ne serait suffisante à elle seule pour modéliser toutes les fonctions d'une cellule. Nous avons donc décidé de créer un modèle de cellule entière constitué de 28 modules distincts, dont chacun utilise l'algorithme qui correspond le mieux au processus biologique concerné et à la connaissance que nous en avons (voir l'encadré page ci-contre). Cette stratégie a conduit à un patchwork de procédures mathématiques. Restait à assembler les morceaux.

J'ai repensé à un cours de conception d'usine chimique que j'avais suivi à l'université. Pour le projet de fin d'année, nous avions utilisé un outil de simulation puissant nommé HYSYS pour ébaucher les plans d'une grande raffinerie. HYSYS permettait d'effectuer chaque réaction chimique principale dans un module séparé et de connecter ces modules. En d'autres termes, il offrait un cadre permettant de réunir de nombreux types d'opérations chimiques en un système ordonné et prévisible.

28 modules interconnectés

Avec quelques modifications, cette approche pouvait fonctionner pour notre simulateur de cellule, à condition de faire une hypothèse simplificatrice importante : supposer que, bien que tous les processus biologiques aient lieu simultanément dans une cellule vivante, leurs actions sont indépendantes sur des périodes inférieures à la seconde. Si cette hypothèse était viable, nous pourrions diviser la vie de la cellule en tranches d'une seconde et faire tourner chacun des 28 modules, dans l'ordre, pendant une seconde avant de mettre à jour l'ensemble des variables de la cellule. Le modèle rendrait compte de toutes les interdépendances biochimiques (par exemple, le fait que la transcription des gènes et la synthèse de l'ADN dépendent de l'énergie et des nucléotides

Un tournant de la biologie des systèmes

Le modèle « pleine-cellule » de *Mycoplasma genitalium* proposé par Markus Covert et ses collègues marque un tournant dans la jeune histoire de la biologie des systèmes.

Si l'idée de modéliser un organisme monocellulaire n'est pas nouvelle, le modèle de M. Covert est le premier à s'appuyer sur un inventaire exhaustif de gènes et à donner une représentation aussi complète que possible de leurs fonctions et des processus cellulaires dans lesquels ils sont impliqués.

Pour construire ce modèle, qui agrège une quantité considérable de connaissances, M. Covert et son équipe ont dû vaincre plusieurs difficultés. Ils ont dû créer un cadre original pour combiner des sous-modèles fondés sur des techniques de modélisation très différentes. Par ailleurs, ils ont su trouver des jeux de paramètres pour lesquels le modèle produit

les bonnes proportions des différents ingrédients qu'une cellule doit accumuler pour compléter son cycle et faire une copie d'elle-même. Leur modèle produit des temps réalistes de division cellulaire et prédit correctement la viabilité d'un certain nombre de mutants.

Ce modèle est en rupture avec la tradition en ce qu'il n'est pas destiné à évaluer une hypothèse préalable spécifique. L'enjeu est d'organiser la totalité des interactions qui composent le système de sorte qu'on puisse en observer le comportement collectif. Ce n'est que dans un second temps que l'on cherchera à exploiter cet outil « panoptique » pour tester de nouvelles hypothèses.

Abandonner ainsi une modélisation d'explication pour une modélisation de navigation ne va pas de soi. On passe des principes clairs et concis d'un système à quelques variables à un système plus complexe difficile à appréhender intuitivement. Le modèle est une carte qui permet d'explorer cette complexité. De ce point de vue, il se manipule comme on le ferait du système réel, et les « faits » qu'il produit réclament une explication au même titre que des données expérimentales. Mais comme le modèle ne reflète que la connaissance imparfaite que nous avons de cet organisme, si simple soit-il, et qu'il ne la reflète qu'imparfaitement, le passage du système réel, avec ses difficultés propres, à son substitut numérique vient sans garantie de qualité. Le passage « au numérique » peut ainsi sembler une opération au bénéfice incertain dans l'état de nos connaissances.

Quoi qu'il en soit, la possibilité d'interroger le système dans sa totalité permet d'aborder de nouvelles questions. On peut par exemple étudier l'impact du niveau des ressources énergétiques sur les composants du système, ou comment offre et demande se répondent sur le « marché » de la traduction des gènes, et si des gènes répondent différemment aux conditions de cette compétition ; ou encore examiner l'influence de la densité des molécules associées à l'ADN sur l'activité transcriptionnelle du système. Le modèle pleine-cellule de M. Covert ouvre un nouveau champ d'enquête en biologie des systèmes, où l'on étudie l'interdépendance des différents processus qui composent une cellule et leur intégration fonctionnelle à l'échelle cellulaire.

– Vincent Danos
CNRS et Université
d'Édimbourg (Écosse)

produits par le métabolisme), mais seulement aux échelles temporelles supérieures à la seconde.

Tout en construisant notre cellule virtuelle, nous avons intégré au logiciel des capteurs pour mesurer l'état de la cellule. Chaque lancement du simulateur, couvrant la totalité d'un cycle de vie d'une cellule, produisait 500 mégaoctets de données. En sortie, on obtenait des dizaines de graphiques et de représentations visuelles.

Au début, les résultats furent frustrants. Pendant des mois, tandis que nous corrigeons le code, précisons les équations mathématiques, ajoutons de nouvelles contraintes sur les paramètres et ajustons certaines (grâce à de nouvelles données de laboratoire), la cellule refusait toujours de se diviser et se comportait de façon imprévisible. Pendant un moment, par exemple, elle a produit d'énormes quantités d'un certain acide aminé (l'alanine) et presque rien d'autre.

Et puis un jour, notre microbe cybernétique est arrivé au terme de son cycle cellulaire et s'est enfin divisé. De plus, son

temps de doublement était proche de neuf heures – celui du mycoplasme réel. De nombreuses mesures étaient encore loin de la réalité, mais le succès était enfin à notre portée.

Quelques mois plus tard, la majorité des données obtenues par simulation correspondait exactement à la croissance observée d'une vraie cellule. Et le reste était fascinant : inattendu, mais biologiquement plausible. Le premier modèle numérique d'un organisme vivant entier était fonctionnel. Qu'allait-il nous apprendre ?

Modéliser la vie d'une cellule humaine ?

Aujourd'hui, après environ un an d'utilisation de ce nouvel outil, nous découvrons encore des informations passionnantes à chaque fois que nous scrutons le fonctionnement de notre micro-organisme virtuel, tandis qu'il gère des millions de détails participant à sa vie et à sa réplication. Par exemple, la fréquence à laquelle les protéines participant au décodage des gènes se

décrochent de l'ADN est impressionnante : environ 30000 fois par cycle de vie de neuf heures. Nous avons aussi découvert que la période de doublement du matériel biologique lors de la division cellulaire, très stable, est en fait une propriété émergente qui découle des interactions de deux phases distinctes et indépendantes de la réplication, toutes deux de durée très variable. Et les enregistrements seconde par seconde du comportement de la cellule nous ont permis d'expliquer pourquoi sa division s'arrête net quand on inactive certains gènes, mais se poursuit encore une dizaine de fois si l'inactivation porte sur d'autres gènes essentiels : ces divisions supplémentaires se produiraient à chaque fois que la cellule a en réserve davantage de copies d'une protéine essentielle qu'elle n'en a besoin durant un cycle de vie : l'excédent est légué à ses descendantes, qui ne meurent que quand le stock est épuisé. Il faudra sans doute encore des années pour comprendre tout ce que ces simulations révèlent sur le fonctionnement des mycoplasmes et des cellules en général.