Représentation systémique multi-échelle des processus biologiques de la bactérie

Vincent J. Henry, Arnaud Ferré, Christine Froidevaux, Anne Goelzer, Vincent V. Fromion, Sarah Cohen-Boulakia, Sandra S. Derozier, Marc Dinh, Ghislain Fiévet, Stephan Fischer, et al.

Représentation systémique multi-échelle des processus biologiques de la bactérie

Vincent Henry[1], Arnaud Ferré[1], Christine Froidevaux[1], Anne Goelzer[2], Vincent Fromion[2], Sarah Cohen-Boulakia[1], Sandra Dérozier[2], Marc Dinh[2], Ghislain Fiévet[1], Stephan Fischer[2], Jean-François Gibrat[2], Valentin Loux[2], Sabine Peres[1,2]

[1] BioInfo, LRI, CNRS UMR 8623, Université Paris Sud, Université Paris-Saclay, France
{prénom.nom}@lri.fr
[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
{prénom.nom}@jouy.inra.fr

Résumé : La production à haut-débit de données biologiques de nature hétérogène nécessite une exploitation et une intégration particulières de celles-ci. Malgré le développement de nombreuses bio-ontologies, l'organisation de ces données dans un cadre structuré et adaptatif reste perfectible. Nous émettons l'hypothèse qu'une approche systémique multi-échelle de la représentation des processus cellulaires permettrait de progresser dans cette problématique. Pour valider cette démarche, nous avons conçu une modélisation ontologique des processus bactériens nécessaires à l'expression génique. Les relations entre ces processus et leurs molécules participantes ou leurs sous-processus ainsi que leurs modèles ont été formellement décrites. Cette description s'accompagne d'axiomes et de relations supplémentaires sur lesquels un raisonnement automatique est effectué. La représentation des processus réalisée permet leur mise en relation avec leurs modèles et paramètres par inférence. Parallèlement, le raisonnement apporte de nouvelles informations contextuelles de séquentialité, agrégation ou compétition. Notre contribution s'appuie sur les bio-ontologies existantes pour une meilleure interopérabilité.

Mots-clés *: Conception d'ontologie, modélisation multi-échelle, biologie systémique, raisonnement, processus cellulaire*.

1 Introduction

Les progrès constants dans les technologies à haut-débit, dites « omiques » (génomique, transcriptomique, métabolomique,…) ont permis de franchir une étape critique dans la production de données biologiques (Metzker, 2010). Cette révolution a des implications sur la masse et l'hétérogénéité des types de ces données à gérer à différentes échelles (moléculaire, cellulaire, phénotypique, *etc* ; Nekrutenko & Taylor, 2012). Ainsi, l'interprétation de la biologie à travers les sciences omiques nécessite la gestion, l'intégration et le partage de données de nature diverse. Dans ce contexte, il s'avère nécessaire de relier ces données aux mécanismes biologiques sous-jacents dans un cadre structuré et adaptatif pour conserver et analyser au mieux l'information.

Cette problématique a été identifiée dans le cadre du projet interdisciplinaire de l'Institut de Modélisation des Systèmes Vivants (IMSV) qui regroupe des biologistes, des modélisateurs et des informaticiens autour d'une question commune relative à l'intégration systémique des organismes. Afin d'y répondre, nous avons avancé l'hypothèse qu'une approche systémique multi-échelle d'un système d'organisation des connaissances (SOC) en biologie permettrait de progresser vers un modèle utile à l'intégration et l'exploitation de données hétérogènes. En effet, en biologie systémique, les données issues d'expériences omiques permettent de caractériser les paramètres de modèles représentant les processus biologiques. Ainsi ces

données, bien qu'hétérogènes, sont organisées intrinsèquement en fonction des processus biologiques et intégrées à travers les modèles mathématiques capables de les simuler et leurs paramètres associés.

L'approche systémique multi-échelle consiste à représenter chaque phénomène à l'échelle la plus pertinente. Cette représentation est rendue possible grâce à l'intégration des connaissances à différentes échelles, dans des briques (ou systèmes) élémentaires qu'il s'agit d'assembler et de manipuler de façon modulaire. L'emboîtement de ces briques offre une vue globale du système et la possibilité de zoomer sur des zones particulières permettant de représenter le comportement du système à différentes granularités. Depuis 2001, il est acquis que les fonctions cellulaires majeures peuvent être représentées par des systèmes (Kitano, 2001). S'appuyant sur ce principe, le comportement général des bactéries a pu être modélisé et simulé à différents niveaux (Goelzer & Fromion, 2011 ; Karr *et al.*, 2012 ; Goelzer *et al.*, 2015). De plus, les processus biologiques se prêtent naturellement à une représentation systémique multi-échelle, qui à son tour peut être organisée sous forme ontologique de manière méthodique (figure 1). Avec *Cell Molecular ONtology* (C'MON), nous nous proposons de tester la faisabilité et la cohérence d'une représentation ontologique des processus en lien avec leurs paramètres à partir de leur définition fonctionnelle biologique.

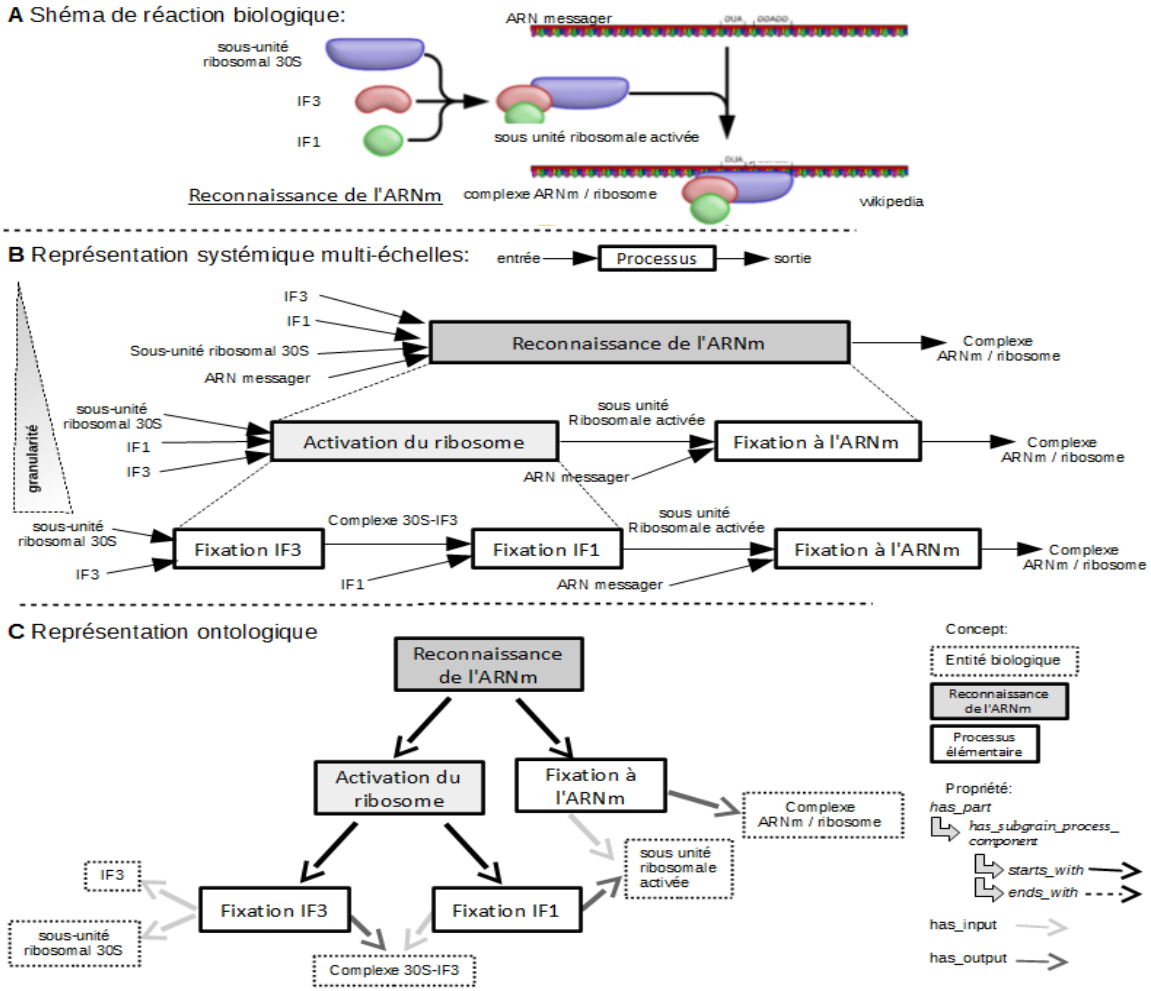

FIGURE 1 – Exemple de représentation biologique schématique, systémique multi-échelle et ontologique correspondant à un même processus cellulaire : la reconnaissance de l'ARN par le ribosome.

Des SOC existent déjà dans le domaine de la Biologie. Des bio-ontologies ont été conçues pour : l'annotation fonctionnelle des gènes et de leurs produits (*Gene Ontology* GO , GO

Consortium, 2000) ; le recensement des molécules d'intérêt biologique et de leur modification chimique en fonction de leur structure (*Chemical Entities of Biological Interest* ChEBI, De Matos *et al.*, 2010) ; la classification des séquences biologiques en fonction de leur implication dans des processus biologiques (*Sequence Ontology* SO, Eilbeck *et al.*, 2005) ; ou la hiérarchisation des principaux modèles de réactions biologiques et des paramètres associés (*Systems Biology Ontology* ; SBO ; Courtot *et al.*, 2011).

De même, des bases de connaissances comme les bases *Kyoto Encyclopedia of Genes and Genome* (KEGG) ou *Metabolic Pathway Database* (MetaCyc, Altman *et al.*, 2013) décrivent finement les réseaux métaboliques : des voies principales aux réactions enzymatiques élémentaires. Ces bases apportent aussi des informations globales et structurées sur les principaux processus cellulaires (transcription, réplication, traduction, …) et les complexes protéiques impliqués.

L'ensemble de ces SOC propose une vision assez large et relativement complète de la biologie moléculaire. Cependant, ces travaux sont développés indépendamment les uns des autres avec des objectifs différents. Néanmoins, le développement de GO-plus a démontré la possibilité d'intégrer au moins deux de ces bio-ontologies standards, GO et ChEBI (Hill *et al.*, 2013). Par ailleurs, leur étude approfondie révèle un déséquilibre d'intérêt en faveur des eucaryotes. Enfin, l'hétérogénéité des processus cellulaires et la nouveauté des connaissances fines dans ces domaines n'ont pas permis une représentation aussi précise que celle des réactions enzymatiques métaboliques.

Si aucun de ces SOC ne répond à lui seul à notre besoin, ils représentent une source de connaissance conséquente et sont assidûment utilisés en biologie. Afin de profiter au maximum de cette base de travail et de son implantation, nous attachons un soin particulier à les intégrer. A titre de preuve de concept de la possibilité de cette intégration avec une approche systémique multi-échelle, nous présentons dans cette article un modèle ontologique des processus cellulaires bactériens nécessaires à l'expression génique, C'MON.

2 Modèle ontologique

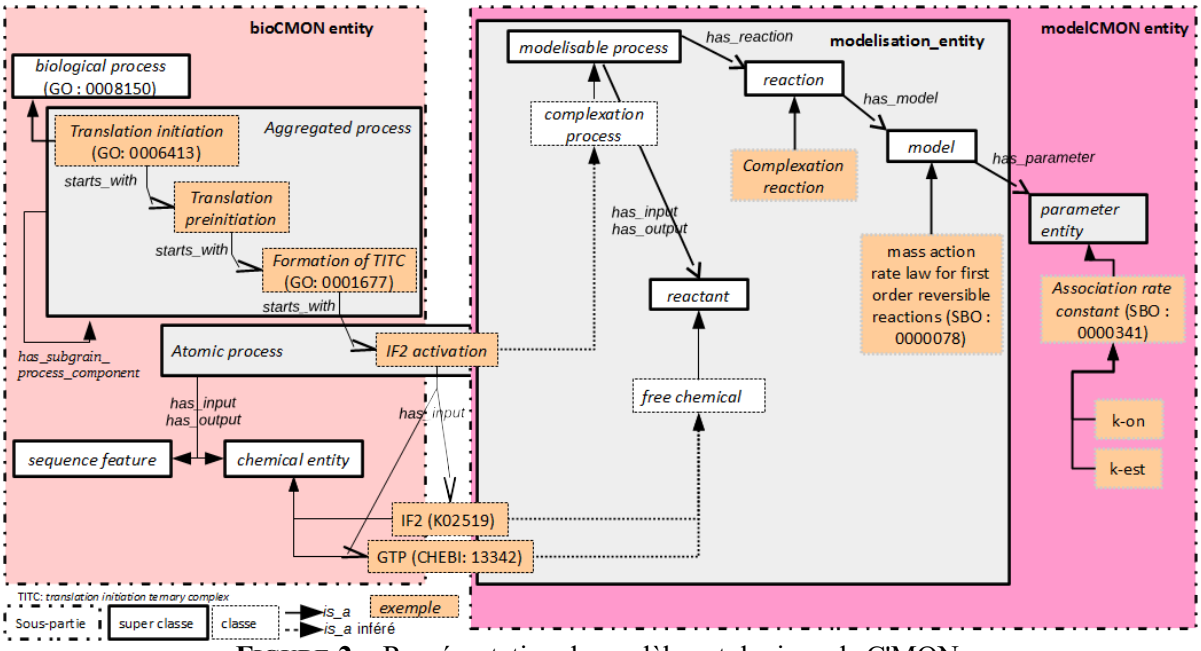

FIGURE 2 – Représentation du modèle ontologique de C'MON.

C'MON est éditée sur Protégé 5.0.0. C'MON contient 2 parties indépendantes (figure 2) : '*bioCMON entity*' (bioCMON) et '*modelCMON entity*' (modelCMON) reliées par une racine commune. bioCMON organise les processus cellulaires atomiques finement définis en fonction

de leurs entités biologiques ou les processus agrégés, hiérarchisés entre eux en sous-propriétés de *has_part.* modelCMON structure les types de paramètres reflétant les données biologiques en fonction de leur modèle. Ceux-ci sont reliés à des processus modélisables définis en fonction de leur type de réactif (reflétant les entités biologiques de bioCMON). Ainsi les processus cellulaires concepts de bioCMON sont automatiquement réorganisés dans modelCMON comme *is_a* des processus modélisables. Ils se retrouvent ainsi reliés aux modèles mathématiques qui les représentent et aux types de paramètres associés.

2.1 bioCMON

bioCMON a été conçue à partir de connaissances biologiques issues de schémas de réaction compilés par les experts biologistes de l'IMSV (figure 1A). Les informations de processus et d'entités biologiques sont extraites en suivant l'approche systémique multi-échelle. Cette étape permet d'obtenir une représentation standardisée, granulaire et centrée sur les processus encadrés par les entités biologiques en entrée et sortie (figure 1B). Cette représentation est ensuite modélisée sous forme ontologique (figure 1C). Un processus de faible granularité, agrégé (en grisé) est formellement défini par ses processus de granularité supérieure par une sous-propriété non transitive de *has_part : has_subgrain_process_component*. Celle-ci est spécifiée par des sous-propriétés *starts_with, ends_with* ou *has_intermediate*. Ces dernières reflètent les changements de granularité et apportent une information plus précise sur l'ordre relatif de réalisation des processus de granularité supérieure. Les processus de granularité maximum, élémentaires ou atomiques (en blanc) sont définis par des propriétés d'entrée(s)/sortie(s) (*input_of, output_of*) des entités biologiques (molécules, complexes macromoléculaires ou séquences) nécessaires à la réalisation du processus ou résultant de son traitement (figures 2 et 3A*).

Dans un souci d'interopérabilité, après l'extraction des connaissances, les concepts ont été prioritairement importés de bio-ontologies pré-existantes permettant la conservation des labels, *Internationalized Resource Identifier* (IRI) et identifiants (Id). Ainsi, les concepts de processus cellulaires bactériens participant à l'expression génique décrits dans GO de leur plus haute hiérarchie à la plus fine possible reliés par des *part_of* ont été importés. Les *part_of* ont été spécialisés en sous-propriétés *starts_with*, *ends_with* ou *has_intermediate*. De même, les concepts de molécules et séquences participant aux processus cellulaires et décrites dans ChEBI ou SO ont été importés. Les concepts correspondant aux complexes macromoléculaires d'intérêt décrits dans KEGG ont été créés et annotés avec l'Id correspondant grâce à la propriété standard *hasDbXref.* Les concepts d'entités biologiques ont été reliés aux processus cellulaires par des relations *input_of* ou *output_of.* Ensuite, les concepts des processus manquants par rapport au modèle systémique et leurs participants ont été ajoutés suite à un effort de curation. Si nécessaire, les processus de plus bas niveau ont été segmentés par l'ajout de nouveaux concepts de processus jusqu'à ce qu'ils impliquent au maximum trois entités biologiques (processus élémentaires).

Au final, bioCMON contient trois super-classes disjointes : '*biological process'* (contenant 40 % de concepts communs avec GO, principalement aux niveaux agrégés) qui hiérarchise les concepts de processus, '*chemical entity*' (contenant 20 % de concepts communs avec ChEBI et 15 % avec KEGG) qui hiérarchise les concepts de molécule et où les classes de mêmes parents sont disjointes, et '*sequence feature'* (contenant 85 % de concepts communs avec SO) qui hiérarchise les concepts de séquences biologiques (figure 2).

2.2 modelCMON

Les concepts modelCMON sont formalisés en lien avec bioCMON. modelCMON comprend 4 super-classes dans '*modelisation entity*' (figure 2):

i) '*reaction*' : *reaction* classifie dix réactions modélisables et est défini en relation avec la super-classe '*model*' par la relation *has_model*. Ces réactions ont été choisies car elles présentent toutes des spécificités en nombre ou en nature (libres, liés ou à motifs) de leurs réactifs.

ii) '*reactant*' : *reactant* est une reclassification des classes de bas niveau de *chemical entity* et *sequence feature* sous forme de réactifs. Ces classes sont automatiquement sélectionnées telles que :

```
reactant ≡ (input_of some 'biological process' or output_of some 'biological process')
```

En fonction de leur nature (séquence ou molécule) et de leur rôle biologique, les *reactant* sont automatiquement répartis entre réactifs libres, réactif à motifs et réactifs liés.

iii) '*modelisable process*' : les processus modélisables sont formellement décrits en fonction du nombre et de la nature de ces réactifs (figure 3A** : *complexation model*). Cette superclasse est donc une nouvelle hiérarchisation des classes de *biological process* qui peuvent être reliées à des réactions modélisables par inférence. Il existe 10 processus modélisables en lien avec les 10 réactions par la relation *has_reaction* (exemple figure 3A*** : *Complexation*).

iv) '*model*' : les modèles hiérarchisés dans cette super-classe ont été sélectionnés et intégrés à partir de SBO. On y retrouve les modèles capables de modéliser les 10 réactions. model est défini en relation avec la super-classe '*parameter entity*' par la relation *has_parameter*. *parameter entity* contient une liste de paramètres intégrés à partir de SBO.

Suite à cette formalisation, modelCMON est construit par inférence à partir de bioCMON. Le raisonneur (HermiT 1.3.8 ; Glimm *et al.*, 2014) s'appuie sur la sélection de réactifs d'intérêt puis des processus biologiques simulables. Sur ce principe, les axiomes sont utilisés pour faire le lien entre un contexte biologique et un contexte de modélisation pour une même connaissance (figure 2).

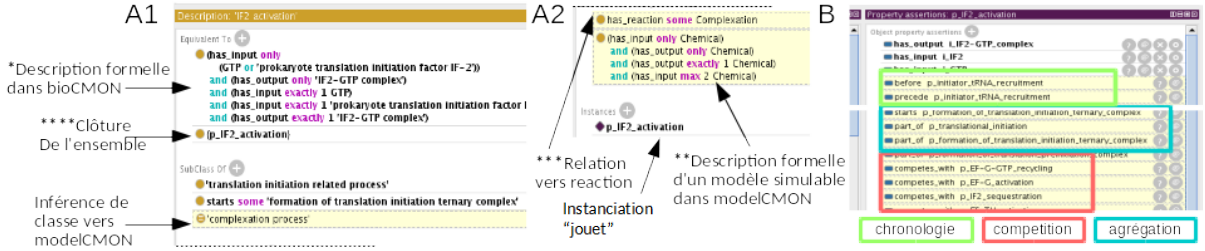

FIGURE 3 – Capture d'écran de Protégé explorant l'activation de l'IF2 (A : classe, B : instance).

2.3 Enrichissement de la connaissance de bioCMON par inférence

Les classes de C'MON ont été conçues dans le cadre d'une expressivité en logique de description SROIQ et certaines inférences au niveau des concepts peuvent être de grande complexité. Pour apporter automatiquement des informations complémentaires à notre modèle, chacune des classes feuilles de bioCMON a été instanciée par des éléments "jouets" et les ensembles ont été clôturés aux seules instances de ces classes (figure 3A****). Dans un second temps, profitant de la formalisation de C'MON en OWL2 (Cuenca Grau *et al.* 2008), des règles en *Semantic Web Rule Language* (SWRL) ont été conçues (Krisnadhi *et al.* 2011). Par exemple :

```
macromolecule(?x), input_of(?x,?p1), input_of(?x,?p2), DifferentFrom (?p1,?p2)
    -> competes_with(?p1,?p2)
```

Ces règles formalisent de nouvelles relations qui permettent de faire ressortir des informations de régulation de processus au travers de compétitions (*compets_with*) ou de chronologies relatives (*before* ou *precedes*) entre des processus et enfin de filtrer les entités biologiques participantes en fonction de l'agrégation des processus (*has_filtered_input*) profitant ainsi pleinement de l'approche systémique multi-échelle (figure 3B).

3 Conclusion et perspectives

C'MON a pour objectif une vision complète à différentes granularités des processus cellulaires. Actuellement C'MON comprend 1384 concepts qui permettent la modélisation des

processus impliqués dans l'expression génique (incluant les processus de régulation). Cette organisation se structure le long d'une échelle de processus de haut niveau (*e.g.* : traduction) aux processus les plus élémentaires (*e.g.* : activation de l'IF2) reliés par des sous-propriétés de *part_of*. Les processus, entités biologiques participantes, modèles et paramètres associés ont été formellement décrits et reliés entre eux. Chacun des processus est relié à un type de réaction et des paramètres [comme un type de constante d'affinité obtenue expérimentalement (k-on) ou estimée à partir de données biologiques (k-est) ; Figure 2]. De plus, C'MON s'enrichit automatiquement d'informations biologiques fonctionnelles complémentaires (chronologie, agrégation, compétition) suite à une inférence sur les instances (figure 3B). En plus d'un effort de curation, ce travail s'appuie sur l'intégration d'ontologies préexistantes (GO, ChEBI, KEGG, SO et SBO). Dans cette démarche, la conservation des IRI et Id des concepts intégrés vise à proposer un modèle interopérable. Ainsi l'utilisation de C'MON peut permettre la réexploitation de travaux préexistants sur l'expression génique en microbiologie avec les nouveaux apports de notre modèle.

Le modèle ontologique C'MON décrit dans cette article montre donc que l'approche systémique multi-échelle est bien pertinente pour décrire les processus cellulaires et les relier à leurs paramètres. C'MON ouvre ainsi de nouvelles perspectives pour l'intégration des données biologiques hétérogènes dans l'objectif d'une utilisation interdisciplinaire.

La preuve de concept étant présentée ici, les prochaines étapes consistent à compléter C'MON avec l'ensemble des processus cellulaires bactériens et leur régulation, comme ceux de la réplication de l'ADN, puis de l'étendre aux processus cellulaires eucaryotes.

A terme, C'MON sera associée à un entrepôt de données biologiques omiques dans le cadre d'un système d'information.

Remerciement : Ce travail a été financé par le projet LIDEX IMSV de Paris-Saclay.

Références

ALTMAN T. *et al.* (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics 14 :112.

COURTOT M *et al.* (2011). Controlled vocabularies and semantics in systems biology. Molecular Systems Biology 7 :543.

CUENCA GRAU B. *et al.* (2008). OWL 2: The Next Step for OWL. Journal of Web Semantics 6.

DE MATOS P *et al.* (2010). Chemical Entities of Biological Interest: an update. Nucleic Acids Res 38, p. 249–254.

EILBECK K. *et al.* (2005). The Sequence Ontology: A tool for the unification of genome annotations. Genome Biol. 6:R44 .

GLIMM B. *et al.* (2014). HermiT : An OWL 2 Reasoner. J. of Automated Reasoning 53, p. 245–269.

GOELZER A. & FROMION V. (2011). Bacterial growth rate reflects a bottleneck in resource allocation. Biochim Biophys Acta. 1810, p. 978-988.

GOELZER A. *et al.* (2015). Quantitative prediction of genome-wide resource allocation in bacteria. Metab Eng. 32, p. 232-243.

HILL *et al.* (2013). Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics 14 :513.

KITANO H. *et al.* (2001). Foundations of Systems Biology. MIT Press.

KRISNADHI A. *et al.* (2011). OWL and Rules. Reasoning Web. 6848, p. 382-415.

METZKER M.L. (2010) Sequencing technologies-the next generation. Nat. Rev. Genet. 11, p. 31–46.

NEKRUTENKO A. &, TAYLOR J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat. Rev. Genet. 13, p. 667–672.

THE GENE ONTOLOGY CONSORTIUM (2000). Gene ontology: tool for the unification of biology. Nat Genet 25, p.25–29.