HAL
open science

Contributions à l'étude de la dynamique des systèmes biologiques et aux systèmes de calcul en biologie synthétique

Patrick Amar

To cite this version:

Patrick Amar. Contributions à l'étude de la dynamique des systèmes biologiques et aux systèmes de calcul en biologie synthétique. Bio-informatique [q-bio.QM]. Université Paris Sud - Paris XI, 2013. tel-00929785

HAL Id: tel-00929785

https://theses.hal.science/tel-00929785v1

Submitted on 13 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS SUD
Spécialité : Bioinformatique

Contributions à l'étude de la dynamique des systèmes biologiques et aux systèmes de calcul en biologie synthétique

Document de synthèse présenté pour l'obtention d'une

Habilitation à Diriger des Recherches

par

Patrick AMAR

Laboratoire de Recherche en Informatique

soutenue le 19 décembre 2013 devant le jury composé de :

Mme. Pascale Le Gall	Professeur, École Centrale Paris
M. Gilles Bernot	Professeur, Université de Nice
M. Philippe Dague	Professeur, Université Paris Sud, Orsay
M. Jacques Demongeot	Professeur, Université Joseph Fourier, Grenoble
M. Alain Denise	Professeur, Université Paris Sud, Orsay
M. Victor Norris	Professeur, Université de Rouen

au vu des rapports de :

Mme. Pascale Le Gall	Professeur, École Centrale Paris
M. Jacques Demongeot	Professeur, Université Joseph Fourier, Grenoble
M. Andre Levchenko	Professeur, Yale University

Table des matières

Table des matières **3**

1 Introduction **7**
1.1 Motivations 8
1.2 Processus dynamiques en biologie 9
1.3 Biologie de synthèse 10
1.4 Plan du mémoire 11

2 Modélisation des réactions biochimiques **13**
2.1 Modélisation continue à équations différentielles 13
2.2 Modélisation discrète 19

3 Le simulateur HSIM **25**
3.1 Introduction 25
3.2 Présentation générale 26
3.3 Fondements physiques 27
3.4 Fonctionnement de HSIM 29
3.5 Conclusion 35

4 Systèmes Complexes **37**
4.1 Introduction 37
4.2 Cytosquelette et interactions avec la membrane 38
4.3 Construction et analyse d'un réseau de réactions 40
4.4 Détection automatique de phénomènes émergents 40
4.5 Études de cas 45
4.6 Conclusion 47

5 Hyperstructures **49**
5.1 Functioning-dependent Structures 49
5.2 Couplage glycolyse-PTS 52
5.3 Couplage cytosquelette et réseaux métaboliques 57
5.4 Life on the scales 58

6 Biologie de synthèse **61**
6.1 Projet CompuBioTic 63
6.2 Projet BS^2 64
6.3 Calculer avec des bactéries 67
6.4 La Réaction en Chaîne Mimétique (Mimic Chain Reaction) 68

7 Métabolisme des streptomyces **73**
7.1 Outils biotechnologiques 73
7.2 Courbe sigmoïde de résistance à la neomycine 75

8 Conclusion et perspectives **81**

Bibliographie **85**

Remerciements

Je remercie vivement Pascale Le Gall, professeur, École Centrale Paris, Jacques Demongeot, professeur, responsable de l'axe *e-santé* au laboratoire AGIM, Université Joseph Fourier à Grenoble et Andre Levchenko, professeur, directeur du *Yale Systems Biology Institute*, Université de Yale, d'avoir accepté d'être rapporteurs de ce document.

Je remercie également Alain Denise, professeur à l'Université Paris Sud, Gilles Bernot, professeur à l'Université de Nice et Victor Norris, professeur à l'Université de Rouen de participer à mon jury. Un clin d'oeil amical à Philippe Dague, professeur à l'Université Paris Sud, directeur du LRI, pour notre collaboration intense sur le rapport d'activité du laboratoire ; merci Philippe de faire partie de mon jury.

Les travaux présentés dans ce mémoire ne sont pas le fruit du travail d'une personne isolée, mais le résultat de collaborations avec divers scientifiques, informaticiens, biologistes et physico-chimistes qui ont partagé avec moi cette démarche pluri-disciplinaire et que je tiens à remercier individuellement.

François Képès en premier lieu, qui est l'initiateur de ma *conversion* à la biologie des systèmes, bien qu'en 1998 ce terme n'existait pas encore, et à qui je dois la majorité de mes connaissances en biologie, acquises notamment à l'occasion des *Introductions Avancées à la Biologie* qu'il a organisées. Cette aventure pluri-disciplinaire a été rendue possible grâce aux *Ateliers du Programme d'Epigénomique* que François a mis en place au Genopole d'Évry dès 2000. Ce sont ces rencontres entre mathématiciens, informaticiens, biologistes et physiciens qui nous ont permis d'acquérir un langage commun autour de la modélisation en biologie.

C'est à l'occasion de ces ateliers que j'ai rencontré Vic Norris, Camille Ripoll et Michel Thellier de l'université de Rouen, avec lesquels j'ai eu (et j'ai encore) à de nombreuses occasions, des discussions scientifiques aussi animées qu'intéressantes et productives. Philippe Tracqui du laboratoire TIMC à Grenoble, qui m'a initié aux interactions entre les propriétés mécaniques des cellules et les processus intra-cellulaires. Une pensée aussi pour mes collègues informaticiens, alors à l'université d'Évry, Pascale Le Gall, Marc Aiguier, Gilles Bernot et Jean-Paul Comet avec lesquels j'ai passé un certain temps à essayer de comprendre ce que disaient nos amis biologistes. Depuis je mesure le chemin parcouru : je suis maintenant autant dans mon élément à un exposé de biologie qu'à un exposé d'informatique !

Merci encore à Gilles Bernot pour son amitié indéfectible depuis l'époque de sa première année de doctorat, pour ses conseils toujours pertinents et pour le don qu'il a de donner une ambiance détendue et propice à un travail de qualité dans les structures qu'il a dirigées.

Un grand merci à Franck Molina, directeur du laboratoire Sysdiag à Montpellier. Dès notre rencontre en 2002, lors de la première école thématique sur la modélisation en biologie, faisant suite aux ateliers d'Evry, Franck a tout de suite vu ce que la modélisation et l'informatique pouvaient apporter à la biologie. Je lui suis infiniment reconnaissant pour son amitié, l'accueil chaleureux dans son laboratoire et les discussions scientifiques très enrichissantes que nous avons à chaque occasion de nous voir.

Merci aussi à Marie-Joelle Virolle de l'Institut de Génétique et Microbiologie à Orsay, grâce à elle je sais (presque) tout ce qui est actuellement connu du métabolisme secondaire des *streptomyces*.

Tous mes remerciements aux membres de l'équipe *Bioinfo* du LRI pour la bonne ambiance qui y régne : les co-fondateurs Christine et Alain, Sarah et Jérome, et enfin les *nouveaux* Sabine et Loïc, avec lesquels j'avais déjà eu le plaisir de travailler avant leur recrutement. Merci aussi à nos doctorants : leur optimisme et leur dynamisme contribuent pleinement à l'ambiance amicale de l'équipe.

Enfin, je profite de cette occasion pour remercier chaleureusement les personnels administratifs et techniques du LRI et de Sysdiag pour leur dévouement, leur gentillesse et leur compétence.

Introduction 1

Jusqu'à la fin des années 1990, le principal apport de l'informatique à la biologie a été en génomique, pour les traitements de séquences, notamment dans les recherches de motifs et de similitudes (alignements) de séquences. Le terme *bioinformatique* est encore employé par certains biologistes pour désigner les outils issus de ces recherches.

Aujourd'hui, les travaux en bioinformatique sur la génomique sont complétés par les travaux sur le transcriptome et sur le protéome. Le terme de bioinformatique post-génomique est employé pour qualifier les projets cherchant à déterminer la partie du génome qui est transcrite en ARN messagers (le transcriptome) et ceux cherchant à connaître les protéines exprimées à partir de ces ARN messagers (le protéome). A travers ces travaux, c'est la prédiction de la fonction des différents gènes qui est attendue. La compréhension de la fonction biologique des gènes est devenue le "nouveau Graal" de la biologie moléculaire, elle est sensée par exemple permettre de concevoir un traitement adapté à un malade, la médecine *personnalisée*.

La connaissance des gènes et des protéines produites à partir de ces gènes est un point important mais qui ne suffit pas pour comprendre dans sa totalité le fonctionnement des organismes vivants. En effet, connaître les constituants élémentaires est un premier pas, mais il faut ensuite comprendre leurs interactions. La biologie des systèmes répond à ces besoins en associant l'informatique, les mathématiques et la biologie. La compréhension des systèmes biologiques dans leur ensemble est un projet à long terme sur lequel travaillent énormément de laboratoires de recherche à travers le monde. Au delà de leur compréhension, le développement de cellules et d'organes virtuels constitue probablement l'un des plus grands défis de la biologie des systèmes.

La biologie des systèmes (Systems Biology) se focalise sur l'étude, de façon intégrative, des interactions complexes dans les organismes biologiques. Elle consiste, d'après Hiroaki Kitano [1, 2], en la compréhension de la structure de systèmes biologiques, de leur dynamique et des processus de régulation les contrôlant.

La modélisation des systèmes biologiques est devenue indispensable ; elle permet, entre autres un meilleur dialogue entre biologistes, informaticiens et mathématiciens en leur fournissant des outils facilitant l'interdisciplinarité comme par

exemple CellDesigner [3] ou SBML [4]. L'utilisation de l'informatique en biologie est aujourd'hui devenue incontournable, ses champs d'application sont très variés et couvrent une large partie des recherches en biologie.

De plus, les informaticiens impliqués dans ces recherches se sont aperçus qu'ils pouvaient trouver dans la biologie une source d'inspiration pour résoudre des problèmes liés uniquement à l'informatique comme par exemple les méthodes d'optimisation avec les réseaux de neurones ou les algorithmes génétiques.

1.1 Motivations

La biologie des systèmes peut se présenter sous la forme d'un cycle partant d'un modèle du processus biologique étudié amenant à proposer des hypothèses, suivi d'une validation expérimentale, celle-ci étant utilisée pour valider ou invalider le modèle de façon à le raffiner et recommencer le cycle.

Je me suis concentré sur la partie conception de modèles et sur leur résolution par des méthodes mathémathiques et informatiques. Ceci permettant de *court-circuiter* le cycle avant la validation expérimentale, dans le but de (ne) proposer (que) des expériences biologiques pertinentes. Lorsque j'ai commencé à m'intéresser à la biologie et à ce que pourrait y apporter l'informatique, le terme de *biologie des systèmes* n'existait pas encore. Il n'est apparu qu'après 2000.

Dans un passé déjà lointain, je m'étais intéressé à la conception de circuits intégrés VLSI. On cherchait à implémenter dans le silicium des algorithmes efficaces de calculs arithmétiques. Pour des raisons de temps et de coûts, il était difficilement concevable de construire réellement les circuits pour les tester. On avait déjà besoin de modèles pour ces circuits et de méthodes efficaces de résolution de ces modèles pour valider les circuits avant leur fabrication. C'est cette première expérience de la pluri-disciplinarité qui m'a incité à m'intéresser à la modélisation de systèmes biologiques.

La biologie, probablement encore plus que la physique, est une science où on est confronté aux aléas des expériences portant sur des systèmes complexes dont on ne connaît pas tous les tenants et les aboutissants : les comportements observés peuvent varier du tout au tout à cause d'une infime variation des paramètres, principalement ceux qu'on contrôle mal.

En d'autres termes, les expériences ne sont pas facilement reproductibles car les conditions extérieures peuvent avoir une grande influence sur les phénomènes observés. Par exemple, des souches bactériennes peuvent avoir mutées, des impuretés peuvent contaminer des produits, etc. C'est pourquoi on utilise quasi systématiquement dans les expériences des *"contrôles"* dont on connaît le comportement pour réduire les risques d'erreurs.

Malgré toutes ces difficultés (inexistantes dans les sciences exactes comme les mathématiques ou l'informatique) on doit bâtir des modèles, sinon des théories, expliquant les mécanismes mis en jeu et qui peuvent prédire les comportements qui seront vérifiés lors de futures expériences.

C'est pour cela qu'une approche rationnelle de modélisation peut apporter beaucoup à la compréhension des organismes vivants. Le fait de pouvoir faire des expériences virtuelles *in silico* avec des outils théoriques et des logiciels éprouvés fait gagner beaucoup de temps à l'expérimentateur. En effet, en testant des hypothèses sur un modèle on peut fréquemment éviter de faire des expériences inutiles. En simplifiant à l'extrême, on peut distinguer trois cas :

1. soit le modèle est faux, et le résultat de l'expérience virtuelle est donc inutile.
2. si le modèle est suffisamment correct, et l'expérience virtuelle confirme l'hypothèse testée, cela ne prouve rien, mais "conforte" la validité de l'hypothèse (on a trouvé **un** mécanisme plausible, pas forcément **le** mécanisme réellement utilisé).
3. si enfin, l'expérience virtuelle invalide l'hypothèse, alors on a **prouvé** que cette hypothèse était fausse et il est complètement inutile de la tester expérimentalement.

1.2 Processus dynamiques en biologie

Parmi les différents processus qui entrent en jeu dans les organismes vivants, je me suis particulièrement intéressé aux réseaux de régulation géniques, aux réseaux métaboliques et aux interactions entre ces deux types de réseaux.

Ces réseaux, ainsi que quasiment tout ce qui entre dans le *vivant*, sont constitués de composés chimiques participant à des réactions. D'où l'idée d'étudier des modèles décrivant l'évolution dans le temps des quantités de réactifs mis en jeu dans ces réactions.

Dans les cellules procaryotes, il n'y a pas de sous-compartiments mais l'intérieur de la cellule n'est pas pour autant homogène. Bien au contraire il est très structuré, avec des régions où l'encombrement est très différent (par exemple le nucléoïde), menant à des vitesses de diffusion elles aussi très différentes et à des concentrations locales de certaines espèces biochmiques plus élevées que leur concentration moyenne. La différence de viscosité est une, mais pas la seule, raison de l'inhomogénéité du milieu qui permet à certaine réactions de se produire alors que les quantités de réactifs sont très faibles.

Une autre raison de l'inhomogénéité du milieu est la capacité de certaines protéines à former des complexes avec d'autres de la même espèce ou d'espèce différente, ceci pouvant avoir pour conséquence de d'élever la concentration locale de ces protéines et donc d'augmenter le nombre de réactions auxquelles elles participent.

On commence à se rendre compte que les choses ne sont pas aussi simples qu'elles pourraient apparaître : la définition de zones de concentration élevée (ou basse) de biomolécules n'est pas forcément fixée par l'environnement (compartimentalisation par exemple) mais peut être une conséquence directe ou indirecte des réactions qui mettent en jeu ces mêmes biomolécules. Quelques exemples de ce phénomène d'*auto organisation* seront montrés par la suite.

Qu'elles soient explicites, comme dans les réseaux de régulation géniques, ou implicites comme avec l'exemple des concentrations locales hors moyenne mentionné précédemment, on voit que les réseaux biologiques sont très rebouclés. Ces boucles de rétroaction pouvant mener à des régulations de type homéostasie ou multistationnarité, ainsi qu'à des comportements oscillatoires.

L'un des buts de la biologie des systèmes est, à l'aide de données expérimentales, de proposer un modèle si possible quantitatif, permettant d'inférer les causes microscopiques (interactions moléculaires) menant aux conséquences macroscopiques qui sont observées par expérimentation. Un tel modèle est dit *explicatif*. S'il est suffisamment complet, il peut être aussi *prédictif*, c'est-à-dire permettre de fournir des résultats qui seront confirmés par des expériences futures, et donc dans certaines limites permettre de faire des expériences *in silico* pertinentes.

Une partie de mes travaux de recherche au cours des dix dernières années a été de réaliser un systéme de simulation intégré, HSIM, le plus complet possible, offrant à la fois une simplicité d'utilisation pour le modélisateur et un grand pouvoir d'expression permettant de prédire la dynamique de modèles issus de domaines très variés de la biologie. Le langage de description de HSIM permet de décrire de façon générique des modèles aussi divers que des réseaux métaboliques, des réseaux d'interaction géniques, ainsi que le couplage de ces deux types de réseaux ; le mécanisme de réplication de l'ADN couplé à la transcription et la traduction des gènes pour étudier l'influence de la réplication sur la dynamique de réseaux d'interaction entre ces gènes ; des modèles permettant de montrer des phénomènes d'auto-organisation spatiale et leur influence sur la dynamique globale du système, etc.

1.3 Biologie de synthèse

La biologie de synthèse est un axe de recherche en plein essor dans le domaine de la conception et de l'ingéniérie de systèmes basés sur des règles fonctionnelles biologiques dans le but d'obtenir de nouvelles fonctionnalités qui ne sont pas présentes dans la nature.

L'ingéniérie génétique est une exemple de biologie de synthèse : pour optimiser dans un organisme vivant une voie métabolique produisant un composé chimique d'intéret thérapeutique ou industriel, on va modifier le génome de cet organisme pour y introduire une machinerie moléculaire spécifique. Cette machinerie sera répliquée lors de la reproduction de l'organisme.

Un autre exemple est la conception de composants artificiels pour des applications de nano technologies où ces composants sont conçus et déployés en dehors de cellules vivantes. L'un des intérêts par rapport à la méthode précédente est d'éviter les éventuelles mutations risquant soit d'affaiblir la production du composé désiré, soit de rendre pathogène les organismes utilisés.

Soit enfin, ce qui est le genre le plus extrème de nano technologie bio-inspirée, la synthèse de systèmes biologiques *vivants* à partir d'assemblages de composants artificiels. En 2010, une équipe de recherche menée par Craig Venter a réussi à

synthétiser le génome complet de la bacterie *Mycoplasma mycoides* (un peu plus d'un million de paires de bases) et à introduire ce chromosome artificiel dans une cellule de *Mycoplasma capricolum* préalablement vidée de son matériel génétique, pour obtenir une bactérie viable [5].

En collaboration avec l'équipe de Franck Molina du laboratoire Sysdiag à Montpellier, je me suis intéressé à la conception et à la réalisation de bio-calculateurs artificiels utilisant des composants logiques implémentés à l'aide de réseaux métaboliques permettant de détecter les marqueurs d'une pathologie particulière (cancer colorectal, néphropathie diabétique) et de faire un calcul programmé pour fournir une réponse intégrée, par exemple sous forme colorimétrique. Des premiers résultats ont été obtenus lors la thèse de Stéphanie Rialle, à laquelle j'ai contribué de façon informelle.

Je continue mes recherches dans cette voie en co-encadrant avec Franck Molina la thèse de Marc Bouffard qui porte sur la définition et la conception de composants logiques enzymatiques, et sur la conception d'outils informatiques permettant de réaliser et de tester *in silico* des réseaux métaboliques artificiels réalisant un calcul donné.

1.4 Plan du mémoire

Ce mémoire est organisé autour de deux parties principales :

1. la première est centrée sur la modélisation et la simulation des systèmes de réactions biochimiques et leur utilisation pour le test d'hypothèses sur le fonctionnement des processus biologiques (chap. 2). Cette partie commence par un rappel des différents types de modélisation de systèmes de réactions, puis par une description détaillée du système de simulation intégré HSIM que j'ai réalisé (chap. 3).

   Puis, considérant de façon plus générale les cellules en tant que systèmes complexes faisant émerger des phénotypes cohérents, je me suis intéressé à la détection automatique de phénomènes émergents dans des simulations multi-agents, où j'ai co-encadré la thèse de Thomas Moncion (chap. 4).

   Viennent ensuite des cas d'étude, notamment autour du concept d'Hyperstructure ou HSIM a permis de faire des avancées notables (chap. 5).

   Enfin dans le cadre du projet de biologie de synthèse *Compubiotic*, nous avons utilisé HSIM pour valider les réseaux métaboliques artificiels implémentant des fonctions logiques spécifiques (chap. 6).

2. la deuxième partie du mémoire est plus axée sur la biologie théorique et la modélisation. En premier lieu, une étude théorique sur l'utilisation de bactéries artificiellement modifiées pour exécuter des calculs informatiques. Puis une autre approche, toujours utilisant des bactéries artificiellement modifiées, permettant de détecter de très faibles concentrations de peptides et de fabriquer des protéines artificielles ayant une structure donnée (chap. 6).

   Enfin, des travaux de biologie des systèmes portant sur la modélisation du mécanisme de résistance des *streptomyces* aux antibiotiques de type

aminoglycocides, donnant pour la première fois une explication a la forme sigmoïde de la courbe de survie en fonction du taux d'expression du gène de résistance (chap. 7).

Après la conclusion sur mes recherches passées, je donnerai la direction de mes travaux futurs, incluant notamment la poursuite du développement de HSIM et son inclusion dans un outil plus général de modélisation pour la biologie de synthèse (chap. 8).

Modélisation des réactions biochimiques 2

La première étape pour modéliser les organismes vivants est d'en modéliser leurs composants : les cellules. Cette modélisation doit se faire autant du point de vue de leur structure morphologique, que leur comportement dynamique, en tenant compte de leurs interactions avec les cellules voisines et avec l'environnement extra-cellulaire.

Si on se place au niveau intra-cellulaire, dans une vision moléculaire de la cellule, on peut se dire qu'aussi bien la forme que le comportement de la cellule sont des conséquences directes ou indirectes des interactions entre les molécules qui la composent. Donc, si on parvient à faire un modèle suffisament fidèle et suffisament complet au niveau moléculaire, on est en bonne voie pour faire émerger le phénotype de la cellule. En extrapolant, on pourrait faire, *in silico* avec un tel modèle, toutes les expériences nécessaires pour tester les effets d'un médicament par exemple.

C'est pourquoi on s'est très tôt intéressés à concevoir des méthodes quantitatives de modélisation et de simulation de réactions biochimiques.

2.1 Modélisation continue à équations différentielles

Modélisation des réactions biochimiques

La plupart des réactions biochimiques dans un organisme vivant sont des réactions catalysées par des protéines ayant des capacités enzymatiques, des réactions d'association ou de dissociation de protéines avec d'autres protéines ou ADN ou ARN, ou enfin des réactions de changement de conformation dans la structure des molécules. Chacun de ces processus peut être décrit à un certain niveau comme des réaction chimiques simples entre deux ou plusieurs molécules/états du système. Nous allons nous intéresser à l'évolution dans le temps des concentrations/états à partir d'une population initale donnée.

Réactions mono-moléculaires

Une réaction mono-moléculaire met en cause une seule molécule qui peut avoir plusieurs états ; Le changement spontané de la structure d'une protéine est un exemple de réaction mono-moléculaire. Le cas le plus simple de réaction mono-moléculaire ou une molécule fluctue entre deux états ou conformation, A et B, peut être noté sous la forme :

$$\mathrm{A} \underset{\mathrm{k_r}}{\overset{\mathrm{k_f}}{\rightleftharpoons}} \mathrm{B}$$

où k_f et k_r sont les constantes cinétiques pour la réaction directe $A \rightharpoonup B$ et la réaction inverse $B \rightharpoonup A$. Ces constantes ont comme dimension s^{-1}.
Les équations qui décrivent la variation des populations de A et de B en fonction du temps sont :

$$\begin{cases} \dfrac{dA}{dt} = -k_f A + k_r B & (2.1) \\ \dfrac{dB}{dt} = k_f A - k_r B & (2.2) \end{cases}$$

Après un certain temps le système atteint un stade où la vitesse à laquelle A est transformé en B devient identique à la vitesse à laquelle B se transforme en A. Dès ce moment les populations relatives de A et de B ne changent plus dans le temps, d'où le terme d'*équilibre* :

$$\frac{B_{eq}}{A_{eq}} = \frac{k_f}{k_r} = K$$

où A_{eq} et B_{eq} sont les populations de A et de B à l'équilibre et K la constante d'équilibre de la réaction.

Pour ces réactions mono-moléculaires les populations à l'équilibre sont indépendantes des concentrations initiales des molécules. Nous pouvons donc normaliser toutes les concentrations à l'unité, de façon à ce que dans les équations précédentes A et B fassent référence à la proportion de chaque population dans les états A et B avec la contrainte qu'à tout moment $A + B = 1$. Dans ces conditions, à l'équilibre nous avons :

$$\begin{cases} A_{eq} = \dfrac{1}{1+K} = \dfrac{k_r}{k_f + k_r} \\ B_{eq} = \dfrac{K}{1+K} = \dfrac{k_f}{k_f + k_r} \end{cases}$$

Si le système est initialement hors équilibre, les populations de A et B vont varier en fonction du temps jusqu'à ce que le système atteigne l'équilibre. Cette variation en fonction du temps est modélisée par les équations différentielles (2.1)

et (2.2). Ces équations sont intégrables (voir démonstration en annexe) la solution est :

$$\boxed{\begin{cases} A(t) = \dfrac{k_f}{k_f + k_r} e^{-(k_f+k_r)t} + \dfrac{k_r}{k_f + k_r} \\ B(t) = \dfrac{k_f}{k_f + k_r} \left(1 - e^{-(k_f+k_r)t}\right) \end{cases}}$$

On voit donc que la concentration de A décroît exponentiellement depuis sa valeur initiale jusqu'à sa valeur d'équilibre alors que la concentration de B croît exponentiellement depuis 0 jusqu'à sa valeur à l'équilibre (fig. 2.1). On peut aussi noter qu'à tout instant t, $A(t) + B(t) = 1$.

Pour résumer, dans cet exemple de réaction mono-moléculaire mettant en jeu seulement deux états, la variation temporelle des populations dans chacun de ces deux états est décrite par une simple exponentielle ayant une constante cinétique caractéristique donnée par la somme des cinétiques directe et inverse.

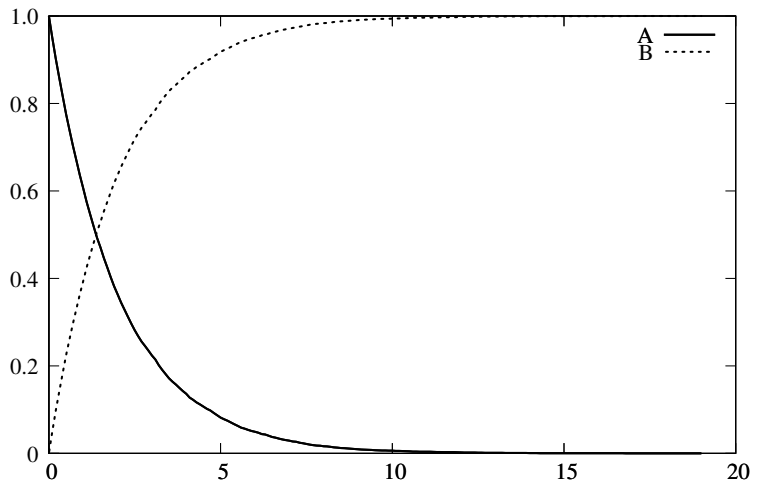

FIG. 2.1: Réaction mono-moléculaire

Réactions bi-moléculaires

Les réaction bi-moléculaires introduisent une complication supplémentaire : la vitesse de la réaction dépend de la concentration de chacune des deux espèces moléculaires.

La réaction se fait en deux étapes : les molécules qui réagissent doivent d'abord être suffisamment proches pour entrer en collision, puis si l'énergie du choc est suffisante, elles peuvent réagir pour former le produit ou sinon continuer à diffuser.

La forme la plus simple de réaction bi-moléculaires est la suivante :

$$\mathrm{A} + \mathrm{B} \underset{k_\mathrm{r}}{\overset{k_\mathrm{f}}{\rightleftharpoons}} \mathrm{A}B \xrightarrow{k_\mathrm{c}} \mathrm{P}$$

où k_f est la constante cinétique bi-moléculaire de formation du complexe AB, k_r est la constante cinétique mono-moléculaire de dissociation du complexe avant que la réaction ne se fasse et k_c est la constante cinétique mono-moléculaire de l'étape de réaction. La dimension de k_r et de k_c est s^{-1} alors que celle de k_f est $M^{-1}s^{-1}$.
Du fait de leur diffusion, des molécules d'espèce A vont entrer en collision avec des molécules d'espèce B, certaines de ces collisions auront une énergie suffisante pour déclencher la formation du complexe AB. C'est ce nombre de collisions *efficaces* qui est capturé par la constante cinétique directe k_f. Le nombre de collisions entre deux molécules d'espèces A et B est proportionnel au produit de leur concentration $[A][B]$. La preuve intuitive est la suivante : si on suppose qu'il n'y a qu'une seule molécule de l'espèce A dans le volume considéré, on voit qu'il y a $\alpha[B]$ chances de collisions avec les molécules de l'espèce B, α étant une constante de proportionnalité. Si maintenant il y a plusieurs molécules de l'espèce A, ce nombre est multiplié par $[A]$, c'est la *loi d'action de masse*, énoncée en 1864 par Guldberg et Waage [6].
Nous pouvons modéliser la vitesse de formation du complexe AB avec l'équation suivante :

$$\frac{d[AB]}{dt} = k_f[A][B] - k_r[AB] - k_c[AB] \tag{2.3}$$

Réaction catalysée par une enzyme

Pour l'étude des réseaux métaboliques nous allons nous intéresser principalement aux réactions catalysées. Certaines réactions n'utilisent qu'une enzyme, d'autres nécessitent l'association de plusieurs enzymes pour fonctionner. Nous allons ici utiliser comme exemple une réaction n'utilisant qu'une enzyme E qui catalyse la transformation d'un substrat S en un produit P.

Toutes les réactions biochimiques sont réversibles, mais souvent les cinétiques dans les sens direct et inverse sont très inégales et l'équilibre de la réaction décalé dans un sens. On peut raisonnablement considérer que la réaction est *irréversible*. On peut donc modéliser cette réaction de catalyse enzymatique sous la forme d'une réaction bi-moléculaire :

$$\mathrm{E} + \mathrm{S} \underset{k_\mathrm{r}}{\overset{k_\mathrm{f}}{\rightleftharpoons}} \mathrm{E}S \xrightarrow{k_{cat}} \mathrm{E} + \mathrm{P}$$

Le système d'équations qui décrit cette réaction est le suivant :

$$\begin{cases} \dfrac{d[ES]}{dt} = k_f[E][S] - k_r[ES] - k_{cat}[ES] \\ \dfrac{d[P]}{dt} = k_{cat}[ES] \end{cases}$$

avec les contraintes de conservation de la matière suivantes :

$$\begin{cases} [E] + [ES] = [E]_0 \\ [S] + [ES] = [S]_0 \end{cases}$$

Si on suppose que la concentration en enzymes est très faible devant la concentration en métabolites, $[E] \ll [S] + [P]$, on peut appliquer l'approximation de l'état quasi-stationnaire, $\frac{d[ES]}{dt} = 0$ et le système se résoud :

$$\frac{dP}{dt} = \frac{k_{cat}[E][S]}{\frac{k_r + k_{cat}}{k_f} + [S]}$$

en posant : $K_m = \frac{k_r + k_{cat}}{k_f}$ et $V_{max} = k_{cat}[E]$, on obtient l'équation de *Michaelis-Menten* :

$$\frac{dP}{dt} = \frac{V_{max}[S]}{K_m + [S]}$$

V_{max} est la vitesse maximale de production de P, atteinte quand les enzymes sont *saturées*. La dimension de K_m est celle d'une concentration (M : moles par litre), on peut voir que K_m est la concentration du substrat S qui permet de produire P à la vitesse $\frac{V_{max}}{2}$ (fig. 2.2).

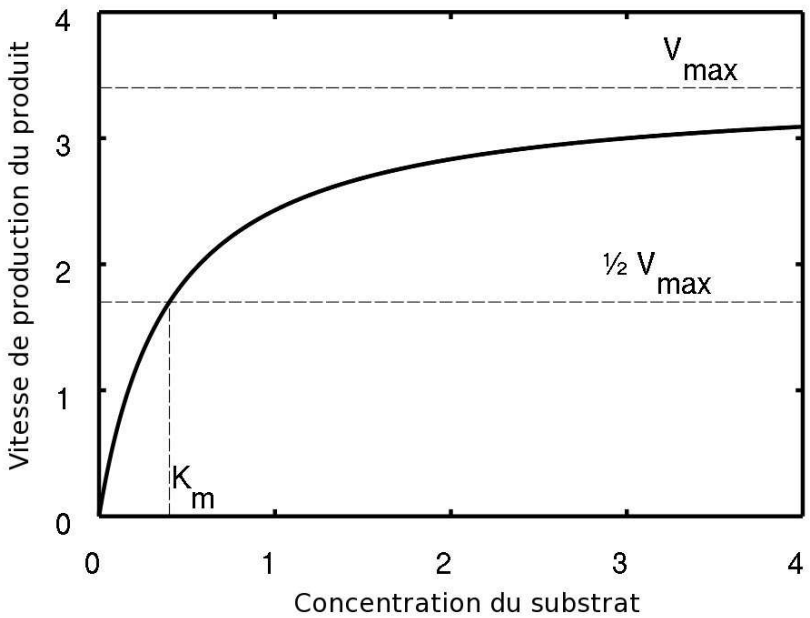

FIG. 2.2: Vitesse de production de P en fonction de [S]

Conditions de validité des modèles à équations différentielles

Il est important de savoir avec quelles hypothèses les modèles à équations différentielles ont été conçus. Le principe de base est que ces modèles sont des modèles *macroscopiques*, ils rendent compte du comportement *d'ensemble* des molécules des réactants mis en jeu. Ces modèles ont été conçus pour approximer de la façon la plus exacte le résultat de nombreuses expériences réelles en tube à essai.

Quand le système d'équations différentielles est intégrable le calcul des concentrations à tout instant est immédiat, dans le cas contraire (hélas très fréquent) on peut néanmoins résoudre le système par des méthodes numériques approchées, assez efficaces en termes de temps calcul.

Il y a trois hypothèses implicites dans ce type de modélisation qu'il convient d'avoir toujours présentes à l'esprit :

1. Le modèle suppose que toutes les molécules des réactants sont *en permanence* réparties de façon homogène dans le volume considéré. Cela revient aussi à supposer que la vitesse de diffusion des molécules est infinie (et de fait elle ne fait absolument pas partie des paramètres de ces modèles). La température, qui influe sur l'agitation moléculaire et donc sur l'énergie des collisions est intégrée dans les *constantes* cinétiques des réactions.
2. Le modèle est *continu* en ce sens qu'il utilise des concentrations traitées comme des nombres réels pour rendre compte des quantités de chaque réactant. Dans beaucoup de cas, ces modèles à équations différentielles décrivent des situations irréelles où la concentration d'un réactant est tellement faible qu'elle représente une fraction de molécule ! Les conséquences sur l'évolution dans le temps du système biologique modélisé qui sont alors prédites dans ce cas peuvent être catastrophiques. Par exemple, un système biologique dont on sait qu'il fait osciller périodiquement la concentration de tel réactant n'oscille plus dans un modèle continu parce que celui-ci ne fait pas la différence entre une molécule et zéro molécule, et donc *rate* un événement clef du système.
3. Le modèle est *déterministe* en ce sens qu'il présuppose que *toutes* les molécules ont le comportement de la *molécule moyenne*. En bref, cette méthode de modélisation ne prend pas en compte la *stochasticité* inhérente au mouvement brownien des molécules en solution.

Ces trois hypothèses imposent comme condition de validité du modèle, qu'à chaque instant les concentrations des réactants soient suffisamment importantes.

Un autre défaut de la modélisation par équations différentielles ordinaires est qu'elle ne tient pas compte de la localisation spatiale des molécules, c'est une conséquence de l'hypothèse 1. Or on sait maintenant que même les bactéries ne sont pas des tubes à essai miniatures, et que bien qu'elles n'aient pas de compartiments délimités par des membranes comme les cellules eucaryotes, elles sont en fait très structurées. Ces modèles ne capturent donc pas les effets souvent très importants de la concentration *locale* élevée de certains réactants (alors que leur concentration globale est quasi-nulle). De plus, ces concentrations locales élevées sont fréquemment des conséquences du fonctionnement même du système

biologique. Il faut donc dans ces cas utiliser des modèles à *équations aux dérivées partielles* qui sont beaucoup plus compliqués à mettre en oeuvre et à intégrer.

Contre-exemple

Le répresseur tétramérique de l'opéron *lac* chez *E.coli* est un bon exemple de ce cas. La concentration de ce répresseur est quasi-nulle chez *E. coli* : environ 10 copies par bactérie, et pourtant son action est déterminante du fait de sa forte concentration locale près de ses sites de fixation[7, 8]. En effet, il suffit que deux copies du répresseur se fixent respectivement sur le promoteur de deux des opérons présents dans le génome, que ces promoteurs soient proches l'un de l'autre dans l'espace (à cause de boucles dans l'ADN) pour qu'ils forment un tétramere ; alternativement chaque dimere va se détacher de son promoteur cible, mais va rapidement s'y réattacher car il est maintenu proche par l'autre dimere (la probabilité que les deux dimeres se détachent en même temps est très faible).

La cause de la co-localisation des deux dimeres est la boucle que fait l'ADN, cette boucle peut se former parce que les sites de fixation sont placés le long de l'ADN à une distance adéquate. Elle est maintenue dans le temps par la tétramérisation des répresseurs, c'est donc un effet coopératif qui crée et maintient la co-localisation des répresseurs près de leur cible.

Ce type de modèle est donc particulièrement inadapté à une résolution par équations différentielles ordinaires en raison (i) de la concentration infinitésimale du répresseur et (ii) de l'inhomogénéité du milieu consécutif au fonctionnement même du système.

Malgré tous ses défauts, la modélisation par systèmes d'équations différentielles a été et est encore très utilisée car elle a fait ses preuves, notamment en physique, et que les méthodes alternatives, les modèles informatiques, ont longtemps soufferts de temps de calcul rédhibitoires.

2.2 Modélisation discrète

Pour pallier certaines des limitations des modélisations continues déterministes, d'autres méthodes calculatoires ont été inventées. Celles qui vont nous intéresser sont des méthodes de *simulation* : on va utiliser un système calculatoire *analogue*, en un certain sens, au système biologique étudié, qu'on va faire évoluer dans le temps à l'aide d'un programme informatique. Le système de simulation va *mimer* le comportement des molécules présentes dans le système biologique. Ce mimétisme peut-être plus ou moins poussé selon la méthode utilisée. Nous allons nous intéresser à deux types de modélisations discrètes stochastiques :

- une méthode qui suppose l'homogénéité des populations de molécules, la méthode de *Monte Carlo Cinétique* et son implémentation par le *SSA* (Stochastic Simulation Algorithm) de D.T. Gillespie [9] et ses améliorations [10].
- une méthode qui ne fait aucun présupposé sur l'homogénéité des populations de molécules et rend compte de leur localisation spatiale, une méthode *Entité-Centrée*, apparentée aux *Systèmes Multi-Agents*.

Algorithme de simulation stochastique de D.T. Gillespie

La méthode SSA s'applique à un système de réactions chimiques où tous les réactants sont répartis de façon homogène dans un volume fixé et à température constante. De ce point de vue les conditions d'utilisabilité de cette méthode ressemblent à celles des équations différentielles, mais ajoute la prise en compte correcte de faibles concentrations, d'une part parce que c'est une méthode discrète et d'autre part parce que c'est une méthode stochastique.

C'est une méthode de simulation à *événements discrets* de réactions entre des molécules. Comme il a déjà été mentionné à la section 2.1, une réaction R_n, par exemple la réaction bi-moléculaire $A + B \longrightarrow C$, peut se produire quand une molécule de l'espèce A et une molécule de l'espèce B entrent en collision avec suffisamment d'énergie. La probabilité $P(n, dt)$ que la réaction R_n se produise pendant l'intervalle de temps infinitésimal dt est proportionnelle à la durée de l'intervalle de temps dt, au nombre de combinaisons moléculaires h_n pouvant causer la réaction (ici $h_n = \alpha[A][B]$) et à la cinétique c_n propre à cette réaction R_n. On a donc $P(n, dt) = h_n c_n dt$ où le produit $a_n = h_n c_n$ est appelé la *tendance*, ou *propension* de la réaction R_n.

Le système à simuler met en jeu N espèces moléculaires $\{S_1, \ldots, S_N\}$ représentées par un vecteur d'état dynamique $X(t) = (X_1(t), \ldots, X_N(t))$ où $X_i(t)$ est le nombre de molécules de l'espèce S_i dans le système à l'instant t, et M réactions chimiques $\{R_1, \ldots, R_M\}$. Chaque réaction R_j est caractérisée par sa propension a_j et un vecteur de changement d'état $\nu_j = \{\nu_{1j}, \ldots, \nu_{Nj}\}$, où ν_{ij} est la variation du nombre de molécules de l'espèce S_i due à la réaction R_j.

Soit $p(\tau, j)$ la probabilité qu'étant donné l'état $X(t) = (X_1(t), \ldots, X_N(t))$ du système à l'instant t, la prochaine réaction soit R_j et qu'elle arrive pendant l'intervalle de temps infinitésimal $[t + \tau, t + \tau + d\tau[$. D.T. Gillespie a montré dans [9] que :

$$p(\tau, j) = a_j e^{-a_0 \tau}$$

où $a_0 = \sum_{k=1}^{k=M} a_k$ est la propension combinée de toutes les réactions du système.

En faisant la somme des probabilités pour les M réactions possibles on obtient la probabilité qu'une quelconque réaction se produise dans l'intervalle $[t + \tau, t + \tau + d\tau[$

$$pr(\tau) = \sum_{k=1}^{k=M} p(\tau, k) = \sum_{k=1}^{k=M} a_k e^{-a_0 \tau} = a_0 e^{-a_0 \tau}$$

L'algorithme calculant l'état du système au cours du temps consiste à initialiser l'état du système : $X(t_0) = (X_1(t_0), \ldots, X_N(t_0)), t = t_0$ et à itérer les étapes suivantes jusqu'à ce que $t > t_{stop}$:

1. partant de l'état du système au temps t, $X(t)$, déterminer le numéro j de la prochaine réaction et le temps τ à attendre pour qu'elle se produise.

2. mettre à jour l'état en fonction du résultat de la réaction :
   $R_j : X(t+\tau) = X(t) + \nu_j$
3. faire avancer le temps : $t = t + \tau$

La méthode *directe* pour calculer le pas de temps τ et le numéro de la réaction qui est déclenchée est fondée sur la probabilité conditionnelle $P(j|\tau)$ que la prochaine réaction soit R_j sachant qu'elle est déclenchée à $t + \tau$:

$$p(\tau, j) = pr(\tau).P(j|\tau)$$

soit :

$$P(j|\tau) = p(\tau, j)/pr(\tau) = a_j/a_0$$

Le pas de temps τ et le numéro de réaction j sont obtenus en utilisant les règles d'inversion : on tire deux nombres aléatoires r_1 et r_2 uniformément distribués dans l'intervalle $[0, 1]$, τ est donné par :

$$\tau = \frac{1}{a_0} ln \left(\frac{1}{r_1} \right)$$

l'indice j de la réaction sélectionnée est le plus petit entier de l'intervalle $[1, M]$ tel que :

$$\sum_{k=1}^{k=j} a_k > r_2 a_0$$

Cet algorithme donne la trajectoire exacte du système, mais est très coûteux en temps calcul car il doit simuler chaque événement (le déclenchement de chaque réaction) et doit recalculer les propensions à chaque itération, car elles sont fonction des concentrations des réactants qui ont pu changer au cours de l'étape courante.

Simulation stochastique approchée, *tau-leaping*

Le SSA étant une méthode exacte qui effectue beaucoup de calculs, il est clair que pour l'accélérer il va falloir tirer parti d'une approximation pour réduire ce nombre de calculs. La méthode du *tau-leaping* est l'une des améliorations les plus connues du SSA.

L'idée de base du tau-leaping est de faire avancer le temps d'une durée τ prédéterminée, durée pendant laquelle plusieurs réactions vont s'effectuer. Pour faire cela de façon précise, on doit choisir τ suffisamment petit pour qu'aucune propension ne change "notablement" durant ce temps. Si c'est le cas, étant donné l'état x du système au temps t, le nombre de réactions R_j qui se seront faites dans l'intervalle $[t, t + \tau[$ sera approximativement $\mathcal{P}_j(a_j(x)\tau)$, la variable aléatoire de Poisson de moyenne $a_j(x)\tau$.

Cela conduit à la formule fondamentale de mise à jour :

$$X(t+\tau) = x + \sum_{j=1}^{j=M} \mathcal{P}_j(a_j(x)\tau)\nu_j$$

Des considérations pratiques font que ce n'est pas si simple à implémenter. Un premier problème est de choisir la valeur de τ la plus grande, telle que les propensions ne varient pas de plus qu'une tolérance donnée. Un deuxième problème est d'assurer qu'aucune population de réactants ne devienne négative durant un saut.

Cette méthode fonctionne bien quand l'échelle de temps de toutes les réactions est du même ordre de grandeur. Dans le cas contraire, la méthode reste très lente car on doit choisir une valeur de τ compatible avec la réaction la plus rapide.

Simulation stochastique *entité-centrée*

Les méthodes *entité-centrée* sont similaires aux systèmes multi-agents en ce sens que chaque molécule est une sorte d'agent dont le programme calcule l'évolution dans le temps et dans l'espace en fonction de ses caractéristiques propres (l'espèce chimique), de ses interactions avec les autres agents et l'environnement, principalement la membrane qui délimite le compartiment qui la contient.

À la différence des modèles à équations différentielles, qui sont des approches *top-down* où c'est le fonctionnement macroscopique qui est modélisé, les approches entité-centrée sont *bottom-up* : c'est le fonctionnement microscopique qui est modélisé et le comportement macroscopique *émerge* des interactions à l'échelle moléculaire.

Dans ces systèmes, les entités représentent les molécules, les différentes espèces chimiques sont codées par le *type* d'entité, celui-ci en déterminant le comportement. Du point de vue conceptuel, une entité est un *objet* au sens de la programmation objet, auquel est associé un *processus* qui l'anime. Les attributs communs à toutes les entités comprennent, entre autres, leur position dans l'espace et leur type.

La principale différence avec les systèmes multi-agents est que tous les types d'entités ont le même comportement, certes paramétré par leur type, mais identique. C'est une simplification dont on va pouvoir tirer parti pour obtenir une implémentation efficace du système. Ce comportement commun comporte deux volets :

1. **diffusion** : chaque molécule diffuse, dans le compartiment où elle se trouve, selon un mouvement *brownien* fonction de la température, de la viscosité du milieu et de l'encombrement moléculaire.
2. **réaction** : selon l'espèce chimique de la molécule, du fait de sa diffusion et de l'occurrence d'une collision d'énergie suffisante avec une autre molécule de l'espèce adéquate, une réaction se produit et l'une, l'autre ou les deux molécules peuvent changer de type, disparaître, former un complexe, etc.

Dans le chapitre suivant nous allons décrire HSIM, un système de simulation stochastique hybride, entité-centrée d'une part et intégrant un algorithme de type SSA optimisé d'autre part, traitant les espèces chimiques très abondantes et réparties de façon homogène, pour lesquelles la méthode entité-centrée n'apporte rien de plus mais est coûteuse en temps calcul.

Le simulateur HSIM

3

3.1 Introduction

Les systèmes entité-centrés combinés avec des méthodes population-centrées ont été de nombreuses fois utilisés dans des applications liées à la biologie, comme par exemple l'étude de la croissance de tumeurs [11]. Ce qui c'est souvent fait, c'est l'implémentation spécifique d'un modèle biologique d'intéret sous forme d'un système multi-agents dédié, couplé ou non à une méthode globale.

D'autres ont utilisé des plateformes générales de systèmes multi-agents pour implémenter leur modèle comme par exemple NetLogo (`http://ccl.northwestern.edu/netlogo/`), SWARM (`http://savannah.nongnu.org/projects/swarm/`), Repast (`http://repast.sourceforge.net/`), etc.

Il existe des systèmes spécifiquement dédiés à la simulation de processus biologiques au niveau cellulaire et intra-cellulaire comme par exemple Smoldyn [12] pour modéliser la cinétique biochimique, ou plus récemment les travaux de Klann *et. al* sur un système de simulation hybride entité-centré et stochastique [13].

HSIM intègre quasiment l'ensemble de toutes les spécificités des systèmes existants, comme la possibilité d'avoir une simulation spatiale hybride entité-centrée / population-centrée *a la* Gillespie, d'être capable de générer le système d'équations différentielles correspondant au modèle d'entrée et d'appeler un solveur numérique, de faire un rendu spatial en temps réel avec son interface OpenGL et l'edition temps réel des courbes de concentrations des réactants.

HSIM a des spécificités supplémentaires comme la possibilité d'avoir des compartiments multiples, éventuellement imbriqués, d'avoir des types de molécules membranaires pouvant réagir différemment d'un coté et de l'autre de la membrane, de modéliser des assemblages macro-moléculaires et de pouvoir *suivre* individuellement ces assemblages le long de la simulation, et aussi de pouvoir spécifier la *géometrie* de ces assemblages (aléatoire, linéaire, hélicoïdaux). L'ensemble de ces caractéristiques permettent de modéliser de façon générique quasiment tous les mécanismes présents dans les cellules depuis les simples réactions biochimiques à la formation d'assemblages complexes, comme des nano tubes ou des nano vesicules, pouvant emprisonner des proteines, jusqu'au transport actif le long de microtubules.

3.2 Présentation générale

Le logiciel de simulation HSIM, pour *Hyperstructure SIMulator*, est un système de simulation qui reproduit les interactions entre biomolécules dans les compartiments d'une cellule virtuelle. Il a été initialement développé pour rendre compte de l'agrégation et dissociation de grands assemblages moléculaires dynamiques appelés *Hyperstructures* [14] qui permettent de structurer la cellule (principalement procaryote) et peuvent jouer un rôle important dans la régulation du métabolisme, de la division cellulaire, de la différenciation cellulaire, etc.

C'est un système stochastique discret où chaque molécule est représentée individuellement et dont le comportement est géré par des règles de réécritures stochastiques locales.

Parmi les caractéristiques principales de HSIM, outre celles de prendre en compte la formation et la dissociation de grands complexes protéiques, on trouve aussi celles de modéliser l'interaction entre la membrane plasmique de la cellule, les récepteurs membranaires et les diverses protéines qui participent à une voie de signalisation. Il permet enfin de modéliser des transporteurs actifs ou passifs entre compartiments.

HSIM permet donc de modéliser des processus intra-cellulaires complexes mettant en jeu beaucoup de types de protéines (enzymes, recepteurs membranaires, etc.) en utilisant un langage de description uniforme. La localisation spatiale de chaque molécule permet de rendre compte de divers phénomènes topologiques tels que l'apparition de structures auto-organisées, des oscillations spatiales ou des gradients de concentration.

Le modèle à étudier est exprimé sous forme règles représentant les réactions biochmiques ; pour cela un langage de description *ad hoc* a été développé. Dans une première phase, le simulateur compile la description du système à simuler dans une représentation interne, puis l'utilise pour calculer l'évolution dans le temps du système. Une représentation 3D de la cellule simulée est affichée ainsi que les courbes représentant les concentrations des diverses espèces moléculaires en fonction du temps. Quand des agrégats se forment, le programme les repère et suit leur évolution dans le temps. On peut donc d'une part suivre l'évolution du système en termes de concentrations aussi bien en phase initiale (transitoire) qu'en phase stationnaire, si le système se stabilise, et d'autre part suivre cette évolution en termes de localisation topologique des macromolécules dans la cellule.

HSIM a été étendu depuis les premières versions, il a été calibré et comparé à d'autres systèmes de simulation tels que les équations différentielles ordinaires, les méthodes stochastiques globale type Monte Carlo, etc. et donne les mêmes résultats quand on le restreint aux les mêmes modèles et paramètres. Parmi les évolutions récentes on a ajouté la prise en compte de très petites molécules qui, étant présentes en très grande quantité et diffusant très rapidement dans les compartiments, sont considérées comme étant en concentration homogène et simulées de façon globale à l'aide d'une variante très efficace de l'algorithme de

Gillespie que nous avons développée. Cette méthode hybride permet de rendre compte efficacement en termes de temps de calcul de très grands nombres de molécules tout en conservant la possibilité de localisation spatiale des protéines.

3.3 Fondements physiques

Pour estimer la faisabilité de cette approche il convient en premier lieu d'avoir une idée des échelles de taille des cellules et de leurs constituants. Les cellules ont des tailles qui sont de l'ordre du micron ($1\mu = 10^{-3}mm = 10^{-6}m$) :

- *Escherichia coli*, un batonnet de $.65\mu$ de diamètre par 3μ de long environ avant division.
- *Saccharomyces cerevisiae*, un sphéroide de 5μ à 10μ de diamètre
- La plupart des cellules humaines font 20μ de diamètre avec un noyau d'environ 10μ de diamètre.
- Certaines cellules eucaryotes peuvent être beaucoup plus grandes, par exemple l'axone d'un neurone peut avoir une longueur de l'ordre du mètre...

Beaucoup de biomolécules sont faites d'unités chimiques (nucléotides, acides aminés, etc.) d'une taille de l'ordre d'un nanomètre ($1nm = 10^{-3}\mu = 10^{-9}m$), d'autres *petites molécules* (lipides, sucres, etc.) sont aussi de cet ordre de taille. La plupart des protéines globulaires ont un diamètre de quelques nanomètres à 10 nanomètres. La capacité des biomolécules à s'agréger font qu'elles peuvent, à partir d'unités de quelques nanomètres, former de très long polymères (ADN, filaments d'actine, microtubules, etc.).

Diffusion et mouvement brownien

L'un des comportements que doit implémenter le simulateur est la diffusion des molécules. Si on observe des petites particules (d'une taille de l'ordre $100nm$ à 1μ) dans de l'eau, on voit qu'elles ont une trajectoire erratique, quasiment *sautant* d'un endroit à l'autre de façon discontinue. Cette caractéristique, le *mouvement brownien*, est due aux collisions avec les molécules d'eau, qui fait que les particules adoptent une *marche aléatoire*.

Quand on observe une de ces particules pendant un intervalle de temps t on la voit se déplacer selon un vecteur r. Si on répète cette observation un grand nombre de fois on peu calculer le déplacement moyen durant la période t. On se rend compte alors que $< r >= 0$, c'est-à-dire qu'il n'y a pas de direction préférée (la notation $<>$ représente la moyenne de beaucoups de mesures non corrélées). Le mouvement brownien est caractéristisé non pas par la moyenne du déplacement, mais par la moyenne du déplacement au carré :

$$< |r|^2 >= 6Dt$$

où D est la constante de diffusion de la particule. Cette formule est connue sous le nom de *loi de la racine carrée* du mouvement brownien. Le facteur 6 correspond à la dimensionnalité, ici 3 dimensions (il vaut 4 pour une diffusion dans le plan

et 2 pour une diffusion linéaire). La constante de diffusion a les dimensions d'une longueur au carré par unité de temps ($m^2 s^{-1}$).

Cette loi ainsi que d'autres caractéristiques du mouvement brownien ont été établies initialement par Robert Brown [15] en 1828, beaucoup de travaux quantitatifs ont été faits par Jean Perrin au début du $XX^{\text{ième}}$ siècle.

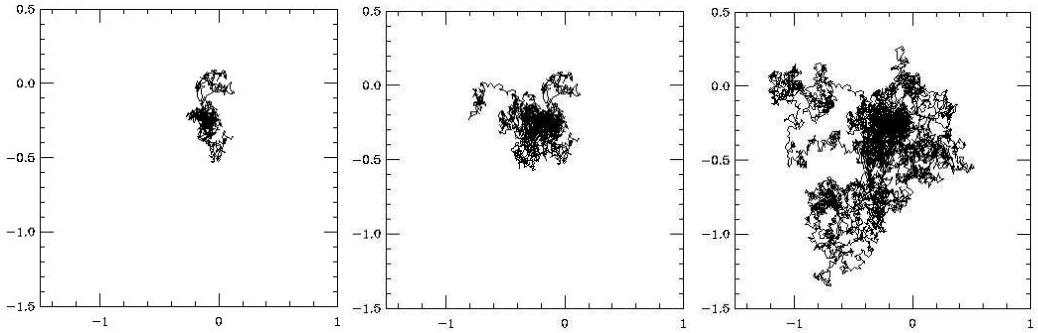

FIG. 3.1: Trajectoire simulée d'une particule suivant un mouvement brownien projeté sur un plan. Avec $D = 0.16\mu^2/sec$, les axes gradués en microns, la position à $t = 0$ est $(0, 0)$. Les figures sont prises après 1 seconde, 3 secondes et 10 secondes.

La constante de diffusion est fonction de la taille et de la forme de la particule, de la viscosité et de la température du fluide dans laquelle elle se déplace. En 1905 A. Einstein [16] à montré que la constante de diffusion D d'une particule était simplement reliée à son coefficient de friction et à la température absolue :

$$D = \frac{k_B T}{f}$$

avec k_B la constante de Boltzmann ($1.38065 \times 10^{-23} J/K$), T la température absolue et f le coefficient de friction. Cette formule est quelquefois appelée *relation d'Einstein* ou *formule de fluctuation-dissipation*.
Pour une particule sphérique le coefficient de friction est minimal et vaut : $f = 6\pi\eta r$, avec η la viscosité du milieu et r le rayon de la sphère.
Dans ce cas la constante de diffusion est :

$$D = \frac{k_B T}{6\pi\eta r}$$

Marche aléatoire et mouvement brownien

Nous allons montrer ici que la marche aléatoire en trois dimensions correspond exactement au mouvement brownien car elle obéit à la loi de la racine carrée.

Supposons qu'à chaque intervalle de temps τ une particule se déplace linéairement d'une distance a dans une direction aléatoire. Alors le déplacement après

N étapes sera :

$$r = a\overrightarrow{n_1} + a\overrightarrow{n_2} + \ldots + a\overrightarrow{n_N}$$

où $\overrightarrow{n_i}$ sont des vecteurs unitaires ($\|\overrightarrow{n_i}\| = 1$) de direction aléatoire. Donc le carré de la distance parcourue par notre particule en N étapes est :

$$|r|^2 = a^2 \left(\overrightarrow{n_1} + \overrightarrow{n_2} + \ldots + \overrightarrow{n_N}\right)^2$$

soit :

$$|r|^2 = a^2 \left(|\overrightarrow{n_1}|^2 + |\overrightarrow{n_2}|^2 + \ldots + |\overrightarrow{n_N}|^2 + \sum_{i \neq j} \overrightarrow{n_i}.\overrightarrow{n_j} \right) = Na^2 + a^2 \sum_{i \neq j} \overrightarrow{n_i}.\overrightarrow{n_j}$$

Si maintenant on répète un grand nombre de fois cette marche aléatoire de N pas, on trouve que la moyenne des produits scalaires est nulle, $< \overrightarrow{n_i}.\overrightarrow{n_j} >= 0$, puisque les déplacements élémentaires se font dans des directions aléatoires. D'où

$$< r^2 >= Na^2 \tag{3.1}$$

Le temps qui s'est écoulé après N étapes est $t = N\tau$, d'où en remplaçant N par t/τ dans l'équation 3.1 :

$$< r^2 >= \frac{a^2}{\tau} t$$

Rappelons nous que le mouvement brownien dans l'espace est caractéristisé par la loi de la racine carrée, $< r^2 >= 6Dt$, cela démontre que cette marche aléatoire définit un mouvement brownien de constante de diffusion $D = a^2/6\tau$.

Dans HSIM, la diffusion des molécules sera donc implémentée par une marche aléatoire dans l'espace. On a choisi le pas de temps et la valeur du saut de telle façon à mimer correctement la diffusion observée de protéines dans une bactérie. La constante de diffusion qui en résulte est bien plus faible que celle d'une protéine diffusant dans de l'eau pour intégrer la viscosité due à l'encombrement moléculaire (fig. 3.2) du cytosol d'une bactérie.

3.4 Fonctionnement de HSIM

Les entités dans HSIM représentent les molécules, ces molécules sont assimilées à des sphères dont le centre indique la position dans l'espace. Les attributs qui caractéristisent une molécule sont :

- sa position (x, y, z) dans l'espace
- son diamètre.
- son type, codant son espèce chimique
- sa liste de liens avec les autres molécules d'un complexe
- si elle est membranaire ou cytoplasmique

ainsi que de nombreux autres attributs de nous ne détaillerons pas ici.

Ces molécules évoluent dans une cellule virtuelle en forme de batonnet, en fait un cylindre terminé par une demi-sphère à chaque extrémité, délimitée par une

membrane. Selon les dimensions données dans le modèle, la cellule peut-être de forme allongée ou sphérique. Ce compartiment principal peut contenir des sous-compartiments, imbriqués ou pas et de même morphologie.

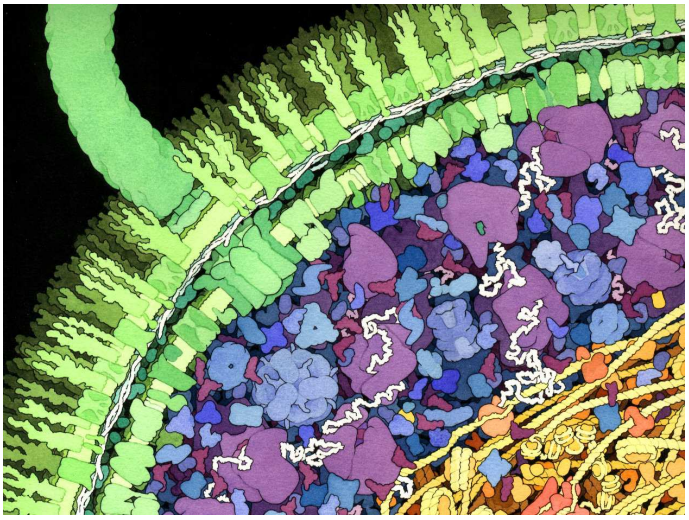

FIG. 3.2: Représentation de l'encombrement moléculaire du cytosol d'une bactérie, *E. coli.* (origine : *The Machinery of Life*, David S. Goodsell)

Comme nous l'avons dit précédemment, la diffusion est implémentée avec une marche aléatoire, le pas de temps est de $100\mu s$, soit le temps moyen pour une protéine de $10nm$ de diamètre de diffuser de son diamètre dans le cytosol d'une cellule. Les molécules, de par leur diffusion, peuvent entrer en collision soit avec la membrane, soit avec d'autres molécules. Ces collisions sont considérées comme des interactions locales et peuvent avoir comme conséquence un changement de type des molécules, leur assemblage, l'adsorption d'une molécule par la membrane, etc. Ces conséquences possibles sont spécifiées dans le modèle à simuler à l'aide de règles de réécriture probabilistes.

Ces règles ont la forme générique suivante :

$$M_A + M_B \rightarrow M_C + M_D \quad [p_r]$$

M_A et M_B sont deux molécule de type A et B qui quand elles entrent en collision, déclenchent la réaction qui transforme M_A en M_C et M_B en M_D avec la probabilité p_r. Cette probabilité code le fait que le choc doit avoir assez d'énergie pour déclencher la réaction. Son interprétation macroscopique est directement corrélée à la cinétique de la réaction. On peut aussi omettre la deuxième molécule à gauche pour modéliser une réaction mono-moléculaire, et/ou à droite pour indiquer que la deuxième molécule de la partie gauche est soit détruite, soit agrégée à la molécule restante. On peut donc par exemple modéliser une réaction

enzymatique Michaelienne avec les trois règles suivantes :

$$\begin{aligned} E + S \rightarrow &\quad ES && [p_f] \\ ES \rightarrow &\quad E + S && [p_r] \\ ES \rightarrow &\quad E + P && [p_c] \end{aligned}$$

La première règle change le type de l'enzyme E en ES et détruit la molécule S. C'est raisonnable car le substrat est beaucoup plus petit que l'enzyme et le complexe est du même ordre de taille que l'enzyme seule. La deuxième règle implémente l'opération inverse, la *molécule* ES représentant le complexe reprend le type de l'enzyme E et une nouvelle molécule de type S est synthétisée à proximité de E. La troisième règle est semblable à la deuxième, mais synthétise le produit P.
Nous verrons dans la section suivante comment faire le lien entre ces probabilités et les paramètres cinétiques standard (K_m et K_{cat}).

Calcul de l'évolution du système

HSIM modélise donc l'espace de façon continue, les coordonnées des molécules étant codées par des nombres réels, et le temps de façon discrète, il est découpé en tranches de $100\mu s$ pendant lequel un pas de diffusion est effectué. En ne tenant compte pour l'instant que des molécules implémentées par des *entités*, en ignorant celles implémentées de façon globales, l'algorithme de calcul de l'évolution du système est le suivant.

À chaque pas de temps, ou *itération*, toutes les molécules sont examinées une et une seule fois dans un ordre aléatoire pour éviter d'introduire des biais statistiques et simuler correctement un processus par entité. Examiner la molécule S, consiste premièrement à vérifier qu'elle n'a pas déjà été examinée précédemment au cours de la même itération, puis choisir de façon aléatoire une direction où elle va tenter de diffuser. Si le long de cette direction elle ne rencontre ni d'autres molécules, ni la membrane elle va y être déplacée (marche aléatoire). Si par contre, elle entre en collision avec une autre molécule, T, qui n'est pas déjà marquée, le programme va chercher à appliquer une éventuelle règle de réaction entre S et T. Ensuite T est marquée comme étant examinée (i.e. S et T ne seront plus jamais considérées avant la prochaine itération) et le programme continue avec une autre molécule jusqu'à ce qu'elles aient été toutes examinées. Puis le temps est avancé d'un pas, $100\mu s$, et la prochaine itération peut commencer (fig. 3.3).

Comme nous l'avons mentionné précédemment, HSIM est un simulateur hybride : en plus des molécules traitées individuellement par des entités, il inclut des espèces qui sont traitées de façon globale. Ce sont fréquemment des petites molécules qui sont présentes en très grand nombre de copies et qui par conséquent sont statistiquement distribuées de façon homogène dans le compartiment. Ces classes de molécules, que nous appellerons *non-entité*, seront représentées uniquement par leur type et leur nombre de copies dans le compartiment. Pour

ces non-entités, il est plus intéressant en terme de temps calcul de les traiter de façon globale en utilisant une méthode semblable à celle de Gillespie.

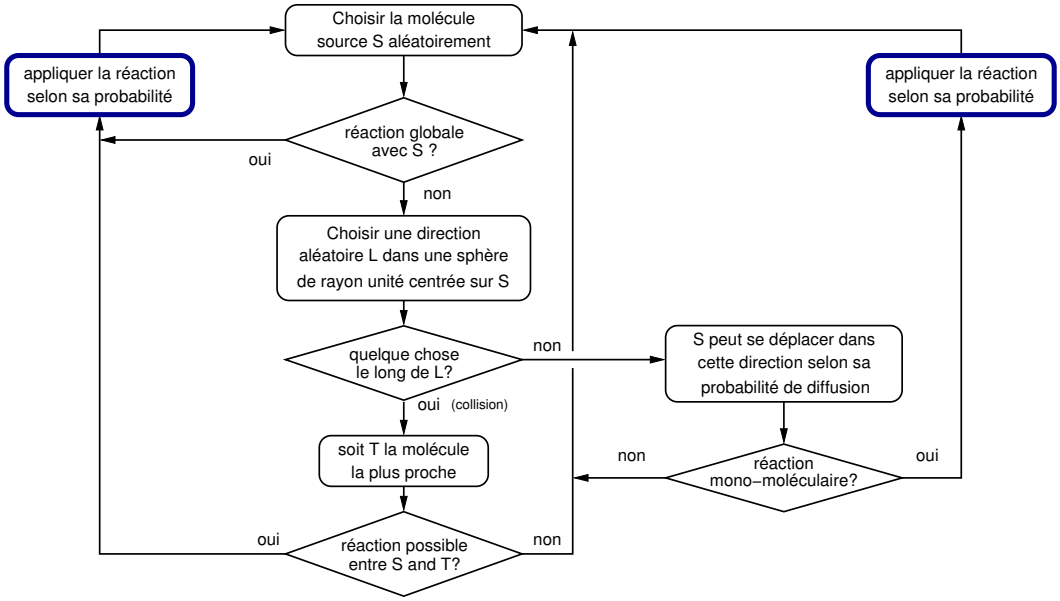

FIG. 3.3: Algorithme de calcul de l'évolution de l'état du système à chaque pas de temps.

Cela ajoute à l'algorithme décrit précédemment deux points :

1. lors de l'examen d'une molécule-entité, si elle peut réagir avec espèces non-entités, on va calculer le nombre moyen de collisions avec celles-ci pendant le pas de temps et appliquer les réactions éventuelles.
2. à chaque itération, après l'examen de toutes les molécules-entités, on va traiter les interactions spécifiques entre molécules non-entités. C'est-à-dire calculer le nombre moyen de leurs collisions pendant le pas de temps et appliquer les réactions correspondantes.

Il est à noter qu'avec cette méthode, il est possible d'avoir des réactions entre entités, entre entité et non-entité, et aussi entre non-entités. La syntaxe de description des réactions est identique quelles que soient les classes de molécules mises en jeu dans les réactions.

Dans l'exemple de réaction enzymatique précédent, les trois règles sont strictement les mêmes quand par exemple on a décidé :

- que les enzymes E et les complexes enzymes-substrat ES seront modélisés par des entités et les substrat S et produit P par des non-entités
- de tout implémenter avec des entités.
- de tout implémenter avec des non-entités.

Calcul du nombre moyen de collisions

Pour calculer le nombre moyen de collisions pendant un pas de temps entre une molécule-entité et une espèce de classe non-entité, il suffit de calculer le volume

parcouru par la molécule-entité pendant le pas de temps et de calculer le nombre moyen de copies de l'espèce non-entité qui se trouve dans ce volume. Comme par définition les espèces non-entité sont homogènes dans le compartiment, il suffit de rapporter le nombre total de copies dans le volume du compartiment au volume parcouru par la molécule-entité.

Il ne reste plus ensuite, pour chacune des collisions, qu'a appliquer ou pas la réaction selon sa probabilité en tirant un nombre aléatoire dans l'intervalle $[0-1]$ et en le comparant à cette probabilité.

Le nombre moyen de collisions entre espèces non-entités est encore plus simple à calculer puisqu'il ne dépend uniquement que du produit des concentrations de ces deux espèces. La constante qui multiplie ce produit de concentrations a été choisie de façon à ce que ce nombre de collisions soit statistiquement identique à celui produit par des molécules-entités dans les mêmes conditions.

Bien sûr, ces nombres moyens de collisions entre non-entités et entités ou non-entités, sont des nombres *réels*, il n'y a aucune raison que compte tenu de la valeur du pas de temps, du volume parcouru par l'entité, du volume du compartiment et du nombre (entier) de copies des non-entités, le nombre moyen de collisions calculé soit entier. La partie entière de ce nombre est prise en compte comme candidats potentiels à la réaction, alors que la partie fractionnaire est stochastiquement promue au rang de collision.

Cette façon de faire a comme conséquence de diminuer la stochasticité du calcul quand le nombre de réactions pendant le pas de temps est élevé, puisque la partie entière de ce nombre de réactions étant elle aussi élevée, ces réactions sont appliquées systématiquement. Cela conduit à un comportement auto-adaptatif de l'algorithme qui est de plus en plus stochastique quand les concentrations sont faibles et de plus en plus moyen quand les concentrations sont fortes.

Cet algorithme est très efficace car il n'a pas à s'adapter sur les réactions rapides : elles sont automatiquement approximées par un calcul *en moyenne*.

Types de réactions

Au début de cette section, nous avons donné une forme générique des réactions que traite HSIM, en fait HSIM en fait plus car on peut spécifier l'assemblage de molécules-entités, éventuellement selon une certaine géométrie, la dissociation d'un assemblage et une réaction spécifique à deux molécules assemblées (qui serait différente, voire n'existerait pas si elles étaient libres). Ces types de réactions ne s'appliquent bien sûr qu'aux espèces modélisées par des entités. Leur syntaxe en est la suivante :

$$
\begin{array}{lrcll}
(1) & M_A + M_B & \rightarrow & M_C * M_D & [p_r] \\
(2) & M_C * M_D & \rightarrow & M_A + M_B & [p_r] \\
(3) & M_A + M_B & \rightarrow & M_C = M_D & [p_r] \\
(4) & M_A + M_B & \rightarrow & M_C / M_D & [p_r] \\
(5) & M_A * M_B & \rightarrow & M_{A'} * M_{B'} & [p_r]
\end{array}
$$

La réaction (1) indique que les molécules M_A et M_B initalement non liées entre elles vont former un complexe avec la probabilité p_r, en changeant de type : M_A devient de type C et M_B de type D. La réaction (2) est sa réciproque, le complexe M_C-M_D se défait avec la probabilité p_r alors que les molécules reprennent les types A et B respectivement. Les réactions (3) et (4) sont des variantes de la réaction (1) où la molécule M_B s'assemble avec M_A de façon colinéaire pour (3) et hélicoïdale pour (4) avec l'éventuelle molécule M_X à laquelle serait déjà reliée M_A. Ces types d'assemblages sont particulièrement utiles pour modéliser des polymères en forme de filaments (fig. 3.4). Les paramètres, diamètre et pas de l'hélice, des assemblages hélicoïdaux sont une propriété caractéristique de la protéine M_B et sont spécifiés dans le modèle.

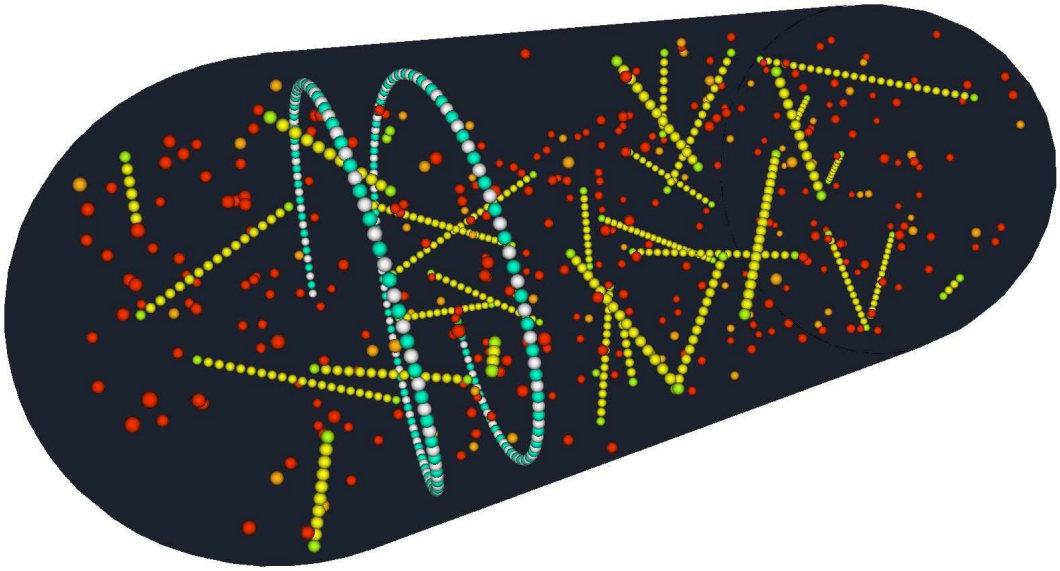

FIG. 3.4: Assemblage de filaments linéaires et hélicoïdaux.

Les molécules-entités sont de deux natures (classes) cytosoliques ou membranaires. Jusqu'à présent on n'a implicitement parlé que des molécules cytosoliques. En fait, tout ce qui s'applique aux unes peut s'appliquer aux autres, les molécules membranaires ayant cela de plus qu'on peut indiquer si elles sont *actives* seulement à l'intérieur du compartiment, seulement à l'extérieur ou bien des deux côtés de la membrane. Cette propriété est une caractéristique de l'espèce moléculaire qui est spécifiée dans le modèle avec la déclaration de son type.

Cette possibilité d'indiquer où une molécule membranaire est active permet d'implémenter très facilement des transporteurs. On veut par exemple, modéliser le transport actif du glucose chez *E. coli* par le récepteur membranaire EII^{BC}, sachant que le glucose importé est transformé en glucose-6-phosphate et que le transporteur n'est actif que lorsqu'il est phosphorylé (fig. 3.5).

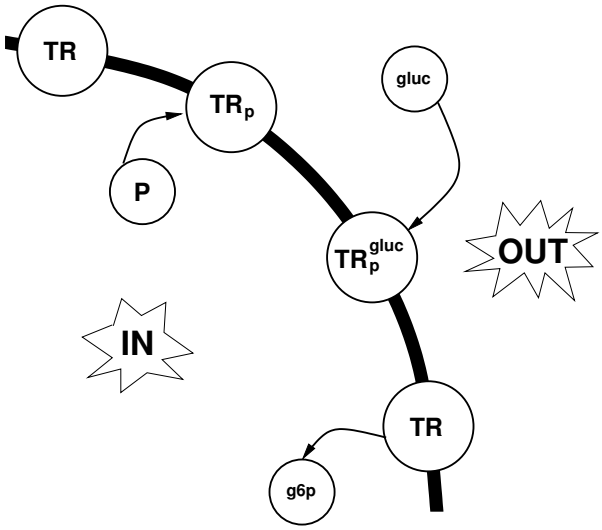

FIG. 3.5: Transporteur membranaire actif EII^{BC} chez *E. coli*.

On va utiliser trois variantes du récepteur membranaire : TR le récepteur non phosphorylé qui est actif à l'intérieur du compartiment, TR_p le récepteur phosphorylé, actif à l'extérieur du compartiment et TR_p^{gluc} le récepteur phosphorylé qui a lié le glucose externe, actif à l'intérieur du compartiment, ligne (1). La ligne (2) déclare les molécules de glucose et de glucose-6-phosphate. La ligne (3) est la réaction qui phosphoryle le récepteur, avec la cinétique associée à la probabilité p_1. La ligne (4) représente la réaction qui fixe le glucose extérieur sur le récepteur phosphorylé. La ligne (5) libère le glucose-6-phosphate dans le compartiment et remet le récepteur dans son état initial pour permettre un nouvel import de glucose.

$$
\begin{array}{ll}
\text{(1)}\ membrane\ TR <, TR_p >, TR_p^{gluc} <; & \\
\text{(2)}\ molecule\quad gluc, g6p; & \\
 & \\
\text{(3)}\ TR + P & \rightarrow TR_p \qquad [p_1] \\
\text{(4)}\ TR_p + gluc & \rightarrow TR_p^{gluc} \qquad [p_2] \\
\text{(5)}\ TR_p^{gluc} & \rightarrow TR + g6p \quad [p_3]
\end{array}
$$

3.5 Conclusion

HSIM est un outil très performant en terme de ressources de calcul : aussi bien son algorithme pour traiter les entités que son algorithme stochastique sont les plus rapides à ce jour. De plus il permet de mixer les deux types de résolutions et de profiter des avantages des deux méthodes sans en supporter les inconvénients (localisation spatiale et rapidité de calcul), cela de façon transparente pour l'utilisateur.

Ces travaux ont été présentés lors de nombreux séminaires :

- 2003 : à l'université de Californie à San Diego, au laboratoire de biologie
- 2005 : à l'université de Bordeaux 1, au LABRI
- 2007 : au CGM (Centre de Génétique Moléculaire) à Gif sur Yvette
- 2008 : à l'INRA de Jouy-en-Josas
- 2008 : à l'université de Versailles, au PRISM
- 2010 : au laboratoire I3S à Sophia Antipolis
- 2009 et en 2011 : au laboratoire SysDiag à Montpellier.

et workshops :

- 2006 : orateur invité à la journée satellite *Multi Agents Systems and molecular biology* de la conférence IPG (Integrative Post-Genomics) à Lyon.
- 2009 : orateur invité au *Workshop on MAS in Biology at meso or macroscopic scales* à Paris.
- 2012 : orateur invité au séminaire de la Société Francophone de Biologie Théorique

ainsi qu'à la conférence internationale *Static Analysis and Systems Biology* [17] et publiés dans le *Journal of Biological Physics and Chemistry* [18] et dans *BMC Systems Biology* [19].

Systèmes Complexes 4

4.1 Introduction

Une des définitions d'un système complexe est celle-ci : *un système complexe est un système composé de nombreux acteurs, interagissant les uns avec les autres de manière multiple et variée.* Ces interactions produisent une dynamique non linéaire. Les modifications continuelles des éléments d'un système complexe, ainsi que leurs interactions, rendent son évolution difficile à prédire et à expliquer.

A ce titre, l'ensemble des biomolécules intervenant dans un organisme vivant en fait un système complexe. D'un certain point de vue, ce qu'on en observe au niveau de l'organisme est la résultante des interactions entre ces biomolécules. Le comportement global de l'organisme, son phénotype, *émerge* des comportements locaux des acteurs qui le composent.

En se restreignant à des espèces simples, les bactéries par exemple, on peut dire qu'au niveau d'un individu-bactérie, l'ensemble des biomolécules qui le constitue est en majeure partie fixé par son génome, et ce sont les interactions entre ces biomolécules qui définissent le phénotype de cette bactérie pendant sa durée de vie.

Si on se place au niveau de l'espèce bactérienne et sur un intervalle de temps beaucoup plus long, on peut aussi dire que le génome de l'espèce possède son propre phénotype qui évolue en fonction de l'environnement dans lequel les colonies sont placées. On pourrait tout à fait considérer que le génome de chaque individu-bactérie est un acteur qui possède des interactions nombreuses et variées avec le génome des autres bactéries de la colonie et avec l'environnement. A ce moment, la colonie devient un système complexe à un niveau d'échelle supérieur.

L'une des propriétés caractéristiques du vivant est son grand potentiel d'adaptation au milieu ou il évolue, à ce titre, le fait de considérer le vivant comme un système complexe est tout à fait approprié.

On peut définir un phénomène émergent comme étant une conséquence macroscopique des interactions microscopiques des acteurs d'un système complexe donné. L'un des points très important à noter est que le phénomène émergent est

à un niveau d'échelle supérieur aux entités du système. Il définit souvent de nouvelles entités ayant des comportements et des interactions propres, introduisant un niveau de structuration au delà des composants initiaux. Ces phénomènes ne peuvent être déduits de façon évidente du comportement individuel des différentes entités du système et se produisent sans qu'aucun élément extérieur n'intervienne.

On s'est donc tout naturellement intéressé à quantifier ce que serait un phénomène émergent et à fournir des moyens automatiques de détection de ces phénomènes à partir de traces de simulations de systèmes complexes donnés. C'est la majeure partie du sujet de la thèse de Thomas Moncion, que j'ai co-encadrée avec Guillaume Hutzler et qui a été soutenue en décembre 2008.

4.2 Cytosquelette et interactions avec la membrane

L'un des premiers modèles qui nous a servi d'étude de cas pour HSIM a été l'auto assemblage et la désagrégation de filaments d'actine et de microtubules dans une cellule. On sait que ces filaments entrent dans la composition du cytosquelette de la cellule ; il interagit avec la membrane en la stabilisant [20]. Ces filaments, qui se forment spontanément dès que la concentration d'actine libre est suffisamment élevée, sont *polarisés* : ils n'ont pas les mêmes cinétiques de polymérisation / dépolymérisation à chaque bout. Le réseau d'actine est dynamique en ce sens qu'il est en perpétuel renouvellement, chaque filament le constituant gagnant et perdant des monomères à chaque extrémité.

La capacité de HSIM à prendre en compte la géométrie des assemblages de protéines et à les localiser dans l'espace nous a permis d'utiliser un modèle biochimique très simple de la polymérisation de l'actine et d'en faire émerger par simulation un phénomène inattendu, mais complètement explicable *a posteriori*.

Dans notre expérience virtuelle, le compartiment est initiallement rempli avec de l'actine globulaire libre qui va rapidement polymériser en des filaments. Le processus est réversible et les filaments peuvent aussi se dépolymériser, relachant ainsi des monomères d'actine libre. Bien entendu, ce processus nécessite de l'énergie, elle est fournie par la conversion d'ATP en ADP (qui n'est pas pris en compte dans cet modèle simplifié où on supposera présente de l'ATP en très grande quantité).

La vitesse de diffusion des filaments d'actine est très faible par rapport à celle des monomères libres ; toujours par souci de simplification, notre modèle la suppose nulle. De ce fait, l'orientation dans l'espace d'un filament est complètement déterminé par celle du dimère initial, qui est aléatoire puisqu'elle dépend de la position relative des deux monomères d'actine au moment du choc qui a formé le dimère.

Le modèle est défini avec seulement trois paires de règles ("$M_A * M_B$" signifiant que les deux molécules sont liées, et "$M_A = M_B$" indique un assemblage rectiligne).

(1) $Act_g + Act_g$	$\rightarrow M_e * P_e$	$[P_1^f]$	$M_e * P_e$	$\rightarrow Act_g + Act_g$	$[P_1^r]$
(2) $Act_g + P_e$	$\rightarrow P_e = Act_f$	$[P_2^f]$	$P_e * Act_f$	$\rightarrow Act_g + P_e$	$[P_2^r]$
(3) $Act_g + M_e$	$\rightarrow M_e = Act_f$	$[P_3^f]$	$M_e * Act_f$	$\rightarrow Act_g + M_e$	$[P_3^r]$

La première paire de règles modélise l'association / dissociation du dimère, la deuxième paire modélise la polymérisation / dépolymérisation au bout *plus* et la troisième paire, la même chose au bout *moins*. Pour tenir compte de la cinétique plus favorable à la polymérisation au bout plus qu'au bout moins, les probabilités sont choisies telles que : $P_2^f/P_2^r > P_3^f/P_3^r$. Les vitesses de diffusion de tous les éléments constituant un filament, M_e, P_e et Act_f, sont mises à zéro.

Après une période transitoire, les concentrations d'actine libre et liée atteignent un état d'équilibre et restent statistiquement stationnaires. La surprise vient de l'arrangement géométrique des filaments : ils deviennent plus ou moins alignés le long de l'axe principal du compartiment (fig. 4.1).

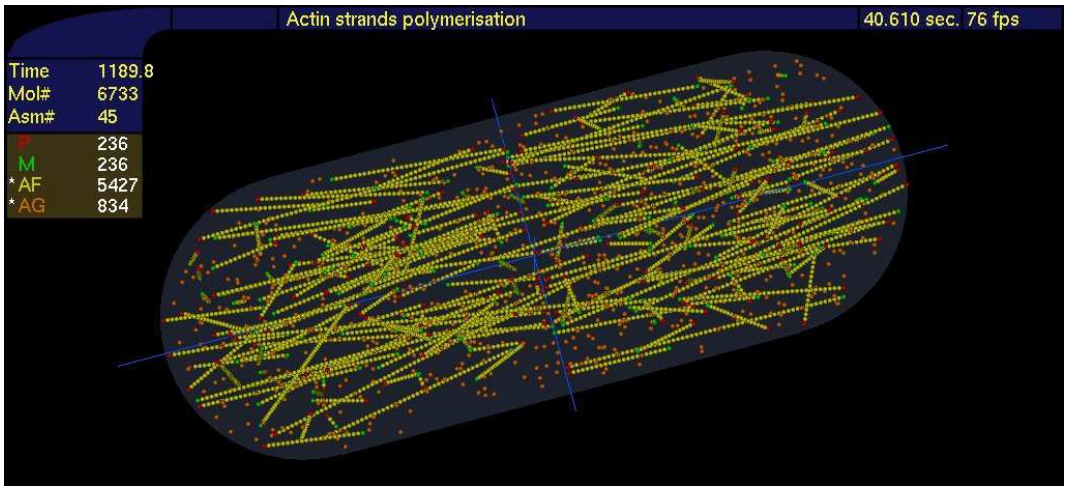

FIG. 4.1: Auto alignement de filaments d'actine le long de l'axe principal du compartiment.

Pourquoi cet alignement se fait-il, alors qu'on a absolument rien précisé dans le modèle ? C'est une simple conséquence de la presssion de sélection exercée sur les filaments courts. Ceux-ci sont moins résistants à une longue période de dépolymérisation que les longs, et disparaissent avant eux. Réciproquement, les filaments plus longs vont, dans les mêmes conditions, se raccourcir, mais rester, et donc leur orientation aussi. En conséquence, les filaments qui vont perdurer seront ceux qui sont orientés de façon à pouvoir grandir le plus, c'est-à-dire le long de l'axe principal du compartiment.

Ce phénomène émerge donc de la dissymétrie du compartiment, en effet, quand on refait une simulation du même modèle avec un compartiment sphérique, aucun alignement n'apparaît.

C'est cet exemple simple, et en soi complet, qui a initialement motivé mon interêt pour étudier des méthodes automatiques de détection de phénomènes émergents et proposer ce sujet de thèse à Thomas Moncion.

4.3 Construction et analyse d'un réseau de réactions

Lors de son stage de DEA et dans la première partie de sa thèse, nous avons étudiés avec Thomas Moncion des algorithmes permettant d'énumérer et de caractériser tous les réactifs mis en cause dans un ensemble de réactions biochimiques (i.e. un modèle pour HSIM) et d'exhiber le réseau de réactions correspondant.

La méthode consiste à créer un réseau de Petri correspondant aux réactions du modèle, en utilisant un algorithme glouton. En partant des espèces présentes initialement, modélisées comme les places du réseau, on va chercher à appliquer toutes les réactions mono-moléculaires possibles pour chaque espèce, ainsi que toutes les réactions bi-moléculaires possibles entre ces espèces. Chaque fois qu'on aura trouvé une réaction, on va la modéliser par une transition du réseau, cette transition permettant d'obtenir soit des molécules d'une espèce déjà existante, soit de nouvelles espèces. Ce processus est itéré tant qu'on peut trouver des réactions pouvant s'appliquer aux espèce moléculaires présentes.

Si l'ensemble des réactions du modèle n'est pas récurrent, ce qui est le cas de quasiment tous les réseaux métaboliques, cet algorithme terminera et fournira la liste (finie) des types moléculaires mis en jeu dans le modèle, ainsi que le réseau de Petri correspondant. Cela permet d'exhiber plusieurs caractéristiques du modèle :

1. montrer si le modèle étudié peut, même rarement, fabriquer des espèces moléculaires qui n'ont pas de sens biochimique (dans ce cas le modèle est probablement faux),
2. par l'étude des propriétés topologiques du réseau de Petri associé au modèle, on peut voir si certaines espèces ne sont que consommées, ou que produites, ou bien, en étudiant les invariants de place dans le réseau, montrer que la somme des concentrations de telles espèces reste constante.

Dans l'exemple du système de réactions de la figure 4.2, on peut constater que la place P_2, correspondant au substrat s_1 est une source et que la place P_9, correspondant à p_3 est un puit. On peut démontrer aussi que quelle que soit le marquage initial, la somme des jetons dans les places contenant les enzymes e_1 ou e_2 est constante. Traduit dans les termes du modèle, cela signifie que les concentrations d'enzymes (libres ou en complexes) restent constantes, et que le réseau métabolique correspondant à ces réactions consomme le métabolite s_1 et produit p_3.

4.4 Détection automatique de phénomènes émergents

Dans la deuxième partie de sa thèse, on s'est intéressés avec Thomas Moncion à analyser les interactions entre les composants d'un système complexe au cours de la simulation d'une trajectoire de ce système. On s'est focalisé sur les simulations

de type *entité-centrée* ou les acteurs sont soit les agents d'un système multi-agents, soit les entités représentant les molécules dans HSIM.

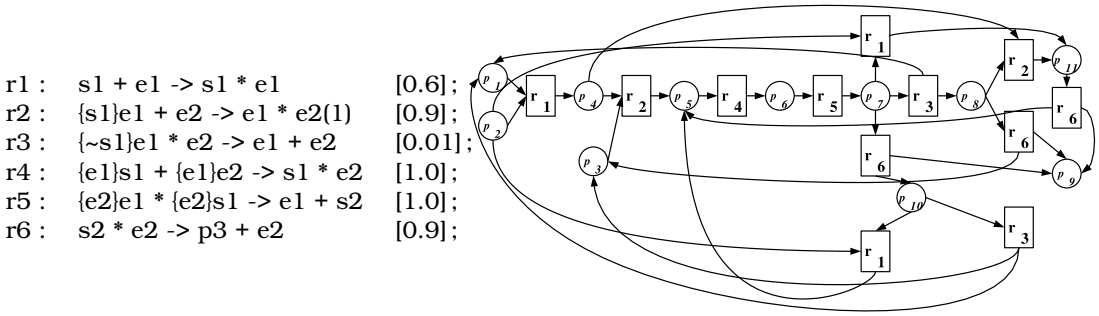

FIG. 4.2: Réseau métabolique montrant la catalyse par deux enzymes successives d'un substrat s_1 en produit p_3, la réaction intermédiaire étant canalisée. Dans le réseau de Petri à droite, les places P_1 à P_{11} représentent respectivement les espèces : e_1, s_1, e_2, e_1s_1, $e_1s_1e_2$, $e_1s_1e_2s_1$, $e_1e_2s_2$, e_2s_2, p_3, e_1e_2, et $e_1s_1e_2s_2$.

On a utilisé un réseau dynamique où les entités ainsi que leurs interactions sont représentées par l'intermédiaire d'un graphe.

Ce graphe évolue à chaque pas de temps en fonction de l'ensemble des données fournies par la simulation. Ne voulant pas faire d'hypothèses quant au type d'interactions impliquées, on s'est basé, non pas sur des interactions explicitement décrite comme telles dans un modèle particulier, mais plutôt sur le repérage, pendant la simulation, d'indices laissant penser qu'il peut y avoir interaction entre deux agents, comme par exemple le fait qu'ils soient proches l'un de l'autre, ou que leurs états internes évoluent de manière similaire.

Il a été montré que certains réseaux modélisant des systèmes complexes, comme les réseaux biologiques, possédaient des particularités topologiques issues de l'évolution de ces systèmes (réseaux de type *small world* ou *scale free*, etc.). Certaines de ces particularités ont été mises en évidence par l'étude des propriétés statistiques du graphe des interactions.

On est parti du principe que la construction du réseau d'agents, déterminé notamment à partir de leurs attributs et de leurs comportements, entraîne également ce type de particularités topologiques. Ces particularités se manifesteront en particulier si des interactions préférentielles se mettent en place entre certains agents. La détection des phénomènes émergents, et principalement ceux d'auto-organisation, passe alors par l'analyse du graphe à chaque pas de temps. Cette analyse comprend l'étude des propriétés classiques des réseaux complexes mais également la détection et le suivi des communautés au sein du graphe.

Graphe des entités en interaction

Le principe consiste à construire et à faire évoluer parallèlement avec la simulation, un graphe dont les sommets représentent les entités du système et les arêtes leurs interactions. Chaque arête est étiquetée par le critère d'interaction

correspondant (proximité, similitude de valeur d'un attribut, etc.). Ainsi, le sous graphe dont les arêtes sont étiquetées selon le critère de proximité (distance entre deux entités inférieure à une valeur donnée) représentera à un moment donné l'ensemble des entités proches les unes des autres. Si ce graphe évolue *peu* dans le temps, on peu inférer qu'un *groupe* d'entités s'est formé et est resté stable. Dans le cas de simulations de réactions biochimiques cela pourra indiquer la formation de complexes protéiques, ou bien de concentrations localement élevées de molécules.

L'évolution du graphe des interaction peut faire apparaître ou disparaître des sommets et les arêtes associés, ou des arêtes entre des sommets pré-existants. A chaque pas de temps, l'algorithme d'évolution du graphe va *renforcer* ou *affaiblir* une interaction entre deux entités en changeant le poids de l'arête correspondant. Deux entités déjà en interaction au temps t pour lesquelles l'interaction demeure au temps suivant, aura le poids de son arête incrémenté au temps $t+1$. Réciproquement, si au temps suivant l'interaction a disparu, le poids de l'arête est décrémenté. Quand le poids d'un arête est à zéro, l'arête est supprimé avec éventuellement les sommets qu'il reliait. Cette méthode de variation du poids des arêtes au cours du temps permet de quantifier la persistance d'une interaction et de lisser leurs changements rapides.

Le fait d'avoir les arêtes étiquetées par des critères d'interactions spécifiques permettra par exemple de comparer les éventuelles similarités de co-évolution des graphes restreints à tel ou tel critère d'interaction, montrant ainsi une corrélation lors de l'évolution temporelle de plusieurs types d'interactions. Ceci peut être un premier critère d'apparition ou un facteur de confirmation d'un phénomène émergent.

Analyse de la complexité des simulations entité-centrées

L'analyse des propriétés du graphe à chaque pas de temps va permettre la détection de propriétés émergentes survenant au cours de la simulation. Cette phase se divise en deux étapes, le calcul de l'évolution de certaines propriétés, pour trouver celles qui se démarquent de l'aléatoire, et la détection des groupes se formant au cours de la simulation.

Propriétés des réseaux complexes

Plusieurs méthodes d'analyse de la complexité d'un réseau d'interactions existent : longueur moyenne d'un chemin, mesure d'efficacité globale, coefficient de *clustering*, efficacité locale, etc.

La longueur moyenne du chemin est la moyenne des plus courts chemins de chacun des sommets du graphe pris deux à deux. Cette notion n'ayant de valeur que pour des graphes connexes, on utilise plutot la *global efficiency* défine par :

$$E_{glob}(G) = \frac{1}{n(n-1)} * \sum_{i \neq j} \frac{1}{d(v_i, v_j)}$$

Quand il n'existe pas de chemin entre v_i et v_j, alors $d(v_i, v_j)$ est infini, et donc $\frac{1}{d(v_i,v_j)} = 0$.

Le coefficient de clustering est défini par :

$$C(G) = \frac{1}{n} \sum_i c_i$$

n étant le nombre de sommets du graphe, et c_i le coefficient de clustering du sommet i :

$$c_i = \frac{2e_i}{k_i(k_i - 1)}$$

e_i étant le nombre de liens entre les voisins de v_i et k_i le degré de v_i.

Une autre méthode permet de déterminer la force des liaisons entre les sommets du graphe. Elle utilise le même formalisme que celui utilisé pour le calcul de la *global efficiency*. Soit G_i le sous graphe constitué des voisins du sommet v_i on peut déterminer ce qu'on appelle la *local efficiency* pour le graphe G :

$$E_{loc}(G) = \frac{1}{n} \sum_i E(G_i)$$

$E(G_i)$ étant la *global efficiency* du sous-graphe composé des sommets voisins de i.

Ces mesures permettent de savoir si le graphe adopte une topologie particulière, qui peut apporter des renseignements sur la dynamique du système. Dans un réseau *small world* la plupart des sommets ne sont pas connectés directement les uns aux autres, mais il est possible de relier tous les sommets deux à deux en utilisant un faible nombre d'arêtes. Une des principales caractéristiques de ce type de réseau est sa robustesse. En effet, même si un sommet du réseau est supprimé, cela n'interférera pas sur son fonctionnement.

Pour déterminer si un réseau G est de type *small world*, considérons un réseau aléatoire G_{alea} possédant le même nombre de sommets et le même nombre d'arêtes que G, et dont la distribution des arêtes est répartie équiprobablement entre les sommets. G est de type *small world* quand :

$$L(G) \approx L(G_{alea}) \quad \text{et} \quad C(G) \gg C(G_{alea})$$

avec $L(G)$ étant la longueur moyenne de chemin et $C(G)$ le coefficient de clustering du graphe G. Les réseaux biologiques d'interactions entre protéines sont très souvent de ce type.

Dans un réseau de type *scale free* quelques sommets ont un degré élevé alors qu'une grande majorité de sommets ont un faible degré. Une des caractéristiques les plus connues de ce type de réseau est sa grande vulnérabilité aux attaques de sommets à fort degré. Beaucoup de réseaux métaboliques de divers organismes présentent des caractéristiques comparables aux réseaux *scale free*.

Détection de groupes d'entités

Les phénomènes d'auto-organisation se caractérisent par l'apparition de groupes d'entités se formant sans principe organisateur, il est intéressant de reconnaître ces groupes et d'étudier leur devenir dans le temps. Pour reconnaître un groupe d'entités des algorithmes de clustering spécifiques ont été développés.

Ces algorithmes sont basés sur une définition simple de *communauté* : les liens entre les membres d'une communauté sont plus nombreux et plus forts que ceux de l'ensemble de la population.

Degré du graphe

Considérons un graphe $G(V, E)$ ou V correspond à l'ensemble des sommets d'une partition donnée et E correspond à l'ensemble des arêtes possédant une étiquette donnée. Soit d le degré moyen de ce graphe $d = 2 * \frac{|E|}{|S|}$.

Un groupe, au sein de ce graphe est un sous-graphe $C(S, A)$ où, pour tout sommet S_i de C, nous avons $d(S_i) > d$.

Un premier algorithme glouton basé sur le calcul du degré a permis de déterminer les groupes à chaque pas de temps avec une complexité en $O(m + n)$ où m correspond au nombre de sommets et n au nombre d'arêtes.

Poids des arêtes

Considérons le même graphe $G(V, E)$ et le poids moyen des arêtes pour une étiquette donnée. Un groupe, au sein de ce graphe est un sous-graphe $C(S, A)$ où pour chaque arête de ce sous-graphe A_i, nous avons $P(A_i) > \bar{P}$ où $\bar{P}$ est le poids moyen des arêtes contenant l'étiquette.

L'algorithme est analogue à celui utilisé pour la détection des groupes en fonction des degrés. Il permet de déterminer les groupes à chaque pas de temps avec une complexité en $O(m + 2n)$ où m correspond au nombre de sommets et n au nombre d'arêtes.

Nommage et suivi des groupes

Les groupes sont re-détectés à chaque pas de temps, sans utiliser la connaissance des groupes reconnus au pas de temps précédent. Comme bien entendu tous les

groupes ne changent pas à chaque pas de temps, il convient de reconnaître les groupes qui perdurent, ceux qui ont disparus et les nouveaux créés. Pour cela, chaque groupe est nommé pas un identifiant unique.

Pour reconnaître qu'un groupe existait au temps précédent, on cherche pour ce groupe au temps t, le plus grand nombre d'éléments communs dans chacun des groupes au temps t_{-1}. Le groupe du temps t prend ainsi le nom du groupe du temps t_{-1} possédant le plus grand nombre d'éléments communs.

Analyse de la topologie du réseau

La simple observation des courbes d'évolution des mesures de topologie du réseau nous fournit de bons renseignements sur les types d'interactions ayant un lien avec la présence des phénomènes émergents. En effet, lorsque par exemple un phénomène d'auto organisation se produit, la topologie du réseau d'interactions entre entités change fortement, reflétant la nouvelle organisation des entités. En croisant le changement de topologie du réseau avec la formation de groupes d'entités, on a une très bonne indication qu'un phénomène d'auto organisation apparaît.

On peut fréquemment trouver des corrélations entre la variation de divers paramètres de la simulation et la survenue du phénomène. Dans ce cas, cela permet souvent d'apporter un début d'explication, voire une chaine de causalité menant au phénomène émergent.

4.5 Études de cas

Dans les différents cas étudiés, on a utilisé la plateforme de simulation multi-agents *NetLogo*, simple à mettre en oeuvre et possédant une large bibliothèque de modèles. Pour automatiser le processus d'analyse d'une simulation, les algorithmes précédents ont été implementés, et une interface générale permet à l'utilisateur de les lancer en précisant les paramètres d'analyse.

Parmi les divers critères définissant une interaction, le système va systématiquement prendre en compte ceux de proximité et de direction de déplacement des agents. L'utilisateur pourra aussi donner d'autres critères d'interaction basés sur les divers attributs des agents.

Dans tous les cas le système va fournir les courbes d'évolution dans le temps pour chaque critère sélectionné :

- des mesures globales sur le graphe d'interactions : efficacité globale, efficacité locale, somme des poids et degré moyen,
- des mesures caractéristiques des groupes détectés : longueur moyenne du plus court chemin, coefficient de clustering, somme des poids, degré moyen et nombre d'agents.

Colonie de fourmis

L'objectif du modèle ANTS de *NetLogo* est de modéliser le comportement de fourragement, c'est-à-dire la récupération de nourriture présente dans l'environnement

par des agents fourmis. Les fourmis ont des règles de comportement simples :

1. dans le cas où elles ne transportent pas de nourriture, elles remontent un gradient de phéromones signalant la présence de nourriture. Si aucune phéromone n'est présente, elles se déplacent aléatoirement ;
2. dans le cas où elles transportent de la nourriture, elles retournent à la fourmilière en déposant des phéromones dans l'environnement (le retour à la fourmilière est également guidé par un gradient d'une autre phéromone)

On observe au cours des simulations, la création successive d'un chemin de phéromones entre la fourmilière et chacune des zones de nourriture. C'est généralement le chemin le plus court entre la fourmilière et la zone de nourriture. On observe également que les fourmis forment des colonnes se déplaçant sur ce chemin de phéromones. Grâce au partage d'information lié au dépôt de phéromones, l'acheminement de la nourriture vers la fourmilière peut alors s'effectuer très rapidement.

Parmi les nombreux critères étudiés, le critère de proximité a donné de très bons résultats comme on pouvait s'y attendre puisque, lorsque les fourmis trouvent une zone de nourriture, elle cessent de se déplacer de façon aléatoire pour se regrouper en colonne entre le nid et la zone de nourriture.

La courbe de la mesure de global efficiency pour le graphe d'interactions sur le critère de proximité montre très clairement les phases de recherche de nourriture par exploration aléatoire ou les fourmis sont très peu regroupées et quand elles ont trouvé une zone de nourriture et l'exploitent en formant des colonnes où elles sont très proches les unes des autres.

A contrario les mêmes mesures appliquées au critère de direction ne montrent pas de changement lors de la formation des colonnes.

Vols d'oiseaux

Le modèle Flocking simule la formation de nuées d'oiseaux, c'est-à-dire de vol groupé d'un nombre d'individus potentiellement élevé. Cette formation est possible à partir de trois règles de base pour les agents :

1. l'alignement : les oiseaux tendent à aller dans la même direction que celle des oiseaux proches,
2. la séparation : les oiseaux vont tourner pour éviter les oiseaux trop proches,
3. la cohésion : les oiseaux se déplacent pour se diriger vers les oiseaux situés à proximité.

Les agents *oiseaux* sont répartis de façon aléatoire initialement. Au cours de la simulation, on observe la formation de plusieurs groupes d'oiseaux. Ces groupes possèdent des trajectoires différentes au début mais adoptent au fur et à mesure une direction commune, tout en restant séparés en plusieurs groupes plus ou moins étendus.

Les analyses faites avec le critère de direction, en utilisant la *global efficieny* montrent que les agents adoptent une trajectoire commune. Ce constat n'est pas

une découverte en soi puisqu'il ne fait que corroborer ce que l'on savait déjà de par l'observation visuelle du système, mais il constitue un élément de validation de la pertinence de l'approche. Les analyses sur les groupes montrent par ailleurs que ces derniers n'évoluent pas énormément dans leur ensemble. Deux groupes peuvent en former un plus gros ou bien un groupe peut se scinder en deux mais il est rare qu'un agent isolé quitte un groupe.

On a constaté que les critères de proximité ou de direction utilisés isolément ne donnaient pas des résultats complètement satisfaisants. Lorsque le critère de proximité est utilisé seul, cela conduit en effet à considérer que deux groupes d'oiseaux se croisant n'en forment plus qu'un seul, alors que l'observation nous montre que chacun des deux groupes peut ensuite continuer sa route de manière indépendante. Lorsque le critère de direction est utilisé seul, deux groupes d'oiseaux distincts mais volant dans la même direction sont cette fois-ci considérés comme un seul et même groupe, ce qui est bien sûr contraire à l'intuition. C'est en combinant ces deux critères qu'on obtient un suivi des groupes satisfaisant. Cela n'est pas une grande surprise puisque le modèle de comportement des oiseaux est basé précisément sur la recherche d'un alignement avec les individus voisins, tout en conservant une distance aussi faible que possible avec ceux-ci. Il s'agit à nouveau, d'une confirmation de la pertinence de la méthode.

4.6 Conclusion

Au cours de sa thèse, Thomas Moncion a étudié d'autres modèles, notamment de type *proie-prédateur* où ses mesures de topologie du graphe d'interactions ont données de très bons résultats, montrant en particulier des corrélations fortes avec les oscillations des populations de proies et de prédateurs.

Il est important de noter que même si dans les exemples étudiés les règles locales de comportement des agents étaient connues, cette connaissance n'a pas été utilisée pour montrer qu'un phénomène émergeait. Ceci valide d'autant notre approche.

Une nouvelle implémentation de ces travaux est en cours d'intégration dans la prochaine version du système de simulation HSIM.

Les travaux issus de la thèse de Thomas Moncion ont été présentés à trois conférences [21], dont deux internationales [22, 23], et ont donné lieu à la publication d'un article dans un journal [24].

Hyperstructures 5

En collaboration avec des microbiologistes et physico-chimistes, Vic Norris, Michel Thellier et Camille Ripoll du laboratoire AMMIS de l'université de Rouen, on a utilisé HSIM comme outil d'investigation permettant d'évaluer la validité d'hypothèses sur la façon dont certaines voies métaboliques pourraient être réalisées dans les cellules vivantes. De nombreuses études ont montré que les protéines impliquées dans les voies métaboliques ou de signalisation sont souvent distribuées de façon non aléatoire, dans des assemblages multi-moléculaires. Ces assemblages vont de complexes multi-enzymes quasi statiques (comme la synthétase des acides gras) jusqu'à des associations transitoires, dynamiques de protéines. Ces assemblages multi moléculaires sont non seulement composés de protéines, mais aussi d'acides nucléiques, de lipides, de petites molécules et d'ions inorganiques. Ces assemblages ont été appelés *metabolons*, *transducons* et *réparosomes* dans le cas des voies métaboliques, transduction du signal et reparation de l'ADN respectivement, ou, plus généralement *hyperstructures* [14].

5.1 Functioning-dependent Structures

Un cas particulier d'hyperstructure est la *Functioning-dependent Structure* qui expliquerait comment les cellules métabolisent des substrats efficacement bien qu'utilisant un faible nombre d'enzymes : même si la concentration globale des enzymes impliquées dans une voie métabolique est très faible, la concentration *locale* peut être très élevée autorisant un phénomène de *canalisation* des réactions.

Plus formellement, une *Functioning-dependent Structure* (ou FDS) est un assemblage de macromolécules qui se forme parce que l'assemblage effectue une fonction biochimique et se désassemble quand la fonction cesse.

Le cas le plus simple de FDS est l'auto assemblage de deux enzymes successives, E_1 et E_2, dans une voie métabolique. Cet assemblage se produit parce que lorsque E_1 a fixé son substrat, E_1 et E_2 prennent de l'affinité l'une pour l'autre et forment un complexe protéique. Une raison de cette affinité accrue peut être le changement de conformation de E_1 quand elle a fixé son substrat.

Lorsque cet assemblage $E_1 - E_2$ est formé, du fait de la proximité des enzymes, la réaction peut être *canalisée* : le produit intermédiaire est directement transféré du site actif de E_1 vers celui de E_2.

Par rapport au cas où les deux enzymes sont libres, les avantages de la canalisation sont multiples :

1. le produit intermédiaire peut-être toxique pour la cellule et il est vital pour la cellule de ne pas le relâcher dans le milieu,
2. le produit intermédiaire peut-être labile, mais il est protégé de la dégradation par sa localisation dans le complexe protéique.
3. la réaction canalisée est plus rapide à fournir le produit final que dans le cas d'enzymes libres où il faut *attendre* que le produit intermédiaire *trouve* la deuxième enzyme par simple diffusion.

Ce mécanisme a été expérimenté *in-silico* avec HSIM et a permis de montrer que quand des enzymes en faible concentration se sont auto agrégées en présence de leur substrat et que la réaction est canalisée, la chaîne enzymatique a une efficacité supérieure à celle observée en simulant le même système, mais sans introduire les affinités qui induisent les assemblages [25].

Toujours dans ce cas simple de deux enzymes successives, on s'est intéressé aux différences de comportement de la cinétique de la réaction globale $s_1 \rightleftharpoons s_3$, avec et sans FDS, en fonction de la concentration de s_1 [26].

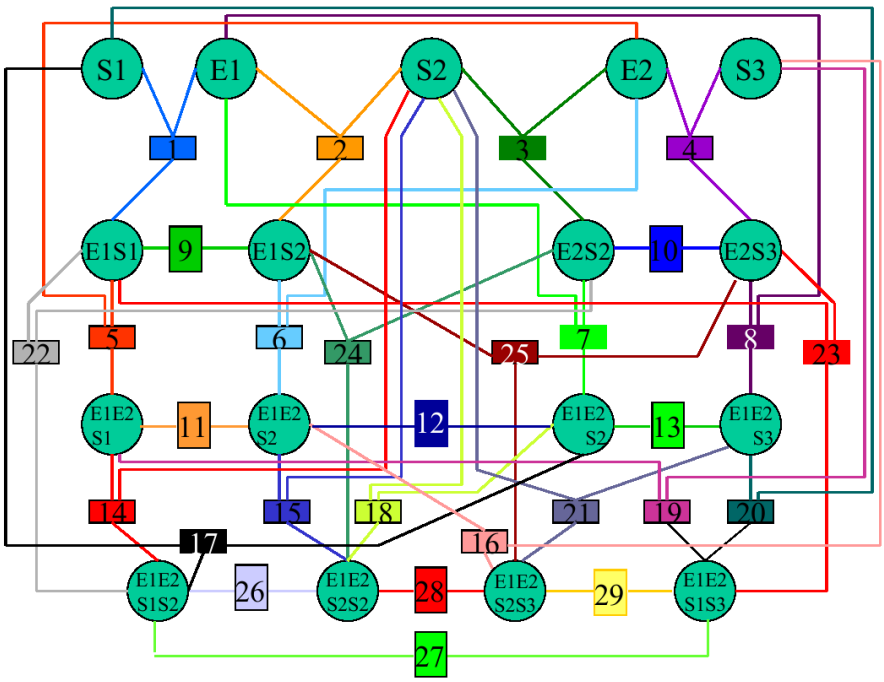

FIG. 5.1: Toutes les réactions possibles avec deux enzymes successives pouvant former une FDS. Le système comprend 17 espèces chimiques (enzymes libres, métabolites et complexes) représentées dans des cercles. Ces espèces sont reliées entre elles par 29 réactions représentées dans des rectangles.

On peut voir que même avec ce système très simple, la complexité en terme de nombre de chemins possibles dans le réseau des réactions menant de s_1 à s_3 est très grande (fig. 5.1).

Selon les paramètres des réactions d'assemblage / désassemblage des enzymes et les paramètres cinétiques de catalyse des réactions enzymatiques selon que les enzymes sont assemblées ou pas, le comportement de l'ensemble peut varier dans de grandes proportions. En effet, si on suppose que l'assemblage des enzymes est favorable à la canalisation de la réaction, la FDS apporte un avantage par rapport aux enzymes libres ; mais si on suppose au contraire que l'assemblage des enzymes nuit à la capacité de catalyse de E_2 par exemple, la FDS devient inhibitrice et ralentit la réaction globale en séquestrant les enzymes.

En choisissant des valeurs de paramètres adéquats on a obtenu des comportements non linéaires très intéressants permettant de penser que déjà avec ce système très simple, on peut obtenir des régulations très efficaces (fig. 5.2).

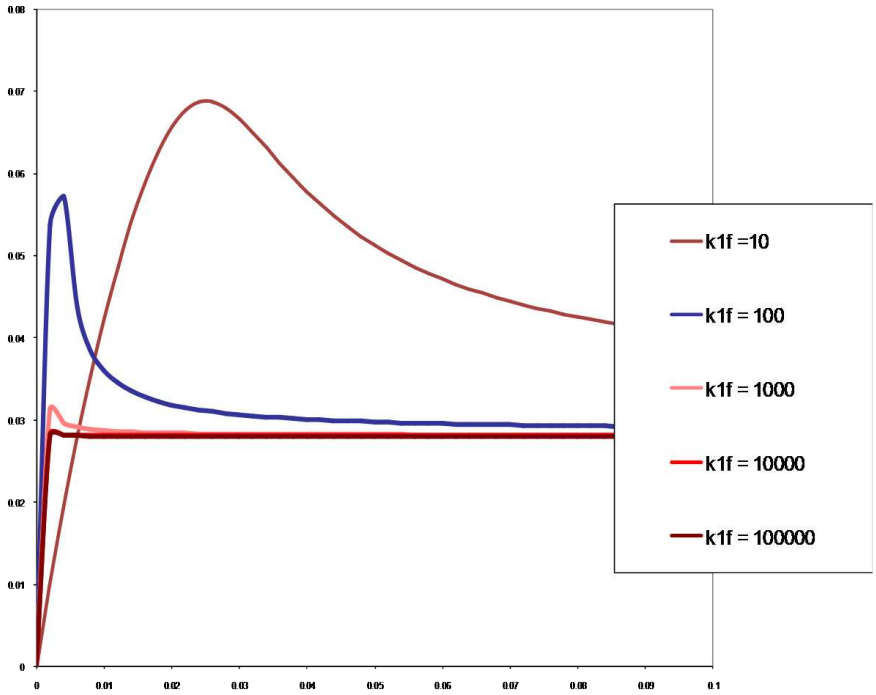

FIG. 5.2: Comportements non-linéaires d'une FDS à deux enzymes selon les paramètres cinétiques.

Par exemple, l'*overshoot* de concentration du produit final pourrait, par une sorte de phénomène d'amplification, signaler à la cellule que cette FDS s'est formée et déclencher d'autres processus. On continue actuellement de travailler dans cette voie.

5.2 Couplage glycolyse-PTS

Au cours du séjour de recherche que j'ai effectué en 2003 à l'université de San Diego, dans le laboratoire de Milton Saier qui est un spécialiste des systèmes de phosphotranspherase (PTS) dans les cellules procaryotes, nous nous sommes intéressés au couplage de cette voie métabolique à celle de la glycolyse. Ces voies métaboliques ont été très étudiées de façon expérimentale et de nombreuses données sont disponibles dans la littérature [27].

Les avantages de grouper les enzymes d'une voie en metabolons, voire en plus grandes structures ont longtemps été discutés. Il a été montré experimentalement chez *E. coli* que la voie de la glycolyse pouvait être isolée comme un grand complexe multi-enzymes dans lequel les métabolites etaient séquestrés [28, 29]. Mais l'avantage en terme d'efficacité de la voie de tels groupements n'est pas clair. C'est pourquoi nous avons décidé d'étudier *in silico*, à l'aide d'un modèle simulé avec HSIM, l'impact sur l'efficacité de la voie de diverses formes de regroupements des enzymes.

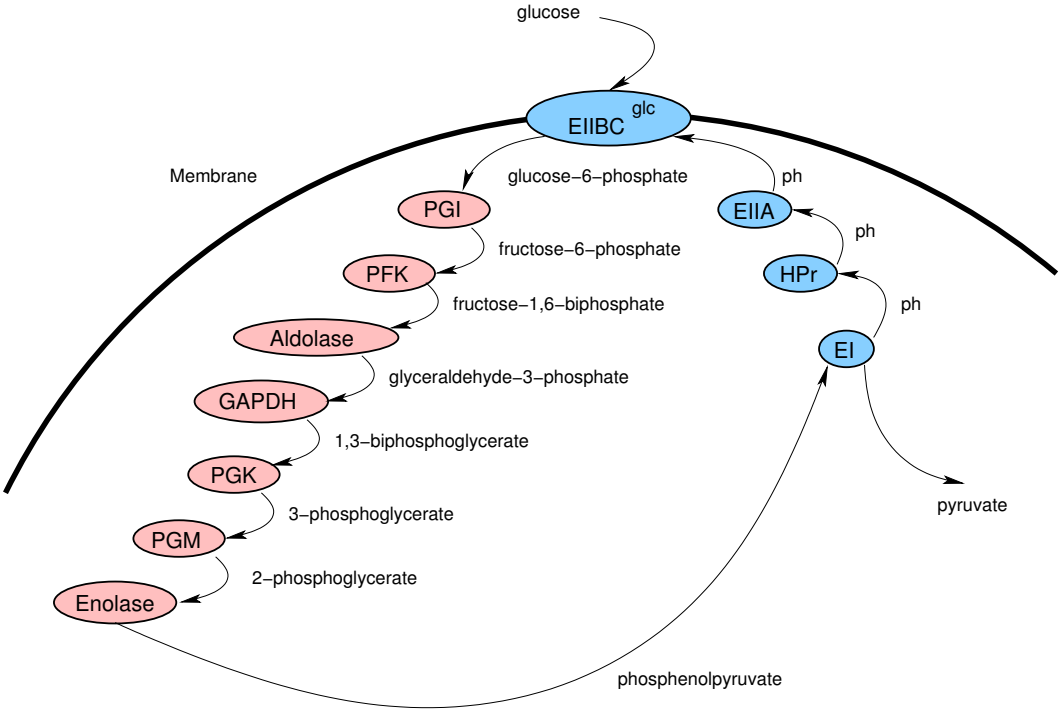

FIG. 5.3: Couplage entre la voie de la glycolyse (enzymes en rose) et la voie de la PTS (enzymes en bleu).

Les deux voies, glycolyse et PTS, sont interdépendantes du fait qu'elles sont connectées l'une à l'autre à leurs deux extremités (fig. 5.3) :

- l'*entrée* de la glycolyse, le récepteur membranaire qui transporte le glucose à l'intérieur de la bactérie est la *sortie* de la PTS. Ce récepteur n'autorise l'importation d'une molécule de glucose que lorsqu'il est phosphorylé par la PTS.

– un sortie intermédiaire de la glycolyse est le phosphenolpyruvate (pep) qui se trouve être l'entrée de la PTS, le rôle de la première enzyme de la PTS étant d'en fixer un groupe phosphate et de libérer du pyruvate.

Cette double dépendance des deux voies forme un circuit à boucle positive car chacune d'entre elles active l'autre. Ce type de circuit est connu pour avoir des comportements non linéaires, avec des phénomènes d'avalanche dans certains cas. Par souci de simplification, on n'a modélisé qu'un partie de la glycolyse, en laissant de coté la duplication de la voie au niveau de la triosephosphate-isomerase.

Hypothèses testées

Dans cette étude, on a supposé que les enzymes successives de chaque voie pouvaient former des FDS, qui donc s'auto assemblaient du fait de leur fonctionnement. Comme c'est la cinétique des voies qui nous intéressait, on a ignoré la phase initiale de formation des metabolons. Nous avons étudié les deux voies dans diverses conditions d'association, en utilisant un nombre identique de copies (64) de chaque type d'enzymes :

1. aucune association entre les diverses enzymes, tout fonctionne uniquement par diffusion dans le milieu ;
2. seule la voie de la glycolyse est assemblée en metabolons (avec canalisation), la PTS fonctionnant toujours par diffusion ;
3. seule la PTS est assemblée en metabolons ;
4. les deux voies sont assemblées en metabolons.

On s'est aussi interessé à l'influence de la colocalisation spatiale des metabolons, on a étudié trois cas : (i) les deux types de metabolons diffusent indépendamment les uns des autres dans la cellule, (ii) chaque chaîne de glycolyse est associée à une chaîne de PTS par attachement au recepteur membranaire (iii) chaque chaîne de glycolyse est associée à une chaine de PTS à la fois par le recepteur membranaire et par l'autre bout, l'enolase étant liée à l'enzyme EI, ce qui permet de *rapprocher* la production de PEP de son utilisation. Dans tous les cas, les metabolons, qu'ils soient attachés deux à deux ou non, sont uniformément répartis dans la cellule.

Comme on l'a mentionné précédemment, on peut voir que ces deux voies métaboliques coopèrent dans un cycle formant une boucle de rétroaction positive. L'enzyme membranaire EIIBC servant de *robinet* pour l'importation du glucose vers la glycolyse. Ce robinet est ouvert quand la PTS lui transfère un groupe phosphate, ce qui arrive quand la glycolyse a dégradé un glucose.

Il faut donc démarrer le système en fournissant du PEP initialement. On a étudié la dynamique du système en faisant varier la concentration initiale de PEP. Dans toutes nos simulations on a mesuré la production de pyruvate après 40 secondes de temps simulé (on a noté que seulement 2 à 3 secondes étaient suffisantes pour atteindre un régime stationnaire).

Résultats

Avec initialement 300 copies de PEP, le système montre que l'assemblage en metabolons de la glycolyse n'offre aucun avantage, 1059 contre 1028 copies de pyruvate (table 5.1), par contre dès que la PTS est assemblée elle accélère la glycolyse non associée d'un facteur 2.5 et d'un facteur 8 quand elle est associée en metabolons.

Association	
Pas d'association	1028
Glycolyse associée	1059
PTS associée	2384
PTS et Glycolyse associées	8053

TAB. 5.1: Production de pyruvate après 40 secondes temps simulé avec 300 copies de PEP initialement.

Ça laisse à penser que la voie de la glycolyse est sous-alimentée en glucose-6-phosphate et que c'est donc la PTS qui est déterminante. Ce résultat correspond aux experiences *in vivo*.

Influence de la concentration initiale de PEP

On a ensuite étudié l'impact de la concentration initiale de PEP sur l'efficacité globale. On a observé que jusqu'à 2000 et probablement au delà, l'efficacité de la glycolyse est identique, ce qui corrobore le fait que c'est la PTS qui est déterminante. Quand la voie de la PTS est associée en metabolons, son efficacité propre croît avec la concentration initiale de PEP et corrélativement la glycolyse aussi. L'accroissement d'efficacité peut atteindre un facteur proche de 50 dans le cas ou les deux voies sont associées (table 5.2 et fig. 5.4).

Concentration initiale de PEP							
PEP initial	**30**	**50**	**100**	**300**	**500**	**1000**	**2000**
Pas d'association	167	279	479	1028	1036	1045	1076
Glyc. associée	265	405	694	1059	1096	1054	1083
PTS associée	245	401	819	2384	3990	7982	15616
PTS+Glyc. associées	550	981	2381	8053	13699	26752	47707

TAB. 5.2: Production de pyruvate après 40 secondes temps simulé en faisant varier la concentration initiale de PEP.

En utilisant de faibles concentrations initiales de PEP on voit que le système a du mal à démarrer, ce qui est tout à fait normal pour un système à circuit de rétroaction positif. On observe aussi que la valeur plateau de production de PEP quand la PTS n'est pas associée est atteinte avec 300 PEP initialement.

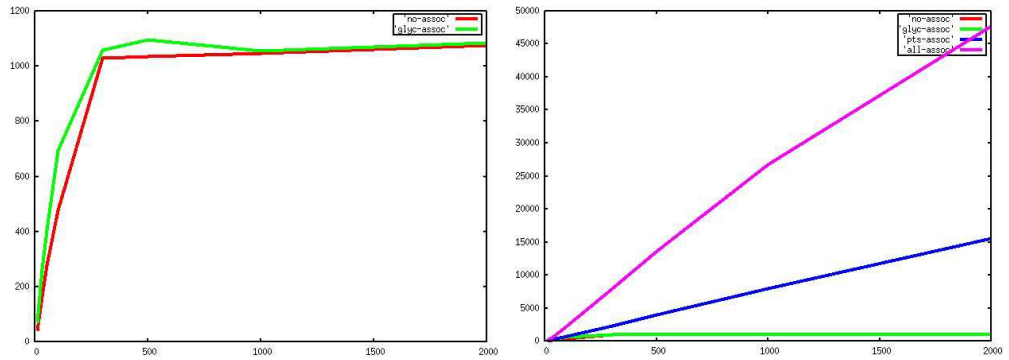

FIG. 5.4: Courbes de production de pyruvate en fonction de la concentration initiale de PEP.

Influence de la colocalisation

On s'est ensuite intéressé à l'influence de la colocalisation spatiale des deux types de metabolons sur l'efficacité globale. Les résultats montrent que la proximité du producteur de PEP (la glycolyse) de son consommateur (la PTS) n'a pas d'influence notable sur l'efficacité de la production de pyruvate (table 5.3). C'est probablement du à la grande vitesse de diffusion du PEP et à sa relativement forte concentration globale qui diminue l'avantage de la proximité spatiale.

Colocalisation				
PEP initial	**300**	**500**	**1000**	**2000**
PTS+Glyc associées	8053	13699	26752	47707
Attachement par EIIBC	8742	14129	26629	46583
Double attachement	7722	12708	23575	42847

TAB. 5.3: Influence de la colocalisation des paires de metabolons de PTS et de glycolyse sur la production de pyruvate.

Metabolons groupés en hyperstructures

Toujours dans le but de tester l'efficacité de la voie de la glycolyse, on a augmenté la concentration locale des enzymes en colocalisant tous les metabolons en hyperstructures. Les metabolons sont attachés à la membrane par les recepteurs membranaires EIIBC que l'on a plus répartis dans la cellule, mais disposés sous forme d'un patch dense sur un pôle de la membrane.
Dans ces expériences les recepteurs membranaires ne sont plus saturés en glucose comme précédemment, mais alimentés à une vitesse fixée, qui devient un des facteurs limitants de la voie.

Pour des taux d'importation de glucose variant sur deux ordres de grandeur, on a constaté que tant qu'on n'utilisait que la diffusion des métabolites pour trouver l'enzyme dont ils sont le substrat (sans canalisation) il n'y avait pas d'avantages en

termes d'efficacité par rapport à la version "enzymes libres" (table 5.4). Par contre l'utilisation de réactions canalisées augmente d'un facteur 2 l'efficacité de la voie.

Hyperstructures			
Glucose par seconde, par récepteur	**0.1**	**1**	**10**
60 jeux d'enzymes, sans metabolons	16	216	2131
60 metabolons, sans canalisation	25	253	2178
60 metabolons, avec canalisation	55	428	4097

TAB. 5.4: Influence de la colocalisation en hyperstructures des metabolons de PTS et de glycolyse sur la production de pyruvate, pour des taux fixes de glucose,

On voit donc que quand c'est l'alimentation en glucose qui est limitante, le fait de beaucoup concentrer les enzymes n'offre pas d'avantage notable. C'est *a priori* contre intuitif, on se serait attendu à ce que les métabolites, plongés dans une 'forêt' d'enzymes trouvent plus rapidement l'enzyme dont ils sont le substrat que quand celles-ci diffusent dans toute la cellule. Il y a deux raisons qui peuvent expliquer cette intuition erronée : (i) on est ici en trois dimensions et les degrés de liberté des métabolites sont beaucoup plus grands qu'en deux dimensions (où on à l'habitude de penser : déplacement dans une forêt) et (ii) les enzymes (les 'arbres') ne sont pas fixes, elles ont une diffusion restreinte par leurs attachements à la précédente et à la suivante, mais elles bougent et de ce fait, reduisent encore la probabilité d'une collision avec les métabolites.

Un effet secondaire de la version "hyperstructure" localisée à un endroit de la cellule est qu'elle produit un gradient de concentration du produit final s'il est dégradé (ou consommé) uniformément dans la cellule par exemple par des protéines de dégradation uniformément réparties. Ce phénomène ne se produisant pas quand il n'y a pas colocalisation de la production.

Conclusion

Ces expériences *in silico* utilisant HSIM nous ont permis d'apporter des réponses quantitatives à des questions fondamentales telles que : "est-ce que les hyperstructures apportent un avantage en termes d'efficacité des voies métaboliques ?" Dans le cas ou l'alimentation en glucose est limitante, seule la canalisation des réactions augmente l'efficacité de la voie. Cet effet peut être encore amplifié quand, dans le cas de la PTS par exemple, cette voie en contrôle une autre, ici la glycolyse.

Dans cette étude on s'est limité à un système simplifié et fonctionnant en régime stationnaire. L'un des intérêts de HSIM est qu'on peut tout aussi bien modéliser et simuler des systèmes hors équilibre et éventuellement observer d'autres comportements intéressants des hyperstructures.

Ces travaux ont été présentés au 4th World Congress of Cell and Molecular Biology [30] et publiés dans BMC Systems Biology [19].

Il est à noter, qu'ultérieurement à la publications de ces travaux, le groupe de Jörg Stülke de l'université de Göttingen a démontré expérimentalement chez

Bacillus subtilis l'existence de complexes d'enzymes de la glycolyse (phosphofructokinase, phosphoglyceromutase et enolase) qui favoriserait la canalisation des réactions [31].

5.3 Couplage cytosquelette et réseaux métaboliques

Lors de la huitième édition de l'école thématique *Modelling Complex Biological Systems in the context of genomics* que j'ai organisée en 2009, Judit Ovàdi, de l'institut d'enzymologie de Budapest a fait un exposé sur les effets de la protéine TPPP/p25 sur la dynamique et la stabilité du réseau de microtubules [32].

À cette occasion, nous avons, avec Vic Norris, commencé à travailler sur les corrélations entre la voie de la glycolyse et les microtubules constituant le cytosquelette de la cellule eucaryote. Le fait que certaines enzymes de la glycolyse se fixent sur les microtubules a comme conséquence d'en changer la dynamique et donc d'une certaine manière le cytosquelette de la cellule "ressent" l'activité métabolique.

L'hypothèse

Les hyperstructures, à un niveau d'organisation intermédiaire entre les macromolécules et la cellule, permettent de réduire le bruit de fond du niveau moléculaire et de faciliter l'émergence d'un motif cohérent au niveau de la cellule, produisant un phénotype adapté. Si nous voulons comprendre la réponse de la cellule à l'information métabolique, cela doit être au niveau des grands assemblages intracellulaires, des hyperstructures.

Un bon exemple est l'hypothèse que l'amplification du signal dans le chimiotactisme bactérienne dépend de la taille de la structure chimiotactique (regroupement de beaucoup de senseurs).

La fixation aux microtubules et aux filaments d'actine d'enzymes catalysant différentes voies métaboliques permet au cytosquelette de sentir et d'intégrer l'activité métabolique. Cela se traduit par des altérations de la stabilité et de la dynamique des filaments du cytosquelette, dans les taux d'hydrolyse de l'ATP et de la GTP par les constituants du cytosquelette et les enzymes associées, et aussi dans les niveaux de métabolites.

Les données expérimentales

Beaucoup d'enzymes de la glycolyse se lient au réseau d'actine et de microtubules, c'est le cas par exemple de l'hexokinase, la phosphofructokinase (PFK), la pyruvate kinase (PK) et l'aldolase. La glyceraldehyde-3-phosphate dehydrogenase (GAPDH) se lie aux microtubules et les organise en faisceaux [33]. Il est connu que la PFK est activée par sa liaison à l'actine filamenteuse, et il a récemment été montré que la signalisation par l'insuline accroît les associations de PFK aux filaments d'actine, et il a été suggéré que cette association joue un rôle dans la stimulation de la glycolyse par l'insuline [34].

Fixation "fonctionnement-dépendant" des enzymes au cytosquelette

Notre interprétation est que la liaison au cytosquelette d'une enzyme accroît sa probabilité de catalyse, par exemple en augmentant l'affinité pour son substrat. Réciproquement, le cytosquelette pourrait avoir une plus grande probabilité d'association avec une enzyme en cours d'activité catalytique. En d'autre termes, si le fait pour une enzyme d'être liée à un filament du cytosquelette lui donne une conformation qui lui permet de lier son substrat, alors l'activation d'une enzyme libre par son substrat pourrait promouvoir la liaison de cette enzyme au filament.

Implications de l'hypothèse

Notre hypothèse de cytosquelette senseur intégratif est une tentative de réponse à la question "*comment la cellule sent et intègre une large diversité d'informations physiques et chimiques de façon à converger vers un phénotype cohérent ?*". Cette hypothèse est insuffisante en soi, mais pourrait, combinée à d'autres fournir un tableau intégré du fonctionnement de la cellule. Une de ces hypothèses pourrait être la condensation d'ions sur le cytosquelette. Une autre hypothèse pourrait inclure la propriété de mécano-transduction du cytosquelette.

Les effets physique et chimiques résultant de la sensibilité du cytosquelette à l'activité métabolique aurait des conséquences majeures sur la forme de la cellule et sur le cycle cellulaire.

Ces travaux ont été publiés dans BMC Biochemistry [35].

Un nouvel article est en cours d'écriture présentant un modèle quantitatif de couplage d'enzymes à des microtubules qui sera simulé avec HSIM.

5.4 Life on the scales

Toujours dans le domaine de la biologie des systèmes, je me suis intéressé à l'impact de la dynamique de la réplication de l'ADN sur le phénotype des bactéries lors de la division cellulaire. A tous les niveaux du vivant, les systèmes évoluent à différentes *échelles d'équilibre*. Au niveau bactérien, chez *Escherishia coli* par exemple, la cellule individuelle doit *choisir* entre deux stratégies opposées : soit prendre des risques et croître (duplication rapide), soit éviter les risques et survivre (spores).

On a initialement proposé un modèle où la différenciation fait partie du cycle cellulaire de façon à donner à l'une des cellules filles le phénotype de croissance et à l'autre le phénotype de survie. Cette situation découle de la régulation des circuits cellulaires basée sur une rétroaction positive locale : un gène qui s'exprime à une plus grande chance de continuer à être exprimé, et une rétroaction négative globale : les gènes peuvent être en concurrence entre eux pour être exprimés. Des preuves de cette propension à se différencier peuvent être trouvées dans la distribution des gènes sur les deux brins de l'ADN, avec les gènes nécessaires à la croissance ayant tendance à être sur un brin alors que ceux de la résistance

aux stress ont tendance à être sur l'autre brin [36]. On a ensuite proposé un autre modèle plus élaboré, où la *variation* dans la vitesse de réplication de l'ADN serait la clef de la différenciation [37].

Plus récemment, on a émis une hypothèse selon laquelle la stratégie de croissance dépend d'hyperstructures à l'équilibre et la stratégie de survie dépend d'hyperstructures hors équilibre. On a aussi proposé que le cycle cellulaire lui-même soit le moyen qu'utilise les cellules pour équilibrer les taux de ces deux types d'hyperstructures de façon à trouver une solution concensus pour évoluer sur les deux échelles. Au cours du cycle cellulaire, la cellule doit pouvoir détecter le rapport entre les structures hors l'équilibre et celles à équilibre (rapport NE/E), cela peut se faire de deux façons :

1. *Intensity sensing* : l'intensité d'utilisation des structures hors équilibre est détectée, elle reflète la disponibilité de l'énergie (ATP, GTP, . . .) et la disponibilité de nutriments (acides aminés, . . .). Quand cette intensité est suffisante pour que la cellule croisse sans risque, des signaux sont émis par ces hyperstructures.
2. *Quantity sensing* : la quantité de structures à l'équilibre est détectée. Quand elle est suffisante pour qu'une cellule fille soit engendrée, un signal est émis. Le fait qu'un mécanisme de *quantity sensing* existe est un avantage évolutionnaire qui permet de s'assurer qu'avant que la cellule s'embarque dans un nouveau cycle cellulaire, elle dispose de quantité suffisantes de membrane, d'énergie et de nucléotides pour achever la division cellulaire.

Ces deux types de signaux, intensité et quantité, sont nécessaires pour que la cellule commence un nouveau cycle et initialise la production de cellules filles. Nous avons proposé divers mécanismes par lequels la cellule pourrait générer ces signaux.

Ces travaux ont été initialement publiés dans les actes de l'édition 2012 de la conférence *Modelling Complex Biological Systems in the Context of Genomics* [38] puis une version longue dans un journal, *Life* [39].

Biologie de synthèse

6

La biologie de synthèse est un axe de recherche en plein essor dans le domaine de la conception et de l'ingéniérie de systèmes basés sur des règles fonctionnelles biologiques dans le but d'obtenir de nouvelles fonctionnalités qui ne sont pas présentes dans la nature (fig. 6.1). D'après le MIT, la biologie de synthèse est focalisée sur la conception raisonnée de systèmes biologiques, plutôt que sur la compréhension de la biologie des organismes naturels, c'est en ce sens qu'elle se distingue de la biologie classique.

On peut distinguer trois types d'approches de la biologie de synthèse :

1. *in vivo* où on utilise l'ingéniérie génétique, par exemple, pour optimiser dans un organisme vivant, une voie métabolique produisant un composé chimique d'intéret thérapeutique ou bien industriel.
2. *in vitro* dans laquelle on conçoit des composants artificiels pour des applications de nano technologies où ces composants sont conçus et déployés en dehors de cellules vivantes.
3. *de novo*, qui est le genre le plus extrème de nano technologie bio-inspirée où des systèmes biologiques complètement artificiels proviennent d'une théorie *top-down* pour obtenir à partir d'assemblages de composants artificiels ce qui peut être considéré comme un système vivant (*Vie Artificielle*).

Depuis cinq ans je collabore avec Franck Molina et Alain Thierry, biologistes au laboratoire SysDiag à Montpellier autour du projet *CompuBioTic*. C'est un projet de biologie de synthèse dans le domaine du diagnostic médical. Le but est de concevoir une vésicule lipidique contenant des récepteurs membranaires et des protéines spécifiques dont les interactions permettent de détecter et rapporter la présence dans l'environnement de molécules impliquées dans des pathologies humaines (cancer colo-rectal, diabète, etc.).

Nous utilisons HSIM pour concevoir et tester qualitativement et quantitativement ces vésicules avant leur réalisation par la division *biologie humide* du laboratoire. Une doctorante, Stéphanie Rialle et une post-doctorante, Sabine Peres, ont travaillé à plein temps avec HSIM sur ce projet. Depuis, Stéphanie Rialle a soutenu

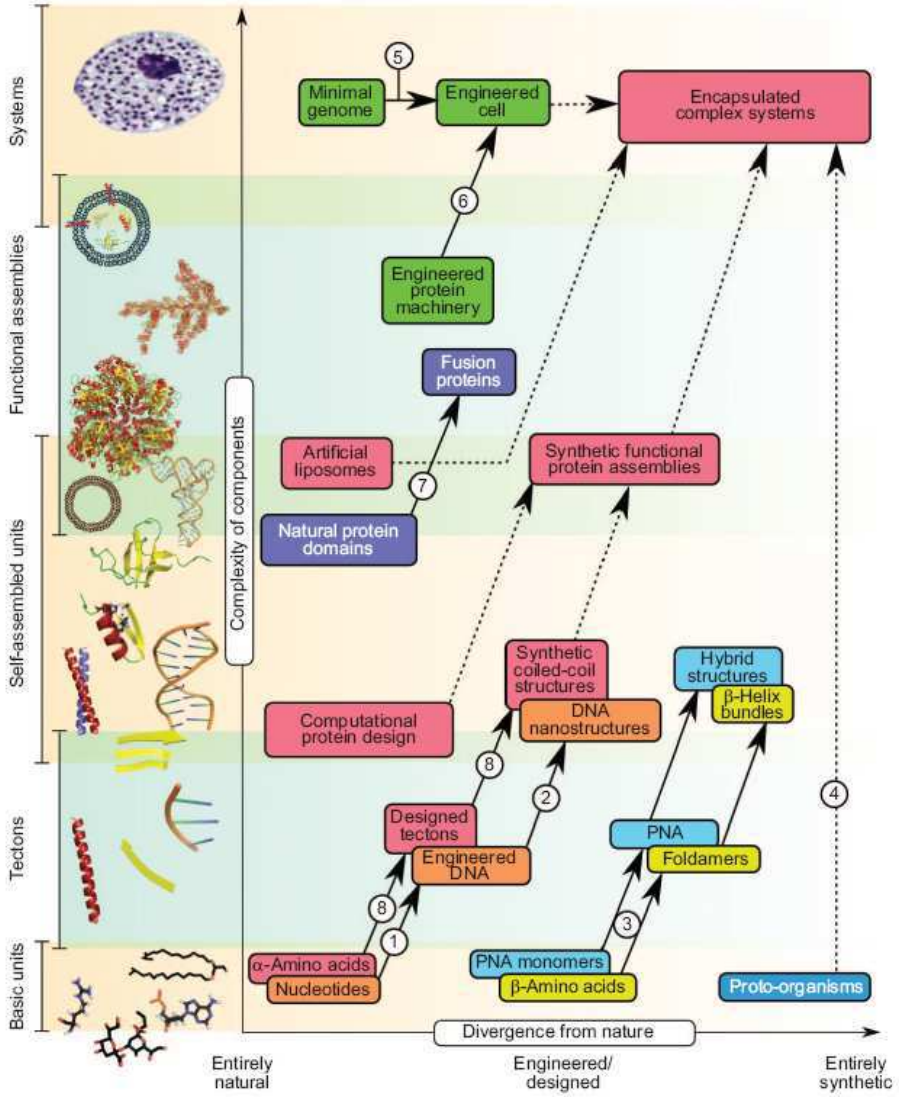

FIG. 6.1: Représentation des différentes approches de la biologie de synthèse selon Bromley *et. al* [40].

sa thèse en octobre 2010 et Sabine Peres a été recrutée MCF à l'IUT d'Orsay et fait partie de l'équipe Bioinformatique du LRI.

C'est dans le cadre de cette thématique que je dirige la thèse de Marc Bouffard. Il s'agit de concevoir et de valider une méthodologie complète de conception de nano systèmes biologiques artificiels qui seront à terme utilisés dans le cadre du diagnostic médical. Ce domaine particulier de la biologie de synthèse est très récent et a des répercussions potentielles dans le champ clinique particulièrement prometteuses.

6.1 Projet CompuBioTic

Le projet CompuBioTic du laboratoire *SysDiag* à Montpellier, utilise la technique de la biologie de synthèse *de novo* dans le but d'obtenir un calculateur biologique embarqué ou non (*bio-computeur*) dédié au diagnostic médical de pathologies ciblées.

Un des buts du projet étant, à terme, de réaliser un système de diagnostic utilisé *in vivo* chez un patient, il est primordial qu'on contrôle ce système parfaitement. Il est donc hors de question d'utiliser une cellule vivante hôte, qui risquerait par exemple de muter et devenir pathogène. Toujours pour des raisons de stabilité et de non-contamination, on n'utilisera pas non plus de systèmes génétiques.

Le nanosystème biologique de diagnostic prendra donc la forme d'une vésicule (liposome) ou d'une micro-goutte (droplet) composée d'un réseau enzymatique biologique conçu artificiellement *ab initio* afin de servir à la fois de senseur des différents biomarqueurs mais aussi de calculateur (algorithme logique) et de révélateur (production d'une coloration par exemple). Le réseau biologique artificiel sera principalement composé d'enzymes et de métabolites d'utilisation courante en biotechnologie, donc stables et robustes.

De façon générale, la conception d'un nanosystème biologique artificiel peut se décomposer en cinq étapes :

1. Spécification du système biologique que l'on désire réaliser.
2. Conception *in silico* du réseau biologique artificiel.
3. Modélisation et simulation dynamique du système pour l'optimisation et les études de robustesse.
4. Construction expérimentale et validation analytique du fonctionnement du système.
5. Validations *in vitro* et éventuellement *in vivo*.

Ces différentes étapes sont intimement liées et nécessitent les compétences et le travail conjoint de biochimistes, modélisateurs (bioinformaticiens) et de cliniciens.

Au cours de sa thèse [41], Stéphanie Rialle a fait un première avancée avec la réalisation d'un prototype d'outil de conception de réseau biologique artificiel encapsulé dans une vésicule lipidique.

Pour cela, elle a en premier lieu réalisé une base de données de composants moléculaires adaptés à la biologie de synthèse utilisant des protéines : CompuBioTicDB. Cette base de données regroupe pour chaque protéine les réactions qu'elle catalyse, ses substrats et produits, les éventuels cofacteurs nécessaires, les conditions de température et de pH où elle fonctionne, ainsi que des liens vers des bases de données externes (BRENDA, PDB, Uniprot).

La base de données comporte aussi des *dispositifs*, ce sont des briques de base abstraites qui serviront à fabriquer des réseaux de calcul plus complexes. Ces dispositifs sont définis par leur rôle et par plusieurs types d'implémentation moléculaires, chacune ayant une description formelle utilisant le langage

Bioψ [42]. Parmi les dispositifs on trouve des *révélateurs colorimétriques* indiquant la présence d'un signal biologique, de *senseurs conditionnels* qui sont sensibles à un signal biologique et engendrent une action quand le signal est détecté, chronomètres, oscillateurs, interrupteurs, portes logiques, etc.

Le deuxième outil réalisé est un *plugin*, BioNetCAD, pour le logiciel de dessin de réseaux CellDesigner [3]. Ce plugin apporte aux fonctionnalités de CellDesigner l'accès à la base de données CompuBioTicDB et la possibilité de simuler le réseau en cours de conception avec HSIM.

Cet outil permet de concevoir un réseau artificiel à l'aide d'enzymes et de métabolites abstraits, puis ensuite, à l'aide de CompuBioTicDB, de trouver une ou plusieurs implémentations avec des biomolécules réelles. Dès la phase de conception du réseau abstrait, on peut lancer des simulations pour avoir une première vue qualitative du comportement du réseau. Une fois une implémentation choisie, on va la simuler de façon quantitative avec HSIM pour (i) valider le comportement du modèle et (ii) quantifier les concentrations initiales des métabolites qui permettent d'obtenir le résultat désiré.

Les travaux sur BioNetCAD ont donné lieu à une publication commune dans *Bioinformatics* [43].

6.2 Projet BS^2

Le projet BS^2 pour "Biologie Systémique et Biologie de Synthèse" est la suite de CompuBioTic. Il explore diverses façons d'obtenir des circuits logiques avec des composants biologiques, toujours dans le but d'obtenir des systèmes de diagnostic médical simples et efficaces.

C'est le contexte dans lequel se place la thèse de Marc Bouffard, co-encadré par Franck Molina et moi-même.

Les premières expériences d'utilisation des outils développés par Stéphanie Rialle nous ont confortées dans cette voie de conception aidée par ordinateur de réseaux biologiques de calcul. À l'issue de cette première expérience diverses faiblesses et manques de l'outil principal, CellDesigner, et aussi le manque de recul sur les éventuelles limitations de l'approche employée pour implémenter des portes logiques nous ont conduites (i) à faire une étude plus poussée des contraintes liées à notre implémentation des portes enzymatiques et (ii) à réaliser à terme un outil intégré de conception / implémentation / simulation de réseaux dédié à la fabrication de bio-calculateurs dans le contexte du diagnostic médical.

Implémentation des portes logiques enzymatiques

Le modèle de portes logiques utilisé est simple et robuste, il utilise comme signal la concentration de métabolites : si elle est faible (en dessous d'un seuil sl_0) le signal est interprété comme un 0 logique, si elle est forte (au dessus d'un seuil $sl_1 > sl_0$) c'est un 1 logique.

Pour implémenter une porte ET à deux entrées s_1 et s_2, il suffit de trouver une enzyme qui catalyse une réaction nécessitant s_1 et s_2 comme substrats et synthétisant un métabolite p. Le produit p sera en concentration suffisante pour être interprété comme un 1 logique si seulement si les deux substrats s_1 et s_2 sont en simultanément présents en concentration suffisante, donc aussi à la valeur logique 1. Dans tous les autres cas, l'enzyme ne synthétisera pas p et sa concentration représentera la valeur logique 0 (fig. 6.2 à gauche).

Pour implémenter une porte OU à deux entrées s_1 et s_2, il suffit d'utiliser deux enzymes E_1 et E_2 qui chacune catalysent une réaction qui transforme son substrat s_1 pour E_1 et s_2 pour E_2 dans le même produit p. Il suffit que l'un des deux métabolites s_1 ou s_2 soit en concentration suffisamment forte, représentant un 1 logique pour que l'enzyme correspondante fabrique le produit p et donc sorte la valeur logique 1. Dans le cas où la concentration des métabolites d'entrées est faible, à la valeur logique 0, aucune des deux réactions ne se fait suffisamment, et la concentration de p restera faible, représentant la valeur logique 0 (fig. 6.2 à droite).

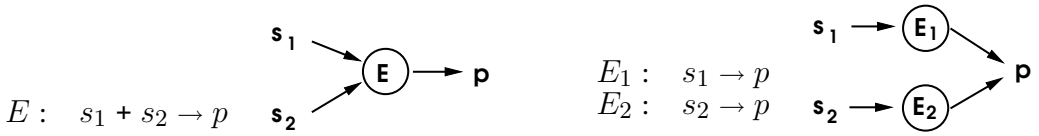

FIG. 6.2: Porte ET (à gauche) et porte OU (à droite) implémentées en mode *mono*.

Lors de son stage de Master 2, Marc Bouffard à beaucoup avancé sur l'étude des limitations inhérentes à l'implémentation de portes logiques par des réactions enzymatiques. Il a mis en évidences des problèmes de synchronicité des signaux d'entrée, d'affaiblissement des signaux intermédiaires et, ce qu'on savait déjà, une possible limitation de la taille du réseau due à l'utilisation d'une espèce moléculaire propre à chaque connexion entre portes.

Il a trouvé des moyens permettant d'apporter des solutions ou de contourner ces limitations et proposé une autre implémentation de portes logiques enzymatiques, les portes en mode *dual* n'ayant aucun des inconvénients cités (fig. 6.3).

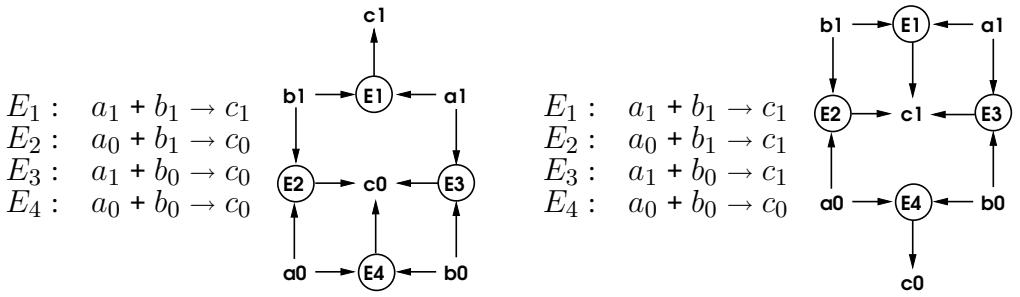

FIG. 6.3: Porte ET (à gauche) et porte OU (à droite) implémentées en mode *dual*.

Le principe de l'implémentation en mode dual est d'avoir un codage différentiel de l'information. Une valeur booléenne est portée par deux métabolites m_0 et m_1. Le 0 logique est codé par une concentration forte de m_0 et une faible de m_1, le 1 logique codé par l'inverse. Si les deux concentrations sont faibles cela signifie "pas de signal" (équivalent aux portes électroniques *tri-state*). Les deux métabolites sont sensés ne jamais être en concentrations fortes simultanément.

L'intérêt de cette méthode est que même en faible concentrations, comme c'est leur différence qui porte l'information, l'affaiblissement est beaucoup moins problématique qu'avec le codage mono. Le codage des portes de base est très générique et il est beaucoup plus aisé de faire des inverseurs. Le point faible est que cette implémentation nécessite plus d'enzymes, et de type plus spécifique que le codage mono.

Si théoriquement tout est très simple, on se rend compte qu'en réalité cette approche est assez restrictive : le fil (ou les fils en mode dual) qui connecte la sortie d'une porte à l'entrée de la suivante est réalisé par un (ou deux) métabolite d'une espèce chimique particulière, qui ne sera donc pas réutilisable ailleurs dans le réseau sous peine de *court-circuit*. Cela veut dire que chaque fil de connection doit être d'une (ou deux) espèce chimique spécifique et qu'on doit trouver les enzymes qui prennent ces espèces chimiques comme substrats et qui fabriquent celles qui seront les substrats des enzymes de la porte suivante.

Quel que soit le codage utilisé, le passage d'un réseau enzymatique abstrait à son implémentation avec des enzymes et des métabolites réels n'est pas immédiat, d'où l'intérêt de se faire aider d'un logiciel adapté.

La *programmation* d'une fonction logique donnée avec un réseau enzymatique abstrait peut aussi se faire automatiquement. Du fait que pour une même fonction logique, plusieurs réseaux peuvent être proposés, il est d'autant plus intéressant de coupler cette phase avec celle d'implémentation qui aura alors plus de latitude pour résoudre les nombreuses contraintes.

Ayant obtenu un ou plusieurs réseaux enzymatiques qui implémentent la fonction logique désirée, il va falloir valider le ou les réseaux avec des simulations quantitatives de leur dynamique pour s'assurer que le calcul se fait correctement. Encore une fois, on a tout intéret à coupler cette phase de validation aux précédentes dans un logiciel intégré dédié à la conception de bio-calculateurs, partant de la formulation d'une fonction logique et allant jusqu'à la liste des composés biochimiques nécessaire à sa réalisation et les informations sur la dynamique du réseau.

Analyse de réseaux métaboliques existants

Le deuxième volet du stage de Marc Bouffard a été de concevoir et d'implémenter un logiciel permettant l'extraction automatique de circuits logiques fonctionnels à partir des réseaux métaboliques qui se trouvent dans des organismes vivants. Le but étant de constituer une bibliothèque réutilisable de composants logiques directement implémentés avec des biomolécules réelles.

Il a implémenté un algorithme qui recherche quelques types de portes (and, or, xor, nand, nor, ...) à deux entrées dans un réseau métabolique. L'algorithme commence par découper le réseau d'entrée en sous réseaux (ensemble de réactions connectées). Il énumère tous les sous-réseaux à 1 réaction, puis 2 réactions, etc. Puis, pour chaque sous-réseau, il trouve toutes les portes potentielles et ne conserve que celles qui sont recherchées.

Pour chaque porte trouvée, le programme sort le type de porte, la liste des réactions biochmiques impliquées et la liste des métabolites devant être présent initialement pour que la porte fonctionne. Le programme fournit aussi des statistiques sur le nombre de portes de chaque type présentes dans le réseau d'entrée.

Conclusion

Ce projet de système biologique artificiel de diagnostic médical est un projet de grande ampleur mettant en relation des informaticiens et des modélisateurs pour la partie conception et des biologistes et biochimistes pour la partie expérimentale (tests *in vitro* et réalisation des vésicules).

La premier volet sur la faisabilité a été validé, entre autres, à l'occasion de la thèse de Stéphanie Rialle. Maintenant, à la fois des parties plus fondamentales comme l'étude d'autres modèles de calcul et de nouvelles implémentations biochimiques, et plus appliquée comme la réalisation d'un prototype d'outil intégré d'aide à la conception de bio-calculateurs, vont constituer le projet de thèse de Marc Bouffard.

6.3 Calculer avec des bactéries

De façon complémentaire à l'approche précédente de conception *ab initio* de calculateurs biochimiques utilisant le paradigme standard de machine de Turing, avec mes collègues microbiologistes et biochimistes, on s'est intéressé à un nouveau paradigme de calcul qu'on a appelé *Bactoputing* (pour Bacterical Computing).

La motivation intiale est d'étudier comment on pourrait utiliser les différentes stratégies développées par des colonies de bactéries pour survivre dans des environnements variables, pour résoudre biologiquement des problèmes calculatoires complexes ("wet computing"). Nos réflexions nous ont menées à la possibilité d'utiliser des bactéries artificiellement modifiées pour concevoir un tel système de calcul.

Nous avons d'une part fait un inventaire de quelques problèmes calculatoires complexes : problèmes NP-complet, problèmes de reconnaissance de formes, problèmes d'optimisation, etc. et d'autre part un inventaire de processus connus chez les bactéries qui pourraient servir de moyens pour la résolution de chacun de ces types de problèmes. Parmi ces processus bactériens, on peut citer la chimiotaxie, la capacité qu'on des bactéries à remonter un gradient d'attractants biochimiques, ou la détection du quorum (phénomène de prise de décision collective d'une population de bactéries).

Divers scénarios ont été étudiés pour montrer sur des cas précis (le problème du voyageur de commerce par exemple) comment il serait possible de les résoudre en détournant certains mécanismes bactériens.

Nos travaux dans ce domaine ont donné lieu à deux publications : dans les actes de l'édition 2008 de la conférence *Modelling Complex Biological Systems in the Context of Genomics* [44] et dans le journal *Theory in Biosciences* [45].

6.4 La Réaction en Chaîne Mimétique (Mimic Chain Reaction)

La *Polymerase Chain Reaction* (PCR) est un moyen de multiplier, d'*amplifier*, un très faible nombre de copies d'un morceau d'ADN pour l'obtenir en grande quantité, soit pour le séquencer, soit quand on connaît sa séquence, pour détecter sa présence dans un gène (pour diagnostiquer une maladie génétique, ou tester une empreinte génétique, etc.).

On donne ici les grandes lignes d'une méthode générique d'amplification, la *Mimic Chain Reaction* (MCR), permettant d'obtenir en grande quantité des peptides ayant une structure tridimentionnelle qui imite celle de la cible. Cette cible pouvant non seulement être une protéine, mais aussi de beaucoup d'autres types de molécules.

Il y a deux grands types d'applications auxquelles cette méthode peut s'appliquer : (i) la détection par amplification d'une cible connue, mais présente en très faible concentration, et (ii) l'obtention d'un ensemble de peptides qui, imitant la structure d'une cible inconnue, en feraient une sorte de *portrait-robot*.

La MCR s'inspire de processus bien connus chez les bactéries : la *détection du quorum* (quorum sensing), les *systèmes régulateurs à deux composants* (two-component systems) et l'*autodisplay*.

La détection du quorum est un système de prise de décision associé à la densité de population. Le mécanisme est le suivant, les bactéries produisent un signal moléculaire appelé *auto-inducteur* qui est excrété. Lorsque la densité de population est faible, ce signal est en faible concentration dans l'environnement. Si la densité de population s'accroît, la concentration de l'auto-inducteur va augmenter et dépasser un seuil au delà duquel il est détecté par chaque bactérie qui, en conséquence surproduisent le signal auto-inducteur. C'est donc une boucle de rétroaction positive qui *informe* chaque bactérie que leur densité a dépassé un seuil. Selon les espèces bactériennes cette information peut mener à diverses réactions : bioluminescence chez *Vibrio fischeri*, contrôle de la division cellulaire chez *Escherichia coli*, etc.

Les systèmes régulateurs à deux composants servent à la bactérie pour détecter et réagir à des changements dans leur environnement. Ils sont constitués d'un récepteur membranaire de type *histidine kinase* qui détecte des signaux dans le milieu et d'une enzyme associée, le régulateur. Quand le récepteur à détecté son signal, il phosphoryle le régulateur qui change alors de conformation et passe dans un état actif lui permettant de réguler un gène cible.

L'*autodisplay* est la diffusion sur la surface extérieure d'une bactérie de peptides ou de protéines à l'aide d'un auto-transporteur. C'est le moyen qu'utilisent les bactéries Gram-négatives pour sécréter des protéines dans le milieu extra-cellulaire. Ce système permet d'avoir plus de 10^5 peptides par cellule sur sa surface.

Principe de fonctionnement

La MCR est composée de quatre acteurs :

1. la *cible* (connue ou inconnue),
2. le *récepteur* qui peut se lier à une partie de la cible, et qui dans ce cas engendre un signal,
3. l'*imitateur* qui a une structure suffisamment similaire à la cible pour être reconnu par le récepteur,
4. l'*amplificateur* qui comporte à la fois le régulateur qui répond au signal du récepteur et toute la machinerie contrôlée par ce régulateur, machinerie qui synthétise l'imitateur.

La cible peut être une protéine ou tout autre type de molécule. L'imitateur peut être n'importe quel type de protéine ou de peptide qui peut être encodé et qui adopte une grande variété de structures. L'amplificateur peut être une cellule, un liposome ou un autre conteneur. Si l'amplificateur est une cellule, le récepteur peut être une protéine membranaire et la machinerie de production régulée un gène codant l'imitateur.

Le principe de fonctionnement de la MCR est le suivant : la cible se lie au récepteur qui en conséquence signale au régulateur d'exprimer la synthèse de l'imitateur. L'imitateur est alors excrété et peut à son tour se lier sur plusieurs récepteurs de même type. Cette boucle de rétroaction positive permettrait en principe de détecter la présence d'une copie unique de virtuellement n'importe quel type de molécule.

Une bibliothèque de bactéries est alors construite de telle façon que chaque bactérie possède un récepteur transmembranaire ayant un domaine extracellulaire différent, et contienne un régulateur qui contrôle un gène codant pour un peptide imitateur, ainsi que la machinerie de transcription, de traduction et d'excrétion de ce peptide.

Toutes ces *bactéries-RM* (RM pour "receptor mimic") ont des récepteurs qui peuvent lier le peptide imitateur qu'elles produisent, c'est à dire que toutes les bactéries-RM possèdent des paires récepteur / imitateur apparentées.

Utilisations principales

Une protéine cible peut être considérée comme un ensemble d'antigènes qui peuvent chacun se lier à un récepteur différent.

Quand la MCR est utilisée pour amplifier une protéine cible connue, la majorité des bactéries-RM utilisées ont un récepteur capable de se lier à la cible (et à

l'imitateur correspondant qu'ils produisent), et très peu de bactéries de cette population auront des récepteur / imitateur différentes.

Quand la MCR est utilisée pour amplifier une protéine cible inconnue, on choisira une bibliothèque de bactéries qui ont des récepteurs capables de se lier à une grande variété de cibles potentielles et probablement seules quelques bactéries-RM auront les récepteurs pouvant effectivement se lier à la protéine cible,

Dans le cas d'une protéine cible inconnue, supposons que seulement quelques-unes des bactéries-RM de la bibliothèque se lient à elle. Si le circuit qui contrôle la production autocatalytique du peptide imitateur contrôle aussi un gène de résistance à un antibiotique, on pourra ségréger ces bactéries car ce seront les seules qui pourront se développer en présence de l'antibiotique.

Vues dans leur ensemble, ces bactéries possèdent les gènes qui codent pour un ensemble de peptides imitateurs qui sont chacun une copie partielle de la protéine cible. Les séquences de ces peptides correspondent à celles de la protéine cible, et les séquences des récepteurs correspondent à celles des anticorps à la protéine cible. Cette information peut être utilisée pour caractériser et identifier la protéine cible. De plus, ce sous-ensemble de la bibliothèque de bactéries-RM originales constitue une population qui pourra à l'avenir être utilisée pour détecter la même protéine ou des protéines ressemblantes.

Autres utilisations

détection de structure

L'analyse structurale d'un ensemble de paires récepteur / imitateur, par exemple par RMN peut donner un éclairage sur la structure de la partie de la cible qui se lie au récepteur. Comme les structures de chaque imitateur et du récepteur associé peuvent être obtenues séparément, ces deux informations peuvent être combinées pour raffiner la structure de la cible. De plus, il peut tout à fait exister plusieurs paires récepteur / imitateur pour le même épitope.

En raccordant les structures des paires récepteur / imitateur de différents épitope on peut concevoir un *portrait-robot* assez précis qui pourrait caractériser la protéine cible.

détection de formes alternatives de la même molécule

Les macromolécules peuvent subir énormément de modifications post traductionnelles, comme des phosphorylations, méthylations, etc. Des changements conformations significatifs peuvent résulter de ces modifications qui peuvent affecter la signalisation par exemple. On pourrait utiliser la MCR pour générer tout un ensemble de paires RM pour un grand intervalle de conformations possibles de la cible.

détection simultanée de différentes cibles

Une extension intéressante de la MCR serait de concevoir différentes bibliothèques de bactéries-RM de façon à ce que le régulateur active aussi une gène codant pour une protéine fluorescente de couleur différente selon la cible (GFP, EBFP ou YFP). De cette façon il serait possible de détecter simultanément jusqu'à trois cibles différentes dans le même échantillon.

Conclusion

La PCR a eu un impact majeur en biologie. Une technique plus générale d'amplification applicable à tous types de molécules cibles, comme la MCR, devrait avoir impact similaire. Ce qu'on peut y gagner potentiellement vaut le travail nécessaire pour surmonter quelques obstacles techniques : bruit, liaisons non spécifiques et la propension qu'ont les bactéries à muter.

Ces travaux ont fait l'objet d'une publication dans le *Journal of Molecular Microbiology and Biotechnology* [46].

Métabolisme des streptomyces

7

Depuis 2010 je collabore avec Marie-Joëlle Virolle, microbiologiste à l'Institut de Microbiologie et de Génétique à Orsay. Marie-Joëlle s'intéresse au métabolisme secondaire des *streptomyces* qui sont des bactérie du sol connues pour produire des antibiotiques et antifongiques. Trois voies de production d'antibiotiques sont actuellement connues mais une vingtaine de voies putatives ont été mises en évidence par des moyens bioinformatiques. On travaille à concevoir des modèles qualitatifs et quantitatifs du métabolisme secondaire de ces bactéries pour, entre autres, trouver des conditions qui activeraient une ou plusieurs de ces voies de production de nouveaux antibiotiques.

7.1 Outils biotechnologiques

Les *streptomyces* sont des bactéries Gram-positive dont le génome à plus de 73% de paires GC. Ces bactéries produisent de nombreuses molécules d'intérêt médical (antibiotiques, anticancereuses, antifongiques) et pour l'agriculture (herbicides, insecticides). Ces composés, issus du métabolisme secondaire, sont produits généralement tardivement à partir de précurseurs du métabolisme primaire. Ces deux métabolismes sont très liés chez les *streptomyces* et ces connexions sont difficiles à révéler.

Le *knock-out* de gènes qui codent des protéines du métabolisme primaire (plus connu) a souvent des effets délétères sur la croissance et la robustesse de la bactérie, ne permettant pas d'obtenir des informations sur le métabolisme secondaire. À l'opposé, la sur-expression massive d'une enzyme peut entrainer un déséquilibre métabolique tout aussi létal.

Par contre, la modulation fine de l'expression de genes codant des enzymes limitantes (*bottleneck)* peut mener à une redirection du flux métabolique vers une voie où la production du composé désiré sera optimale. C'est pourquoi il est intéressant d'avoir des outils biotechnologiques permettant un contrôle fin de l'expression des gènes. Étonnament, pour les *streptomyces*, il n'existe que peu

de promoteurs constitutifs ou inductibles. L'ajout d'inducteurs ayant quelquefois des effets secondaires non voulus sur la physiologie de la bactérie, il est apparu utile de concevoir une bibliothèque de promoteurs constitutifs couvrant un large intervalle de force.

Méthode

La méthode utilisée est basée sur celle de Jensen et Hammer [47, 48], elle consiste à synthétiser des promoteurs ayant une séquence engendrée aléatoirement aux alentours des positions -35 à -10 de séquences promotrices concensus pour les *streptomyces*.

Pour tester la force des promoteurs synthétisés, on les a clonés devant le gène *aphII*, qui rend la bactérie résistante à un antibiotique de type aminoglycocide, la neomycine. Les lignées résultantes ont montré une résistance à différentes concentrations de neomycine, indiquant que différents niveaux d'expression du gène *aphII* ont été obtenus.

Les promoteurs synthétisés ont été classés selon la dose de neomycine à partir de laquelle les colonies de bactéries correspondantes réussissaient à survivre, et répartis en trois groupes, forts, moyens et faibles. Pour valider la méthode des tests de survie, le taux d'expression de quelques uns des promoteurs de chaque classe a aussi été mesuré par RT-PCR. Les résultats se sont révélés cohérents entre les deux méthodes.

Analyse des séquences promotrices

Sur les 38 promoteurs séquencés, les séquences des 14 les plus faibles et des 14 les plus forts ont été comparées, dans le but de trouver les caractéristiques de séquence qui leur confèrent leur activité (fig. 7.1).

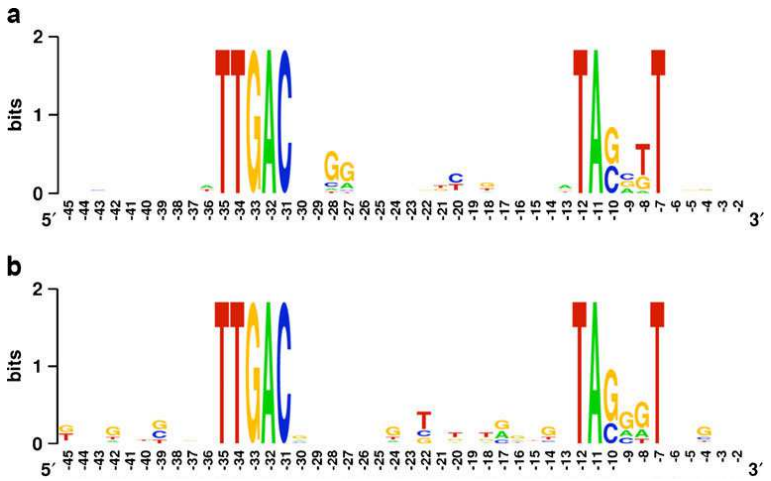

FIG. 7.1: Séquences des 14 plus faibles (a) et des 14 plus forts (b) promoteurs affichées avec WebLogo.

Les caractéristiques importantes des séquences pour la force d'un promoteur se répartissent en deux grandes catégories : les éléments reconnus comme double

brin (E^{ds}) qui sont importants pour la reconnaissance et la liaison de l'ARN-polymérase et les éléments reconnus comme simple brin (E^{ss}) qui sont importants pour l'isomérisation de la polymérase et la formation du complexe ouvert. Les E^{ds} sont la boîte -35 et certains éléments en amont, ainsi que le motif de l'élément étendu -10 ($^{-17}$tgTGnT^{-12}), tandis que les E^{ss} sont les éléments ^{11}ATAAT^{-7} (majorité de la boîte -10) et les éléments $^{-5}$G.

De façon surprenante, notre étude a montré une fréquence élevée de G au niveau des positions 3, 4, et 5 de l'hexamère -10 des promoteurs forts (TAGGGT), alors que chez *E. coli*, cette boîte est généralement riche en AT (TATAAT). De même, l'"élément étendu -10" trouvé par l'analyse WebLogo en amont de la séquence -10 des promoteurs forts, mais pas en amont des faibles ($^{-17}$PuGgGnT^{-12}) est plus riche en G que celui établi pour *E. coli* ($^{-17}$tgTGnT^{-12}). Quand il est étendu (position -18 inclue), ce motif ($^{-18}$Tg/aGgGnT^{-12}) posséde une similitude frappante avec la boîte -10 elle-même (TAGGGT). Cette observation indique que même une répétition imparfaite de la séquence -10 pourrait aider la polymérase à se placer correctement sur la séquence.

On constate que les sequences de la boîte -10, du motif étendu -10, aussi bien que l'espaceur des promoteurs forts, sont plus riches en G que les promoteurs faibles. La richesse / répartition en G ainsi que la présence de séquences courbant l'ADN dans l'espaceur des promoteurs forts pourraient être un facteur déterminant dans la courbure de l'ADN et donc de la force.

En conclusion, cette étude a permis (i) de concevoir une banque réutilisable de promoteurs constitutifs de différentes forces pour les *streptomyces* et (ii) de donner des prémisses d'explications sur les raisons de la forces de ces promoteurs en rapport avec leur séquence.

Ce travail a été publié dans *Applied Genetics and Molecular Biotechnology* [49].

7.2 Courbe sigmoïde de résistance à la neomycine

Lors des travaux ayant menés à la publication précédente, on a été surpris par la courbe obtenue en classant les promoteurs par force en fonction du taux de survie des colonies de bactéries. En regardant les valeurs numériques de ces taux de survie, on se rend compte qu'elles varient dans un intervalle couvrant 6 ordres de grandeurs (de moins de 10^{-4} à 10^{2}).

Mais ce qui est le plus surprenant c'est la forme sigmoïde de la courbe : les 8 premiers promoteurs, jusqu'à *B4-21* sont trop faibles et le taux de survie est inférieur à 0.01%, puis le taux de survie augmente fortement avec les 5 promoteurs suivants, jusqu'à *D1-10* avec un taux de survie supérieur à 75% et va vers 100% sur les derniers (fig. 7.2).

Bien entendu, si les taux d'expression du gène de résistance sous le contrôle des différents promoteurs suivent eux aussi une courbe analogue, rien n'est plus surprenant ! Pour s'en assurer on a mesuré par RT-PCR les taux d'expression du

gène avec 5 promoteurs choisis parmi ceux (i) qui induisent un faible taux de survie, (ii) dans la partie ou le taux de survie augmente de façon abrupte, et (iii) dans les forts taux de survie.

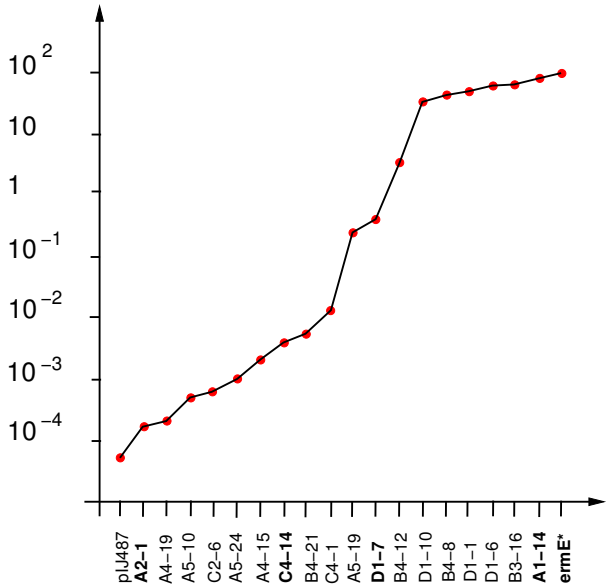

FIG. 7.2: Courbe de survie en fonction de la force des promoteurs dans une culture contenant 100μg/ml de neomycine. Le taux d'expression des promoteurs dont les noms sont en gras à aussi été mesuré par RT-PCR.

Notre intuition à été confirmée par les résultats de RT-PCR, le taux de transcription varie dans un rapport de 1 à 12, et surtout de façon quasi-linéaire (fig. 7.3 à gauche). La forme sigmoïde de la courbe de survie en fonction du taux d'expression du gène de résistance est donc due à une autre cause (fig. 7.3 à droite).

Ce type de comportement, réaction sigmoïde à une action linéaire, est caractéristique de systèmes contenant un cycle de rétroaction positif. C'est en regardant en détail le processus par lequel les antibiotiques de type aminoglycocide empoisonnent les bactéries ainsi que le processus de défense utilisé par les bacteries ayant un gène de résistance de type *aphII* qu'on a pu trouver le mécanisme expliquant ce comportement et en faire un modèle.

Le modèle

Les antibiotiques de type aminoglycocide empoisonnent les bactéries en se fixant de façon irréversible sur les ribosomes, et de ce fait en en empêchant le bon fonctionnement [50]. Si le nombre de ribosomes invalidés est trop important, la synthèse des protéines va être de plus en plus faible et la bactérie n'y survivra pas.

Il est clair que le taux d'invalidation des ribosomes est proportionnel à la concentration d'antibiotique dans la bactérie. Quand la bactérie ne possède pas

de gène de résistance, la concentration d'antibiotique dans la bactérie est proportionnel à sa concentration dans le milieu extérieur.

Les gènes de résistance de type *aphII* fonctionnent en codant une protéine, l'enzyme aphII, qui inactive la neomycine qui ne peut plus se fixer sur les ribosomes [51]. Les bactéries possédant ce gène contrent l'effet létal de la neomycine en en réduisant la concentration interne.

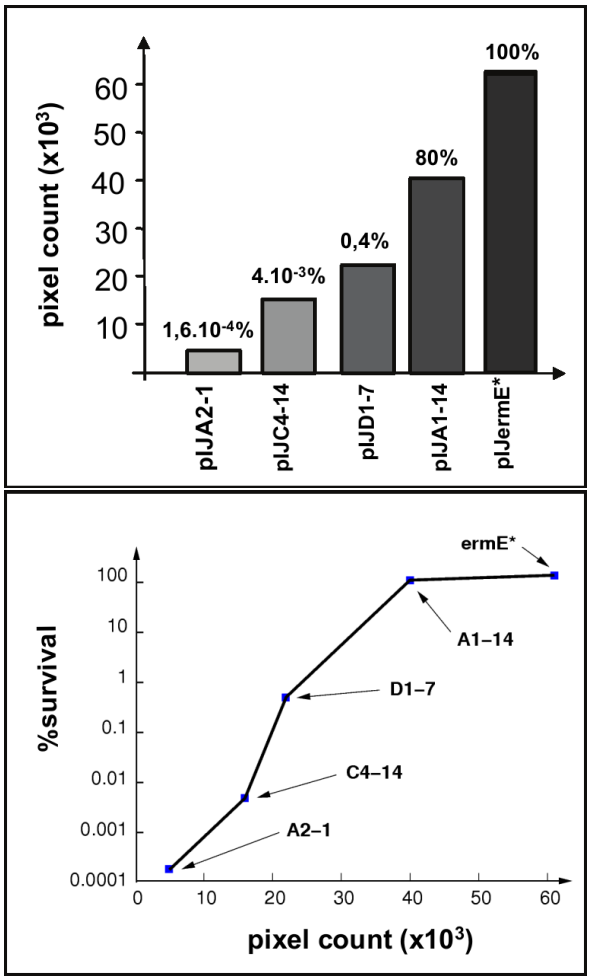

FIG. 7.3: Taux d'expression des promoteurs mesurée par RT-PCR (à gauche) et taux de survie en fonction de ces taux d'expression (à droite).

La boucle de rétroaction positive mentionnée précédemment se trouve dans le fait que la neomycine attaque l'appareil traductionnel et que c'est une enzyme, traduite à partir de l'expression du gène, qui va en diminuer la concentration interne. Cette réflexion nous a conduit a proposer le modèle suivant (fig. 7.4).

Dans ce modèle la transcription du gène *aphII* se fait à un taux constant, fonction du promoteur utilisé, la traduction se faisant en fonction de la disponibilité des ribosomes, la quantité d'enzymes aphII synthétisée y est directement liée. Selon la

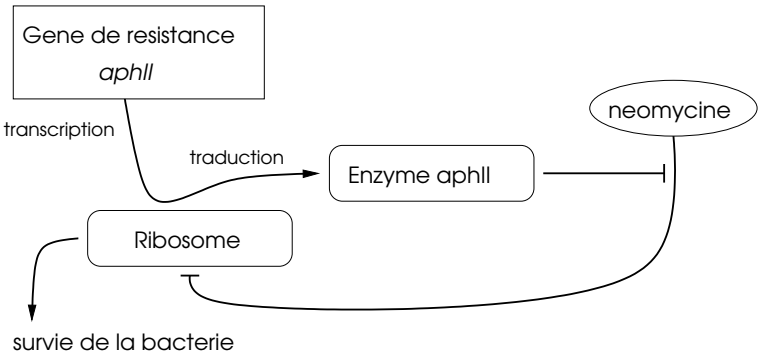

FIG. 7.4: Réseau d'interactions entre le gène *aphII* et la neomycine.

concentration d'enzymes aphII, une quantité de neomycine correspondante sera invalidée, tendant ainsi à diminuer sa concentration interne. Comme la quantité de ribosomes empoisonnés est directement fonction de la concentration interne de neomycine, la traduction des ARNm, et spécifiquement celui d'aphII, sera donc inhibée par celle-ci.

On arrive donc à un modèle où on a trois interactions en boucle qui forment un cycle positif :

1. la quantité de ribosomes fonctionnels influence positivement la synthèse d'aphII ;
2. la concentration d'aphII influence négativement la concentration interne de neomycine ;
3. la concentration interne de neomycine influence négativement le nombre de ribosomes fonctionnels.

Pour une concentration externe constante de neomycine (100μg/ml dans nos expériences), ce type de circuit a une dynamique sigmoïde en terme de nombre de ribosomes fonctionnels en fonction du taux de transcription du gène *aphII*. Pour chaque concentration externe de neomycine il existe un taux de transcription seuil où le nombre de ribosomes fonctionnels est la limite de viabilité pour la bactérie.

De façon stochastique les bactéries vont survivre ou pas. Au niveau macroscopique, la résultante de ce processus est le taux de survie moyen de la colonie qui est statistiquement soit proche de 0 soit proche de 100%, ainsi qu'on l'a observé dans nos expériences.

Conclusion

Le gène *aphII* est largement utilisé comme système rapporteur chez les *streptomyces*, bien que ce système très utile ait été quelquefois décrié à cause de comportements aberrants. La dynamique de ce système, comme expliquée dans notre étude, peut donner un éclairage nouveau à ces comportements paradoxaux.

Une autre conséquence de notre étude est qu'il n'est pas nécessaire d'éteindre, mais de réduire l'expression d'un gène de résistance juste en dessous d'un seuil pour augmenter l'action létale de l'antibiotique correspondant. C'est pourquoi des inhibiteurs de transcription et/ou de traduction sont associés, dans certains cas, à l'antibiotique pour venir à bout de certaines souches bactériennes particulièrement résistantes.

Cette étude a été publiée dans *AMB express* [52].

Conclusion et perspectives

8

À l'occasion de la transition que j'ai opérée au début des années 2000 depuis une science exacte, l'informatique, vers la science expérimentale qu'est la biologie, j'ai eu le plaisir de rencontrer des chercheurs ayant des façons très différentes d'appréhender les problèmes scientifiques que celles auxquelles j'étais habitué en tant qu'informaticien et ancien mathématicien.

Les expériences, qui servent à valider ou invalider une hypothèse, sont parfois longues et difficiles à mener, elles sont sensibles à beaucoup de facteurs extérieurs. Les résultats numériques d'une expérience *réussie* peuvent paradoxalement montrer une grande variabilité. Cela explique les approches différentes de celles des sciences exactes, moins complexes de ce point de vue.

Les organismes vivants, même les plus simples bactéries, sont des systèmes complexes qui ont évolués pendant des millénaires. Leurs phénotypes sont, à un niveau d'échelle supérieur, les conséquences des interactions des acteurs des niveaux inférieurs. De plus, tous les niveaux d'échelle interfèrent et se contraignent entre eux. Ces organismes sont cependant suffisamment adaptables à leur environnement pour que cela puisse expliquer en partie la difficulté à mener des expériences et la variabilité des résultats.

C'est la raison pour laquelle je me suis intéressé aux systèmes complexes, aussi bien du point de vue de l'informaticien qui cherche à les reproduire sur ordinateur, que du biologiste theorique, qui tente de concevoir des modèles pour en predire et expliquer le comportement.

Simulation

En tant qu'informaticien, j'ai commencé par étudier les différentes méthodes et systèmes permettant de calculer l'évolution temporelle des concentrations de composés dans un réacteur virtuel (après tout les organismes vivants sont une sorte de réacteur biochmique !). Comme on l'a vu dans le chapitre 2, il y a basiquement trois types d'approches : (i) l'approche mathémathique avec les systèmes d'équations différentielles ordinaires et aux dérivées partielles, (ii) les simulation stochastiques globales, et (iii) les systèmes de simulation entité-centrés.

J'ai délaissé momentanément les approches *top-down* de type systèmes d'équations différentielles, où par définition on connaît *a priori* les lois globales, pour me focaliser sur les systèmes de simulation de type entité-centrés, en raison de leur forte expresssivité, et de leur aspect *bottom-up* permettant de faire émerger des comportements globaux non prévus.

Avec HSIM, j'ai cherché à obtenir un système de simulation efficace (beaucoup de statistique et d'algorithmique fine expliquent son efficacité en terme de temps calcul) mais simple d'usage, cachant à l'utilisateur sa complexité interne. J'y ai adjoint un nouvel algorithme de simulation stochastique global particulièrement efficace et ne souffrant pas des défauts de ceux existant, et aussi la possibilité de mixer les deux approches, offrant ainsi les avantages de chacune sans leurs inconvénients. Enfin, je suis revenu sur les systèmes continus déterministes, en incluant un module permettant de générer automatiquement le système d'équations différentielles ordinaires équivalent au modèle d'entrée, offrant ainsi à l'utilisateur un panel complet de méthodes de résolution de son modèle.

Des points de vues de la simplicité d'usage, de la grande diversité possible des modèles d'entrée, et de son efficacité en terme de temps de calcul, je pense avoir réussi le challenge de faire de HSIM l'un des systèmes de simulation de processus biochimiques le plus abouti disponible actuellement.

Néanmoins, de nouvelles fonctionnalités restent à lui apporter, comme par exemple, pouvoir fournir le modèle à simuler non plus avec la liste des réactions, mais avec une représentation du réseau biochimique à simuler (HSIM déduisant les réactions à partir du réseau). Cette représentation pourrait bien sûr être textuelle, mais il serait commode d'adjoindre à HSIM un éditeur de réseaux adapté aux réseaux biologiques, permettant à l'utilisateur d'obtenir directement le comportement spatial et temporel d'un réseau particulier.

Une autre optimisation de HSIM, purement informatique, sera d'en faire une version multithreads pour calculateurs multiprocesseurs et/ou une version distribuée, pour pouvoir l'utiliser sur un cluster. Si j'ai déjà expérimenté plusieurs pistes pour la version parallèle multiprocesseurs, la version distribuée est plus ardue à mettre en oeuvre (principalement pour des raisons de synchronisation temporelle). Par contre si on veut simuler le comportement d'une colonie de bactéries avec leurs interactions, le calcul peut se distribuer beaucoup plus facilement.

Systèmes complexes

Les travaux menés par Thomas Moncion au cours de sa thèse ont permis de défricher le terrain de la détection automatique de phénomènes émergents. En plus des études théoriques qu'il a faites sur le sujet, il a développé des prototypes permettant de faire une preuve de concept de son approche.

Une des améliorations à y apporter serait d'intégrer un mécanisme automatique ou semi-automatique de détection de profils 'intéressants' des courbes correspondant à l'évolution dans le temps des nombres d'agents de chaque classe (en utilisant des séries chronologiques par exemple). Cela permettrait de détecter

les profils *oscillatoires* de la concentration de composés lors de la simulation d'un système biologique de type horloge circadienne, ou de détecter le profil sigmoïde caractéristique d'un phénomène de nucléation.

Une autre voie qui me semble être très prometteuse serait de coupler l'évolution temporelle de ces mesures d'interactions avec un système d'apprentissage automatique dans le but d'extraire une information synthétique sur le comportement macroscopique du système en fonction de ses divers paramètres. On pourrait dans une certaine mesure *abstraire* les détails microscopiques du modèle pour passer au niveau supérieur. Un bon exemple serait de pouvoir automatiquement simuler les variations du comportement d'une colonie de bactéries en fonction de paramètres (environnementaux ou autres), sans avoir besoin de simuler le fonctionnement fin de chacune des bactéries.

Enfin, une amélioration envisageable consisterait à complètement automatiser la reconnaissance de phénomènes émergents en détectant par programme le changement d'allure des courbes des différentes mesures et d'intégrer un tel module à HSIM.

Hyperstructures et modélisation

L'hypothèse de grands assemblages macromoléculaire fonctionnels, les hyperstructures, est de plus en plus corroborée expérimentalement. Des couplages entre hyperstructures comme par exemple la glycolyse et les mécanismes contrôlant la réplication de l'ADN ont été démontrés : certaines enzymes de la glycolyse ont des activités liées à l'initiation de la réplication chez *Bacillus subtilis*. De même entre la voie de la glycolyse et le cytosquelette dans les cellules eucaryotes. On est en train de préparer une publication quantifiant l'impact de la fixation sur des microtubules d'enzymes de la glycolyse, sur à la fois l'efficacité de la voie et sur la stabilité des microtubules.

On pourrait envisager le fonctionnement des cellules à un niveau supérieur à celui des macromolécules, celui des hyperstructures dialoguant entre elles par des voies de signalisation internes. Cela constitue une nouvelle vision de la cellule, plus intégrée, mettant en jeu moins de participants qu'au niveau macromoléculaire, mais d'un niveau de complication plus élevé. Ces hyperstructures ne sont pas forcément figées dans le temps, elles sont le résultat d'un équilibre moléculaire dynamique à une période donnée du cycle cellulaire, pendant laquelle elles réalisent une fonction particulière.

Avec les *Functioning-dependent Structures* on a considéré un cas très élémentaire d'hyperstructure, celui d'une partie de voie métabolique (quelques enzymes en cascade) qui s'auto-assemble du fait de la présence du substrat initial, et se défait après avoir terminé de le métaboliser. On a montré que même avec un exemple simple à deux enzymes, la dynamique de l'ensemble offrait un panorama très riche de comportements.

D'autres avantages de l'auto-assemblage d'une partie des enzymes d'une voie métabolique peuvent exister pour la cellule. Chez *Escherichia coli* une partie de

la voie de glycolyse est commune pour la métabolisation de différents sucres (glucose, lactose, galactose, maltose) ; si ces enzymes sont déjà assemblées, quand un sucre vient à manquer mais qu'un autre est présent, la cellule est déjà prête à le métaboliser. Il serait particulièrement intéressant d'étudier le comportement transitoire lorsque la cellule change de type de nutriment, avec et sans FDS.

Métabolisme des streptomyces

L'étude que nous avons menée sur la façon dont le gène *aphII* permet aux *streptomyces* de résister aux antibiotiques de type aminoglycocides nous a permis de montrer à beaucoup de nos collègues biologistes le pouvoir explicatif que peut apporter un modèle.

Ces travaux sur *aphII* ont constitué en fait la première étape d'une collaboration de plus longue haleine pour l'obtention d'un modèle plus complet du métabolisme secondaire des *streptomyces* dans le but de pouvoir activer des voies de production de nouveaux antibiotiques, et aussi de permettre de produire des composés de type *biofuel*. En effet, la première étape de l'utilisation raisonnée de bacteries pour la production de composés d'interêt thérapeutique ou industriel est la connaissance suffisante des mécanismes mis en jeu, avant l'étape suivante qui est plus de l'ordre de la biologie de synthèse.

Biologie de synthèse

Les travaux menés en collaboration avec les collègues du laboratoire Sysdiag, sur l'ingéniérie *ab initio* de portes logiques utilisant des réactions enzymatiques et sur les outils informatiques permettant de les concevoir sont déjà bien avancés avec entre autres la thèse de Stéphanie Rialle.

Des exemples de circuits enchaînant quelques portes ont déjà été réalisés, partant des spécifications du réseau métabolique abstrait, à une implémentation avec des enzymes et métabolites spécifiques, aux simulations de la dynamique du réseau, puis enfin à la validation expérimentale *in vitro*.

Les pistes pour continuer ces travaux sont multiples :

1. approfondir l'étude de toutes les contraintes inhérentes au modèle de portes logiques utilisé (synchronisation, affaiblissement, etc.),
2. établir automatiquement une bibliothèque d'implémentations de composants logiques à partir de réseaux métaboliques d'organismes existants,
3. réaliser un outil intégré d'aide à la conception, à l'implémentation et à la simulation de bio-calculateurs dédiés au diagnostic médical,
4. étudier d'autres approches pour implémenter des portes logiques avec un réseau enzymatique qui n'auraient pas les mêmes contraintes que celle actuellement utilisée et éventuellement intégrer ces nouvelles implémentations dans l'outil d'aide à la conception.

Une large partie de ces points constitue le projet de la thèse de Marc Bouffard que je vais co-encadrer au LRI, avec Franck Molina à Sysdiag.

Bibliographie

[1] H. Kitano, *Foundations of Systems Biology*. MIT Press, Cambridge, MA (USA), 2002.

[2] H. Kitano, "Systems biology : a brief overview," *Science*, vol. 295, pp. 1662–1664, March 2002.

[3] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikushi, and H. Kitano, "Celldesigner 3.5 : A versatile modeling tool for biochemical networks," *Proceedings of the IEEE*, vol. 96, pp. 1254–1265, Aug 2008.

[4] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. L. Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and S. B. M. L. Forum, "The systems biology markup language (sbml) : a medium for representation and exchange of biochemical network models.," *Bioinformatics*, vol. 19, pp. 524–531, Mar. 2003.

[5] D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R.-Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z.-Q. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. H. III, H. O. Smith, and J. C. Venter, "Creation of a bacterial cell controlled by a chemically synthesized genome," *Science*, vol. 329, no. 5987, pp. 52–56, 2010.

[6] C. Guldberg and P. Waage, "Studies concerning affinity," *C. M. Forhandlinger i Videnskabs-Selskabet i Christiana*, p. 1864 :35, 1864.

[7] B. Müller-Hill, "Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator.," *J. Mol. Biol.*, vol. 257, pp. 21–29, 1996.

[8] B. Müller-Hill, "The function of auxiliary operators," *Molecular Microbiology*, vol. 29, no. 1, pp. 13–18, 1998.

[9] D. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[10] Y. Cao, D. Gillespie, and L. Petzold, "Efficient formulation of the stochastic simulation algorithm for chemically reacting systems," *Journal of Chemical Physics*, vol. 121, pp. 4059–4067, 2004.

[11] H. Byrne and D. Drasdo, "Individual-based and continuum models of growing cell populations : a comparison.," *Journal of Mathematical Biology*, vol. 58, no. 4-5, pp. 657–687, 2009.

[12] S. S. Andrews, N. J. Addy, R. Brent, and A. P. Arkin, "Detailed simulations of cell biology with smoldyn 2.1," *PLoS Comp. Biol.*, vol. 6, mar 2010.

[13] M. Klann and H. Ganguly, A.and Koeppl, "Hybrid spatial gillespie and particle tracking simulation," *BMC Bioinformatics*, vol. 28, no. 18, 2012.

[14] V. Norris, T. Den Blaauwen, A. Cabin-Flaman, R. Doi, R. Harshey, L. Jannière, A. Jimenez-Sanchez, D. Jun Jin, P. Levin, E. Mileykovskaya, A. Minsky, M. Saier Jr, and K. Skarstad, "Functional taxonomy of bacterial hyperstructures," *Microbiology and Molecular Biology Reviews*, vol. 71, no. 1, pp. 230–253, 2007.

[15] R. Brown, "On the general existence of active molecules in organic and inorganic bodies," *Philos Mag, Ann Philos*, vol. 4, 1828.

[16] A. Einstein, "Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. (*Sur la théorie moléculaire cinétique de la chaleur requise par le mouvement des particules en suspension dans un liquide au repos.*)," *Annalen der Physik*, vol. 17, pp. 549–560, 1905.

[17] P. Amar and L. Paulevé, "HSIM : an hybrid stochastic simulation system for systems biology," in *The Third International Workshop on Static Analysis and Systems Biology (SASB 2012)*, (Deauville, France), Sept. 2012.

[18] P. Amar, G. Bernot, and V. Norris, "Hsim : a simulation programme to study large assemblies of proteins," *Journal of Biological Physics and Chemistry*, vol. 4, pp. 79–84, 2004.

[19] P. Amar, G. Legent, M. Thellier, C. Ripoll, G. Bernot, T. Nystrom, M. S. Jr, and V. Norris, "A stochastic automaton shows how enzyme assemblies may contribute to metabolic efficiency," *BMC Systems Biology*, vol. 2, no. 27, 2008.

[20] P. Tracqui, E. Promayon, P. Amar, N. Huc, V. Norris, and J.-L. Martiel, "Emergent features of cell structural dynamics : a review of models based on tensegrity and nonlinear oscillations," in *Proceedings of the DIEPPE spring school on Modelling and simulation of biological processes in the context of genomics*, pp. 161–190, Platypus Press, 2003.

[21] T. Moncion, G. Hutzler, and P. Amar, "Validation d'une simulation à base d'agents par l'utilisation d'un réseau de petri," in *Systèmes multi-agents vers la conception de systèmes artificiels socio-mimétiques (JFSMA 2005)* (D. Alexis and R. Eric, eds.), pp. 187–190, Hermès Lavoisier, 2005.

[22] T. Moncion, G. Hutzler, and P. Amar, "Verification of biochemical agent-based models using petri nets," in *International Symposium on Agent Based Modeling and Simulation, ABModSim'06* (T. Robert, ed.), pp. 695–700, Austrian Society for Cybernetics Studies, 2006.

[23] T. Moncion, G. Hutzler, and P. Amar, "Validation of an agent based system using petri nets," in *Proceedings of the Fourth International Joint Conference on Autonoumous Agents and Multi Agent Systems* (F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. Pechoucek, M. Singh, D. Steiner, S. Thompson, and M. Wooldridge, eds.), (New York), pp. 1365–1366, ACM Press, 2005.

[24] T. Moncion, P. Amar, and G. Hutzler, "Automatic characterization of emergent phenomena in complex systems," *Journal of Biological Physics and Chemistry*, vol. 10, pp. 16–23, 2010.

[25] P. Amar, G. Bernot, and V. Norris, "Modelling and simulation of large assemblies of proteins," in *Proceedings of the DIEPPE spring school on Modelling and simulation of biological processes in the context of genomics*, pp. 41–48, Platypus Press, 2003.

[26] M. Thellier, G. Legent, P. Amar, V. Norris, and C. Ripoll, "Steady-state kinetic behaviour of functioning-dependent structures," *The FEBS Journal (Federation of European Biochemical Societies)*, vol. 273, pp. 4287–4299, 2006.

[27] J. Deutscher, C. Francke, and P. Postma, "How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria," *Microbiol Mol Biol Rev.*, vol. 70, pp. 939–1031, Dec 2006.

[28] J. Mowbray and V. Moses, "The tentative identification in escherichia coli of a multi-enzyme complex with glycolytic activity," *European Journal of Biochemistry*, vol. 66, pp. 25–36, 1976.

[29] D. Gorringe and V. Moses, "A multi-enzyme aggregate with glycolytic activity from escherichia coli," *Biochemical Society Transactions*, vol. 6, pp. 167–169, 1978.

[30] P. Amar, V. Norris, M. S. jnr., and G. Bernot, "Metabolons and hyperstructures in cell metabolism and signaling," in *Actes du colloque 4th World Congress of Cell and Molecular Biology*, 2005.

[31] F. Commichau, F. Rothe, C. Herzberg, E. Wagner, D. Hellwig, M. Lehnik-Habrink, E. Hammer, U. Völker, and J. Stülke, "Novel activities of glycolytic enzymes in bacillus subtilis : Interactions with essential proteins involved in mrna processing.," *Mol Cell Proteomics*, vol. 8, pp. 1350–1360, Jun 2009.

[32] J. Ovàdi, "Experiment-based mathematical modelling of energy metabolism in diseases caused by unfolded/misfolded proteins," in *Modelling Complex Biological Systems in the Context of Genomics* (P. Amar, F. Képès, V. Norris, and G. Bernot, eds.), pp. 47–50, EDP Sciences, 2009.

[33] S. Chuong, A. Good, G. Taylor, M. Freeman, G. Moorhead, and D. Muench, "Large-scale identification of tubulin-binding proteins provides insight on subcellular trafficking, metabolic channeling, and signaling in plant cells," *Mol Cell Proteomics*, vol. 3, no. 10, pp. 970–983, 2004.

[34] A. Real-Hohn, P. Zancan, D. Da Silva, E. Martins, L. Salgado, C. Mermelstein, A. Gomes, and M. Sola-Penna, "Filamentous actin and its associated binding proteins are the stimulatory site for 6-phosphofructo-1-kinase association within the membrane of human erythrocytes," *Biochimie*, vol. 92, no. 5, pp. 538–544, 2010.

[35] V. Norris, P. Amar, G. Legent, C. Ripoll, M. Thellier, and J. Ovadi, "Sensor potency of the moonlighting enzyme-decorated cytoskeleton," *BMC Biochemistry*, vol. 14, no. 3, 2013.

[36] E. Rocha, J. Fralick, G. Vediyappan, A. Danchin, and V. Norris, "A strand-specific model for chromosome segregation in bacteria," *Mol. Microbiol.*, vol. 49, pp. 895–903, 2003.

[37] V. Norris, L. Janniere, and P. Amar, "Hypothesis : Variations in the rate of dna replication determine the phenotype of daughter cells," in *Modelling Complex Biological Systems in the Context of Genomics*, pp. 71–82, EDP Sciences, 2007.

[38] V. Norris and P. Amar, "Life on the scales : initiation of replication in escherichia coli," in *Modelling Complex Biological Systems in the Context of Genomics*, pp. 55–77, EDP Sciences, 2012.

[39] V. Norris and P. Amar, "Chromosome Replication in Escherichia coli : Life on the Scales," *Life*, vol. 2, pp. 286–312, Oct. 2012.

[40] E. Bromley, K. Channon, E. Moutevelis, and D. Woolfson, "Peptide and protein building blocks for synthetic biology : from programming biomolecules to self-organized biomolecular systems," *ACS Chemical Biology*, vol. 3, pp. 38–50, Jan 2008.

[41] S. Rialle, *Méthodologie et outils bioinformatiques d'aide à la conception de systèmes biologiques synthétiques pour de nouveaux diagnostics en santé humaine*. PhD thesis, Université Montpellier II, Oct 2010.

[42] S. Peres, S. Rialle, L. Felicori, and F. Molina, "Formal language for detailed structure function annotation based on elementary bricks of action," *Bioinformatics*, vol. 26, no. 12, pp. 1542–1547, 2010.

[43] S. Rialle, L. Felicori, C. Dias-Lopes, S. Peres, S. E. Atia, A. R. Thierry, P. Amar, and F. Molina, "Bionetcad : design, simulation and experimental validation of synthetic biochemical networks," *Bioinformatics*, vol. 26, no. 18, pp. 2298–2304, 2010.

[44] V. Norris, A. Zemirline, P. Amar, P. Ballet, E. B. Jacob, G. Bernot, G. Beslon, E. Fanchon, J.-L. Giavitto, N. Glade, P. Greussay, Y. Grondin, J. A. Foster, G. Hutzler, F. Képès, O. Michel, G. Misevic, F. Molina, J. Signorini, P. Stano, and A. Thierry, "From bioputing to bactoputing : computing with bacteria," in *Modelling Complex Biological Systems in the Context of Genomics*, pp. 123–150, EDP Sciences, 2008.

[45] V. Norris, A. Zemirline, P. Amar, J. N. Audinot, P. Ballet, E. B. Jacob, G. Bernot, G. Beslon, A. Cabin, E. Fanchon, J.-L. Giavitto, N. Glade, P. Greussay, Y. Grondin, J. A. Foster, G. Hutzler, F. Kepes, O. Michel, F. Molina, J. Signorini, P. Stano, and A. R. Thierry, "Computing with bacterial constituents, cells and populations : from bioputing to bactoputing," *Theory in Biosciences*, vol. 130, no. 3, pp. 211–228, 2011.

[46] V. Norris, A. R. Thierry, P. Amar, B. I. Holland, and F. Molina, "The Mimic Chain Reaction," *Journal of Molecular Microbiology and Biotechnology*, vol. 22, pp. 335–343, Dec. 2012.

[47] P. Jensen and K. Hammer, "The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters," *Appl Environ Microbiol*, vol. 64, no. 1, pp. 82–87, 1998.

[48] K. Hammer, I. Mijakovic, and P. Jensen, "Synthetic promoter libraries – tuning of gene expression," *Trends Biotechnol*, vol. 24, no. 2, pp. 53–55, 2006.

[49] N. Seghezzi, P. Amar, B. Koebmann, P. R. Jensen, and M.-J. Virolle, "The construction of a library of synthetic promoters revealed some specific features of strong streptomyces promoters," *Applied Microbiology and Biotechnology*, vol. 90, no. 2, pp. 615–623, 2011.

[50] L. P. Kotra, J. Haddad, and S. Mobashery, "Aminoglycosides : Perspectives on mechanisms of action and resistance and strategies to counter resistance," *Antimicrobial Agents and Chemotherapy*, vol. 44, pp. 3249–3256, Dec 2000.

[51] K. Shaw, P. Rather, R. Hare, and G. Miller, "Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes," *Microbiology and Molecular Biology Reviews*, vol. 57, no. 1, pp. 138–163, 1993.

[52] N. Seghezzi, M.-J. Virolle, and P. Amar, "Novel insights regarding the sigmoidal pattern of resistance to neomycin conferred by the aphii gene, in streptomyces lividans," *AMB Express*, vol. 3, no. 13, 2013.