Avant-propos

« Biologie systémique — Standards et modèles » : les titre et sous-titre de cet ouvrage sont tous deux indispensables pour en qualifier le sujet, l'irruption des standards dans la recherche biologique accompagnant le développement de la biologie systémique.

La biologie systémique est une science nouvelle qui a pour objet de décrire et prédire le fonctionnement des systèmes vivants à partir de la connaissance de leurs composants et des relations entre ces composants en confrontant expériences et modélisations. De ce fait, la biologie systémique est une science interdisciplinaire qui utilise un grand nombre de concepts, de méthodologies et de techniques qui fondent aussi bien les sciences du vivant (avec les disciplines classiques telles que l'anatomie, la physiologie, la biochimie, la biophysique, mais aussi les approches plus récentes : génomique, transcriptomique, protéomique, métabolomique, etc.) que les mathématiques et l'informatique.

Ce sont les développements de la post-génomique qui ont permis de décrire et d'analyser les phénomènes biologiques à une échelle inégalée précédemment. Le concept de post-génomique a été forgé en 1997 par Diane Gershon de la revue scientifique *Nature* pour désigner les développements nécessaires à l'exploration du vivant au-delà du grand projet de séquençage des génomes. Aujourd'hui, on dispose d'outils de haute technologie pour identifier et quantifier non seulement tous les gènes d'un génome, mais encore tous les ARN et les protéines d'une cellule, tous les processus métaboliques d'un tissu, tous les phénotypes d'un organisme, etc. Toutefois, les grandes masses de données produites par ces approches présentent différents niveaux d'hétérogénéité, tant du point de vue de leur structure que de leur signification, et il convient de représenter les objets, les concepts biologiques et leurs relations sous des formats suffisamment précis pour pouvoir les échanger, les comparer, les analyser, les combiner de façons nouvelles et cohérentes en excluant toute ambiguïté d'interprétation.

C'est pour répondre à ces enjeux que des initiatives spontanées ont vu le jour afin de définir les briques d'un langage commun basé sur l'élaboration de standards. Ainsi, dès 1996, les centres de séquençages ont établi, lors de la convention des Bermudes, un standard pour la publication et l'utilisation des séquences génomiques qui a été extrêmement bénéfique à la communauté. En 1999, des biologistes, informaticiens et industriels se sont réunis au sein de groupes de travail pour produire et coordonner des standards dans le domaine du transcriptome. Depuis, l'élaboration de standards par des spécialistes réunis selon leurs domaines d'activité est devenue tellement importante que la création d'un organisme de standardisation pour la biologie systémique est envisagée par différentes instances internationales.

Toutefois, une certaine méconnaissance quant à la nature et à l'élaboration des standards suscite parfois un scepticisme au sein de la communauté scientifique, et certains considèrent les standards comme une tentative pour réduire leur autonomie. Inversement, pour beaucoup d'autres, ils apparaissent comme les instruments d'un consensus qui favorisent l'échange et la communication. En effet, les standards de domaine expriment une représentation du domaine partagée par la communauté correspondante et ne sont en aucun cas la transcription d'une réalité figée et définitive. Par nature, les standards sont des modèles conceptuels au sens de l'organisation des connaissances et partagent le caractère « construit » de la connaissance. Ainsi, lorsqu'on décide de partitionner la représentation du transcriptome en 17 groupes de

concepts (voir chapitre 1), c'est un choix de représentation qui est partagé par les utilisateurs ; un tout autre choix aurait été tout aussi légitime, le critère final étant son adoption ou son… obsolescence !

Si l'importance des standards pour le développement économique est acquise, on en ignore les conséquences au niveau scientifique : les standards engagent sur des types de représentations, en définissant, par exemple, les mots (ontologies, voir chapitres 4 à 6) ou les expressions abstraites (formalismes mathématiques, voir chapitre 9) pour les décrire. Face à cette évolution, la communauté scientifique doit mobiliser les compétences nécessaires pour contribuer de façon significative à l'effort international et s'engager de plus en plus dans les choix et l'élaboration de ces normes. C'est l'initiative encouragée par la revue scientifique *Nature* qui ouvre son site à la communauté internationale pour y débattre lors de la proposition d'un nouveau standard ou de modifications de standards existants.

C'est dans ce contexte que nous avons créé en novembre 2004, sous l'impulsion du département des sciences de la vie du CNRS, le groupe de travail « Métamodèle et langage de modélisation » au sein du pôle de biologie intégrative du domaine « Biotechnologies » de l'association Écrin. L'objet de ce groupe interdisciplinaire était de réunir des biologistes et des informaticiens pour examiner les enjeux et perspectives de l'intégration des données en biologie systémique et proposer des recommandations ; le choix de l'intitulé du groupe référait à un préalable méthodologique fondé sur une architecture de standards (voir chapitre 12). Les nombreux échanges effectués au sein du groupe ont établi un déficit de connaissances en matière de normes. Pour pallier cette difficulté, la réalisation d'un ouvrage collectif s'est imposée pour rassembler les éléments représentatifs du domaine et ouvrir sur les perspectives les plus prometteuses. Un comité scientifique a été constitué pour identifier, valider et solliciter les experts francophones du domaine. C'est dans ce contexte que s'inscrit cet ouvrage, organisé en quatre parties de trois chapitres chacune.

Les préfaces de MM. Alain Pompidou et Emmanuel Canet, au-delà des simple aspects techniques, ouvrent le débat sur une éthique de la standardisation en biologie, tant du point de vue de la recherche de base que de celui de la recherche finalisée.

Dans leur introduction, Pierre Legrain, directeur de l'Institut de biologie et de technologies de Saclay (CEA, France) et Marc Vidal (Dana Farber Cancer Institute, Boston, États-Unis), précurseurs au niveau international de la biologie systémique, rappellent le rôle général des standards et en détaillent les enjeux dans la recherche en biologie.

Une première partie présente deux grands secteurs de productions de données à haut débit, la transcriptomique et la protéomique, les schémas de représentation ainsi que le langage XML, dans lequel ces données sont transcrites pour être échangées, comparées, stockées. Dans le chapitre 1, Catherine Nguyen, responsable du groupe Inserm « Technologies avancées pour le génome et la clinique », pionnière en France de l'utilisation des puces à ADN avec sa collègue Béatrice Loriod, expose le point de vue du biologiste confronté aux nécessités de la standardisation dans un domaine technologique qui comporte bien des difficultés et des écueils. Dans le chapitre 2, Yves Vandenbrouck, directeur du Laboratoire biologie, informatique, mathématiques (LBIM, CEA) en étroite collaboration avec Henning Hermjakob, contributeurs

actifs dans le développement de standards en protéomique, détaille les schémas et explicite le contenu des standards du domaine. Dans le chapitre 3, les langages de description des données XML et OWL, développés par le W3C au sein duquel Jérôme Chailloux assure les fonctions de directeur général de l'ERCIM, sont décrits par Éric Leclercq, Marinette Savonnet, Jean-Claude Simon, et Marie Beurton-Aimar, enseignants-chercheurs en informatique (laboratoires mixtes université-CNRS). Ils en expliquent avec pédagogie les caractéristiques et présentent les principes d'une adaptation à la biologie.

La deuxième partie introduit le domaine des ontologies, dont le développement du web sémantique a favorisé l'expansion. Dans le chapitre 4, Pierre Grenon, chercheur au centre IFOMIS (Institute for Formal Ontology and Medical Information science, Saarland University, Allemagne) et partenaire de la Fonderie OBO (ontologies biomédicales en format ouvert), introduit les différents sens et caractéristiques que peut couvrir une ontologie suivant le domaine de conception, en s'attachant plus particulièrement à la notion d'ontologie formelle indépendante de tout domaine d'application. Dans le chapitre 5, Natalia Grabar et Patrick Ruch, chercheurs des Hôpitaux universitaires de Genève et de l'Inserm, présentent la structure et des exemples d'utilisation de l'ontologie GO, la plus largement utilisée en biologie. Ils complètent cette présentation par une analyse des forces et des faiblesses et les perspectives de développement. Dans le chapitre 6, le professeur Marie-Paule Lefranc, fondatrice et directrice mondialement reconnue de la base de données IMGT en immunogénétique humaine, présente, avec Véronique Giudicelli, un cadre général constitué de sept axiomes pour le développement de tout type d'ontologie en biologie. Sa mise en œuvre dans le cadre de la base de données IMGT est détaillée.

La troisième partie porte sur les méthodes développées pour produire différents types de représentation (abstraction) et d'annotation (spécification) de processus biologiques. Celles-ci s'appuient sur des langages informatiques spécifiques du domaine (SBML), des systèmes de notation graphique (SBGN) ou des formalismes mathématiques. Dans le chapitre 7, Marie Beurton-Aimar (Laboratoire de physiologie mitochondriale, université de Bordeaux) rappelle ce qu'est un modèle dans le contexte des réseaux biochimiques et présente le langage spécifique de domaine, SBML, basé sur le format XML. Pour illustrer cette présentation, l'auteur choisit le cycle de Krebs, dont certaines réactions sont décrites en SBML. Dans le chapitre 8, Franck Molina, directeur de l'unité « Modélisation et ingénierie des systèmes complexes pour le diagnostic » (CNRS-université de Montpellier) et sa collègue Sabine Peres traitent de la représentation graphique des réseaux moléculaires à l'aide du standard SBGN. Celui-ci propose des constructeurs graphiques qui permettent une description intuitive et non ambiguë des phénomènes biologiques. Le cas du cycle de Krebs est à nouveau exploité pour étayer cette présentation. Dans le chapitre 9, François Képès, codirecteur du programme « Épigénomique » (Évry) traite des différents formalismes (langages) mathématiques nécessaires pour capturer les multiples propriétés des systèmes biologiques. Ceci le conduit à constater l'expressivité limitée bien que complémentaire de ces langages, tout en ouvrant sur des perspectives mettant en œuvre le couplage de plusieurs formalismes.

La quatrième partie regroupe différentes approches intégratives des phénomènes biologiques, de la molécule à l'organisme. La phase d'intégration est le but ultime de la biologie sys-

témique, elle assure l'assemblage des données et la production de nouvelles connaissances. Les bases méthodologiques de cette intégration sont en cours d'élaboration et mettent en œuvre des approches expérimentales et théoriques (chapitre 10 et 11), et conceptuelles (chapitre 12). Dans le chapitre 10, S. Randall Thomas (Laboratoire informatique, biologie intégrative et systèmes complexes, CNRS-université d'Évry) présente le projet international Physiome. L'un des buts principaux de cette infrastructure ouverte et collaborative est de développer des bases de données et des modèles mathématiques sur la base de standards et d'ontologies. Parmi les différentes contributions à Physiome, S. Randall Thomas décrit plus particulièrement la contribution française qui vise le développement d'un environnement de modélisation multiorgane, multiéchelle, multiformalisme des fonctions rénales, cardiaques et respiratoires. Dans le chapitre 11, Alan Garny et Denis Noble (Oxford Cardiac Electrophysiology Group, Oxford, UK) décrivent les différents concepts fondateurs du standard CellML à travers l'exemple du premier modèle mathématique d'une cellule cardiaque. Après une rapide présentation des outils et techniques mettant en œuvre CellML, les auteurs montrent l'importance de leur utilisation pour différentes applications pharmaceutiques en modélisant les effets de différents médicaments sur l'électrophysiologie du cœur. Dans le chapitre 12, Marie-Noëlle Terrasse (Laboratoire électronique, informatique, image, CNRS-université de Bourgogne) et Magali Roux (Laboratoire d'informatique de Paris 6, CNRS-université de Paris 6) s'appuient sur l'utilisation des standards pour préconiser des architectures de modèles qui permettent de garantir la validité des données intégrées, de vérifier les similitudes et différences entre modèles, etc., et ainsi de proposer un véritable cadre théorique à l'intégration des données inspiré de l'ingénierie dirigée par les modèles.

Magali Roux

(Laboratoire d'informatique de Paris 6)

Françoise Xavier

(Laboratoire mathématiques appliquées aux systèmes,
École Centrale, Paris)