

Systèmes dynamiques aux interfaces mathématiques/informatique/biologie

Anne Siegel, IRISA, Rennes

Résumé

J'exploite les propriétés de systèmes dynamiques dans des domaines où ces systèmes apparaissent de manière cachée. J'identifie des questions reliées à des trajectoires de systèmes dans des domaines variés, et j'utilise les propriétés de ces trajectoires couplé à des approches algorithmiques pour répondre aux questions initiales.

En informatique théorique, j'ai caractérisé les propriétés d'écritures de nombres en base non entière et proposé une stratégie d'engendrement et d'étude de plans discrets. Pour cela, j'ai utilisé des additions modulo 1 et de la géométrie fractale.

En bioinformatique, en plus d'une activité autour de la construction de modèles au bon niveau d'abstraction, j'ai développé un formalisme de résolution de contraintes pour la confrontation de données large-échelle issues de la biologie moléculaire (transcriptomique, chip-chip, CGH) avec des modèles de connaissances. Ceci permet d'introduire des concepts dynamiques au sein de l'analyse des données en mimant un raisonnement automatique sur des variations.

Mes principaux résultats sont les suivants.

- a. Formalisme de résolution de contraintes pour la confrontation de données grande-échelle (transcriptôme, chip-chip, CGH) avec des modèles de connaissances.
- b. Construction de hiérarchies de modèles pour l'interprétation de données (régulation du métabolisme des acides gras, initiation de la traduction chez l'oursin).
- c. Description par des graphes de la frontière de fractals et algorithmes pour caractériser leurs propriétés topologiques. Application à la construction de partition de Markov et caractérisation des développements périodiques en base non-entière.
- d. Règles de construction de plans discrets à l'aide de modèles de \mathbb{Z}^2 action

Parcours scientifique et mobilités thématiques

Mon parcours s'insère naturellement dans le cadre de l'interdisciplinarité : j'ai suivi une formation en mathématiques pures (thèse sur l'études de systèmes dynamiques symboliques), puis j'ai été recrutée comme maître de conférences au laboratoire de mathématiques de Rennes 1 (IRMAR) pour quelques mois avant d'obtenir un poste au CNRS via le concours CR de la section 07 (informatique), pour réfléchir à des applications de la théorie des systèmes dynamiques dans différents domaines. J'ai ainsi été affectée au laboratoire d'informatique de Rennes (IRISA), où j'ai intégré l'équipe de bioinformatique.

Depuis mon recrutement, mes travaux portent donc sur l'étude et l'utilisation de systèmes dynamiques dans différents domaines. En informatique théorique, il s'est agit de caractériser les écritures de nombres en base non entière et de proposer une stratégie d'engendrement et

d'étude de plans discrets. L'interprétation mathématique de ces travaux repose sur l'exploitation du concept d'autosimilarité et de géométrie fractale.

En bioinformatique, il s'est agi d'intégrer des concepts dynamiques au sein de l'analyse des données en biologie. La biologie moléculaire est un domaine très innovant du point de vue des systèmes dynamiques : les informations y sont partielles et finalement en nombre assez limité. Partant de ce constat, j'ai développé des approches à la fois autour de la construction de modèles, et de l'intégration de données de type variationnelles (plutôt que des trajectoires) au sein des analyses de données issues de la biomoléculaire. Le prolongement naturel de ces travaux repose sur la notion d'aide au raisonnement pour la construction de modèle et l'intégration de données.

Utilisations de systèmes dynamiques

Les systèmes dynamiques sont des objets mathématiques introduits pour prédire l'avenir à partir de règles d'évolution. Utiliser ce concept suppose qu'un système en évolution est contraint par un certain nombre de lois, et que la connaissance de ces lois permet de prédire exactement l'avenir de n'importe quel élément si on connaît ses caractéristiques en un instant donné. Les premiers exemples étudiés, au 18ème siècle, concernaient la mécanique céleste. Ils ont mené au 19ème siècle à de multiples questions sur la résolution d'équations différentielles. On s'est ensuite aperçu au 20ème siècle, avec Poincaré, que certaines équations différentielles ne peuvent pas avoir de solution explicite, ce qui a ouvert un champ énorme de recherche autour de la simulation numérique des solutions de systèmes d'équations différentielles et de l'identification de paramètres en accord avec les données observées. Ces méthodes sont maintenant utilisées et discutées dans de nombreux domaines mathématiques, informatiques ou technologiques.

Cependant, le concept de système dynamique est loin d'être limité à la notion de simulation de trajectoires. En effet, un système dynamique n'est finalement rien d'autre que la donnée d'un ensemble d'éléments et d'une loi de transformation de ces éléments. Or, ces informations (loi sur un ensemble) se retrouvent dans de nombreux domaines scientifiques. Par exemple, en mathématiques, on sait qu'il existe une infinité de nombres dont les développements dans n'importe quelle base sont équidistribués. En informatique, l'analyse en moyenne d'algorithmes n'est rien d'autre que l'étude du comportement moyen de systèmes dynamiques. En physique, la structure atomique des cristaux est associée à une notion de répétition, et donc de système dynamique. En biologie, l'évolution des populations est naturellement décrite à l'aide de systèmes dynamiques.

Mon activité de recherche porte sur l'étude des systèmes dynamiques qui sont cachés au sein de questions mathématiques, informatiques ou biologiques. Je ne cherche pas à simuler le comportement de ces systèmes, mais à utiliser les propriétés globales de leurs trajectoires pour retirer des informations permettant de mieux circonscrire le problème initial. Concrètement, il s'agit d'abstraire et de discrétiser les dynamiques pour capter des propriétés globales.

J'aborde ces questions dans deux domaines distincts : d'abord, l'étude des multiplications et des additions modulo 1, avec des applications pour l'étude des systèmes de numération et la génération de plans discrets. En parallèle, je m'intéresse à l'analyse de données en biologie moléculaire, en vue de proposer des méthodes d'aide à la modélisation, l'expérimentation et la compréhension du fonctionnement des systèmes biologiques.

Informatique théorique et mathématiques : exploiter la "fractalité" de différents objets

Mes travaux en informatique théorique et en mathématiques portent sur les additions modulo 1 : il s'agit de considérer un nombre x entre 0 et 1, de lui additionner un autre nombre fixé β , et de conserver la partie fractionnaire de cette somme. On peut ensuite ajouter à nouveau β au réel ainsi obtenu... L'action de l'addition modulo 1 sur $[0, 1[$ (tore) forme donc un système dynamique. Lorsqu'on additionne des coordonnées de vecteurs pour ne garder que la partie fractionnaire des sommes, on parle plus généralement d'*additions ou d'automorphismes du tore*. Ces additions modulo 1 (ou leurs cousines, des multiplications) sont présentes dans différents domaines : elles servent à écrire des nombres dans des bases non entière (par exemple, en remplaçant une base décimale par le nombre d'or). On utilise aussi ces additions pour décrire des approximations de droites du plan par des escaliers (on parle de droites discrètes). La notion de quasi-cristal en physique théorique (un cristal avec des régularité partielle) s'illustre bien avec ces additions. Le point commun entre tous ces objets mathématiques est leur **auto-similarité** : par un processus de zoom (appelée induction), on peut retrouver au sein de chacun ces objets une structure qui est très proche de la structure originale. Mes travaux consistent à isoler cette structure auto-similaire (on parle aussi de fractal) au sein de ces objets mathématiques, informatiques ou physiques, et à interpréter les propriétés mathématiques de l'objet initial comme des propriétés topologiques de fractals. Ceci permet d'exploiter la répétition induite par ce système pour mieux comprendre les propriétés de l'objet initial.

J'ai en particulier développé une "boîte à outils" algorithmique assez fine pour décrire les propriétés topologiques des fractals concernés (MémoiresSMF, 2010). Cette boîte à outils est basé sur la recherche des écritures redondantes de réels dans différentes bases, et leur description par des automates finis. D'un point de vue topologique, il s'agit de caractériser la frontière d'un ensemble fractal (Ann.Inst.Fourier, 2004). Cette boîte à outils m'a permis de trouver de nouvelles propriétés sur les objets originaux.

Les objets mathématiques que j'ai étudiés ainsi sont les écritures en bases non entière (où la base est la racine d'un polynôme de degré 2 ou 3) (JNT, 2007 ; Integers, 2005), les plans discrets (TCS, 2004), les automorphismes de groupes libres pseudo-Anosov (Ann. Inst. Four., 2006) et les automorphismes et additions du tore (Erg.Th.Dyn.Sys, 2003). Comme mentionné plus haut, ces objets mathématiques sont en fait fortement reliés, et j'ai travaillé sur différents survols pour expliciter leurs constructions et leurs points communs, en ramenant de nombreuses questions à leur sujet à une "simple" question de pavage (CANT, 2010 ; Pytheas-Fogg LNM, 2002 ; TransAMS, 2001 ; JNT Bordeaux, 2001). Ainsi, en théorie des nombres, ces approches ont permis de caractériser les propriétés des développements finis ou purement périodiques de rationnels en base non entière (BulletinLMS, revision ; Mona.Math., 2008). En géométrie discrète, ces propriétés s'interprètent en termes de conditions pour l'engendrement de plans discrets par des méthodes itératives (Journées Montoises, 2006).

Modélisation en biologie moléculaire : croiser les informations dynamiques incomplètes apportées par les données et les connaissances

En biologie, j'ai considéré des questions d'analyses de données avec une approche similaire : isoler, abstraire et exploiter les informations dynamiques au sein des données. Les masses de données produites par la biologie moléculaire explosent littéralement depuis une quinzaine d'années, et on doit constater que ces données ne sont que assez grossièrement exploitées. De nombreuses méthodes, désignées sous le nom de *biologie intégrative*, essaient actuellement de croiser ces données pour en extraire des signaux statistiquement fiables.

Je me suis personnellement intéressée à l'interprétation des données issues des "puce à ADN" : il s'agit de stresser une cellule par une action physiologique ou génétique (activation / inhibition de gène), d'attendre que les éléments au sein de la cellule interagissent et que leurs actions s'équilibrent, puis de comparer la quantité finale d'ARN dans la cellule par rapport à la cellule non stressée. Sur un plan dynamique, on dispose ainsi d'une information sur les variations qualitatives de milliers d'ARN d'une cellule, entre deux états stationnaires.

J'ai proposé d'analyser ces données variationnelles en les confrontant aux variations qu'on attendrait intuitivement voir surgir d'un système théorique reprenant l'ensemble des connaissances sur les interactions à l'intérieur d'une cellule (Biosystems, 2006 ; Roy, Interf. Soc, 2006). Il s'agit d'une vision originale de la domaine de la modélisation dynamique, dans la mesure où on ne cherche pas à comprendre/simuler les évolutions dans le temps des concentrations brutes des éléments du système, mais à simplement faire des comparaisons qualitatives entre le début et la fin de la trajectoire. L'information fondamentale de nature dynamique sur laquelle repose notre intuition est que la variation d'un élément ne doit pas être contradictoire avec celle des ses régulateurs. D'un point de vue informatique, on considère que le système impose des contraintes sur les variations globales de ses éléments, ce qui a mené à voir l'analyse des données transcriptomiques comme une question de résolution de contraintes qualitatives, qui est NP-difficile (ComplexUs, 2006). La résolution de ces systèmes de contraintes a été faite d'abord avec des diagrammes de décision, puis, pour étendre le champ d'application, avec des solveurs basés sur de la programmation par ensembles-réponses.

Cette approche a été validée et exploitée sur différents organismes modèles : correction du réseau transcriptionnel de la bactérie E. Coli (CIBB, 2009), inférence de réseau de régulation pour la levure S. Cerevisiae (BMC bioinfo, 2008), recherche des causes de l'apoptose dans le réseau de régulation de la protéine chimère responsable des tumeurs pédiatriques d'Ewing (TCCB, revision), interprétation du nombre de copies chromosomiques (BMINT, 2009). Cette approche a aussi donné lieu à des développements logiciels pour faciliter son utilisation (BMC genomics, 2010).

En parallèle, ces travaux ont permis de réfléchir aux besoins en modélisation en biologie moléculaire, en fonction des données disponibles. Il apparaît maintenant que les systèmes dynamiques en biologie moléculaire sont par nature incomplets (sous-paramétré et peu observables), et qu'ils le resteront pendant longtemps. Prendre cette incomplétude en compte amène à développer des approches de comparaison de systèmes plutôt que de simulation, à différentes échelles. Des travaux dans cette direction ont été initiés sur la modélisation des régulations du métabolisme des acides gras (JTB, 2009), et sur l'initiation de la traduction chez l'oursin (MRD, 2010). Là encore, le point commun entre ces approches est d'identifier le bon niveau de modélisation en fonction des connaissances et d'intégrer les informations incomplètes à bon escient pour extraire des informations de type dynamique des observations, en particulier pour décider quelles interactions cellulaires sont fondamentales pour décrire le comportement d'une cellule.

Spécificité : combiner des systèmes dynamiques et des approches algorithmiques

Même si ces deux domaines d'application semblent éloignés, le point commun qui les rassemble est que toutes les approches proposées se sont adaptées à l'état des connaissances actuelles, en adoptant une approche algorithmique : la recherche de modèles d'abstraction adaptés s'est révélée fondamentale pour résoudre des questions à la fois biologique, mathématiques et informatique.

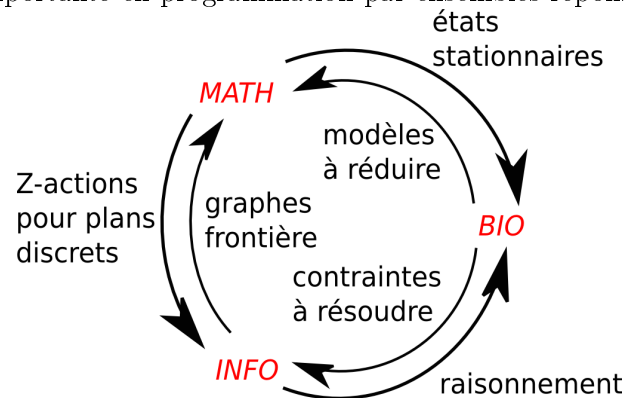
En mathématiques, on souhaiterait idéalement prouver une conjecture générale qui dit que tous les systèmes symboliques engendrés par une substitution Pisot sont mesurablement isomorphes à une addition sur un tore. Pour l'instant, la preuve de ce résultat semble inaccessible ;

en fait, il s'avère que cette conjecture n'a pas été testée de manière intensive, en particulier en grande dimension. Les méthodes que j'ai développées, s'appuyant sur un modèle explicite de la frontière de fractal sous forme de graphe, consistent donc à vérifier cette propriété par un calcul du rayon spectral du graphe décrivant la frontière. Lorsque la propriété est vérifiée, on peut concrètement exploiter cette information pour l'étude des écritures de nombre ou la structure des discrétisation de plans.

De même, en biologie moléculaire, nous sommes partis du constat que même si des informations de plus en plus nombreuses sont mises à disposition sur le comportement des systèmes biologiques, elles ne sont pas en nombre suffisant pour modéliser la complexité de la dynamique des systèmes usuellement considérés par les chercheurs en biologie (en particulier ceux qui travaillent sur des modèles eucaryotes). J'ai donc soutenu l'idée qu'il fallait volontairement se restreindre à la compréhension des déplacements d'états stationnaires. Cette restriction a permis de s'appuyer sur différentes approches de résolution de contraintes, avec l'ambition d'aider au diagnostic et à la mise en place de plans expérimentaux.

Le point caractéristique de mes travaux est donc de s'appuyer sur des interfaces biologie/mathématiques/informatiques pour enrichir les différents domaines

- *Math* → *Info* Les mathématiques apportent des structures de \mathbb{Z}^2 -action pour la caractérisation de plans discrets, et permettent de déduire des règles de construction des plans discrets.
- *Info* → *Math* L'informatique, via des algorithmes pour décrire la frontière de fractals, permettent de calculer explicitement des développements de constantes mathématiques, ou de prouver que des partitions du plans sont parfaitement adéquates pour coder certaines additions modulo 1.
- *Math* → *Bio* Les mathématiques fournissent des critères de réductions de modèles qui permettent de décrire les propriétés de systèmes biologiques et de faire des prédictions à leur sujet.
- *Info* → *Bio* L'informatique fournit des méthodes de résolution de contraintes utiles pour diagnostiquer des incohérences entre les modèles des biologistes et les données grande-échelle à leur disposition.
- *Bio* → *Info/Math* La biologie, en particulier la biologie moléculaire, fournit de nouveau types de modèles qui génère des problématiques nouvelles et très riches, par exemple la réduction de modèles non paramétrés en systèmes dynamiques et l'analyse de contraintes de complexité importante en programmation par ensembles-réponses.



Publications citées dans le texte

- (BulletinLMS, revision) B. Adamczewski, C. Frougny, A. Siegel, W. Steiner, Rational numbers with purely periodic β -expansion, *Bulletin de la London Mathematical Society*, en révision.
- (Memoires SMF, 2010) A. Siegel and J. Thuswaldner, Topological properties of Rauzy fractals, *Mémoires de la SMF*, accepté pour publication, en révision, 2010.
- (MRD, 2010) R. Bellé, S. Prigent, A. Siegel and P. Cormier, Model of cap-dependent translation initiation in sea urchin. A step towards the eukaryotic translation regulation network *Molecular Reproduction and Development*, in press
- (CANT, 2010) V. Berthé, A. Siegel, J. Thuswaldner *Tilings, substitutions and Rauzy fractals : effectivity and complexity*, Chapter of *Combinatorics, Automata and Number Theory*, Cambridge University Press, in press.
- (BMIINT, 2009) S. Blachon, A. Siegel, O. Radulescu, *Relate inter-patient gene copy numbers variations with gene expression via gene networks in Ewing sarcoma*, Biomedical Informatics and Intelligent Methods in the Support of Genomic Medicine, Workshop, AIAI 2009, Greece. CEUR Workshop Proceedings vol. 475, 2009. Pages 72-87.
- (CIBB, 2009) C. Guziolowski, J. Gruel, O. Radulescu and A. Siegel, *Curating a large-scale regulatory network by evaluating its consistency with expression datasets*, CIBB 2008 : Computational Intelligence Methods for Bioinformatics and Biostatistics. Peer-reviewed selected revised papers. Lecture Notes in Computer Science, 5488, Springer-Verlag, p.144-155, 2009
- (JTB, 2009) P. Blavy, F. Gondret, H. Guillou, S. Lagarrigue, P. G. P. Martin, J. van Milgen, O. Radulescu and A. Siegel A minimal model for hepatic fatty acid balance during fasting : application to PPAR alpha-deficient mice, *Journal of Theoretical Biology*, 2009
- (BMC genomics, 2009) C. Guziolowski, A. Bourdé, F. Moreews and A. Siegel, *Bioquali Cytoscape Plugin : Analysing the global consistency of regulatory networks*, Soumis, 2008.
- (Repeat, submitted) J. Nicolas ; C. Rousseau ; A. Siegel ; P. Peterlongo ; F. Coste ; P. Durand ; S. Tempel ; A.-S. Valin and F. Mahé, *Local and Maximal Repeats*, Soumis, 2008.
- (BMC bioinfo, 2008) P. Veber, C. Guziolowski, M. Le Borgne, O. Radulescu, and A. Siegel, Inferring the role of transcription factors in regulatory networks, *BMC BioInformatics* 9, 2008. .
- (Monas. Math., 2008) S. Akiyama, G. Barat, V. Berthé and A. Siegel, Boundary of central tiles associated with Pisot beta-numeration and purely periodic expansions, *Monashefte für Mathematik* 155 (3-4) 2008, 377-419.
- (JNT, 2007) V. Berthé and A. Siegel, Purely Periodic beta-Expansions in the Pisot Non-unit Case *Journal of Number Theory* 153 (2) 2007, 153-172.
- (Ann. Inst. Fourier, 2006) P. Arnoux, V. Berthé, A. Hilion and A. Siegel Fractal representation of the attractive lamination of an automorphism of the free group, *Annales de l'Institut Fourier* 56(7), 2006, 2161-2212.
- (JPBC, 2006) C. Guziolowski, P. Veber, M. Le Borgne, O. Radulescu, and A. Siegel Checking Consistency Between Expression Data and Large Scale Regulatory Networks : A Case Study *Journal of Biological Physics and Chemistry* (7) 2007, 37-43 (Presented at Réseaux d'interaction : analyse, modélisation et simulation. RIAMS'06, Lyon, France.)
- (Biosystems, 2006) A. Siegel, O. Radulescu, M. Le Borgne, P. Veber, J. Ouy, S. Laguarigue Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks *BioSystems* 84, 2006, 153-174

- (Roy. Soc. Inter, 2006) O. Radulescu, S. Lagarrigue, A. Siegel, M. Le Borgne, P. Veber Topology and linear response of interaction networks in molecular biology, *Journal of The Royal Society Interface* 3(6), 2006, pp. 185 - 196
- (Jour. Montoises, 2006) P. Arnoux, V Berthé, and A Siegel, *Finiteness properties for Pisot S-adic tilings* Journées Montoises d'Informatique théorique. JM'06, Rennes, 2006
- (Integers, 2005) V. Berthé and A. Siegel, Tilings associated with beta-numeration and substitutions, *INTEGERS*, 5(3), pp. A2, 2005
- (Complex Us, 2005) P. Veber, M. Le Borgne, A. Siegel, S. Lagarrigue, O. Radulescu Complex Qualitative Models in Biology : a new approach *Complexus* 2, 2004/2005, pp. 140-151 (Presented at of ECCS, Paris, 2005)
- (TCS, 2004) P. Arnoux, V. Berthé and A. Siegel. Two-dimensional iterated morphisms and discrete planes . *Theoretical Computer Science*, 319, 2004, p. 145–176.
- (Ann. Inst. Fourier, 2004) A. Siegel, Pure discrete spectrum dynamical system and periodic tiling associated with a substitution. *Annales de l'Institut Fourier* 2(54), 2004, p. 288-299.
- (Erg. Th. Dyn. Sys., 2003) A. Siegel. Représentation des systèmes dynamiques substitutifs non unimodulaires. *Ergodic Theory and Dynamical Systems*, 23, 2003, 1247-1273
- (EWT, 2003) A. Siegel, *Spectral theory for dynamical system arising from substitutions*, EWM, K. Dajani, J. Von Reis (eds.), CWITract, Marseille, 2003.
- (Pytheas Fogg Intro, LNM, 2002) V. Berthé and A. Siegel. Basic notions on substitutions. In N. Pytheas-Fogg, *Substitutions in Dynamics, Arithmetics and Combinatorics*. Lectures Notes in Mathematics 1794, Springer-Verlag, 2002, pp. 1–34 (Chapter 1).
- (Pytheas Fogg, LNM, 2002) A. Siegel, Spectral theory and geometric representation of substitutions. In N. Pytheas-Fogg, *Substitutions in Dynamics, Arithmetics and Combinatorics*. Lectures Notes in Mathematics 1794, Springer-Verlag, 2002, pp. 199–252 (Chapter 7).
- (Trans. AMS, 2001) V. Canterini and A. Siegel. Geometric Representation of primitive substitution of Pisot type. *Transactions of the AMS*, 353(12), 2001, pp. 5121–5144.
- (JTN Bord., 2001) V. Canterini and A. Siegel. Automate des préfixes-suffixes associé à une substitution primitive. *Journal de théorie des nombres de Bordeaux*, 13(2), 2001, pp. 353–369.
- (Acta Arith, 2001) A. Siegel. Théorème des trois longueurs et suites sturmiennes : mots d'agencement des longueurs. *Acta Arithmetica*, 97 (1-3), 2001, pp. 195–210.