Modélisation cellulaire pour l'émergence de structures multiprotéiques auto-organisées

Antoine Coulon[*] — Hédi Soula[*] — Olivier Mazet[**] — Olivier Gandrillon[***] — Guillaume Beslon[*]

[*] *INSA-Lyon – Dept. Informatique – Laboratoire PRISMa*
20, avenue Albert Einstein, 69621 Villeurbanne cedex

[**] *INSA-Lyon – Centre de Mathématiques – Institut Camille Jordan*
20, avenue Albert Einstein, 69621 Villeurbanne cedex

[***] *UCBL – Centre de Génétique Moléculaire et Cellulaire*
Bât. Gregor Mendel, 43 bd. du 11 novembre 1918, 69622 Villeurbanne cedex

guillaume.beslon@insa-lyon.fr

RÉSUMÉ. *La biologie des systèmes et la simulation cellulaire sont nées conjointement et entretiennent des relations ambigües entre approches globalisantes et approches émergentistes. Nous présentons ici une démarche émergentiste de la modélisation cellulaire, basée sur une simulation multi-agents. Nous décrivons l'ensemble du processus de création du modèle, depuis le projet scientifique jusqu'aux méthodes d'implémentation, en insistant particulièrement sur le modèle d'interactions entre agents qui est la base de notre simulation. Enfin, des résultats préliminaires montrant l'émergence de structures organisées sont présentés pour illustrer l'intérêt de l'approche proposée.*

ABSTRACT. *Systems biology and cell simulation started together and maintain a controvertible relationship within the scope of global approaches and emergence-driven approaches. Here, we present a multi-agent based model for cell simulation. We describe the whole model creation process, from its scientific project perspective to implementation considerations. The agents interactions model will be particularly detailed. Finally, preliminary results showing self-organised structures will be presented to demonstrate the interest of this approach.*

MOTS-CLÉS : *structures cellulaires, simulation cellulaire, auto-organisation, systèmes multi-agents, dynamique moléculaire.*

KEYWORDS: *cellular structures, cell simulation, self-organisation, multi-agent systems, molecular dynamics.*

2e soumission à *RSTI - TSI*, le 20/03/06

1. Introduction

Depuis la fin des années quatre-vingt dix, deux tendances fortes sont apparues en biologie cellulaire. D'une part un courant théorique relevant d'une approche intégrative (la "biologie des systèmes"), d'autre part une approche plus technique, souvent conduite par de grands consortiums scientifiques, visant à produire des modèles cellulaires *in silico*. Ces simulations cellulaires relevant d'une approche globalisante, ces deux approches sont souvent confondues. Pourtant, il ne suffit pas de simuler une entité biologique dans sa globalité pour pouvoir prétendre relever de la biologie des systèmes. La notion de "système" dépasse celle de totalité et intègre celle d'organisation : le projet scientifique de la biologie des systèmes est d'expliquer comment, à partir des propriétés locales des constituants moléculaires, la cellule forme un tout organisé. De façon évidente, une simulation qui intégrerait *ab initio* l'organisation globale de la cellule ne pourrait contribuer à ce projet.

À partir de ce constat, nous proposons dans cet article un outil de simulation cellulaire, 3DSPI (3D Simulator of Protein Interactions), conforme à ce que nous estimons devoir être un modèle cellulaire en biologie des systèmes. Ce modèle est basé sur le formalisme multi-agents, que nous pensons – rejoignant en cela d'autres auteurs (Amar *et al.*, 2002) – être le plus adapté. Nous dressons ici un tableau aussi complet que possible de la démarche suivie, depuis le projet scientifique jusqu'à l'implémentation. Ainsi, dans une première partie, nous présenterons les motivations de ce projet (section 2) avant de présenter le modèle proprement dit. Notre objectif étant la simulation de structures multi-protéiques et l'étude de leur organisation spatiale et dynamique, le modèle sera basé sur la description des énergies d'interaction entre les protéines (section 3). Enfin, bien que ce projet soit encore en phase de développement, nous présenterons les principaux choix d'implémentation (section 4) et les premiers résultats obtenus (section 5). Ceux-ci montrent que la démarche de modélisation adoptée permet d'observer l'émergence de structures organisées. Alors que le modèle moléculaire utilisé est isotrope, nous voyons en particulier apparaître des structures présentant des symétries ou des orientations préférentielles alors qu'aucune de ces propriétés n'est présente au niveau des agents moléculaires.

2. Motivations pour une simulation cellulaire multi-agents

2.1. *De la biologie des systèmes*

La notion de biologie des systèmes a émergé à la fin des années 90 (Ideker *et al.*, 2001, Kitano, 2002) alors que les outils d'acquisition de données à haut-débit fournissaient des résultats dont l'interprétation était manifestement impossible dans le cadre réductionniste. Le paradigme cybernétique de premier ordre (correspondant aux travaux princeps sur l'opéron Lactose) s'est alors peu à peu retrouvé contesté et remplacé par un paradigme de régulation "en réseau". Dans ce dernier, les mécanismes biologiques sont considérés comme totalement intégrés au sens où il n'est pas possible, dans le cas général, d'isoler les éléments de leur contexte. Ce changement

de paradigme accompagne la découverte (ou la redécouverte) d'un certain nombre de propriétés des *systèmes* biologiques. Ces propriétés étaient largement ignorées dans les approches réductionnistes, non par négligence mais plutôt parce qu'elles ne s'expriment qu'à des niveaux globaux qui ne sont pas accessibles par ces approches.

Les systèmes biologiques sont des systèmes complexes, au sens où ils sont composés d'un grand nombre d'éléments en interaction et où ces interactions sont non-linéaires. De ce fait, le fonctionnement global de ces systèmes ne peut être prédit par une décomposition raisonnée en sous-parties. C'est cette caractéristique qui justifie à elle seule la démarche de la biologie des systèmes.

Ce sont néanmoins des systèmes modulaires. La complexité de ces systèmes n'interdit pas une certaine organisation. Ainsi, les réseaux biologiques sont organisés en modules au sein desquels la connectivité est sensiblement plus élevée que dans le reste du réseau (Vespignani, 2003, Barabasi *et al.*, 2004). De plus, l'analyse des nœuds appartenant à un même module révèle souvent des fonctions biologiques communes (Stuart *et al.*, 2003).

Ils sont basés sur des phénomènes stochastiques. La plupart des mécanismes biologiques présentent, aux échelles microscopiques, un caractère fortement probabiliste. C'est par exemple le cas de la transcription (Levsky *et al.*, 2003, Blake *et al.*, 2003). La disparition de ce caractère probabiliste aux échelles macroscopiques est une question ouverte pour la biologie contemporaine (Kupiec, 1997).

Ce sont des systèmes dégénérés. Des structures différentes peuvent remplir une même fonction et, inversement, une même structure peut contribuer à des fonctions différentes (Tononi *et al.*, 1999, Atamas, 1996). Il est fort probable que l'apparente spécificité des voies métaboliques et/ou de signalisation soit plus révélatrice d'une démarche méthodologique que d'une réalité biologique.

À partir de ce(s) constat(s), la biologie des systèmes relève plus du projet que du programme scientifique : il s'agit d'étudier les objets biologiques – la cellule en particulier – en assumant ces spécificités. Or, nous l'avons dit, celles-ci n'ont de sens que si on considère un niveau macroscopique. C'est pourquoi on associe souvent la "biologie des systèmes" à des approches qualifiées de "holistes" ou "intégratives". Cependant, celles-ci recouvrent deux réalités distinctes. D'une part une approche *orientée données*, d'autre part une approche *orientée processus*. La première part du présupposé que la juxtaposition d'un grand nombre de données permettra d'accéder à une connaissance biologique objective. Il s'agit d'étudier les objets biologiques *dans leur contexte* mais celui-ci est caractérisé par des données supplémentaires. La seconde est basée sur les concepts de la "science des systèmes" et recherche les conditions d'émergence de structures organisées à partir des propriétés des éléments (par exemple des molécules). Essentiellement européenne par ses idées (elle emprunte beaucoup de ses concepts à la systémique ou au structuralisme), cette deuxième approche s'est enrichie d'un courant plus pragmatique provenant d'outre-atlantique. C'est elle que nous considererons ici comme relevant de la biologie des systèmes.

Si l'émergence d'une biologie des systèmes est aujourd'hui une évidence dans les textes, elle l'est moins dans les faits. Force est de constater qu'on attend d'elle beau-

coup plus que ce qu'elle s'est révélée jusqu'ici capable de produire. Cette absence de résultats concrets est largement due à son statut épistémologique, souvent mal assumé y compris par ceux-là mêmes qui s'en réclament. La biologie des systèmes est trop souvent mise en demeure de fournir des résultats expoitables dans un cadre épistémologique qui n'est pas le sien, celui du réductionnisme cartésien. Le peut-elle seulement ? Probablement pas mais il reste alors à exprimer *ce que serait* un résultat dans son cadre propre et comment un tel résultat pourrait communiquer avec le paradigme cartésien qui reste le cadre de référence de la pensée scientifique.

2.2. *La simulation : un outil fondamental en biologie des systèmes*

Ce n'est que depuis une trentaine d'année que, grâce à la diffusion massive des outils informatiques, la simulation a fait son apparition en biologie. Dans un premier temps, elle a été utilisée essentiellement pour étudier des modèles mathématiques dont l'analyse se révélait trop complexe. Parallèlement s'est développée, principalement aux États-Unis, une "sciences de la complexité" en grande partie basée sur des simulations informatiques. Au croisement des deux, la biologie des systèmes intègre totalement la simulation parmi ses outils fondamentaux. En outre, la simulation acquière ici une valeur explicative par elle-même : dès lors que l'objet de la biologie des systèmes est l'étude de l'organisme *sachant* ses constituants (et non l'étude de ceux-ci), la simulation permet de montrer comment des propriétés locales peuvent conduire à l'émergence de propriétés globales (Schweitzer, 2003). Cette émergence n'est cependant pas nécessairement une relation causale[1] et la simulation n'a pas un caractère de preuve puisque par définition elle montre ce qu'elle a été construite pour montrer.

Le statut épistémologique de la simulation en biologie des systèmes est donc très différent de ce qu'il peut être, par exemple, en mécanique. Dans ce dernier cas, la simulation est utilisée pour montrer/rechercher la cohérence d'un agencement de parties sachant le tout. Il s'agit donc d'une démarche de conception. En biologie des systèmes, en revanche, la simulation est utilisée pour explorer les propriétés du tout sachant les parties. Contrairement à une démarche de conception, les lois codées dans le simulateur ne régissent pas le comportement du tout, du moins pas son comportement *intéressant* aux yeux de l'observateur[2]. L'émergence n'est ici plus vue comme une propriété intrinsèque du système mais comme une propriété du couple système-observateur (Ronald *et al.*, 2001).

1. Ainsi, dans le jeu de la vie, il n'y a pas de relation causale entre les règles des automates et l'émergence de propriétés globales telles que la propriété de déplacement des "planneurs".
2. Il existe évidemment, en biologie, des simulations relevant d'une démarche de conception, par exemple lorsqu'une voie métabolique est simulée sous la forme d'un ensemble d'équations différentielles. Dans ce cas, le comportement du tout est codé (par exemple : tel dépassement de concentration provoque la mort cellulaire) et la simulation est utilisée pour rechercher, dans les parties (les variables des équations différentielles), les conditions provoquant ce comportement. L'intérêt d'une telle simulation est évident, par exemple en médecine, mais elle ne relève pas, selon nous, de la biologie des systèmes.

2.3. *Approches de la simulation cellulaire*

2.3.1. *Pourquoi simuler à l'échelle cellulaire*

Pour certains organismes simples, tels la bactérie *Escherichia coli* ou l'eucaryote unicellulaire *Saccharomyces cerevisiae*, l'accumulation d'informations rend aujourd'hui envisageable une simulation complète de la cellule. D'une certaine façon, une telle simulation constitue le seul moyen d'établir une réelle intégration de données hétérogènes rassemblées par des équipes différentes, dans des conditions différentes et par des méthodes d'observation différentes. L'objectif d'une telle simulation n'est cependant pas là. Puisqu'il est désormais clair que les données rassemblées n'éclaireront pas à elles seules le fonctionnement des organismes, la simulation apparaît comme le seul recours pour comprendre comment des systèmes aussi complexes et hétérogènes que des cellules vivantes sont organisés (Bork *et al.,* 2005).

Ainsi, l'observation macroscopique d'une cellule montre un tout très organisé, dynamique, composé de structures relativement distinctes. Ces structures peuvent être relativement simples, tels les agrégats situés dans le noyau des euraryotes, ou beaucoup plus complexes (par exemple les filaments d'actine ou les microtubules). Or on ne dispose pas d'outils permettant d'expliquer simplement comment passer des propriétés des composants moléculaires aux propriétés topologiques ou fonctionnelles de ces structures. Les plus grosses d'entre elles sont observables en microscopie optique (par exemple le nucléole), les plus petites en microscopie électronique ou confocale. Cependant, aucune de ces techniques ne permet de détecter individuellement les molécules qui les composent et seules des techniques indirectes basées sur la fluorescence permettent d'appréhender leur dynamique (McNally *et al.,* 2000). Inversement, si on peut accéder aux propriétés individuelles des protéines, nous ne pouvons pas les analyser dans le contexte de la formation d'une structure cellulaire.

C'est donc à l'interface entre ces deux niveaux, moléculaire d'un coté et structures cellulaires de l'autre, que la simulation peut apporter un gain substantiel (Amar *et al.,* 2002). On notera cependant que le transfert de propriétés d'un niveau à l'autre est un processus dans lequel l'organisation spatiale revêt une importance prépondérante, aussi bien au niveau moléculaire (taille et conformation des molécules) qu'au niveau cellulaire (localisation et conformation des structures). En outre, ce processus est aussi fortement inscrit dans une dimension temporelle : la plupart des structures cellulaires présentent des conformations différentes suivant le cycle cellulaire. C'est pourquoi une simulation cellulaire devra nécessairement intégrer les dimensions spatiales et temporelles (Bork *et al.,* 2005, Lemerle *et al.,* 2005).

2.3.2. *Cellules virtuelles et systèmes cellulaires*

Parallèlement à l'émergence de la biologie des systèmes, de nombreux projets de simulation cellulaire ont vu le jour (Lemerle *et al.,* 2005). Cependant ces projets sont souvent basés sur une interprétation trop rapide de la notion même de systèmes : on n'en retient que le principe globalisant, oubliant par là que c'est le projet et non l'objet qui fait d'un modèle biologique un modèle de biologie des systèmes. Or, simuler une

entité cellulaire globale est nécessaire mais certainement pas suffisant. La notion de système *complet* n'a de sens que relativement à un projet scientifique et, en l'absence d'un tel projet, simuler une cellule complète produirait certes un bel objet technologique mais finalement plus illustratif qu'explicatif.

Ainsi, la plupart des outils existants pour simuler des systèmes biologiques sont développés sous un présupposé continu : ils modélisent le comportement moyen de populations *supposées grandes* de molécules semblables. À partir de tels présupposés, plusieurs modèles ont été proposés (Takahashi *et al.,* 2004, Loew *et al.,* 2001, Broderick *et al.,* 2004), avec l'objectif ambitieux de simuler une cellule complète. Or, malgré leur volonté "globalisante", ces modèles ne peuvent pas intégrer les caractéristiques des systèmes biologiques tels que nous les avons définis. Ils se révèlent en particulier inadéquats lorsque le nombre de molécules est trop faible ou leur variété trop importante. Par ailleurs, ils ne sont pas du tout adaptés à l'étude de phénomènes spatio-dynamiques puisqu'ils supposent une décomposition *a priori* en compartiments. Or cette approche ne permet pas d'expliquer l'extrême mobilité intra- et inter-compartiments des molécules biologiques alors même que les données qui montrent cette mobilité s'accumulent (Misteli, 2001, Phair *et al.,* 2004, Misteli, 2005). Dans un tel contexte, il apparaît pertinent de proposer des modèles permettant de rendre compte des propriétés spatio-dynamiques observées (Lemerle *et al.,* 2005).

Pour répondre à ce besoin d'outils de modélisation alternatifs, plusieurs auteurs ont proposé des approches basées sur des paradigmes souvent très variés, allant de la théorie des graphes (Ballet *et al.,* 2004) aux automates cellulaires (LeSceller *et al.,* 2000, Wishart *et al.,* 2004). Cependant, on oublie trop souvent que la simulation est l'entreprise la plus réductionniste qui soit : tous les paramètres doivent être fixés et toutes les interactions doivent être explicitées. L'*absence d'information* sur une interaction – situation courante en biologie – correspond, lorsqu'elle est intégrée à une simulation, à une *absence d'interaction*. Comment, dès lors, ne pas "tout" simuler ? La solution est alors de définir des entités locales – des agents – dont le comportement crée indirectement les interactions. On définit pour cela un cadre général, une "chimie artificielle" (Dittrich *et al.,* 2001), qui permet de calculer les interactions connaissant les propriétés des éléments. Toutes les interactions sont alors indirectement déterminées et on en observe les conséquences macroscopiques. Une telle approche, relevant de la modélisation individu-centrée ou des Systèmes Multi-Agents (Weiss, 1999), a déjà été largement utilisée pour expliquer nombre de phénomènes d'organisation dynamique tels que les essaims d'insectes ou les bancs de poissons (Theraulaz *et al.,* 1997, Schweitzer, 2003)). Elle est encore balbutiante au niveau cellulaire, probablement parce que la démarche de modélisation y est moins intuitive.

2.4. *Problématique de l'agentification*

Le préalable à toute mise en œuvre d'un modèle individu-centré est le choix du niveau de description des agents et de leurs propriétés. Ce choix est crucial dès lors qu'il s'agit de mettre en évidence des relations entre le niveau local et le niveau global.

Celles-ci ne doivent évidemment pas être imposées par les agents mais bien émerger de leurs interactions. Là encore c'est le projet scientifique qui permet de choisir l'agentification. En fonction de ce projet, certaines propriétés globales seront codées dans le simulateur tandis que d'autres seront laissées libres. Or, si des formalismes ont été proposées pour la mise en œuvre d'une approche multi-agents[3], ceux-ci reposent sur une part d'anthropomorphisme et se révèlent donc inopérants dès lors que le système considéré n'est plus soumis aux lois physiques auxquelles nous sommes accoutumés. Dans ce cas, il n'est plus possible de raisonner à l'aide des schémas de pensée phénoménologiques classiques et on doit revenir aux lois physiques pour déterminer les propriétés des agents.

Les modèles cellulaires sont trop souvent dérivés de modèles construits pour des échelles macroscopiques (modèles pluricellulaires ou sociaux) alors que le passage à l'échelle moléculaire entraîne nombre de changements de lois. Or, si l'objectif du système multi-agents est l'étude des mécanismes d'émergence de structures dans un système cellulaire, il est évident que ces structures dépendront des propriétés locales des agents. Celles-ci doivent donc être soigneusement déterminées, en évitant en particulier tout recours à des schemas basés sur des modèles perceptifs ou sur des modèles d'action. La difficulté de produire un résultat concret par une démarche de simulation nous impose, dans le cas d'une simulation à l'échelle moléculaire, de définir très précisément le comportement des agents. Il est en effet nécessaire de s'assurer que les structures produites sont bien des propriétés émergentes et non des artéfacts dûs à la réutilisation, au niveau moléculaire, de comportements initialement définis à une autre échelle ou pour des lois physiques différentes (voir section 3.3.2).

3. 3DSPI, un simulateur cellulaire multi-agents

On le voit, la simulation de systèmes cellulaires se doit, pour produire des résultats exploitables, d'être basée sur une agentification raisonnée du niveau moléculaire. Il nous semble en effet clair que c'est d'une telle démarche, plus que d'une simple logique de simulation d'un tout (la cellule), que viendront les véritables résultats en biologie des systèmes. Dans ce cadre, nous avons entrepris une démarche de modélisation intégrant, au sein d'un même groupe de travail, biologistes, mathématiciens et informaticiens, tous provenant d'un même campus et disposés à travailler régulièrement ensemble pour échanger leurs connaissances. Cette approche nous a permis de développer un simulateur 3D d'interactions protéines-protéines (3DSPI) dont les premiers résultats nous montrent que la démarche de modélisation entreprise conduit bien à l'émergence de propriétés spatio-dynamiques globales.

3. Par exemple les approches Voyelles (Demazeau, 1995) ou "Agents-Groupes-Rôles" (Ferber *et al.*, 1998).

3.1. *Les agents moléculaires dans 3DSPI*

La simulation de structures multi-protéiques résulte nécessairement d'un compromis entre le réalisme physique des agents moléculaires et la capacité à augmenter le nombre d'agents en interaction. Il n'est donc pas envisageable de simuler les propriétés physico-chimiques des protéines impliquées dans les structures. Cependant, il est évident que l'organisation spatiale des structures sera dépendante de la modélisation des interactions protéines-protéines et des capacités de déplacement de celles-ci au sein d'un compartiment cellulaire. C'est pourquoi nous avons choisi une agentification intermédiaire entre le réalisme physique et l'abstraction que représente, par exemple une modélisation par automates cellulaires. Dans 3DSPI, les protéines se déplacent dans un espace tridimensionnel en coordonnées réelles (le calcul des interactions mais aussi des mouvements est alors réaliste, aux approximations numériques près). La dynamique des objets pourra donc être étudiée, pour peu que la résolution temporelle des simulations permette une précision suffisante (voir section 4.1.1). En revanche, les protéines elles-mêmes sont modélisées sous la forme d'un domaine spatial d'interaction correspondant au rayon d'action de la force de Born (section 3.3.1). Dans ce modèle, les protéines sont donc considérées comme isotropes.

Dans 3DSPI, les protéines sont soumises à deux types d'interactions : les interactions protéines-protéines et les interactions avec les autres molécules du milieu intracellulaires. Ces dernières seront modélisées classiquement sous la forme d'un mouvement brownien et d'une force de viscosité. Les interactions protéines-proteines, en revanche, seront modélisées individuellement. Dans un premier temps, nous avons proposé et testé un modèle probabiliste ce qui nous a permis d'étudier les propriétés dynamiques des structures (section suivante). Cependant, un tel modèle ne permet pas l'étude de la conformation spatiale des structures émergentes. En outre, la relation entre la loi d'interaction et la loi de déplacement est difficile à établir. C'est pourquoi nous avons implémenté une deuxième version de 3DSPI dans laquelle les interactions protéines-protéines sont modélisées sous la forme de potentiels d'énergie (section 3.3.1). Cela nous permet de proposer un formalisme unifié, intégrant le déplacement des protéines et leurs interactions. Nous avons ainsi la possibilité d'explorer les relations entre ces différentes composantes et d'étudier, par exemple, l'influence de la température du milieu sur la conformation 3D des structures multi-protéiques.

Tel quel, le modèle d'agent utilisé dans 3DSPI est limité par le caractère isotrope des protéines et une grande partie des structures cellulaires d'intéret nous sont inaccessibles (microtubules, filaments, ...). Cependant, la modélisation des interactions sous la forme de potentiels d'énergie nous permet de dépasser cette limite. Il est en effet possible de modéliser les protéines sous la forme d'un *groupe* de domaines d'interactions[4]. Chacun de ces domaines reste alors isotrope mais les possibilités de combinaisons permettent de créer des structures moléculaires anisotropes (structures orientées, site actif, ...). Cette approche est cependant encore peu avancée et nous n'en sommes

4. Les interactions sont alors calculées en chaque point, c'est-à-dire pour chaque domaine, mais le mouvement résultant est calculé pour l'ensemble du groupe.

qu'au stade de l'expérimentation. Bien qu'elle décuple les possibilités de modélisation de 3DSPI, elle ne sera donc qu'évoquée ici.

3.2. *Modèle probabiliste (3DSPI-V1)*

Dans le modèle probabiliste de 3DSPI, les protéines sont soumises à un mouvement brownien mais les interactions protéiques sont simplifiées de façon à permettre une simulation sur des grandes échelles de temps. Conformément aux principes méthodologiques mis en avant, nous avons choisi ce type de modèle relativement à un objectif scientifique précis : il s'agit ici d'étudier la formation d'agrégats multiprotéiques (similaires aux corps nucléaires) afin d'expliquer leur caractère dynamique tel qu'il est mis en évidence en biologie cellulaire. Nous nous sommes donc intéressés non pas aux conformations 3D des structures mais à leur évolution temporelle. C'est pourquoi nous utilisons le modèle des sphères dures pour représenter les forces de Born tandis que les interactions entre deux protéines sont modélisées sous la forme d'un coefficient d'agrégation (Coefficient Of Stickiness, COS) caractérisant la probabilité que deux protéines ont, lorsqu'elles entrent en contact, de former un agrégat.

Ce premier modèle nous a permis de mettre en évidence des phénomènes de transition de phase liés à la valeur de COS (en deçà de la transition aucun agrégat ne se forme, sauf de façon transitoire, tandis qu'au delà de la transition des agrégats de plus en plus gros apparaissent (Soula *et al.*, 2005)). En outre, nous avons pu mettre en place une expérience de FLIP[5] virtuel ce qui nous a conduit à vérifier, par des méthodes analogues à celles employées en biologie cellulaire, le caractère dynamique des agrégats obtenus (Soula *et al.*, 2005).

Malgré le caractère simpliste de l'agentification, la version probabiliste de 3DSPI a montré l'intérêt de la démarche de modélisation adoptée. Dans le but d'étudier la structure tridimensionnelle des agrégats obtenus, nous avons alors modifié le niveau de description des agents moléculaires pour passer à un modèle énergétique.

3.3. *Modèle énergétique (3DSPI-V2)*

3.3.1. *Le champ d'énergie : une modélisation fi ne des interactions*

À l'échelle moléculaire, les forces d'interaction sont très nombreuses. Elles ont des caractéristiques très différentes en termes d'intensité ou de distance d'action. Elles entrent en jeu dans tous les mécanismes cellulaires, de la conformation spatiale des acides aminés d'une protéine au déplacement de complexes protéiques entre les différents composants de la cellule et au travers des membranes (Lagües *et al.*, 2003).

5. Fluorescence Loss In Photobleaching. Le FLIP permet d'estimer expérimentalement la mobilité intercompartimentale des protéines nucléaires (McNally *et al.*, 2000).

Aux échelles spatiales et temporelles qui nous intéressent, on peut négliger ou approximer une partie de ces forces. En effet, l'entité de base du modèle (l'agent) étant constitué de plusieurs dizaines de milliers d'atomes, il ne convient pas de simuler les petites molécules comme l'eau ou les ions. Il suffit de modéliser leur action moyenne. Ainsi, des forces complexes telles que la liaison hydrophobe peuvent être modélisées simplement : lorsqu'une molécule d'eau (très polaire) se trouve entre deux chaînes lipidiques (très peu polaires), elle est attirée par les molécules d'eau avoisinantes. Ceci tend à regrouper les chaînes lipidiques et équivaut, à l'échelle des domaines protéiques, à une force d'attraction uniforme.

À l'exception de la force à l'origine du mouvement brownien (discutée dans la section 3.3.3), toutes les forces d'interaction entre les domaines protéiques dérivent d'un potentiel scalaire : *l'énergie potentielle*. On peut alors définir nos agents comme des entités ponctuelles engendrant un champ de force isotrope et soumises aux forces engendrées par les autres agents. Du point de vue de la simulation, cette méthode permet de se passer du modèle des sphères dures et élimine ainsi les calculs de contact très couteux en temps. Elle modélise plus fidèlement le comportement des protéines car les forces ne sont plus de classe C^0 (affines par morceau) mais C^∞. La force équivalente à la force de contact du modèle des sphères dures (force de Born) est implicitement rendue par l'allure de l'énergie potentielle pour des distances proches de zéro et sera gérée au même titre que les autres forces.

Les interactions pertinantes pour notre étude se regroupent en trois catégories : les interactions coulombiennes, les forces de Van der Waals et l'énergie de Born. Leur composition donne une force résultante relativement complexe (figure 1) car elles ont des atténuations différentes en fonction de la distance :

Les interactions coulombiennes s'exercent entre charges permanentes. Les protéines en solution s'entourent d'un nuage ionique différent selon le caractère plus ou moins polaire de leur surface. La diminution de la concentration ionique en fonction de la distance à la protéine est généralement décrite comme étant exponentielle : on obtient une variation de l'énergie en $(d + d_c)^{-1}$ avec d la distance par rapport au centre du domaine protéique et d_c une constante dépendant de la répartion du nuage ionique.

Les forces de Van der Waals regroupent toutes les forces qui mettent en jeu des dipôles (hétérogénéité momentanée de répartition des électrons d'une molécule). Elles agissent à beaucoup plus courte distance que les interactions coulombiennes. On peut distinguer la force de Keesom (entre dipoles permanents), la force de Debye (entre dipoles induits) et la force de London (entre dipoles instantanés). Toutes ces énergies sont globalement isotropes et attractives sur des molécules de taille suffisante telles que les protéines. Leur énergie varie en $-d^{-6}$.

L'énergie de Born est celle décrite par le principe de Pauli qui interdit aux cortèges électroniques de deux atomes de s'interpénétrer. Cette force est en d^{-12}. C'est elle qui est à l'origine des forces de contact, même aux échelles macroscopiques. Elle est, à très courte distance, prédominante et extrêmement répulsive.

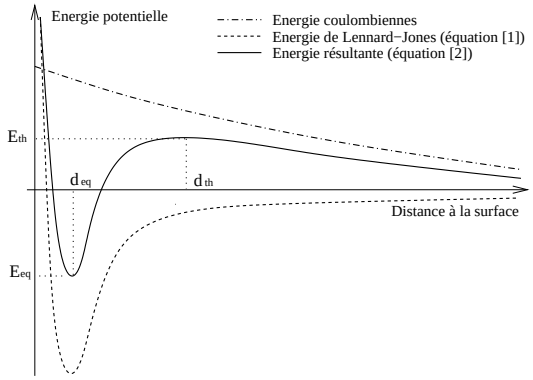


Figure 1. *Energies potentielles d'interaction entre deux domaines protéiques. La force résultante est donnée par la dérivée de l'énergie potentielle.*

Lorsque cette distance (sensiblement le rayon atomique) est négligeable devant la dimension des objets considérés, on peut utiliser le modèle des sphères dures qui considère que cette force est nulle si les objets ne se touchent pas et infinie si les objets tentent de s'interpénétrer. Ce modèle fait appel à la théorie dite *des chocs* qui convient très bien à des situations macroscopiques, mais perd tout son sens dès lors qu'on descend à des échelles microscopiques.

Un modèle connu sous le nom de système de particules de Lennard-Jones utilise un champ d'énergie prenant en compte les forces de Van der Waals et l'énergie de Born. L'energie potentielle est donnée par :

$$E_p = \frac{A}{d^{12}} - \frac{B}{d^6} \qquad [1]$$

Dans ce modèle, les énergies d'interaction possèdent un minimum E_{eq}, ce qui permet de définir une distance d'équilibre d_{eq} au-delà de laquelle il y a attraction et en deçà de laquelle il y a répulsion (figure 1). Lorsqu'on ajoute les interactions coulombiennes, cette distance d'équilibre change peu, mais un troisième changement de pente apparaît : il y a répulsion au-delà d'une distance seuil d_{th} et il faut fournir une énergie E_{th} pour rentrer dans le bassin d'attraction et former un complexe protéique.

$$E_p = \frac{A}{d^{12}} - \frac{B}{d^6} + \frac{E_0 d_c}{d + d_c} \qquad [2]$$

Si on note u le vecteur unitaire, au point considéré, dirigé vers le centre du domaine protéique, le champs de forces est obtenu en calculant le gradient de l'énergie. Dans notre cas, l'energie potentielle étant radiale, cela revient à dire que $F \cdot u$ est la dérivée de E_p. En tout point de l'espace la force est alors le vecteur colinéaire à u tel que :

$$F \cdot u = \frac{\partial E_p}{\partial d} = -\frac{12A}{d^{13}} + \frac{6B}{d^7} - \frac{E_0 d_c}{(d + d_c)^2} \qquad [3]$$

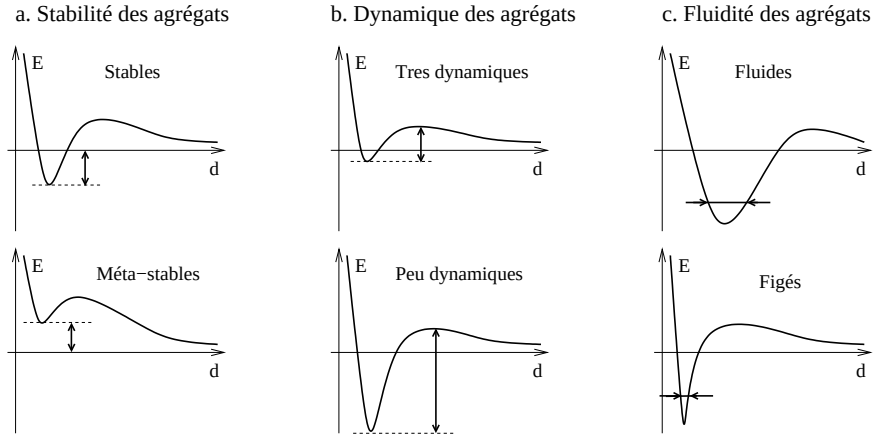


Figure 2. *Interprétation des coordonnées du point d'équilibre et du point seuil.*

A, B, E_0 et d_c sont des constantes qui dépendent des deux espèces protéiques mises en jeu. Ainsi, dans notre modèle, la définition d'une interaction possède quatre degrés de liberté. On montre facilement que la connaissance des coordonnées du point d'équilibre et du point seuil (soit d_{eq}, E_{eq}, d_{th} et E_{th}, voir figure 1) est suffisante pour définir complètement une interaction.

Il est possible d'interpréter *a priori* les coordonnées du point d'équilibre et du point seuil pour tenter de prédire le comportement des différentes espèces de domaines protéiques. En effet, le signe de E_{eq} décrit la stabilité d'un composé : $E_{eq} > 0$ correspond à un composé méta-stable (bien qu'un équilibre y soit possible, l'état agrégé est moins stable que l'état désagrégé) et $E_{eq} < 0$ à un composé stable (figure 2.a). En revanche, E_{th} représente l'énergie qu'il faut fournir pour amener deux domaines à former une liaison. Ainsi, plus ce paramètre est élevé, plus la température nécessaire à la formation de composés sera élevée. On remarque alors que l'énergie à fournir pour rompre une liaison est $E_{th} - E_{eq}$. Elle permet de prévoir à quel degré un agrégat sera dynamique : plus la valeur est faible, plus les domaines pourront facilement sortir de l'agrégat (figures 2.b). Enfin, plus d_{eq} est faible et E_{eq} important, plus le puit d'énergie au point d'équilibre sera fin et réduira la liberté de mouvement des deux domaines l'un par rapport à l'autre, modifiant la fluidité des agrégats (figures 2.c).

3.3.2. *Mécanique à l'échelle microscopique*

La loi fondamentale de la dynamique décrit l'accélération d'un corps comme étant proportionnelle à la force appliquée. Elle s'énonce $\sum_i f_i = m\ddot{x}$, avec m la masse du corps considéré, x sa position et les f_i les forces appliquées. On se permet de négliger les forces de très faible intensité devant les autres, telle la viscosité qui s'exprime $-\lambda\dot{x}$, où λ est le coefficient de viscosité qui dépend de la nature du milieu et de la forme de l'objet. Il est faible pour l'air, ce qui permet souvent de négliger ce terme sauf

si la vitesse est importante ($\dot{x}$ important) ou si la forme est particulièrement sujette au frottement de l'air (λ important). Il n'est plus négligeable dans un milieu liquide, auquel cas la loi fondamentale de la dynamique s'écrit :

$$-\lambda\dot{x} + \sum_i f_i = m\ddot{x} \quad [4]$$

Si on note r la dimension d'un objet, alors sa masse varie en r^3. On peut donc écrire $m \approx \rho_m r^3$, avec ρ_m la masse volumique de l'objet. Le coefficient de viscosité varie en r. Par exemple, pour une sphère il est donné par $\lambda = 6\pi\eta r$, η étant la viscosité dynamique du milieu. On peut donc écrire de façon plus générale $\lambda = \lambda_r r$, avec λ_r constant. Ainsi, l'équation [4] peut se réécrire :

$$\sum_i f_i = (\rho_m\ddot{x})r^3 + (\lambda_r\dot{x})r \quad [5]$$

Dans le cas habituel r est assez grand pour avoir $r = o(r^3)$, ce qui nous amène à éliminer le terme de la viscosité. À l'inverse, dans notre cas r est suffisamment petit pour avoir $r^3 = o(r)$ et négliger l'inertie devant la viscosité. La loi de la dynamique devient alors :

$$\sum_i f_i = \lambda\dot{x} \quad [6]$$

Ce changement n'est pas sans conséquence, il transpose les lois de Newton en remplaçant $\ddot{x}$ par $\dot{x}$. En effet, c'est la vitesse et non plus l'accélération qui est proportionnelle à la force. Ainsi, un corps isolé ou pseudo-isolé (résultante des forces nulle) n'a plus une vitesse constante (principe d'inertie), mais une position constante. La composante traduisant l'inertie a laissé place à celle traduisant l'amortissement, ce qui a tendance à interdire les mouvements d'oscillation. Ceci a deux conséquences majeures : d'une part la stabilisation des structures, et d'autre part l'interdiction de mouvements périodiques ce qui implique une dynamique complexe du système. Le facteur fondamental dans ce rapport entre ordre et désordre est la température car elle détermine l'intensité du mouvement brownien qui est à l'origine des déplacements imprédictibles des particules. D'autres paramètres ont une influence sur la quantification de ces deux phénomènes. On peut citer : les coordonnées des points d'équilibre et des points seuils des interactions, la concentration des protéines et les concentrations relatives des espèces de protéines. C'est en cherchant des valeurs critiques de ces paramètres, pour lesquelles se fait un équilibre entre ordre et désordre que l'on peut s'attendre à découvrir de l'émergence. Ceci n'est pas sans rappeler les travaux de C. Langton (Langton, 1986, Langton, 1991) sur les automates cellulaires. En effet, avec cet outil simple à mettre en œuvre, il a montré que les conditions propices à l'émergence correspondent à des paramètres impliquant des dynamiques à la frontière entre ordre et désordre. D'un côté de cette frontière le système ira vers des attracteurs stables et périodiques, et de l'autre côté vers des attracteurs chaotiques. Dans le cas du modèle énergétique, c'est le mouvement brownien qui jouera le rôle de paramètre de contrôle. Il convient donc, si l'on cherche de l'auto-organisation sur des modèles physiquement plausibles, de porter une grande attention à sa modélisation.

3.3.3. *Modélisation du mouvement brownien*

Le mouvement brownien, observé pour la première fois par Robert Brown en 1827, est souvent décrit comme le mouvement de particules en suspension dans une solution, d'apparence aléatoire et dont l'intensité dépend de la température. Il est dû au choc des molécules de la solution sur les particules observées dont la résultante est généralement non nulle. Il convient donc de prendre en compte cet aspect du mouvement brownien si on souhaite avoir un modèle physiquement realiste (Berg, 1993).

Les modèles mathématiques permettant de modéliser le mouvement brownien sont des processus de Markov à valeur dans $\mathbb{R}^3$. L'un des plus simples est la *marche aléatoire* de pas constant. La règle décrivant la succession de ses états est : $x_{i+1} = x_i + u_i$ avec x_i la position de la particule à la $i^{\text{ème}}$ itération et u_i un vecteur de direction aléatoire indépendante de i et de norme fixe. En d'autres termes la marche aléatoire est le déplacement discret d'un point par des pas de longueur constante. On définie $\mathcal{X}(t)$ la variable alétoire représentant la distance parcourue en un temps t :

$$\mathcal{X}(t) = \| \sum_{i=0}^{t/\tau - 1} u_i \| \qquad [7]$$

avec τ la durée du pas de temps. Si $n = t/\tau$ et n est grand, alors il est montré que l'espérance est :$\langle \mathcal{X}(t) \rangle = u\sqrt{n}$. On obtient donc bien une marche aléatoire avec une trajectoire résultante fractale. De plus, la distribution de la distance parcourue suite à n pas d'une marche aléatoire de longueur constante (avec n grand), est identique à celle définie par les équations de diffusion dont le mouvement brownien est à l'origine. En effet la diffusion est décrite par la densité de probabilité qu'une particule ait parcouru une distance r point à point, durant un temps t :

$$P_D^v(r,t) = \left(\frac{1}{4\pi D t}\right)^{d/2} e^{\frac{-r^2}{4Dt}} \qquad [8]$$

d représente la dimension de l'espace (ici $d = 3$) et D est un coefficient de diffusion qui fait intervenir la température T, le coefficient de viscosité λ (section 3.3.2), la constante des gaz parfaits R et le nombre d'Avogadro $\mathcal{N}_{av}$:

$$D = \frac{RT}{\mathcal{N}_{av}\lambda} \qquad [9]$$

Ainsi la marche aléatoire de pas constant est un outil mathématique simple qui rend compte de beaucoup de propriétés du mouvement brownien. Cependant pour avoir une répartition réaliste de la distance parcourue durant un intervalle de temps t_1, il faut simuler n pas de durée $t_2 = t_1/n$ avec n grand, ce qui implique $t_2 \ll t_1$. Or, dans notre cas, le modèle est destiné à étudier la formation de structures et il est indispensable d'avoir ce réalisme même à l'échelle du pas de temps. En effet, il est nécessaire que l'intensité du mouvement brownien ne soit pas constante pour éviter les effets stéréotypés. Ainsi, une intensité variable permet qu'il y ait en permanence formation et destruction de structures multiprotéiques et qu'un équilibre puisse s'établir.

Il faut donc adapter la marche aléatoire pour en faire un vrai mouvement brownien de pas variable en intégrant de façon précise la distribution de la longueur des pas (c'est à dire de la norme de u_i). La densité de probabilité radiale $P_D^r(r,t)$ s'exprime à partir de la densité de probabilité volumique $P_D^v(r,t)$ par la relation suivante[6] :

$$P_D^r(r,t) = 4\pi P_D^v(r,t) r^2 \qquad [10]$$

On peut alors obtenir la fonction de répartition de $\mathcal{X}(t)$:

$$P(\mathcal{X}(t) \leq r) = \int_0^r P_D^r(r_s,t) dr_s \qquad [11]$$

Elle correspond à une *distribution de Maxwell*, solution de l'équation de Boltzmann dans le cas d'un système homogène, [7] décrivant la répartition des vitesses des molécules d'un fluide en fonction de la température. Cela confirme la pertinence de cette distribution pour le calcul du déplacement sur un temps court de particules animées d'un mouvement brownien.

Si on s'intéresse maintenant à un modèle où l'on utilise des pas de temps de durée variable (voir section 4.1.1), la nature fractale de la trajectoire d'une particule soumise à un mouvement brownien a une grande importance. En effet, le mouvement brownien ne peux pas être considéré comme une simple force car son intensité est fonction de la durée du pas de temps τ du fait de sa non linéarité en t (équation [8]). L'expression[8] de $P(\mathcal{X}(t) \leq r)$ n'est en fait qu'une fonction d'un unique paramètre $r/\sqrt{Dt}$. En utilisant cette fonction comme loi de distribution, on peut déterminer r quel que soit τ et le coefficient de diffusion D. Si on défini la variable aléatoire $\mathcal{V}$ comme étant un vecteur de direction aléatoire uniformément répartie et dont la norme suit la loi de Maxwell, alors on a :

$$\|\mathcal{V}\| = \frac{r}{\sqrt{D\tau}} \qquad [12]$$

On en déduit, en utilisant [6] et en explicitant D, l'équation discrète du mouvement des particules, avec un pas de temps τ potentiellement variable :

$$x_{t+\tau} - x_t = \frac{\sum_i f_i}{\lambda} \tau + \mathcal{V} \sqrt{\frac{RT}{\mathcal{N}_{av}\lambda} \tau} \qquad [13]$$

6. Pour ce faire, rappelons que le volume dV compris entre la sphère de rayon $r+dr$ et la sphère de rayon r s'exprime $dV = \frac{4\pi}{3}((r+dr)^3 - r^3) \approx 4\pi r^2 dr$, et que $P_D^v(r,t)dV = P_D^r(r,t)dr$ exprime la probabilité de présence de la particule dans dV.

7. Elle considère les chocs entre particules tels que (i) les colisions à trois corps et plus soient minoritaires et que (ii) les régions de colision soient négligeables devant le parcours des particules ; approximation correcte pour les gaz, mais également pour certain systèmes denses.

8. L'expression explicitée est $P(\mathcal{X}(t) \leq r) = \mathrm{erf}\left(\frac{r}{2\sqrt{Dt}}\right) - \frac{r}{\sqrt{\pi Dt}}\, e^{\frac{-r^2}{4Dt}}$

4. Mise en œuvre du modèle

Le modèle tel qu'il a été défini précédemment, bien que discrétisé et préparé à un pas de temps variable (équation [13]), est théorique et ne prend pas en compte les aspects pratiques inhérents à la simulation informatique tels que la discrétisation, l'imprécision numérique, la puissance de calcul, ... Nous allons maintenant nous intéresser à l'implémentation et plus particulièrement à l'optimisation tant au niveau du modèle que de l'algorithme.

4.1. *Simulation asynchrone et gestion du temps*

Dans les simulations informatiques basées sur l'itération traitant des problèmes physiques, la gestion du temps est un vrai dilemme : il faut choisir le bon équilibre entre la vitesse de calcul et le réalisme de la simulation. Différents critères concernant la nature du modèle peuvent déterminer si celui-ci est adapté ou non à un pas de temps fixe.

4.1.1. *Le problème du pas de temps fixe*

Dans une gestion classique du temps, la valeur du pas de temps est déterminée *a priori* en fonction de ce que l'on prévoit de simuler. Une façon de la déterminer consiste à spécifier une tolérance ainsi que les bornes des forces et des vitesses. On calcule alors le pas de temps le plus long possible qui maintient l'erreur inférieure à celle tolérée.

La discrétisation du temps implique que le vecteur vitesse d'un objet simulé soit approximé par une constante pendant toute la durée du pas de temps τ. On approche donc la trajectoire par une suite de déplacements rectilignes. Cela revient à dire que la force s'exerce ponctuellement à chaque pas et est nulle sur toute la trajectoire entre deux points. L'objet a alors une position et une vitesse légèrement différentes de celles qu'il devrait théoriquement avoir et cette erreur s'accumule au fil des itérations. Différentes méthodes d'intégration ont été proposées pour répondre à ce problème (Euler, Verlet, Runge-Kutta, ...). Elles permettent plus ou moins de limiter certains effets de bord, mais le problème de base reste présent. Ainsi, plus la vitesse est importante, plus τ devra être petit pour que les pas ne soient pas trop espacés. De même, plus la force est importante, plus l'approximation sur l'application de l'accélération est grossière et implique un τ petit.

Si on choisit la valeur de τ suffisamment faible, on retrouve bien une trajectoire proche de la théorie. Cependant, dans le cas d'une particule évoluant dans un champs de force très variable, le pas de temps fixe (calculé en fonction de la force maximale) n'est pas adapté dans les régions où la force est peu importante. Dans ce cas, il devient intéressant d'utiliser un pas de temps variable, calculé dynamiquement pour chaque agent en fonction de sa vitesse et de l'intensité locale du champs de forces. Le temps de calcul supplémentaire doit cependant être compensé par un gain suffisant ce qui est le cas lorsque les forces ou les vitesses sont très hétérogènes.

L'utilisation d'un pas de temps variable permet de contrôler l'erreur commise. En effet, τ est calculé de façon à ce que cette erreur soit fixe, spécifiée par l'utilisateur. Cela permet d'imposer un degré de réalisme nécessaire à la simulation de systèmes complexes, dont le comportement global résulte des comportements à de petites échelles de temps et d'espace. En effet, lorsque les forces sont très variables (figure 1), l'erreur commise sur un pas, même minime, peut avoir des conséquences très importantes : la particule se trouve juste à côte de là où elle devrait être et la force utilisée pour le calcul du pas suivant peut s'en trouver radicalement modifiée. La trajectoire simulée diverge alors extrêmement rapidement de la trajectoire théorique. Dans un contexte où l'on s'attend à de l'émergence, il est indispensable que les phénomènes chaotiques ne soient dus qu'aux propriétés physiques particulières du système et non à l'imprécision de la simulation.

4.1.2. *Un pas de temps optimal pour l'agent*

Dans le cas de particules microscopiques en suspension dans une solution, la loi qui régie le mouvement (équation [6]) est différente du cas macroscopique. En conséquence, le pas de temps optimal sera, lui aussi, calculé de façon différente.

Comme dans le cas précédent, une erreur maximale tolérée est fixée en début de simulation et le pas de temps est calculé de telle façon que l'erreur sur un pas ne dépasse jamais ce seuil. Afin de déterminer l'erreur, il faut quantifier l'écart entre la théorie et l'approximation faite par la discrétisation du temps. Cette approximation consiste à considérer que la vitesse est constante (en direction et en intensité) pendant toute la durée du pas de temps. Dans notre cas, cela revient à dire que la force appliquée est constante. Il est important de noter ici que ce sont des forces très variables qui imposent un pas de temps court et non des forces très importantes comme précédemment. En effet, si les forces sont très importantes mais constantes, le pas de temps sera long malgré tout (car, aux échelles considérées, nous avons vu que le déplacement est proportionnel à la force). Dans notre cas, l'intérêt d'un pas de temps variable est dû à l'hétérogénéité des forces : la force exercée par une protéine sur une autre est très variable à courte distance tandis qu'elle est quasiment constante (et généralement nulle) à grande distance. Le pas de temps optimal varie donc d'une protéine à l'autre suivant leur contexte environnemental.

La mesure de cette erreur peut donc être donnée par l'intégrale sur tout le trajet simulé de la variation du champs de force. L'expression de ce champs de force en un point x est théoriquement $F(x) = \sum_i f_i$. Or cette formule impose un calcul couteux. Partant du principe d'additivité des forces et du fait que l'encombrement spatial empêche une protéine d'être entourée directement d'un trop grand nombre d'autres protéines, on considère l'interaction avec une seule autre protéine, de coordonnées x' (cette approximation n'est utilisée que pour le calcul de τ et pas pour la simulation elle-même). On ne retient alors que la valeur minimale des τ calculés pour toutes les protéines avoisinantes. L'expression de $F(x)$ est alors donnée par l'équation [3] où $d = \|x' - x\|$. Afin de simplifier et d'accélérer le calcul, on peut faire l'approximation que, pour une valeur de $\|F(x)\|$ fixée, l'erreur est maximale lorsque $F(x_0)$ est di-

rigé vers le centre de la particule en x'. On note alors $F_n(d) = \|F(d\,u)\|$ la norme du champs de force à une distance d de la surface, u étant un vecteur unitaire quelconque. Si on ne considère pas le mouvement brownien, l'expression de l'erreur est :

$$Err(\tau, x_0) = \int_{\|x_0 - x'\| - \frac{\tau}{\lambda} F_n(\|x_0 - x'\|)}^{\|x_0 - x'\|} \left| \frac{\partial F_n}{\partial d}(d) \right| dd \qquad [14]$$

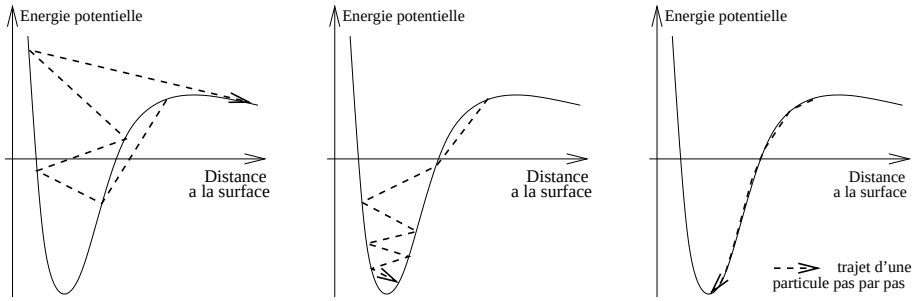
L'introduction du mouvement brownien complique le calcul car le caractère déterministe de la trajectoire est perdu, ce qui amène à considérer un ensemble de chemins possibles sur lesquels faire la sommation de l'erreur. Il est impératif de ne pas calculer le pas de temps *a posteriori* du tirage aléatoire du mouvement brownien car sa distribution spatiale en serait modifiée : le calcul étant basé sur le déplacement de la particule, il prend obligatoirement en compte le mouvement brownien. Si celui-ci était déterminé préalablement, le résultat du calcul serait dépendant du tirage aléatoire et pénaliserait les tirages ayant tendance à rapprocher les particules les unes des autres (puisqu'ils entrainent des pas de temps plus courts). L'isotropie et l'allure de la distribution du mouvement brownien seraient alors perdues. La distribution de Maxwell étant non bornée, l'ensemble des chemins possibles couvre théoriquement tous les points de l'espace, ce qui n'est bien sûr pas possible en pratique. Il faut donc borner la distribution. Une valeur V_{max} est alors définie telle que $V_{max} \geq \mathcal{V}$, avec $\mathcal{V}$ la variable aléatoire définie à la section 3.3.3. En paramétrant la distribution de Maxwell pour notre modèle et en fixant $V_{max} = 6$, la probabilité sur un tirage que cette approximation n'ait aucun effet est $P(\mathcal{V} \leq V_{max}) > 0.9995$.

Pour tenir compte du mouvement brownien (équation [13]), il faut distinguer deux cas selon que son intensité (variant en $\sqrt{\tau}$) est supérieure ou non au mouvement induit par le champs de forces (variant en τ). L'équation [14] doit donc être remplacée par :

$$Err(\tau, x_0) = \int_{\|x_0 - x'\| - \frac{\tau}{\lambda} F_n(\|x_0 - x'\|) - V_{max}\sqrt{\frac{RT}{\mathcal{N}_{av}\lambda}\tau}}^{\|x_0 - x'\| + \max\left(0, V_{max}\sqrt{\frac{RT}{\mathcal{N}_{av}\lambda}\tau}\right)} \left| \frac{\partial F_n}{\partial d}(d) \right| dd \qquad [15]$$

4.1.3. *Une simulation physiquement plus réaliste*

Cette gestion dynamique du temps permet d'avoir un contrôle sur l'erreur et sur la précision de la simulation. Dans beaucoup de simulations à pas de temps fixe, celui-ci est fixé par rapport au résultat observé. Dans le cas d'une simulation de particules avec un pas de temps fixe, celui-ci sera jugé trop long tant que le formation d'agrégats n'est pas observée. Cela correspond au cas de la figure 3.a où la stabilisation dans le minimum de l'énergie est difficilement possible et la formation de liaisons rare et instable. Si la valeur de τ est diminuée, des agrégats se forment car la stabilisation dans le minimum d'énergie est possible (figure 3.b). Cependant, bien que le résultat macroscopique observé paraisse correct, le pas de temps est encore beaucoup trop important car la distance d oscille autour du minimum d'énergie. Cela implique une dynamique artefactuelle et empêche la formation de structures complexes. L'utilisation du pas de temps variable permet de contrôler cet artefact et assure que la stabilisation d'une liaison se fait de façon réaliste (figure 3.c).



a. Pas de temps trop long pour la formation d'agregats

b. Pas de temps trop long pour une dynamique realiste

c. Pas de temps variable

Figure 3. *Effet de la durée du pas de temps sur le réalisme de la simulation : (a) comportement non réaliste, (b) comportement d'apparence réaliste et (c) comportement physiquement réaliste.*

4.2. *Passage à l'échelle cellulaire : optimisation et performances*

Afin de pouvoir observer non seulement des comportements locaux (formation de structures dynamiques et statiques), mais également des comportements globaux (différentiation des structures, auto-organisation), il est nécessaire de simuler une quantité suffisante de protéines. On doit alors rechercher des algorithmes de calcul dont la complexité en fonction du nombre d'agents ou de leur concentration permette un passage à l'échelle cellulaire ou, du moins, à l'échelle des structures étudiées.

4.2.1. *Accès rapide aux voisins*

Dans une simulation multi-agents, l'algorithme doit calculer les interactions entre toutes les paires d'agents. Si on note n le nombre d'agents, la complexité varie alors en $C = n^2$. Cependant les forces, très importantes à courte distance, deviennent rapidement très faibles, à tel point qu'on peut les négliger face au mouvement brownien. Une distance seuil d_s est alors définie, au-delà de laquelle on considère qu'il n'y a plus d'interaction. La complexité dépend alors de la densité des agents dans l'espace puisque chaque agent n'interagit qu'avec ceux situés à l'intérieur de son domaine d'interaction. Si on note R le rayon de la cellule et que l'on considère une répartition uniforme des agents dans l'espace, alors la complexité est multipliée par le rapport entre le volume d'interaction et le volume de la cellule. Elle varie alors en :

$$C = n^2 \left(\frac{d_s}{R} \right)^3 \qquad [16]$$

On remarque que si l'on impose une concentration (n/R^3 fixé) et une valeur de la distance seuil, alors la complexité de l'algorithme varie en n. Cette linéarité montre que simuler k cellules contenant n protéines par k simulations différentes (qu'elles

soient successives ou parallèles) revient, en termes de temps total de calcul, à simuler une unique cellule homogène contenant kn protéines. La difficulté à paralléliser le calcul vient de la forte connectivité des domaines d'interaction des agents.

La technique utilisée pour accéder rapidement aux agents en fonction de leur position dans l'espace consiste à les atteindre à travers un tableau tridimentionnel quadrillant l'espace. Chaque cellule du tableau permet l'accès rapide à tous les agents qui se situent dans une région de l'espace. Le niveau de détail du tableau doit résulter d'un compromis entre le temps perdu avec des agents se trouvant dans le sous-ensemble utilisé mais pas dans le domaine d'interaction de l'agent considéré (le sous-ensemble est une union de cubes et le domaine d'interaction est une sphère) et le temps perdu pour l'actualisation de la structure (faire passer les agents qui se déplacent d'une cellule du tableau à une autre). La taille de cellule permettant cet équilibre varie en fonction de la concentration des agents et de leur déplacement moyen (et donc de la température).

4.2.2. *Optimisation de la gestion des temps*

D'un point de vu algorithmique, la difficulté de la simulation à pas de temps variable vient de ce que le calcul de l'état des différents agents ne peut plus se faire par un balayage exhaustif de la population. On peut imaginer plusieurs méthodes pour déterminer, à un certain instant, le prochain agent à simuler. La solution que nous avons choisie, consiste à choisir l'agent dont le temps est le plus faible. L'algorithme gère une liste triée en fonction du temps, simule le premier agent, actualise son temps propre, puis le replace dans la liste. La complexité de la gestion temporelle des agents est alors en $C = n^2$, tout comme la gestion spatiale des agents, à taille de cellule et de rayon d'interaction fixe (d_s/R fixe). Cependant, si on se place dans le cas de la concentration fixe (n/R^3 fixe), c'est la gestion du temps qui prime lorsque le nombre d'agents devient élevé. Dans ce cas, il devient intéressant de paralléliser la simulation.

5. Expérimentation et résultats

5.1. *Vitesse de simulation*

Pour une simulation donnée, on fixe les valeurs du rayon de la cellule, de la distance d'interaction, du nombre de protéines, des coordonnées du point seuil et du point d'équilibre de chaque interaction. Ainsi, la vitesse de simulation ne dépend que de la position des agents les uns par rapport aux autres et de la température. La durée moyenne du pas de temps peut donc, à température constante, servir d'indicateur pour caractériser le degré d'agrégation des protéines et la stabilité du système.

La figure 4 est le résultat d'une expérience de simulation qui consiste à augmenter la température du milieu par paliers sur une petite quantité de protéines[9]. La figure

9. Le modèle a été testé, dans les mêmes conditions, avec des quantités de protéines allant jusqu'à quelques milliers. Les temps de simulation restent acceptables sur un PC de bureau mais il devient impossible de suivre la simulation en temps réel.

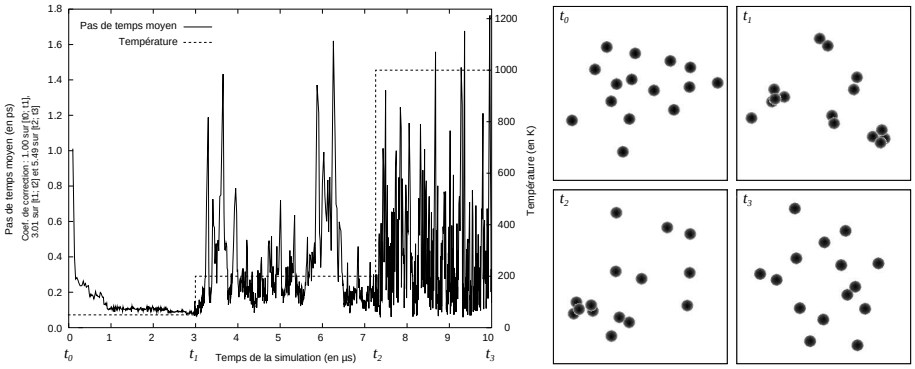


Figure 4. *Expérience de simulation de 15 protéines à trois températures : 50 K, 200 K et 1000 K. La durée du pas de temps est prise comme indicateur du degré d'agrégation des protéines.*

représente l'évolution de l'indicateur et de la température en fonction du temps de la simulation. Au début de l'expérience, la température est de 50 K et les protéines sont placées uniformément dans l'espace cellulaire. Entre les instants t_0 et t_1, la décroissance de l'indicateur témoigne de la formation d'agrégats et sa faible variabilité indique que les structures sont stables. À l'instant t_1, la température est brutalement portée à une valeur de 200 K, ce qui provoque une augmentation de l'amplitude du mouvement brownien, et donc de l'erreur calculée. Cela se traduit par une diminution brutale du pas de temps et entraine une discontinuité de l'indicateur (corrigée sur la figure, cf légende). Suite à cette augmentation de température, on constate que l'indicateur croit en moyenne, puis oscille entre des périodes présentant de fortes fluctuations de durées très variables dans lesquelles des composés instables se forment (diminution de l'indicateur) et se disloquent (augmentation de l'indicateur) et, des périodes durant lesquelles des structures stables perdurent (l'indicateur est alors faible et stable). Ces oscillations entre deux états très différents du système (stable et instable) avec de longues durées de cohérence sont caractéristiques des phénomènes critiques (Lagües *et al.*, 2003). À l'instant t_2, on porte la température à une valeur de 1000 K. On constate alors non seulement une augmentation moyenne de l'indicateur, mais surtout des variations d'amplitude très importantes et très rapides. Ceci est dû à une fréquence élevée de formations et de destructions de liaisons et témoigne de leur caractère transitoire.

Ainsi, par cette simulation très simple, on observe déjà une transition de phase qui met en évidence l'existence de valeurs critiques des paramètres. Cette expérience n'a bien sûr qu'une vocation d'exemple et l'intérêt est d'effectuer les mêmes manipulations sur un ensemble de protéines beaucoup plus important, en cherchant des points critiques sur l'ensemble des paramètres.

5.2. *Formations de structures*

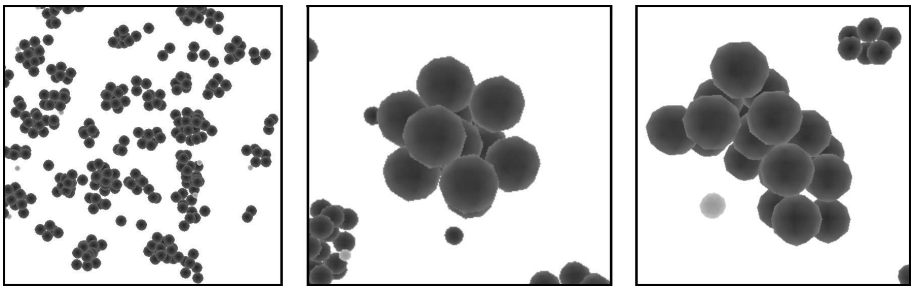
Bien que tout l'espace de paramètres n'ait pas encore été exploré, les premiers résultats se sont révélés encourangeants (figure 5). En effet, malgré les conditions simples dans lesquelles les expériences ont été effectuées (deux types de protéines isotropes), on observe que des structures auto-organisées très différentes émergent assez rapidement. De plus, la forme de ces structures apparaît comme très dépendante des paramètres de la simulation.

Ainsi, dans le cas de la figure 5, la seule augmentation de la concentration relative d'une des protéines provoque l'émergence de structures. Dans le cas 5.a, on constate l'apparition d'agrégats de protéines A, parfois influencés par les protéines B (les agrégats prennent alors une forme hémisphérique). En revanche, du fait de la faible concentration de protéines B, il n'y a pas d'influence réciproque. Dans le cas 5.b, les conditions de concentration provoquent un comportement totalement différent : des agrégats de A et de B se forment et s'influencent mutuellement. On observe alors l'émergence de structures spatialement organisées. La dynamique de ces structures montre un phénomène de croissance lié à la conformation initiale de l'amorce, conduisant à l'apparition de complexes de formes très différentes (feuillets ou hélices imbriquées).

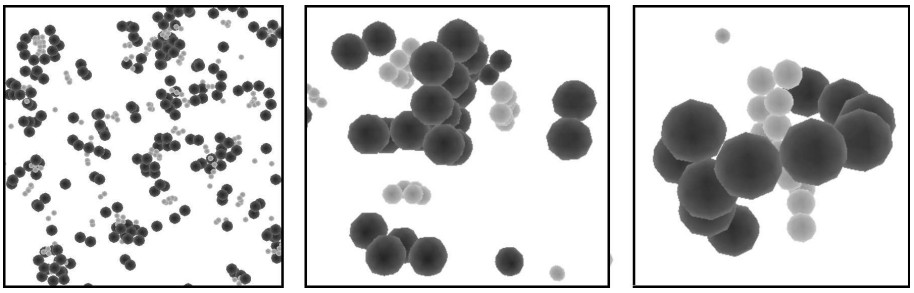
6. Conclusion

À travers la description de toutes les étapes de construction d'un modèle cellulaire, nous avons illustré ce que, selon nous, doit être une démarche de simulation en biologie des systèmes. Partant d'un choix paradigmatique (rechercher les conditions d'émergence d'une structure) et d'une problématique scientifique précise (étudier l'organisation tridimensionnelle de structures multiprotéiques), nous avons défini un cadre de modélisation. Celui-ci est basé sur une description précise des entités locales (ici les protéines), description qui doit être raisonnée entre une nécessaire simplification et la volonté de créer les conditions d'une "vraie" émergence (au sens où les conditions de l'émergence se trouvent dans les paramètres du modèle et non dans ceux de la simulation qui implémente ce modèle). Dans notre cas, cet équilibre a été obtenu en recherchant un bon réalisme des interactions entre agents – ce qui nous a conduit à utiliser un modèle énergétique des interactions – et des déplacements libres des molécules, décrits par un mouvement brownien. En revanche, la structure 3D des agents eux-même a pu être en grande partie négligée puisque nous ne nous intéressons pas aux interactions issues des conformations secondaires des protéines.

Les conditions regroupées dans le modèle permettent la recherche de points d'équilibre entre mouvement spontané (source de désordre) et les forces d'interactions tendant à conduire à la formation de structures stables. Il est alors possible d'étudier les structures multi-protéiques émergeant à proximité de cet équilibre. Les premières expérimentations montrent que ces structures présentent des caractéristiques tridimensionnelles d'autant plus remarquables qu'elles ne sont aucunement présentes dans les



(a) A : 96%, B : 4% – agrégats peu organisés ; structuration partielle due aux protéines B.



(b) A : 50%, B : 50% – complexes multi-protéiques spatialement organisés du fait de la structuration réciproques entre les espèces A et B.

Figure 5. *Résultats préliminaires : dépendance des structures aux concentrations de protéines. Deux types de protéines – A (gris fonçé, rayon = 2nm) et B (gris clair, rayon = 1nm) – sont simulés dans un même environnement (T = 37˚C). La variation de concentration entraine la formation de structures différentes.*

molécules qui les composent. Si ces premiers résultats ont pour l'instant été obtenus hors de toute problématique biologique, ils nous permettent d'envisager des développements dans au moins deux directions. D'une part, nous pouvons utiliser 3DSPI pour modéliser une structure biologique particulière en recherchant des propriétés dynamiques et spatiales similaires à celles observées en biologie (par exemple sur les temps de résidence (Phair *et al.*, 2004)). D'autre part, les premiers résultats nous permettent d'envisager l'exploration de l'espace des paramètres pour rechercher les conditions d'émergence de structures organisées, dans une démarche proche de celles conduites sur les automates cellulaires (Langton, 1986, Langton, 1991).

Enfin, il est clair que ce modèle ouvre de nombreuses perspectives en termes de modélisation. On retiendra en particulier la possibilité de modéliser les protéines sous la forme d'une collection de domaines, ce qui permettrait d'étudier des structures plus complexes. Il s'agit cependant là d'un projet scientifique différent que, fidèles à notre conception de la biologie des systèmes, nous devons formaliser préalablement à toute mise en œuvre.

Remerciements

Nous tenons à remercier Jean-Jacques Diaz et Annick Lesnes ainsi que tous les membres du groupe de Biologie des Systèmes et Modélisation Cellulaire (BSMC, INSA-UCBL) pour les discussions sur ce modèle. Ce projet est soutenu financièrement par l'INSA de Lyon, la région Rhônes-Alpes et le programme ACI IMPBio (projet MOCEME).

7. Bibliographie

Amar P., Ballet P., Barlovatz-Meimon G., Benecke A., Bernot G., Bouligand Y., Bourguine P., Delaplace F., Delosme J.-M., Demarty M., Fishov I., Fourmentin-Guilbert J., Fralick J., Giavitto J.-L., Gleyse B., Godin C., Incitti R., Kepes F., Lange C., Sceller L. L., Loutellier C., Michel O., Molina F., Monnier C., Natowicz R., Norris V., Orange N., Pollard H., Raine D., Ripoll C., Rouviere-Yaniv J., Jr M. S., Soler P., Tambourin P., Thellier M., Tracqui P., Ussery D., Vincent J.-C., Vannier J.-P., Wiggins P., Zemirline A., « Hyperstructures, génome analysis and Icell », *Acta Biotheor.*, vol. 50, n° 4, p. 357-373, 2002.

Atamas S., « Self-organization in computer simulated selective systems », *Biosystems*, vol. 39, p. 143-151, 1996.

Ballet P., Zemirline A., Marcé L., « The BioDyn Language and Simulator. Application to an immune response and E.Coli and Phage interaction », *Journal of Biological Physics and Chemistry*, vol. 4, p. 93-101, 2004.

Barabasi A., Oltvai Z., « Network biology : understanding the cell's functional organization », *Nat. Rev. Genetics*, vol. 5, p. 101-114, 2004.

Berg H. C., *Random Walks in Biology*, Princeton University Press (2nd edition), 1993.

Blake W., M K., Cantor C., Collins J., « Noise in eukaryotic gene expression », *Nature*, vol. 422, p. 633-637, 2003.

Bork P., Serrano L., « Towards cellular systems in 4D », *Cell*, vol. 121, p. 507-509, 2005.

Broderick G., Ru'aini M., Chan E., Ellison M., « A life-like virtual cell membrane using discrete automata », *In Silico Biology*, 2004.

Demazeau Y., « From interactions to collective behaviour in agent-based systems », *European Conference on Cognitive Science*, p. 117-132, 1995.

Dittrich P., Ziegler J., Banzhaf W., « Artifi cial chemistry - a review », *Artificial Life*, vol. 7, n° 3, p. 225-275, 2001.

Ferber J., Gutknecht O., « A meta-model for the analysis and design of organizations in multi-agent systems », *Proc. of 3rd Int. Conf on Multi-Agent Systems*, p. 128-135, 1998.

Ideker T., Galitski T., Hood L., « A new approach to decoding life : systems biology », *Annu. Rev. Genomics Hum. Genet.*, vol. 2, p. 343-372, 2001.

Kitano H., « Systems biology : a brief overview », *Science*, vol. 295, p. 1662-1664, 2002.

Kupiec J.-J., « A Darwinian theory for the origin of cellular differentiation », *Mol Gen Genet*, vol. 255, p. 201-208, 1997.

Lagües M., Lesnes A., *Invariances d'échelle, des changements d'états à la turbulence*, Belin, 2003.

Langton C., « Life at the edge of chaos », *Artificial Life II*, Addison-Wesley, 1991.

Langton C. G., « Studying Artifi cial Life with cellular automata », *Physica D*, 1986.

Lemerle C., Ventura B. D., Serrano L., « Space as the fi nal frontier in stochastic simulations of biological systems », *FEBS Letter*, vol. 579, p. 1789-1794, 2005.

LeSceller L., Ripoll C., Demarty M., Cabin-Flamand A., Nystrom T., Saier M., Norris V., « Modelling bacterial hyperstructures with cellular automata », *InterJournal*, 2000.

Levsky J., Singer R., « Gene expression and the myth of the average cell », *Trends Cell Biol.*, vol. 13, p. 4-6, 2003.

Loew L., Schaff J., « The Virtual Cell : a software environment for computational cell biology », *Trends Biotechnol*, vol. 19, n° 10, p. 401-406, 2001.

McNally J.-G., Muller W.-G., Walker D., Wolford R., Hager G.-L., « The glucocorticoid receptor : rapid exchange with regulatory sites in living cells », *Science*, vol. 287, p. 1262-1265, 2000.

Misteli T., « Protein dynamics : implication for nuclear architecture and gene expression », *Science*, vol. 291, p. 843-847, 2001.

Misteli T., « Concepts in nuclear architecture », *BioEssays*, vol. 27, p. 477-487, 2005.

Phair R., Scaffi di P., Elbi C., Vecerová J., Dey A., Ozato K., Brown D., Hager G., Bustin M., Misteli T., « Global nature of dynamic protein-chromatin interactions in vivo : three-dimensional genome scanning and dynamic interaction networks of chromatin proteins », *Molecular and Cellular Biology*, vol. 24, n° 14, p. 6393-6402, 2004.

Ronald E., Sipper M., « Surprise versus unsurprise : Implications of emergence in robotics », *Robotics and Autonomous Systems*, vol. 37, p. 19-24, 2001.

Schweitzer F., *Brownian Agents and Active Particles, Collective Dynamics in the Natural and Social Sciences*, Springer, 2003.

Soula H., Robardet C., Perrin F., Gripon S., Beslon G., Gandrillon O., « Modeling the emergence of multi-protein dynamic structures by principles of self-organization through the use of 3DSpi, a multi-agent-based software », *BMC Bioinformatics*, 2005.

Stuart J., Segal E., Koller D., Kim S., « A gene-coexpression network for global discovery of conserved genetic modules », *Science*, vol. 302, p. 249-255, 2003.

Takahashi K., Kaizu K., Hu B., Tomita M., « A multi-alforithm, multi-timescale method for cell simulation », *Bioinformatics*, vol. 20, n° 4, p. 536-546, 2004.

Theraulaz G., Spitz F., *Auto-organisation et comportement*, Hermès (Paris), 1997.

Tononi G., Sporns O., Edelman G., « Measures of degeneracy and redundancy in biological networks », *Proc. Natl. Acad. Sci. USA*, vol. 96, p. 3257-3262, 1999.

Vespignani A., « Evolution thinks modular », *Nat. Genet.*, vol. 35, p. 118-119, 2003.

Weiss G., *Multiagent Systems, a Modern Approach to Distributed Artificial Intelligence*, MIT Press (Cambridge), 1999.

Wishart D., Yang R., Arndt D., Cruz J., « Dynamic cellular automata : an alternative approach to cellular simulation », *In Silico Biology*, 2004.