Une ontologie pour organiser les données de processus biologiques: la contribution des modèles mathématiques

O. Inizan[1], V. Fromion[1], A. Goelzer[1], F. Saïs[2], D. Symeonidou[3]

[1] Université Paris Saclay, INRAE, MaIAGE
[2] Université Paris Saclay, LISN, CNRS UMR9015
[3] Université de Montpellier, INRAE, SupAgro, UMR MISTEA

olivier.inizan@inrae.fr

Résumé

La biologie est un domaine connu pour sa production massive de données. Ces données sont souvent qualifiées d'hétérogènes et de fragmentées, et les biologistes ne disposent pas d'une représentation formelle, qui, à l'échelle de l'organisme, permettrait de les représenter et de les organiser. Depuis une dizaine d'années les modèles mathématiques systémiques se sont révélés être des outils utiles pour comprendre le comportement de la cellule. Nous montrons dans ce travail qu'une ontologie construite sur les principes qui régissent la conception de ces modèles peut aider à organiser les données biologiques. Nous présentons ici un choix de concepts et relations compatibles avec les principes à l'oeuvre dans les modèles systémiques.

Mots-clés

Ontologies, modèles mathématiques, données biologiques.

Abstract

Biology is a research field well known for its huge quantity and diversity of data. These data are recognized as heterogeneous and fragmented. Biologists do not have a formal representation that, at the level of the entire organism, can help them to tackle such diversity and quatity. Recently, the systemic mathematical models have proven to be a powerful tool for understanding the bacterial cell behavior. We advocate that an ontology built on the principles that govern the design of such models, can help to organize the biological data. In this article we present a choice of concepts and relations compliant with principles at work in the systemic mathematical models.

Keywords

Ontology, Mathematical Models, Biological Data.

1 Introduction

En biologie, l'avancée récente des technologies de séquençage a permis une production rapide et peu onéreuse de données [15]. Aujourd'hui, les biologistes et bioinformaticiens manipulent une grande quantité et une grande diversité de données dites omiques (la génomique, la transcriptomique, la protéomique, la métabolomique et la métagénomique) [9]. Ces données sont principalement obtenues dans le contexte d'expérimentations conçues pour répondre à des questions précises. D'un point de vue plus général, les données produites sont hétérogènes et fragmentées [2]. Ainsi, malgré l'abondance de données disponibles pour un organisme particulier, la capacité de lier ces données entre elles représente un défi majeur [11]. Une telle démarche présente de nombreux intérêts comme celui d'élucider des mécanismes métaboliques en vue de traitements thérapeutiques [14]. Bien que la recherche en représentation des connaissances soit très active en biologie [13], il n'existe pas de représentation formelle destinée à organiser les données pour l'intégralité d'un organisme, aux échelles moléculaires. Une telle représentation permettrait de mieux exploiter tout le potentiel des données produites par les expériences. Depuis une dizaine d'années l'approche de modélisation "*cellule entière*" a montré que des modèles mathématiques systémiques représentent un outil important pour comprendre et décrire le comportement de la cellule bactérienne. Plus précisément, lorsque ces modèles sont calibrés à l'aide de données biologiques, il est possible d'identifier des principes organisateurs conduisant à la prédiction de comportements qui n'avaient pas été observés expérimentalement [10, 3]. Il existe donc un réel besoin de développer une nouvelle représentation formelle (i) qui a pour objectif de représenter de manière sémantique les liens entre données biologiques et (ii) qui s'inspire des principes à l'oeuvre dans la modélisation de processus biologiques.

Les travaux que nous présentons dans cet article ont été déjà publiés dans [8]. Dans ce travail nous décrivons les premières étapes du développement d'une représentation formelle destinée à l'organisation de données biologiques et conçue selon les concepts présents dans les modèles mathématiques. Ce travail est en cours de réalisation, les tâches effectuées sont principalement conceptuelles et nous évaluons l'ontologie sur un exemple simple. L'article est organisé comme suit : la section 2 présente l'état de l'art en relation avec ce travail et sa principale motivation. Les concepts et relations de l'ontologie sont présentés en section 3 et illustrés à travers l'exemple de la section 4. La section 5 présente les conclusions et perspectives.

2 État de l'art et motivation

Deux points de départ permettent de comprendre le travail présenté dans cet article : les ontologies BiPOm et BiPON [6, 5] et les contraintes relatives à la construction de modèles mathématiques.

2.1 BiPON et BiPOm

La biologie est un domaine où différentes communautés peuvent travailler sur les mêmes objets mais avec des buts différents. Il est donc crucial d'être en mesure d'éviter les ambiguïtés alors que l'on se réfère au même objet. Les bio-ontologies couramment utilisées, par exemple l'ontologie *Gene Ontology (GO)* [1] décrivent une hiérarchie de concepts utilisés comme vocabulaire contrôlé. D'autres bio-ontologies sont aussi utilisées à des fins d'échange de données : c'est le cas de BioPax ([2]) qui permet de décrire les voies métaboliques. En 2017 et 2020, deux ontologies au format OWL [1], BiPON [6] et BiPOm[5], ont proposé de nouveaux usages pour les bio-ontologies. Elles ont tout d'abord introduit l'approche systémique comme principe pour organiser la connaissance relative aux objets biologiques. Cette approche émane des sciences de l'ingénieur et consiste à découper un système en un ensemble de modules et sous-modules interconnectés [4]. Un module systémique est défini par ses entrées, ses sorties et la fonction qu'il remplit. Cette fonction, avec les entrées et les sorties sont regroupées dans un modèle mathématique. Ainsi, le comportement du module est représenté de façon formelle. Les auteurs de BiPON et BiPOm ont montré que la cellule bactérienne peut être considérée comme un système et organisée en modules et sous modules systémiques. Ces modules sont représentés par des concepts OWL nommés *processus biologiques*. D'autre part, BiPON et BiPOm sont des ontologies plus expressives que les bio-ontologies courantes, puisqu'en plus des concepts, relations et les axiomes OWL, elles contiennent également un ensemble de règles de Horn exprimées en SWRL [2]. Elles exploitent en effet les capacités de raisonnement fournies par la sémantique logique des axiomes OWL, et des règles SWRL déclarés afin d'inférer de nouvelles relations entre les individus. A titre d'exemple, un raisonnement sur l'ontologie BiPOM a montré qu'un vaste ensemble de processus biologiques pouvait être décrit par un ensemble restreint de concepts mathématiques.

2.2 Les contraintes et les modèles mathématiques

La figure 1.a illustre l'association entre un module systémique (ici le processus biologique) et son modèle mathématique. Nous présentons ici les contraintes qui permettent de construire de tels modèles. Afin de mieux comprendre ces contraintes, il faut d'abord détailler un peu plus le concept de processus biologique tel qu'il est défini dans les ontologies BiPON/BiPOm. Un processus biologique a une ou plusieurs molécules comme entrée et une ou plusieurs molécules en sortie. Nous dirons qu'un processus *consomme* ses entrées et *produit* ses sorties. De plus et comme évoqué ci dessus, un processus remplit une fonction. Le processus possède finalement un moyen de transformer les entrées en sorties et ce moyen est exprimé au travers d'un modèle mathématique. La forme générale d'un processus biologique est détaillée dans la figure 1a. La figure 1b présente une simple réaction biochimique (la conversion d'une molécule 'A' en molécule 'B') et le processus biologique correspondant.

Un point important est que quel que soit le modèle mathématique construit, trois contraintes sont toujours respectées. Nous considérons donc que (i) ces contraintes sont majeures et (ii) qu'elles pilotent la construction de modèles mathématiques. Dans la suite nous nommerons ces contraintes les *contraintes des modèles*. Ces trois contraintes sont :

1. *La causalité physique*. En physique la causalité indique que si les entrées d'un modèle produisent les sorties du modèle alors les entrées précèdent les sorties. Dans la représentation que nous construisons nous ne considérons pas le temps et nous reformulons la causalité ainsi : si les entrées sont présentes en quantité suffisante alors le processus peut consommer les entrées et produire les sorties.
2. *La conservation de la masse* est une contrainte importante pour la construction de modèles. Elle assure leur consistance.
3. *La compétition pour l'accès aux ressources*. Dans la cellule les processus biologiques sont en compétition pour l'accès aux ressources. Plus précisément, les mêmes molécules peuvent être consommées par des processus différents. Ainsi, la molécule d'ATP fournit l'énergie nécessaire à la cellule et est, par conséquent, consommée par de nombreuses réactions biochimiques.

Malgré le fait que le concept de processus biologique soit présent dans les ontologies BiPON/BiPOM et que les modèles mathématiques sont représentés dans BiPON, aucune de ces deux ontologies ne considère ces contraintes.

2.3 Motivation

Nous envisageons les contraintes des modèles comme un moyen de valider la consistance de la connaissance biologique et des données associées à un organisme. Si nous souhaitons utiliser ces contraintes dans une représentation formelle, nous devons d'abord fournir des concepts et des relations qui nous permettent de *compter* les molécules produites ou consommées par les processus. Si l'on considère l'exemple de la figure 1b : la causalité indique qu'il faut au moins une molécule A disponible pour produire une molécule B. La compétition pour l'accès aux ressources nécessite aussi de compter les molécules. Imaginons qu'un processus P' consomme également de la molécule A. S'il y a seulement *une seule* molécule A dans toute la cellule, P et P' sont en compétition. Comme évoqué dans la section 2.2 BiPON et BiPOm ont validé l'approche systémique pour

1. https ://www.w3.org/OWL/
2. https ://www.w3.org/Submission/SWRL/

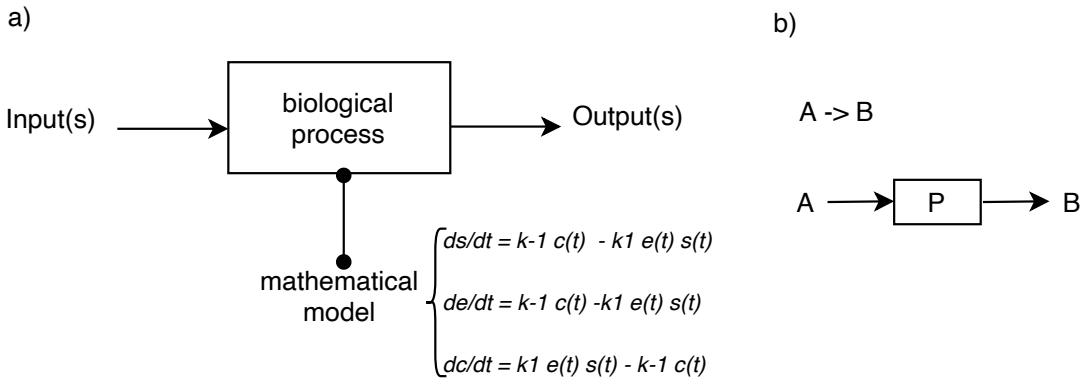

FIGURE 1 – a) Un processus biologique et le modèle mathématique associé. b) Une réaction biochimique et le processus P qui la représente.

représenter la connaissance biologique. Cependant elles ne permettent pas de compter les entités consommées ou produites par les processus et par conséquent de représenter les contraintes des modèles. Nous avons donc entrepris la construction d'une nouvelle ontologie.

3 Eléments d'une bio-ontologie pour l'organisation des données et des connaissances

Nous souhaitons donc construire une représentation formelle en nous basant sur les contraintes qui pilotent la construction de modèles mathématiques. Pour ce faire, nous avons montré qu'il doit être possible de compter les molécules consommées ou produites par les processus. Nous proposons tout d'abord un ensemble de concepts et de relations qui vont nous permettre de compter les molécules (section 3.1). Ces concepts et relations nous permettront ensuite de définir de façon formelle un processus biologique (section 3.2).

3.1 Concepts et relations afin de compter les molécules

Nous reprenons dans ce travail l'approche formelle et le concept de processus biologique (nommé *process*) tel qu'il est défini dans les ontologies BiPON/BiPOm. Afin de prendre en compte les contraintes des modèles (la causalité, la conservation de la masse et la compétition pour les ressources), nous utilisons les concepts fréquemment utilisés par les modélisateurs [16]. Nous créons tout d'abord le concept de *pool* qui permet de regrouper toutes les molécules de la même entité chimique. Par exemple toutes les molécules d'eau seront regroupées dans un pool nommé H2O. Ensuite, le pool ayant un volume fini, le nombre de molécules présentes est donné par la *concentration* du pool. Nous décrétons aussi que les processus communiquent uniquement par les pools. Pour ce faire nous imaginons trois opérations (lecture, consommation et production) : (i) un processus peut lire (*reads*) la *concentration* de molécules d'un *pool* et (ii) un processus peut consommer les molécules d'un *pool* et/ou produire les molécules d'un *pool*. Nous représentons ces opérations de consommation/production avec la relation *triggers* et le concept de flux (*flow*).

La figure 2.b reproduit l'exemple de la figure 1.b où le processus P convertit la molécule A en molécule B. Cette figure peut être détaillée ainsi : le processus (*process*) P lit (*reads*) la *concentration* de molécule du *pool* A (flèche grise pointillée). S'il y a suffisamment de molécule (ici une seule molécule est requise), P consomme cette molécule (on dira que P déclenche (*triggers*) un flux (*flow*) de molécule A (première flèche noire)) et produit un flux de molécule B dans le *pool* B (P déclenche un flux de molécule B, deuxième flèche noire).

3.2 Définition formelle du processus biologique

L'ensemble de concepts et de relations décrits ci-dessus nous permet d'affiner la définition de processus biologique proposé par BiPON/BiPOm. Dans ces ontologies le processus est décrit à travers les relations qu'il entretient avec les molécules qui participent à la réaction : un processus a comme entrée (*has_input*) des molécules et comme sortie (*has_output*) des molécules. Nous proposons de faire évoluer cette description. Cette évolution est expliquée à travers les exemples des figures 1 et 2. Tandis que BiPON/BiPOm décrivent le processus P avec comme entrée la molécule A et comme sortie la molécule B, nous déclarons que le processus P lit la *concentration* de molécules A et (s'il y a suffisamment de molécules) déclenche un flux de molécules A et un flux de molécules B. Cette nouvelle description nous permet d'être plus en adéquation avec la contrainte de causalité : l'information donnée par la lecture de la concentration (i.e. le fait qu'il y ait suffisamment de molécules) est la cause du comportement du processus alors que le flux de molécules est considéré comme son effet. Avec ces considérations nous pouvons fournir une définition du processus biologique. Un processus biologique est

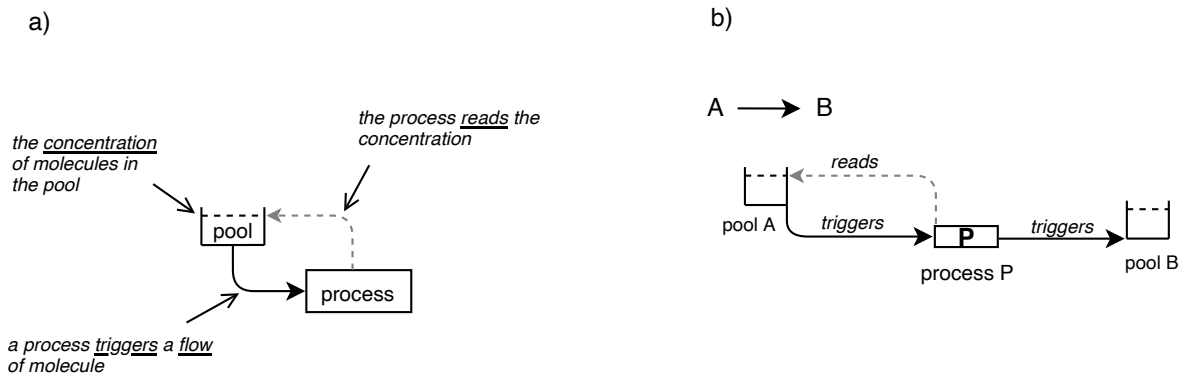

FIGURE 2 – a) Les concepts et relations de la nouvelle ontologie. b) Une réaction biochimique représentée par ces concepts et ces relations.

caractérisé par ses relations avec ses entrées et ses sorties :

$BiologicalProcess \equiv$
$\exists has\_input.Input \sqcap \forall has\_input.Input \sqcap$
$\exists has\_output.Output \sqcap \forall has\_output.Output$

Une entrée est la concentration lue par un processus :

$Input \equiv Concentration \sqcap \exists is\_read\_by.BiologicalProcess$

Une sortie est le flux de molécules déclenché par le processus :

$Output \equiv Flow \sqcap \exists triggered\_by.BiologicalProcess$

Il faut noter que la définition du processus est cyclique : il est défini par ses entrées et ses sorties et elles même sont définies par le processus. Ces situations sont communes dans la mise au point d'ontologies. Les cycles peuvent être résolus lors du peuplement en précisant l'ordre selon lequel les individus sont créés.

4 Exemple

Nous illustrons l'utilisation de l'ontologie avec l'exemple d'une réaction catalysée par une enzyme. Cette classe de réaction est représentative d'une large part des processus métaboliques à l'oeuvre dans la cellule bactérienne. (Il faut aussi noter qu'un tiers des gènes de la bactérie sont impliqués dans la synthèse d'enzymes.) Par conséquent, si le processus de catalyse peut être représenté par les concepts et relations décrits dans la section 3.1, nous aurons accompli une première étape dans le processus d'évaluation de l'ontologie. Le modèle chimique qui décrit la catalyse a été proposé par Michaelis et Menten [12]. Ce modèle comprend 2 réactions :

$$E + S \leftrightarrows [ES] \rightarrow E + P$$

Dans la première réaction l'enzyme *E* se lie au substrat *S* pour former le complexe *ES*. Cette réaction est réversible : le complexe *ES* peut se dissocier pour relâcher l'enzyme *E* et le substrat *S*. La deuxième réaction est irréversible : le complexe *ES* se dissocie pour relâcher l'enzyme *E* et le produit *P*.

Afin de représenter ce modèle chimique avec les concepts et relations proposés ci-dessus nous construisons tout d'abord deux processus P1 et P2, pour la première et la seconde réaction. Pour chaque type de molécule nous construisons quatre pools nommés *S*, *E*, *ES* et *P* qui correspondent respectivement au substrat, à l'enzyme, au complexe et au produit. Les processus, les pools et leurs relations sont présentés sur la figure 3.b. Cette figure peut être lue comme suit : P2 lit la *concentration* du *pool* ES et (si la quantité de molécules ES est suffisante) déclenche un flux de molécules E, P et ES. Le processus P1 représente une réaction réversible. Pour la première réaction élémentaire (E+S $\longrightarrow$ [*ES*]) P1 lit la concentration du *pool* E et S et déclenche un flux de E, S et ES. Pour la seconde réaction élémentaire ([*ES*] $\longrightarrow$ E+S) P1 lit la concentration du *pool* ES et déclenche un flux de ES, E et S. Ainsi, dans l'ontologie les deux réactions élémentaires sont agrégées dans un seul processus. Il faut noter que suite à cette agrégation la causalité est toujours respectée. En effet, les sorties du processus P1 (les flux de ES, S et E) sont bien causées par le niveau de concentration des pools ES, S et E.

5 Conclusion et perspectives

Nous avons décrit dans cet article les premières étapes du développement d'une ontologie dédiée à l'organisation des données biologiques. Cette ontologie a été construite en fonction des contraintes qui régissent la construction des modèles mathématiques. L'ensemble de concepts et relations qui en résulte (i) rend possible la représentation des quantités, (ii) a été validé sur un exemple représentatif et (iii) nous a conduits à donner une nouvelle définition formelle du processus biologie. Nous prévoyons maintenant de peupler l'ontologie avec un réseau complet de réactions [5] au format SBML [7]. Lors de cette opération nous pourrons associer des quantités aux concentrations et aux flux. De plus, nous évaluerons la capacité du langage SHACL à exprimer les contraintes des modèles. Ce travail s'inscrit dans un contexte où il s'agit de rendre les ontologies plus expressives avec l'idée de représenter de la connaissance quantitative. Nous pourrions ainsi vérifier la consistance et

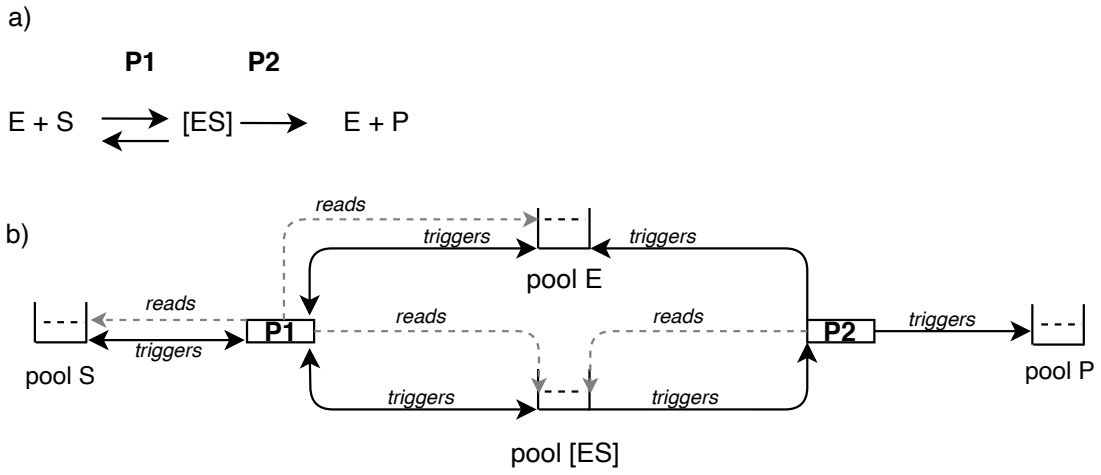

FIGURE 3 – a) Le modèle de catalyse de Michaelis et Menten b) La représentation de ce modèle avec les processus, les pools et les relations.

la validité des données biologiques organisées par cette ontologie.

Références

[1] Gene Ontology Consortium. The gene ontology resource : 20 years and still GOing strong. *Nucleic acids research*, 47(D1) :D330–D338, 2019.

[2] Emek Demir, Michael P Cary, Suzanne Paley, et al. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9) :935–942, 2010.

[3] Anne Goelzer, Jan Muntel, Victor Chubukov, et al. Quantitative prediction of genome-wide resource allocation in bacteria. *Metabolic engineering*, 32 :232–243, 2015.

[4] Leland H Hartwell, John J Hopfield, Stanislas Leibler, et al. From molecular to modular cell biology. *Nature*, 402(6761) :C47–C52, 1999.

[5] Vincent Henry, Fatiha Saïs, Olivier Inizan, et al. BiPOm : a rule-based ontology to represent and infer molecule knowledge from a biological process-centered viewpoint. *BMC bioinformatics*, 21(1) :1–18, 2020.

[6] Vincent J Henry, Anne Goelzer, Arnaud Ferré, et al. The bacterial interlocked process ONtology (BiPON) : a systemic multi-scale unified representation of biological processes in prokaryotes. *Journal of biomedical semantics*, 8(1) :1–16, 2017.

[7] Michael Hucka, Andrew Finney, Herbert M Sauro, et al. The systems biology markup language (SBML) : a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4) :524–531, 2003.

[8] Olivier Inizan, Vincent Fromion, Anne Goelzer, Fatiha Saïs, and Danai Symeonidou. An ontology to structure biological data : thecontribution of mathematical models. In *Metadata and Semantic Research - 15th International Conference, MTSR 2021*, Communications in Computer and Information Science. Springer, 2021.

[9] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system : integrating'omics' data sets. *Nature reviews Molecular cell biology*, 7(3) :198–210, 2006.

[10] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, et al. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2) :389–401, 2012.

[11] Evangelina López de Maturana, Lola Alonso, Pablo Alarcón, et al. Challenges in the integration of omics and non-omics data. *Genes*, 10(3) :238, 2019.

[12] Leonor Michaelis, Maud L Menten, et al. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369) :352, 1913.

[13] Jacques Nicolas. Artificial intelligence and bioinformatics. *A Guided Tour of Artificial Intelligence Research*, pages 209–264, 2020.

[14] Charlotte Ramon, Mattia G Gollub, and Jörg Stelling. Integrating–omics data into genome-scale metabolic network models : principles and challenges. *Essays in biochemistry*, 62(4) :563–574, 2018.

[15] Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4) :586–597, 2015.

[16] Eberhard O Voit. *Computational analysis of biochemical systems : a practical guide for biochemists and molecular biologists*. Cambridge University Press, 2000.