Mathématiques et biologie

Bruno Goffinet
Chef du département Mathématiques et Informatique Appliquées, INRA

Les enjeux de la biologie sont fascinants et les mathématiques sont de plus en plus nécessaires aussi bien pour proposer des concepts qui permettent de modéliser les processus extrêmement complexes qui régissent le vivant que pour analyser le nombre considérable de données que produit aujourd'hui la biologie. Au-delà de la compréhension du vivant, il s'agit d'être capable de prédire le comportement de systèmes vivants dans certains environnements comme par exemple, la résistance à la sécheresse d'un nouveau génotype (nouvelle variété) de blé.

Les mathématiques n'ont commencé à jouer un rôle décisif dans la biologie que tardivement. C'est d'abord avec Fischer que la variabilité des résultats d'expérience et la diversité entre les individus ont trouvé un cadre de formalisation pertinent avec les statistiques et les probabilités. La génétique quantitative, la génétique des populations, et l'analyse des résultats expérimentaux en agronomie ont même été des moteurs essentiels du développement des statistiques au début du 20ème siècle. En parallèle, les modèles de type *prédateur-proie*, à base d'équations différentielles ont permis de poser les bases de l'écologie quantitative. Ce qui est nouveau aujourd'hui c'est la formidable croissance des données, associée avec celle des moyens de calcul, qui rend possible de modéliser les processus intégrant des niveaux d'échelles multiples aussi bien sur le plan spatial, que temporel ou des niveaux d'organisation de la cellule à l'individu, puis à la communauté d'individus. La construction de ces modèles offre de nouveaux défis pour les mathématiques et l'informatique.

Avec l'écologie sont apparus des concepts opérationnels comme la plasticité, la compétition, etc... Un enjeu est donc de construire avec les biologistes qui travaillent à l'échelle de la cellule des concepts du même type, et de les confronter à la réalité. Il s'agit bien de concepts qui mobilisent des mathématiques, de l'informatique ou de la physique. On peut penser par exemple que les notions d'optimisation sous contrainte ou de *feed-back*, dans un cadre d'objets en interactions, devraient permettre de *mathématiser* la biologie et ainsi conduire à des lois générales de fonctionnement. On peut arriver ainsi à des problématiques génériques irriguant de nombreuses questions biologiques.

Une question en particulier est l'identification de l'ensemble d'échelles perti-

nentes pour travailler sur un problème. Il ne suffit pas d'affirmer que l'on peut intégrer les échelles inférieures d'organisation du vivant pour répondre à une question à une échelle donnée, ou au contraire d'être trop prudent en refusant systématiquement ce que l'on appelle des *usines à gaz*, mais de se donner une méthode pour savoir jusqu'où il est pertinent de modéliser, en fonction des données disponibles, des connaissances sur les processus, de l'objectif du modèle, etc. A quel point l'intégration des connaissances à l'échelle n-1 est elle nécessaire si l'on veut prédire à l'échelle n au-delà du domaine de validité des modèles statistiques basés sur l'observation ?

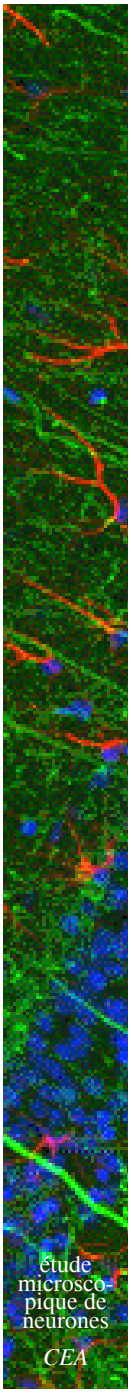
Un modèle complexe peut avoir un comportement chaotique (au sens où un changement minime dans un paramètre ou donnée d'entrée conduit à des résultats très différents), mais avoir un comportement beaucoup plus stable et prédictif si l'on s'intéresse à des résultats synthétiques. Par exemple, la prédiction de la position précise d'un cyclone par des modèles météorologiques est très instable, mais le nombre de cyclones est une quantité relativement prédictible.

Un exemple de problématique biologique mobilisant des mathématiques : les gènes et leur régulation.

Chaque cellule possède un bagage génétique contenu en grande partie dans son ADN qui porte une quantité considérable d'informations. Les gènes qui se trouvent sur l'ADN sont traduits et conduisent à la synthèse de protéines qui jouent un rôle décisif dans le métabolisme des cellules. En effet la capacité de ces protéines à accélérer ou catalyser des réactions biochimiques va permettre à la cellule de produire tel ou tel métabolite, d'être capable d'exploiter telle ou telle molécule pour pourvoir à ses besoins énergétiques et donc de s'adapter à des conditions extérieures changeantes.

Pour analyser ces phénomènes, les mathématiques jouent un rôle à de multiples niveaux. Tout d'abord, elles offrent une modélisation de l'ADN pour identifier les zones codantes pour des gènes. L'ADN est une suite de nucléotides de 4 types, A (pour la base azotée Adénine), C (Cytosine), G (Guanine) et T (Thymine). En théorie des graphes, on peut modéliser la suite des lettres (A, C, G et T) par des chaînes de Markov plus ou moins complexes. Ceci a conduit à des méthodes parfois sophistiquées qui s'avèrent extrêmement efficaces.

Après cela, on dispose avec les puces à ADN d'un indicateur de l'intensité de l'expression des gènes qui conduit à la synthèse des protéines, à un moment donné de la vie de la cellule et le cas échéant sous différentes conditions expérimentales. L'évolution dans le temps



étude microscopique de neurones

CEA

de ces quantités permet d'inférer des liens entre ces gènes à l'aide de méthodes statistiques d'analyse de la co-variation entre ces quantités.

Ayant identifié des réseaux de gènes, l'enjeu devient alors d'analyser leurs fonctionnements. On peut construire un graphe qui réunit les différents gènes impliqués dans un processus et modéliser les flux entre ces nœuds du réseau à partir d'un système d'équations différentielles. La difficulté vient alors de la taille qui peut être énorme de ces systèmes et des informations très incomplètes sur le réseau. L'enjeu est d'en tirer cependant des connaissances profondes sur le comportement du système en terme de cycles, de contrôle, de rétroaction, etc. Des techniques de discrétisation du processus peuvent s'avérer pertinentes. D'autre part, les méthodes à la frontière des statistiques et de l'informatique pour extraire et gérer des connaissances à partir d'un nombre considérable de données de nature hétérogène (données de flux, d'expression des gènes, connaissances expertes, résultats d'expérimentations spécifiques, etc.) se développent et apparaissent cruciales pour répondre aux enjeux.

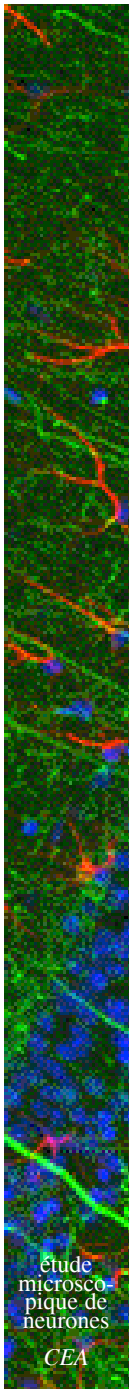
Des enjeux de recherche pour les mathématiques.

Plusieurs fronts de recherche s'ouvrent actuellement face à l'accroissement du nombre de données.

Le premier est lié à la massification de la production actuelle des connaissances au niveau mondial et aux difficultés qu'il y a à les exploiter de façon systématique et efficace. Le deuxième est lié aux possibilités technologiques nouvelles d'observation et de récolte des données à des échelles de résolution spatiale et temporelle sans précédent. Le troisième est le fruit des deux premiers et concerne l'intégration dite verticale, où il s'agit de combiner/explorer/expliquer des données diverses acquises sur un même objet mais à des échelles différentes.

Sur le plan de la recherche, le premier front conduit à poursuivre le développement de méthodes et d'algorithmes pour extraire automatiquement des connaissances à partir des données, en particulier celles issues de la littérature scientifique.

Le second front de recherche touche lui à de nombreux domaines et constitue un enjeu central dans l'acquisition de nouvelles connaissances. L'exemple emblématique est celui de l'imagerie 3D ou 4D, qui constitue clairement à ce jour l'un des outils clés pour améliorer la compréhension de l'organisation des mécanismes cellulaires. Des enjeux aussi importants se retrouvent au niveau de l'acquisition de données à l'échelle des paysages pour citer une autre facette du problème. Tout cela introduit des besoins très importants de *modélisation à façon* par ces nouveaux moyens d'acquisitions, qui conduisent à examiner et développer des modèles entremêlant aspects biologiques, physiques et biophysiques.



étude microscopique de neurones

CEA

Le troisième front est en quelque sorte le fruit des deux précédents et constitue sans doute l'un des enjeux les plus difficiles à relever car il nécessite de combiner des objets/méthodes mathématiques différents et de collaborer activement entre plusieurs disciplines comme l'intelligence artificielle, l'informatique, les statistiques, l'automatique, etc.

Bien au-delà de la gestion, du stockage des données (qui pose bien sûr des problèmes d'ingénierie importants) et de l'accès à des ressources informatiques (grilles de calcul et de données), il faut souligner que l'ensemble de ces fronts de recherche, et spécifiquement le dernier, renouvelle profondément la pratique actuelle des différentes disciplines, en leur adressant non seulement un nombre important de nouvelles questions (conséquence des spécificités des systèmes biologiques), mais aussi en les obligeant à collaborer davantage ensemble pour traiter les différentes facettes des problèmes de façon synergique et non concurrente.

Quelques exemples pris dans les très nombreux domaines où les mathématiques développent des concepts et des méthodes pertinents pour les enjeux de la biologie.

- Meta analyse

Dans de nombreuses situations, en écologie en particulier, les chercheurs accumulent des données expérimentales diverses et hétérogènes, à des degrés de résolution variés. Il s'agit de les organiser, de les analyser, de les modéliser pour en extraire des relations signifiantes et structurantes. Pour exploiter pleinement de tels modèles, il y a souvent besoin de les mettre en relation avec des données ou méta données à une échelle plus large dans l'espace ou dans le temps : données ou modèles météo, cartographies diverses et variées, etc. Il faut recueillir ces données et les adapter aux objectifs. C'est à ce niveau que se situe l'intérêt potentiel des générateurs de climat ou de paysages. D'autre part, il est souvent envisageable d'utiliser des données issues de multiples et diverses expérimentations déjà effectuées sur des problématiques scientifiques voisines, pour analyser une nouvelle problématique scientifique. Il s'agit alors de faire une *Meta analyse* de ces informations, avec des difficultés liées à leur potentielle grande hétérogénéité.

- Méta modélisation et calculs intensifs

Les différentes disciplines des mathématiques et de l'informatique envisagent la problématique de l'analyse de *grands modèles* avec des méthodologies variées : apprentissage de structures, émer-

gence de propriétés à différentes échelles, agrégation de variables, réduction de modèles par méthodes de régression, à l'aide de systèmes intégro-différentiels, etc. Ces différentes méthodologies visent à se donner une représentation manipulable d'un *grand modèle* de façon à pouvoir d'une part en analyser le comportement et d'autre part rendre les calculs praticables pour comparer des scénarios, optimiser des variables de décision, ou prédire en tenant compte des incertitudes des entrées des modèles. Cet aspect *calculatoire* est à rapprocher des problématiques numériques et calculatoires pour des simulations intensives, posant aussi les enjeux de parallélisation des calculs.

- Méthodes d'inférence des systèmes dynamiques

Les nouvelles technologies et moyens d'acquisition de données conduisent de plus en plus à observer les phases transitoires des systèmes biologiques/agro-écologiques. Cela va de l'observation à toutes les échelles de l'adaptation d'une population de cellules bactériennes à une modification de son environnement, en passant par les problèmes du développement de l'embryon à celle des plantes. Dans ce contexte, les questions de l'exploitation systématique des données posent des problèmes qui sont à la frontière entre les statistiques (fonctionnelles) et l'automatique (identification/estimation de paramètres). Au delà des aspects touchant à l'inférence de modèle, il s'agit bien ici de développer des outils permettant d'intégrer les aspects temporels dans l'analyse des données.

- Méthodes de simulation stochastique pour l'inférence des paramètres des modèles

Les modélisations statistiques réunissant des données multiples à des échelles et à des précisions variées, utilisent de plus en plus des modèles avec un très grand nombre de paramètres où le problème à résoudre est en grande partie algorithmique. Les méthodes ABC (pour Approximate Bayesian Computing) apparaissent comme extrêmement prometteuses pour l'inférence et les tests dans le cadre des modèles pour lesquels la vraisemblance n'est pas accessible. Elles sont apparues dans certains domaines (génétique des populations) il y a quelques années, et commencent à investir d'autres domaines (épidémiologie par exemple). Les méthodes particulaires constituent une autre alternative pour l'estimation dans le cadre des modèles multi-échelles. Ces deux classes de méthodes reposent sur la mise en oeuvre de simulations intensives.

- Optimisation de modèles complexes

La conception de stratégies de gestion innovantes pour les agroécosystèmes implique la résolution de problèmes d'optimisation multicritères dans l'incertain sur la base de modèles complexes. Il s'agit là d'un domaine méthodologique en pleine expansion, fédérant les approches informatiques et mathématiques.

- Validation des modèles

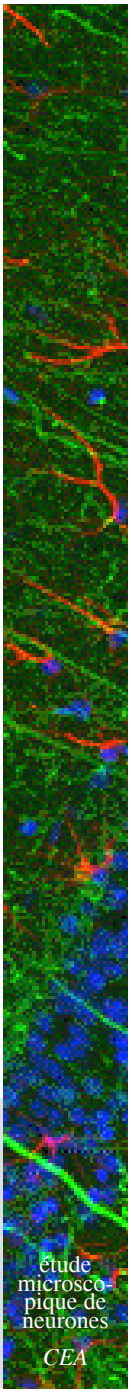
La validation pose à la fois un enjeu théorique et un enjeu pratique car il

nécessite de disposer de jeux de données sur lesquels on confronte la prédiction aux données réelles. Il est courant en statistique de réaliser des exercices de validation, soit sur des données synthétiques (données simulées selon un certain modèle), soit sur une partie du jeu de données écartée de la phase d'estimation. Il s'agit là de validation interne, dans le sens où on reste à l'intérieur d'un problème donné. Plus difficile est la validation externe, lorsqu'on cherche à utiliser un modèle dans un contexte différent de celui pour lequel il a été construit. Notons que c'est malheureusement dans ce cadre là que seront utilisés bon nombre de modèles qui servent en effet à l'aide à la décision, bien que chaque situation ait ses singularités propres. La question de la validation rejoint donc la question de la robustesse du modèle et donc la question de la connaissance du comportement qualitatif des modèles. Remarquons également que très souvent ce sont les modèles *simples* ayant peu de paramètres qui ont les meilleures prédictions en extrapolation. Dans ce sens la modélisation dans un objectif de prédiction peut se révéler différente en pratique d'une modélisation fine cherchant à représenter fidèlement l'ensemble des processus impliqués.

Conclusion.

Face aux enjeux de la biologie, il apparaît clairement qu'il faut pouvoir mobiliser des méthodologies et des compétences multiples. Il est donc particulièrement important que les mathématiciens impliqués dans ces problématiques aient une culture mathématique plus large et que les diverses communautés scientifiques en mathématiques et en informatique communiquent entre elles. Il est essentiel aussi que les biologistes aient un bagage suffisant pour communiquer avec les mathématiciens car de nombreuses avancées ne se feront qu'à l'interface entre les champs disciplinaires.

B. G.



étude microscopique de neurones

CEA

Pour en savoir (un peu) plus

http://images.math.cnrs.fr/Des-mathematiques-dans-nos.html

J.D. Murray : *Mathematical Biology.* Biomathematics Texts. Volume 19. Springer
G. Israel : *La mathématisation du réel. Essai sur la modélisaton du réel*. Seuil.