NOUVEAUX ENJEUX D' INGENIERIE DES CONNAISSANCES EN BIOLOGIE

Magali ROUX-ROUQUIE
Directeur de recherche-CNRS
mroux@pasteur.fr

Du Flot à l'Avalanche de Données

Les avancées de la biologie à *haut-débit*[1] rendent désormais possible l'élaboration de cartes d'identités moléculaires d'une cellule ou d'un tissu biologique, par le dénombrement de ses composants biochimiques (gènes, protéines, *etc.*). Ces approches largement adoptées par l'industrie des biotechnologies se heurtent à des problèmes de gestion et de traitement des données, conséquence d'un accroissement exponentiel des quantités d'informations produites.

La figure ci-dessous (figure 1) illustre une des approches mises en œuvre pour catégoriser des tumeurs du poumon (disposition en colonnes) en fonction des niveaux d'expression de deux séries (A et B) de gènes (disposition en lignes) : le groupe 1 illustre la carte moléculaire d'un échantillon de tissus normaux, alors que le groupe 2 caractérise une variété de tumeurs. Ces résultats, associés à des données cliniques (ages des patients, sexe, antécédents familiaux, traitements, taux de guérison ou de rechute, décès, *etc.*), sont analysés par des méthodes statistiques d'analyse de données (méthodes paramétriques et non-paramétriques, analyses multivariées, ACP, AFC, *etc.*) qui permettent de déduire des scénarios d'évolution et de survie des patients.

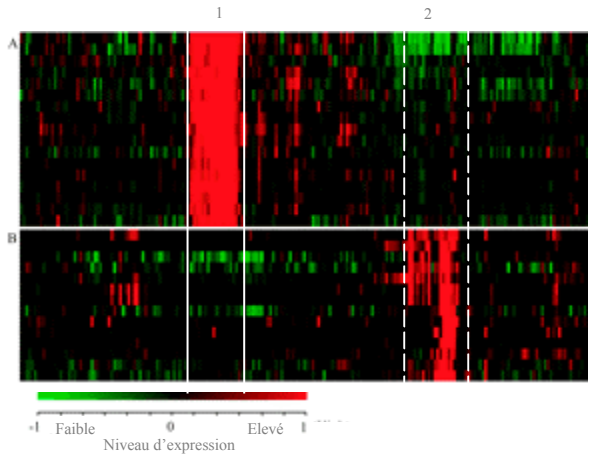

Figure 1. Catégorisation de tumeurs du poumon [adapté de PNAS 98 : 13790-13795 (2001)]

De plus, toutes sortes d'informations interdépendantes concernant la séquence des gènes, la structure de leurs produits, leurs rôles et fonctions cellulaires, paramètres biochimiques, *etc.*, peuvent être associées à ces résultats de sorte qu'un programme de criblage du type de celui présenté ci-dessus génère aujourd'hui plus de 50 millions d'informations distinctes alors qu'il n'en produisait "seulement" 200.000, il y a dix ans[2] !

Ainsi, un des enjeux majeurs de la biologie post-génomique[3] vise l'*intégration sémantique* des données produites par des approches à haut-débit, dans des représentations cohérentes des systèmes vivants. Ce faisant, l'objectif est de raisonner, *in silico*, sur les propriétés de ces systèmes (état, robustesse, *etc.*) comme il est possible de raisonner, *in vitro*, sur les propriétés de leurs composants (par exemple, des réactions biochimiques) ou, *in vivo*, sur des propriétés phénotypiques (apparition ou perte d'un caractère génétique par croisement d'espèces et sélection).

[1] On désigne par biologie à haut-débit, les études conduites à partir de stratégie globale et portant sur l'ensemble des gènes (génomique), des transcripts (transcriptome), des protéines (protéomes) d'un organisme.
[2] Knowledge Management. MindBranch Market Analysis (2001) AdvanceTech Monitor, Ed.
[3] La biologie post-génomique marque la transition d'une approche analytique (décomposition en éléments simples) à une approche intégrative (organisation de ces composants en ensembles fonctionnels).

Méta-modèle(s) d'Intégration Sémantique des Données Biologiques.

Un méta-modèle est un modèle de modèles utilisé pour décrire les données et servir de cadre à leur intégration. Dans le cas d'un domaine correctement spécifié, il est alors possible de prévoir le comportement du système par un arbre de décision conduisant à une solution optimale en faisant, notamment, des choix dans un domaine multi-attributs. Au contraire, dans des domaines complexes et/ou mal spécifiés comme c'est généralement le cas en biologie du fait de connaissances incomplètes, il sera nécessaire d'expliciter des modèles alternatifs. Les solutions retenues pour palier ces difficultés passent notamment par un couplage des modèles, tant au niveau de leur conception que de leur simulation, à un modèle de haut niveau (méta-modèle).

Cette stratégie générale est mise en œuvre par nos partenaires de la société SHARING KNOWLEDGE pour le développement d'une plateforme de partage des connaissances ; elle a inspirée l'adaptation suivante à la biologie post-génomique (figure 2).

En effet, les problèmes soulevés par la diversité et l'interdépendance des données biologiques augure du développement de plateformes informatiques du type de celles mises en œuvre dans les systèmes d'aide à la décision pour structurer de grandes quantités de données et transformer information en connaissance.

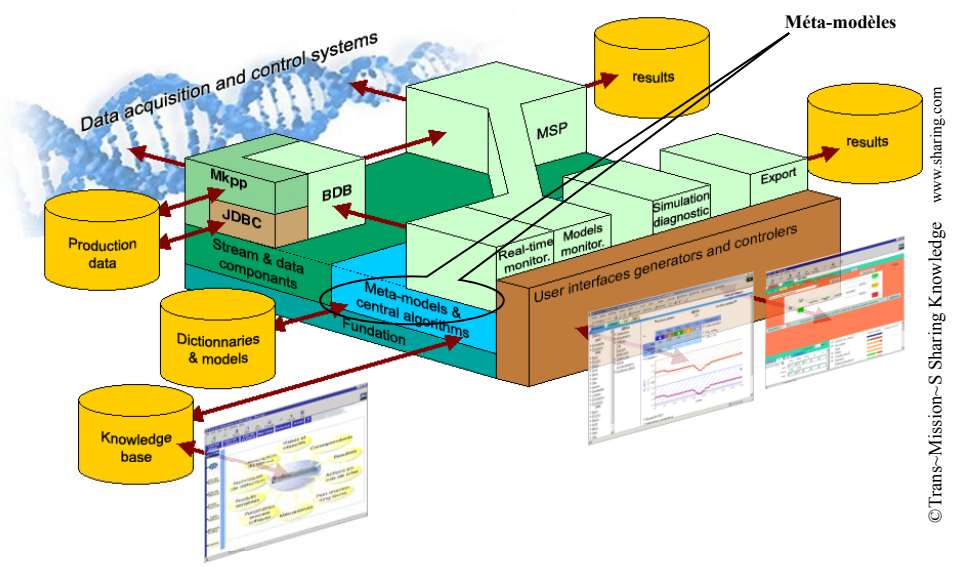

Figure 2. Plateforme d'acquisition, modélisation et simulation des données en Biologie Systémique.

Tout l'enjeu consiste donc à établir un méta-modèle explicatif de la diversité et de l'interdépendance des données. C'est le défi que la *Biologie Systémique* entend relever en s'inspirant des principes opérationnels de la *modélisation systémique* introduits par Herbert A. Simon (1990, 1996, p.115-128; 1977, p.214-5, voir aussi : J.-L. Le Moigne, 1977-1994, p. 153-167; 1999, p.46-47) qui décrit un système comme un ensemble de processus et couple la description des états à la description des processus *(…mapping the state description to the process description).*

Méta-modèle Systémique et Profil UML pour la Spécification des Systèmes Biologiques

Notre équipe a développé un méta-modèle systémique fondé sur une distinction générique entre les entités actives, fonctionnelles qui composent les systèmes biologiques et qui se transforment au cours du temps (tel organe, telle population cellulaire, telle molécule) et les données persistantes sur ces entités.

La nécessité de cette dualité entre entités fonctionnelles (forcément dynamiques) et données persistantes (forcément statiques) s'impose lorsqu'on considère l'activité de *tel* individu, à *tel* instant et à *tel* endroit (par exemple, son activité professionnelle) dont la description s'établira sur des données contextualisées, différentes et complémentaires de celles qui l'identifient (par exemple, son numéro de sécurité sociale).

La référence de ce(s) méta-modèle(s) au paradigme systémique s'établit sur une spécification des entités fonctionnelles par leurs occurrences spatio-temporelles et morphogénétiques. Cette notion d'occurrence morphogénétique est centrale au paradigme systémique puisqu'elle explique la transformation des entités (notamment, biologiques) par leur fonctionnement. La figure 3 donne une représentation objet des entités

fonctionnelles (a) et de leur occurrence morphogénétique (b), elle-même spécifiée par la composante persistante de ces entités (BioComponent) et les tranformations qu'elles réalisent au cours du temps (Biotransformation). Des ontologies sont associées à ces occurrences pour en détailler les spécificités.

Une autre référence majeure au paradigme systémique se fonde sur les notions d'environnement interne et externe. Nos méta-modèles systémiques satisfont à cette condition en décrivant une unité biologique fonctionnelle (par exemple, un complexe moléculaire) (i) incluse dans l'unité fonctionnelle dans laquelle elle fonctionne (supraFunctionalUnit ; par exemple, une cellule), (ii) en interaction avec d'autres unités fonctionnelles de son environnement local (neighborFunctionalUnit ; par exemple, d'autres complexes moléculaires), et (iii) composée d'unités fonctionnelles (infraFunctionalUnit ; par exemple, les molécules composant un complexe moléculaire).

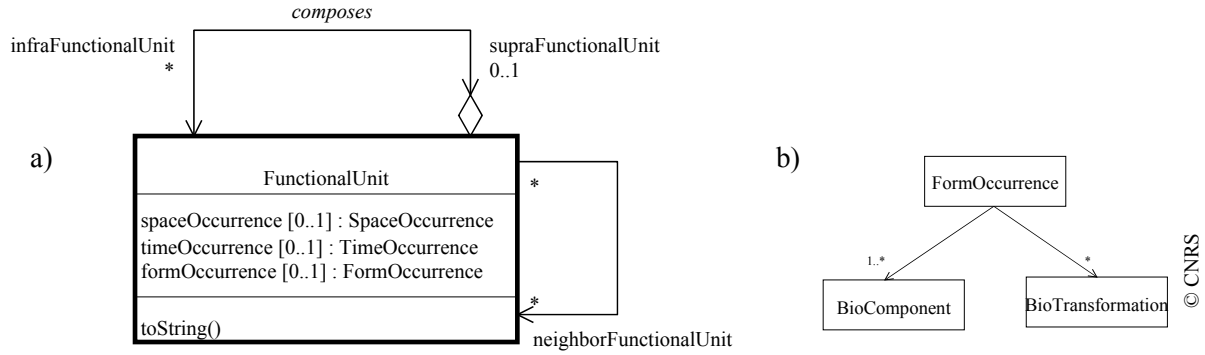

Figure 3. Modélisation orientée objet des processus biologiques à l'aide d'unités fonctionnelles (FUn)

Des langages de modélisation sont développés dans cette double perspective de spécifier « structure et comportement », dont le plus répandu est le langage de modélisation unifié UML (Unified Modeling Language), standard de modélisation orientée objet pour la spécification, la modélisation et l'implémentation de tous types de systèmes. La version UML 2.0 en cours d'adoption par l'OMG[4] a développé la notion d'*objets actifs*, c'est-à-dire d'objets dotés de comportements autonomes, pour spécifier des processus concurrents.

Une problématique clé en biologie concerne la correspondance entre les faits (données) et l'expressivité des langages de spécifications, c'est pourquoi l'utilisation d'un modèle de haut niveau doit permettre de réduire ces différences. Toutefois, l'expressivité d'UML 2.0 étant adaptée à la description systémique des processus biologiques, nous avons récemment proposé à l'OMG, un projet de « *profil* » UML pour les systèmes biologiques, c'est-à-dire une extension du langage adaptée aux spécificités des systèmes biologiques (SB-UML).

Un des objectifs poursuivi vise à réaliser la transformation des spécifications semi-formelles établies dans SB-UML vers des langages formels permettant leur traitement mathématique.

Perspectives: Ingénierie Transdisciplinaire

Ce rapide cheminement, de la biologie moléculaire *classique* qui identifie individuellement les structures moléculaires, aux développement de la biologie à haut-débit qui préjuge des processus dans lesquels ces structures sont impliquées, conduit au projet de spécification formelle des processus biologiques et révèle l'effort transdisciplinaire à accomplir.

Pour progresser dans la compréhension des systèmes biologiques, de leur structure et de leur comportement, la pratique de la biologie évolue, en incluant aux approches technologiques des phénomènes biologiques, des approches formelles inspirées non seulement des mathématiques mais aussi des sciences informatiques, associées à des approches opérationnelles de management des connaissances. Rapidement un langage de communication devrait s'avérer le liant nécessaire aux multiples apports disciplinaires pour dégager de nouveaux paradigmes. C'est l'évolution que notre équipe s'efforce d'anticiper en développant le profil SB-UML.

[4] Object Management Group, consortium international comprenant plus de 600 entreprises et chargé de la standardisation des spécifications logicielles (http://www.omg.org).