

de concevoir des dizaines de modules à partir des données scientifiques. Concrètement, cela signifiait passer au peigne fin les données biochimiques de près d'un millier d'articles scientifiques, puis d'utiliser les valeurs collectées pour contraindre et ajuster des milliers de paramètres, tels que l'affinité des enzymes avec leurs cibles ou la fréquence de liaison et de décrochage des protéines qui « lisent » l'ADN. Même avec l'aide de collaborateurs, le projet prendrait des années ! Mais j'avais bon espoir.

Les 400 gènes essentiels du mycoplasme

Nous nous sommes inspirés des premières recherches visant à modéliser la vie. En 1984, Harold Morowitz, alors à l'Université Yale, avait ouvert la voie. Pour lui, il fallait commencer par modéliser les bactéries les plus simples que l'on savait cultiver, les mycoplasmes. Minuscules et relativement

simples, ces bactéries ont un intérêt particulier : deux espèces déclenchent des maladies humaines, le parasite à transmission sexuelle *M. genitalium*, qui colonise les voies urinaires et génitales, et le germe *Mycoplasma pneumoniae*, qui peut entraîner une pneumonie atypique. Un modèle de l'une ou l'autre espèce aurait un intérêt tant biologique que médical.

La première étape, selon H. Morowitz, devait être le séquençage du génome du microbe choisi. En 1995, J. Craig Venter et ses collègues de l'Institut pour la recherche génomique ont mené cette tâche à bien pour *M. genitalium*, qui ne compte que 525 gènes (une cellule humaine en a plus de 20000).

Quatre ans plus tard (j'étais alors étudiant en thèse à San Diego), l'équipe de J. C. Venter concluait que seuls environ 400 de ces gènes sont essentiels à la vie de cette espèce (tant que les mycoplasmes sont cultivés dans un milieu riche). J. C. Venter et

ses collègues ont alors synthétisé les gènes essentiels d'une espèce de mycoplasme et montré qu'ils étaient fonctionnels dans une cellule.

Pour moi et d'autres jeunes biologistes de la fin des années 1990, ce groupe était comme Led Zeppelin : des personnalités iconoclastes, remarquables, jouant une musique que personne n'avait entendue jusque-là. Clyde Hutchinson, un biologiste du groupe, a prédit que le test ultime de notre compréhension des cellules simples serait la modélisation informatique de l'une d'elles : on peut construire une cellule fonctionnelle en laboratoire en combinant des éléments sans comprendre dans le détail comment ils s'assemblent. Ce n'est pas le cas en informatique.

H. Morowitz aussi avait appelé de ses vœux la construction d'un simulateur de cellule fondé sur les données du génome de *M. genitalium*. Selon lui, toute expérience réalisable en laboratoire l'est aussi sur un

Les grandes étapes de la modélisation cellulaire

Le long parcours de l'auteur pour aboutir à la première modélisation d'une cellule unique d'une simple bactérie, *Mycoplasma genitalium*, a été nourri des travaux théoriques, des études génétiques et des essais de modélisation d'autres chercheurs, dont voici les principaux. Concevoir un modèle numérique de cellule humaine sera encore plus difficile : une cellule humaine contient presque 40 fois plus de gènes qu'une bactérie, regroupés en chromosomes dont la structure et les informations contenues sont bien plus complexes. Quelques étapes essentielles pour y parvenir sont proposées à la fin de cette chronologie.

1967

Francis Crick et Sydney Brenner proposent le « Projet K : la solution complète d'*Escherichia coli* », une initiative visant à dresser le « patron » de cette bactérie digestive commune, y compris des détails fins de sa génétique, de ses processus énergétiques et de sa reproduction.

1984

Harold Morowitz, alors à l'Université Yale, présente les grandes lignes d'un plan de séquençage, puis de modélisation d'un mycoplasme.

1984

Une équipe menée par Michael Schuler de l'Université Cornell présente un modèle informatique utilisant des équations différentielles pour traduire la plupart des processus biologiques majeurs intervenant dans la croissance d'une cellule *E. coli*. Le modèle n'est cependant pas en mesure d'inclure l'activité des gènes, car le génome d'*E. coli* n'a pas encore été séquencé.

1989-1990

Bernhard Palsson de l'Université du Michigan publie un modèle complet du métabolisme d'un

globule rouge humain qui inclut les effets des variations de l'acidité et de l'hypoglycémie.

1995

J. Craig Venter, de l'Institut TIGR, et ses collègues achèvent le séquençage du génome de *M. genitalium*.

1999

Masaru Tomita et ses collaborateurs de l'Université Keio au Japon construisent *E-Cell*, un système de modélisation de cellule fondé sur des équations différentielles et incluant

127 gènes, la plupart issus de *M. genitalium*.

2002

L'Alliance pour la signalisation cellulaire, une vaste collaboration d'une cinquantaine de chercheurs, lance un projet ambitieux évalué à dix millions de dollars sur dix ans : modéliser certaines cellules de la souris (les lymphocytes B du système immunitaire et les cellules du muscle cardiaque). Le projet produit des données intéressantes, mais se heurte aux difficultés de culture des lymphocytes B.

2002

B. Palsson, George Church de l'Université Harvard, l'auteur et d'autres conçoivent un modèle à l'échelle génomique du métabolisme d'*Helicobacter pylori*, une bactérie qui infecte l'homme et peut entraîner des ulcères gastriques et un cancer de l'estomac.

ordinateur. Le degré de correspondance entre les résultats expérimentaux et les simulations est une mesure de la validité de la biologie moléculaire actuelle, c'est-à-dire de la théorie expliquant comment l'ADN et les autres biomolécules de la cellule interagissent et donnent naissance à la vie telle que nous la connaissons. En d'autres termes, à mesure que nous assemblons le puzzle de la vie apparaissent de façon plus claire les pièces et les interactions qui manquent à notre théorie.

Les séquenceurs à haut débit et les équipements robotiques des laboratoires ont accéléré la recherche des pièces manquantes. Toutefois, les masses récoltées de séquences d'ADN et d'informations sur l'activité des gènes n'étaient pas (et ne sont toujours pas) accompagnées d'explications sur la façon dont ces activités se combinent. C'est en partie pour cela que, malgré l'annonce régulière de la découverte de nouveaux gènes associés au cancer, à

l'obésité ou au diabète, les traitements curatifs de ces maladies restent hors d'atteinte. Seule la compréhension des nombreuses interactions moléculaires à l'origine de ces maladies conduira à ces traitements.

Les pionniers de la modélisation cellulaire ont compris que les simulations de cellules entières, avec leurs composants et leur réseau d'interactions, seraient de puissants outils pour comprendre ces données incomplètes et désordonnées. Un simulateur de cellule entière transformerait un jeu complet d'hypothèses sur le fonctionnement d'une cellule en un ensemble d'algorithmes mathématiques rigoureux.

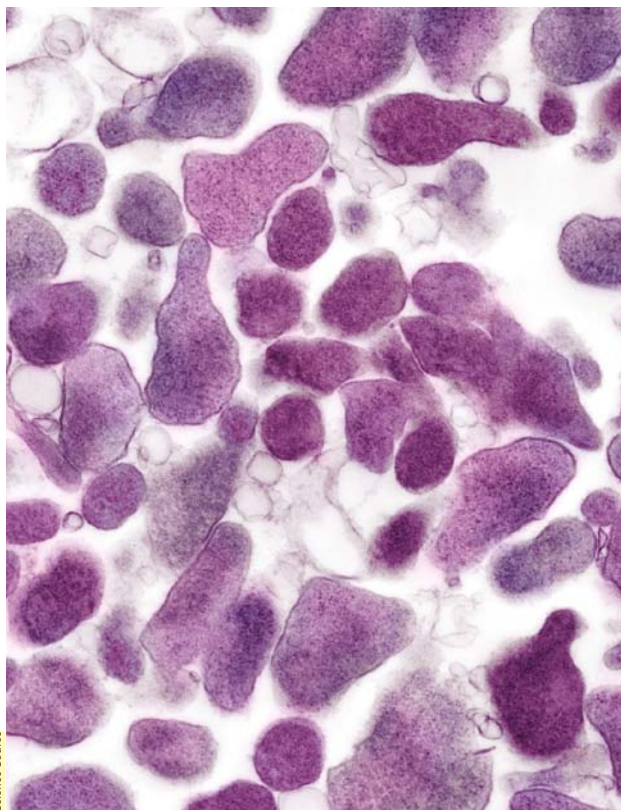
Certes, la route était encore longue : il ne suffisait pas de savoir que tel facteur X régule le gène Y pour traduire ce processus en langage de programmation. Il manquait des précisions. Les programmeurs expriment les processus sous forme d'équations, dont l'une des plus simples est $Y = aX + b$. De telles équations nécessitent d'avancer

■ L'AUTEUR



Markus COVERT est professeur assistant de génie biologique à l'Université Stanford, où

il dirige un laboratoire dédié à la biologie des systèmes.



Science Source

LA BACTÉRIE UNICELLULAIRE *Mycoplasma genitalium* (en violet) est l'un des organismes vivants les plus simples. Cela ne signifie pas pour autant qu'il soit facile de modéliser son cycle de vie...

2004

B. Palsson et l'auteur publient, avec trois autres chercheurs, un modèle informatique des 1010 gènes intervenant dans la régulation du métabolisme et de la transcription de l'ADN d'*E. coli* et montrent que le modèle prédit avec précision les résultats des expériences réalisées sur les bactéries réelles.

2012

L'auteur et ses collègues publient un modèle de cellule entière de *M. genitalium* qui, pour la première fois, simule tous les gènes et processus chimiques connus dans un organisme qui s'autoréplique.

2013

L'auteur et ses collègues montrent que le modèle prédit avec précision l'activité de plusieurs enzymes.

ET APRÈS ?

- Concevoir un modèle de cellule entière pour une bactérie plus classique et mieux étudiée, telle *E. coli*.
- Modéliser un organisme eucaryote unicellulaire tel que la levure *Saccharomyces cerevisiae*. Dans une telle cellule, l'ADN est contenu dans un noyau entouré d'une membrane, et non libre comme dans une bactérie.
- Construire un modèle de cellule animale facile à cultiver en laboratoire, tel un macrophage (un type de cellule immunitaire) de souris.
- Construire un premier modèle de cellule humaine, sans doute, là encore, un macrophage.
- Modéliser d'autres types de cellules humaines, en particulier celles jouant un rôle important dans les maladies communes.

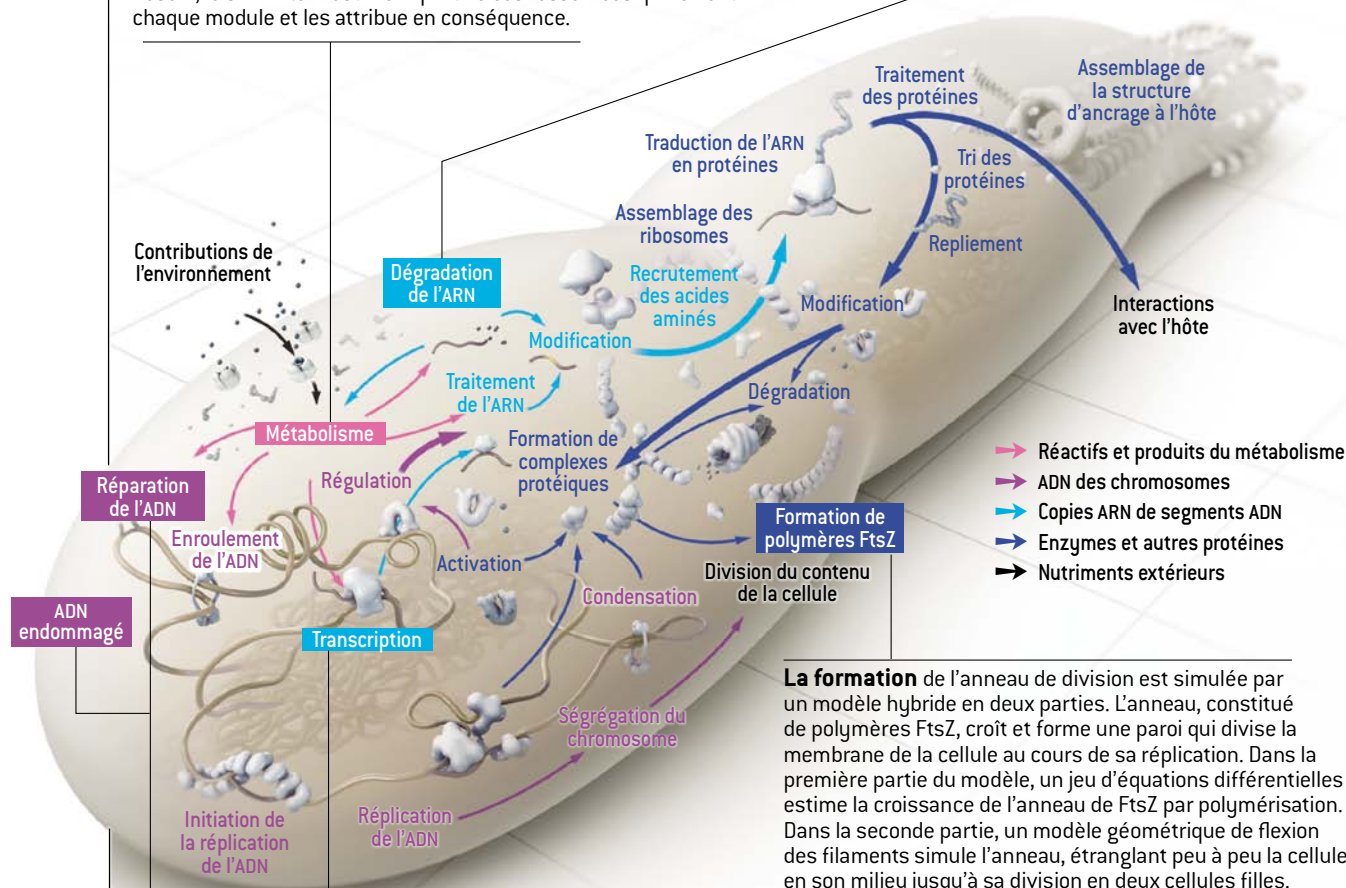
LE PRINCIPE DU SIMULATEUR

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.