INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet SYSDYS

Systèmes Dynamiques Stochastiques

Sophia Antipolis

THÈME 4B

Table des matières

SYSDYS *est un projet commun à l'INRIA (Unité de recherche de Sophia Antipolis), au CNRS (LATP) et à l'Université de Provence.* SYSDYS *est localisé à Marseille au sein du LATP (Technopole de Château–Gombert)*

1 Composition de l'équipe

Responsable scientifique

Fabien Campillo

Assistante de projet

Sylvie Blanc

Personnel Université de Provence

Etienne Pardoux

Bruno Torrésani

Personnel Université Joseph Fourier

Marie–Christine Roubaud

Chercheurs doctorants

François Delarue [bourse AMN]

Lorie Dudoignon [depuis Octobre, bourse Génopole]

Antoine Lejay [jusqu'en Juin, bourse AMN]

Elisabeth Remy [1er au 31 Janvier]

Chercheur post-doctorant

Elisabeth Remy [1er au 30 Septembre]

Antoine Lejay [1er au 30 Septembre]

Enrico Formenti [depuis Octobre, bourse INRIA]

2 Présentation et objectifs généraux

En modélisation, analyse ou simulation numérique «l'aléa» transparaît à différents niveaux.

Le phénomène étudié peut intrinsèquement comporter des composantes stochastiques (coefficients aléatoires, entrées aléatoires etc.). Ou bien la modélisation précise qui en est faite peut s'avérer trop «inextricable» ; il est alors préférable d'en «simplifier» certaines difficultés, en introduisant des termes aléatoires, afin de le rendre accessible à l'analyse puis à la simulation. Enfin, depuis quelques années, les probabilités apparaissent comme un outil d'analyse à part entière.

Sur le plan algorithmique, on peut également faire appel à des approches stochastiques. Le phénomène que l'on souhaite simuler numériquement peut comporter des entrées aléatoires dues à la physique même du problème. Mais l'aléatoire apparaît également comme un outil numérique (méthodes de Monte Carlo, gradient stochastique, recuit simulé, algorithmes génétiques etc.).

Le projet cible deux thèmes prioritaires. Le premier dans le cadre des milieux hétérogènes — modélisés comme des milieux aléatoires – on s'intéresse à l'analyse de phénomènes d'homogénéisation, ainsi qu'à l'analyse et au calcul de coefficients effectifs.

La candidature de Marseille dans le programme national «Génopole» a été retenue. SYSDYS a été l'un des groupes de recherche à l'initiative de cette candidature. Le projet s'implique plus particulièrement dans l'utilisation des modèles markoviens pour l'identification de paramètres (comme les paramètres de déscendance, de mutation etc.).

3 Fondements scientifiques

3.1 Analyse stochastique

Participants : François Delarue, Antoine Lejay, Etienne Pardoux.

Mots clés : équations différentielles stochastiques, équations différentielles stochastiques rétrogrades, équations différentielles stochastiques progressives-rétrogrades.

L'étude des milieux aléatoires et de leurs applications ne peut se faire sans l'aide de l'analyse stochastique. L'environnement scientifique du pôle marseillais nous fournit l'assise nécessaire, notamment au travers de l'équipe de Probabilités du LATP.

Dans ce domaine, le projet étudie plus particulièrement les EDSR (équations différentielles stochastiques rétrogrades) et les EDPS (équations aux dérivées partielles stochastiques). Les EDSR, introduites par Étienne Pardoux et Peng Shige [1], ont engendré un mouvement de recherche important et proposent de nouvelles modélisations dans différentes applications (comme les mathématiques financières). On applique ces travaux pour établir des résultats d'homogénéisation pour des EDP semi ou quasi–linéaires.

3.2 Homogénéisation et milieux aléatoires

Participants : Fabien Campillo, François Delarue, Antoine Lejay, Etienne Pardoux,

Elisabeth Remy.

Mots clés : milieux aléatoires, homogénéisation.

Il s'agit d'étudier les propriétés mécaniques ou physiques des milieux hétérogènes en utilisant des méthodes probabilistes. Ceci conduit à l'étude qualitative et quantitative des solutions d'équations aux dérivées partielles, ou équations aux différences, à coefficients aléatoires. Toute une théorie de ces équations s'est développée au cours des dernières années.

Un premier axe de recherche concerne la théorie de l'homogénéisation des solutions d'EDP dont les coefficients sont des champs aléatoires. On développe des méthodes d'analyse s'appuyant, soit sur des approches « analytiques », soit sur des approches « probabilistes ». Une des applications importantes est le calcul des coefficients dits effectifs, i.e. décrivant le modèle à une échelle macroscopique.

4 Domaines d'applications

4.1 Méthode de Monte Carlo pour les milieux aléatoires

Participants : Fabien Campillo, Antoine Lejay.

Le sous-sol est constitué de matériaux ayant la propriété d'emmagasiner (réservoirs), de laisser s'écouler et de restituer l'eau souterraine. Le débit souterrain, régi par la loi de Darcy, est fonction de la perméabilité.

Il existe deux types de roches réservoirs:

- les pores sont des vides de petites dimensions existant entre les grains de formes et de grosseurs variables,
- les fissures sont des fentes allongées (à une échelle inférieure, les micro–fissures ont les mêmes caractéristiques que les pores).

Un réservoir est homogène lorsque ses caractéristiques physiques sont constantes dans la direction de l'écoulement des eaux souterraines. Or, cette condition n'est pas réaliste : le sous-sol est de nature hétérogène et présente de nombreuses fissures de différentes natures.

À ce niveau, les outils de modélisation/analyse/simulation par milieu aléatoire que se propose d'étudier SYSDYS prennent toute leur dimension.

Le problème d'homogénéisation provient du fait que les lois de l'hydrodynamique souterraine ne s'appliquent qu'aux seuls milieux homogènes. Ainsi le problème est d'identifier un volume de réservoir considéré comme homogène dans son ensemble.

4.2 Génomique

Participants : Fabien Campillo, Lorie Dudoignon, Enrico Formenti, Bruno Torrésani.

En terme de méthodologie, il s'agit, en quelque sorte, d'un retour à « l'identification de systèmes probabilistes avec observations bruitées» (un des thèmes de l'ancien projet Mefisto) mais avec une problématique et une approche sensiblement différente.

Il y a une convergence entre généticiens (les seuls à pouvoir dire si tel ou tel thème est pertinent, si tel ou tel résultat est significatif) et probabilistes/statisticiens qui peuvent proposer une méthodologie dans la modélisation et l'algorithmique (en plus des apports des informaticiens et des physiciens).

Il a été clair que les probabilistes (et autres mathématiciens, informaticiens et physiciens) ne doivent pas servir de simple « boite à outils» pour les généticiens, mais doivent réellement apporter quelque chose de nouveau. Ceci implique un investissement intellectuel minimal des mathématiciens/informaticiens en génétique.

5 Résultats nouveaux

5.1 Analyse stochastique

5.1.1 Représentation probabiliste des équations quasi–linéaires avec un opérateur sous forme divergence

Mots clés : Équations différentielles stochastiques rétrogrades (EDSR), équations aux dérivées partielles (EDP) quasi–linéaires.

Participant : Antoine Lejay.

On s'intéresse à une représentation probabiliste des solutions d'EDP quasi–linéaires avec un opérateur sous forme divergence, dans un cadre où la représentation par les équations différentielles stochastiques progressives rétrogrades ne tient plus.

Lorsque L est un opérateur différentiel du second-ordre, la solution u d'une EDP semi–linéaire de la forme

$$\partial u(t,x) + Lu(t,x) + h(t,x,u(t,x),\nabla u(t,x)) = 0 \text{ et } u(T,x) = g(x) \tag{1}$$

peut être représentée à l'aide de l'EDSR

$$Y_t = g(X_T) + \int_t^T h(s,X_s,Y_s,Z_s)\mathrm{d}s - \int_t^T Z_s \mathrm{d}M_s, \tag{2}$$

où X est le processus stochastique engendré par L et M sa partie martingale, par la relation

$$Y_0 = u(s,x) \text{ si } X_s = x. \tag{3}$$

Dans le cas d'une équation quasi–linéaire du type

$$\partial_t u(t,x) + \partial_{x_i}(a_{i,j}(t,x,u(t,x),\nabla u(t,x))\partial x_j) + h(t,x,u(t,x),u(t,x)) = 0, \tag{4}$$

si a est dérivable alors la solution est donnée par la solution d'une équation différentielle progressive rétrograde (EDSPR), c'est-à-dire que l'équation différentielle stochastique (EDS) donnant le processus X fait elle-même intervenir les processus Y et Z.

Lorsque a n'est pas régulière et ne dépend pas de la solution, alors le processus X de générateur $\partial_{x_i}(a_{i,j}\partial x_j)$ n'est pas en général solution d'une EDS. Ainsi, dans le cas d'une

équation quasi–linéaire avec un coefficient non régulier, la représentation par les EDSPR n'est plus valable.

En utilisant les résultats de [17], il est toutefois possible de donner une certaine représentation, unique en un certain sens, de la solution u à l'aide des EDSR. Mais ces résultats requièrent *a priori* l'existence et l'unicité de la solutions des EDP.

5.1.2 Interprétation probabiliste d'une classe d'EDP semi–linéaires

Mots clés : Processus de Dirichlet, opérateur sous forme divergence.

Participant : Etienne Pardoux.

Il s'agit d'une collaboration avec Vlad Bally (Université Paris 6) et Lucreţiu Stoica (Institut de Mathématiques, Bucarest).

V. Bally, Paris 6 et L. Stoica, Bucharest

On donne une interprétation probabiliste d'une classe d'EDP semi–linéaires, où l'opérateur du second ordre est sous forme divergence, avec des coefficients peu réguliers. Le versant probabiliste de ce résultat combine les théories des EDSR et des processus de Dirichlet.

Le but de ce travail est de donner un interprétation probabiliste de l'EDP semi–linéaire

$$\frac{\partial u}{\partial t}(t,x) = Au(t,x) + f(t,x,u(t,x),\nabla u(t,x)),$$

avec une condition initiale en $t = 0$. Ici A est un opérateur elliptique du second ordre sous forme divergence, avec des coefficients peu réguliers. La formule probabiliste pour la solution u fait intervenir une « équation différentielle stochastique rétrograde », dont le coefficient f et la condition finale sont fonction d'un processus markovien de Dirichlet, de générateur infinitésimal A. Ce résultat généralise, au cas où le terme non linéaire dépend de ∇u, un résultat obtenu par A. Lejay.

5.2 Homogénéisation et milieux aléatoires

5.2.1 Homogénéisation d'EDP paraboliques non linéaires

Mots clés : Homogénéisation, EDP stochastique.

Participant : Etienne Pardoux.

On homogénéise des EDP paraboliques non linéaires à coefficients périodiques, qui sont perturbés par une diffusion ergodique, et contiennent un terme fortement oscillant. L'équation limite est une EDPS à coefficients constants.

Il s'agit d'une collaboration avec Andrey Piatnitski (Lebedev Physical Institute, Moscou).

On homogénéise l'EDP parabolique non linéaire :

$$\frac{\partial u^\epsilon}{\partial t}(t,x) = \frac{\partial}{\partial x_i} a_{ij}(\frac{x}{\epsilon}, \xi_{t/\epsilon^2}) \frac{\partial u^\epsilon}{\partial x_j}(t,x) + \frac{1}{\epsilon} g(\frac{x}{\epsilon}, \xi_{t/\epsilon^2}, u^\epsilon(t,x)),$$
$$(t,x) \in (0,T) \times \mathbf{R}^n \quad u^\epsilon(0,x) = u_0(x).$$

Tous les coefficients sont fonction périodique de leur première variable, et ξ est une diffusion ergodique à valeurs dans $\mathbf{R}^d$. On peut considérer que cette équation généralise l'EDP linéaire étudiée récemment par Campillo, Kleptsina, Piatnitski (Rapport INRIA 3520), et aussi à la fois les résultats de Bouc–Pardoux des années 80, et celui récent de Pardoux (JFA 1999). On est amené à utiliser des correcteurs «non locaux» fonction de la trajectoire future du processus ξ. L'unicité de la loi limite (solution «faible» d'une EDPS) reste à ce jour un problème ouvert.

5.2.2 Homogénéisation d'un opérateur parabolique aléatoire

Mots clés : opérateur aléatoire, homogénisation, moyennisation.

Participant : Fabien Campillo.

Il s'agit d'une collaboration avec Andrey Piatnitski (Lebedev Physical Institute, Moscou) et Marina Klepstina (Institute for Information Transmission Problems, Moscou)

On considère le comportement asymptotique de la solution de l'équation de Cauchy suivante lorsque $\epsilon \downarrow 0$:

$$\frac{\partial}{\partial t}u^\epsilon(t,x) = \operatorname{div}\left[a\left(x, \tfrac{x}{\epsilon}, \xi_{\frac{t}{\epsilon^\alpha}}\right)\,\nabla u^\epsilon(t,x)\right] + \tfrac{1}{\epsilon^{1\wedge\frac{\alpha}{2}}}\,c\left(x, \tfrac{x}{\epsilon}, \xi_{\frac{t}{\epsilon^\alpha}}\right)\,u^\epsilon(t,x)\,, \tag{5}$$

avec $u^\epsilon(0,x) = u_0(x)$ où $x \in \mathbf{R}^n$, $t \in [0,T]$, α est un paramètre positif, $T > 0$ est fixe et $u_0 \in L^2(\mathbf{R}^n)$.

Les coefficients $a(x,z,y)$ et $c(x,z,y)$ sont périodiques en z (ou de façon équivalente z appartient au tore unité $\mathbf{T}^n = \mathbf{R}^n/\mathbf{Z}^n$) et $\{\xi_t\}_{t\geq 0}$ est un processus de diffusion stationnaire, ergodique à valeurs dans $\mathbf{R}^d$, défini par $d\xi_t = B(\xi_t)\,dt + \sigma(\xi_t)\,dW_t$. On suppose que la mesure invariante de cette diffusion admet une densité $\rho(\cdot)$.

Pour $\alpha \leq 2$, on obtient une convergence faible de la loi de la solution $\{u^\epsilon(t)\}_{0\leq t\leq T}$ vers la solution d'un problème de martingale (la solution faible d'une EDPS), alors que dans le cas $\alpha > 2$, la loi limite est de type Dirac centrée sur la solution du problème de Cauchy limite associé à (5). Dans tous les cas on obtient l'existence et l'unicité de la loi limite.

Dans une précédente étude[CKP] nous avions considéré le cas où les coefficients ne dépendaient que des variables z et y. Dans la cas présent l'analyse (la détermination des correcteurs) a été beaucoup plus fine).

5.2.3 Inégalité isopérimétrique locale sur l'amas de percolation infini

Mots clés : inégalité isopérimétrique, amas infini de percolation, marche aléatoire, inégalité de Cheeger, inégalité de Nash, tempos de convergence vers l'équilibre.

Participant : Elisabeth Remy.

Nous trouvons le comportement asymptotique de la constante isopérimétrique de la partie de l'amas de percolation infini contenu dans la boite $[-n, n]^2$. Ce résultat nous permet, par exemple, d'estimer le

[CKP] F. Campillo, M. Klepstina, A. Piatnitski, «Homogenization of random parabolic operator with large potential», *Stochastic Processes and their Applications*, à paraître.

temps de convergence vers l'équilibre d'une marche aléatoire évoluant sur cette partie de l'amas infini, et d'établir une inégalité de Nash ainsi que l'inégalité de Cheeger pour le trou spectral du semi–groupe associé à cette marche.

Il s'agit d'une collaboration avec Pierre Mathieu (Université de Provence)

On considère un modèle de percolation de site en dimension 2, en régime sur-critique (i.e. $p > p_c$, où p_c est la probabilité critique).

Soit $\mathcal{C}^n$ la partie connexe de l'amas infini qui contient 0, contenue dans la boite $[-n,\, n]^2$. On définit la constante isopérimétrique de dimension ϵ par :

$$I_\epsilon = \inf_{A \subset \mathcal{C}^n} \left\{ \frac{\pi(\partial A)}{\pi(A)^{\frac{\epsilon-1}{\epsilon}}} \,;\, \pi(A) \leq \frac{1}{2}\,,\, A \text{ connexe} \right\}$$

où π est la mesure uniforme sur $\mathcal{C}^n$ et ∂A le bord de l'ensemble A. Pour tout $\epsilon > 2$, on montre que $I_\epsilon = \beta(p,\, \epsilon)\, \frac{1}{n}$, où la constante β ne dépend pas de la réalisation du milieu ω. Pour montrer ce résultat, nous nous appuyons sur un résultat de Kesten[Kes82, Théorème 11.1] qui nous donne des indications sur la structure géométrique de l'amas : il existe une « grille » contenue dans $\mathcal{C}^n$.

Considérons maintenant une marche aléatoire évoluant sur $\mathcal{C}^n$. En utilisant les résultats énoncés par Saloff-Coste[SC96], on peut alors estimer le trou spectral de P_t, le semi–groupe de la marche (inégalité de Cheeger), établir une inégalité de Nash et ainsi estimer le temps de convergence vers l'équilibre de la marche :

$$\forall \epsilon > 2,\ \forall t > 0\,, \qquad \sup_{x,y \in \mathcal{C}^n} P_t(x,\, y) \leq \frac{C}{t^{\epsilon/2}} n^{\epsilon/2}\,.$$

5.2.4 Méthodes de Monte Carlo en milieu fissuré

Mots clés : milieu fissuré, modèle double porosité, méthode de Monte Carlo.

Participants : Fabien Campillo, Antoine Lejay.

Nous proposons un algorithme de simulation d'une diffusion dans un milieu poreux fissuré aléatoire borné de $\mathbf{R}^2$ (la *matrice*). Cet algorithme donne le temps et la position d'atteinte du réseau de fissure (ici un ensemble de segments de droites) ou du bord par une particule brownienne. Une des applications est de calculer le coefficient d'échange dans un modèle de double porosité. Il n'est pas nécessaire de connaître le comportement de cette particule dans la fissure si on suppose que le fluide dans le milieu fissuré est capté par les fissures et que les fissures sont fines. Pour cela nous utilisons une formule donnant le coefficient d'échange en fonction de la moyenne du premier temps de sortie de la matrice par des particules initialement uniformément réparties dans la matrice.

Nous étudions actuellement le comportement de la particule dans le réseau de fissures qui nous amène à proposer également un algorithme de Monte Carlo.

[Kes82] H. Kesten, *Percolation Theory for Mathematicians*, *Progress in Probability*, *2*, Birkhäuser, Boston, 1982.

[SC96] L. Saloff-Coste, «Lectures on finite markov chains», *in : Ecole d'été de probabilité de Saint-Flour XXVI*, P. Bernard (éditeur), *Lecture Notes in Mathematics*, *1665*, Springer, p. 301–413, 1996.

Lorsque l'on utilise un algorithme de Monte Carlo pour calculer l'espérance de certaines fonctionnelles de processus de diffusion données à des temps d'arrêt, on a tendance à simuler le processus de diffusion (des trajectoires) au « sens faible ». Dans notre cas, on a établi une formule explicite de la loi de cette fonctionnelle et c'est selon cette loi que l'on produit les tirages aléatoires nécessaires à la méthode de Monte Carlo. On évite ainsi un détour qui nuit à la qualité de l'approximation.

Ce travail permet de calculer des approximations par méthode de Monte Carlo du problème de Cauchy dans $\mathbf{R}^2$ sur des domaines bornés polygonaux : dans ce cas on obtient une expression analytique (et utilisable en pratique) de la loi de la variable aléatoire utilisée dans l'algorithme de Monte Carlo.

Ce travail est à rapprocher de celui de Milstein-Tretyakov[MT99] pour la simulation d'équations différentielles stochastiques à coefficients non nécessairement constants. Mais du fait qu'ils considèrent des domaines bornés quelconques ils n'ont pas exactement le même algorithme et, en pratique, doivent faire face au problème du bord : quand doit-on considérer que l'on a atteint le bord. Avec des domaines polygonaux et un algorithme légèrement différent on peut répondre précisément à cette question.

5.3 Identification et filtrage

5.3.1 Filtrage non linéaire

Mots clés : Filtrage non linéaire, grandes déviations.

Participant : Etienne Pardoux.

Collaborateur extérieur O. Zeitouni (Haifa).

On étudie les grandes déviations en filtrage non linéaire unidimensionnel, avec petit bruit d'observation.

On a repris les travaux de l'année antérieure. Une nouvelle approche, qui consiste à étudier globalement tous les termes ensemble, et à faire un changement adéquat de probabilité, donne un principe de grandes déviations avec une fonctionnelle qui cette fois a un sens intuitif.

5.3.2 Calcul stochastique fractionnaire et applications statistiques

Mots clés : Mouvement brownien fractionnaire, statistique des processus stochastiques, filtrage optimal, formule de Girsanov, processus d'innovation, filtrage optimal, maximum de vraisemblance.

Participant : Marie–Christine Roubaud.

Cette étude est effectuée en collaboration avec M.L. Kleptsyna (Université de Communication de Moscou) et A. Le Breton (LMC–Université Joseph Fourier).

Depuis ces dernières années un engouement croissant est suscité par la modélisation stochastique des phénomènes à *longue mémoire*. Ceci est justifié notamment par le fait que les

[MT99] G. Milstein, M. Tretyakova, « Simulation of a space–time bounded diffusion », *The Annals of Applied Probability 9*, 3, 1999, p. 732–779.

phénomènes naturels ou économiques présentant des comportements dynamiques de ce type sont fréquents. Cependant, pour ces modèles, beaucoup de problèmes restent ouverts aussi bien du point de vue théorique que numérique.

Le mouvement brownien fractionnaire (mBf), qui est une généralisation du mouvement brownien usuel, est un des processus les plus simples pour traduire la longue dépendance. En particulier, il apparaît très intéressant d'étendre la théorie classique des systèmes stochastiques gouvernés par des mouvements browniens usuels, aux systèmes gouvernés par des mouvements browniens fractionnaires. Plusieurs travaux ont déjà été faits dans ce sens pour des modèles plus ou moins spécifiques (voir par exemple les travaux de Gripenberg et Norros[GN96], Le Breton[Bre98a] [Bre98b], Kleptsyna *et al.*[KKA98], Coutin et Decreusefond[CD99]).

C'est dans cette problématique que s'inscrit notre étude. La principale difficulté de ce travail réside dans le fait que le mBf n'est pas une semi–martingale et donc les résultats classiques d'intégration stochastique pour les semi–martingales ne peuvent s'appliquer tout au moins directement. En utilisant des outils assez sophistiqués tel que le calcul des variations stochastiques, une analyse stochastique du mBf a été développée par Decreusefond et Üstünel, mais il est difficile pour les non spécialistes d'en extraire des outils.

L'approche que nous avons choisie est élémentaire. Nous considérons uniquement l'intégration des fonctions déterministes par rapport à un mBf ; ceci étant suffisant pour un grand nombre d'applications intéressantes. Nous avons complété l'approche élémentaire proposée par Norros *et al.* [NVV99] dont l'idée de base est d'associer au mBf une martingale appropriée appelée *martingale fondamentale* par une transformation intégrale simple. L'analogue de la formule de Girsanov et des théorèmes de représentation ont été établis.

Ces résultats ont été appliqués principalement à des problèmes de filtrage linéaire et non linéaire mais aussi à un problème statistique de base tel que l'estimation de paramètre. La méthode du maximum de vraisemblance a été utilisée pour résoudre un problème d'estimation de paramètre dans un modèle de régression linéaire avec un bruit brownien fractionnaire.

Dans le cadre du filtrage optimal de systèmes linéaires et non linéaires les bruits de dynamique et d'observation sont représentés par des mouvements browniens fractionnaires. Dans le cas d'un système linéaire gaussien les équations du filtre ont été établies lorsque les paramètres de Hurst des mBfs appartiennent à $(0,1)$.

Dans le cas non linéaire, la solution a été obtenue lorsque le signal est une variable aléatoire et dans le cas dit *semi–linéaire* où la non linéarité se situe uniquement dans l'équation de

[GN96] G. Gripenberg, I. Norros, «On the prediction of fractional Brownian motion», *Journal of Applied Probability 33*, 2, 1996, p. 400–410.

[Bre98a] A. L. Breton, «Filtering and parameter estimation in a simple linear model driven by a fractional Brownian motion», *Statistics and Probability Letters 38*, 3, 1998, p. 263–274.

[Bre98b] A. L. Breton, « Une approche de type Girsanov pour le filtrage dans un système linéaire simple avec bruit brownien fractionnaire », *Comptes Rendus à l'Académie des Sciences 326*, Série I, 1998, p. 997–1002.

[KKA98] M. Kleptsyna, P. Kloeden, V. Ahn, « Linear filtering with fractional Brownian motion », *Stochastic Analysis and its Applications 16*, 16, 1998, p. 907–914.

[CD99] L. Coutin, L. Decreusefond, « Abstract nonlinear filtering theory in the presence of fractional Brownian motion», *The Annals of Applied Probability 9*, 4, 1999, p. 1058–1090.

[NVV99] I. Norros, E. Valkeila, J. Virtamo, « An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions», *Bernoulli 5*, 4, 1999, p. 571–587.

dynamique du signal qui est une diffusion.

Cette étude a fait l'objet de trois articles acceptés dans des revues internationales [6, 7, 8].

5.3.3 Filtres approchés pour des systèmes avec bruits colorés

Mots clés : Filtrage non linéaire, bruits colorés, filtres approchés, efficacité asymptotique.

Participant : Marie–Christine Roubaud.

Collaboration avec A. Le Breton (LMC-Université Joseph Fourier).

Dans cette étude nous avons proposé des filtres approchés pour des systèmes stochastiques semi–linéaires et non linéaires avec des bruits colorés. Ces filtres sont définis comme étant identiques aux filtres optimaux lorsque les bruits sont blancs. Nous avons étudié leur comportement en temps long et montré leur efficacité asymptotique dans deux situations sous des hypothèses raisonnables. Dans le cas d'un système où la dynamique du signal et celle de l'observation sont linéaires, le filtre approché considéré est un filtre de Kalman et dans l'étude asymptotique une représentation du filtre optimal est donné. Cette formule étend à des systèmes avec bruits colorés celle connue pour des systèmes avec bruits blancs et condition initiale non gaussienne. Dans le cas d'un système non linéaire sous des hypothèses d'ergodicité, le filtre approché proposé correspond au filtre optimal du système initialisé avec une distribution erronée. Pour l'étude asymptotique nous avons utilisé les résultats d'Ocone et Pardoux sur la stabilité asymptotique du filtre optimal par rapport à la condition initiale[OP96].

Cette étude a fait l'objet d'un article à paraître [8].

5.4 Génomique

5.4.1 Phylogénie moléculaire

Participants : Fabien Campillo, Lorie Dudoignon.

Mots clés : processus de Markov, modèle de Markov caché, protéine, structure secondaire des protéines.

En phylogénie, il s'agit d'étudier comment le vivant a évolué pour aboutir aux espèces que nous connaissons aujourd'hui. Sous l'hypothèse d'évolution divergente à partir d'un ancêtre commun, on peut représenter l'histoire évolutive sous forme d'arbre phylogénétique. Le problème que nous nous posons est comment reconstruire cet arbre, en utilisant comme information ce que nous connaissons du monde vivant actuel. Pour cela, il existe plusieurs méthodes. Celle qui a retenu notre attention est la méthode du maximum de vraisemblance[Fel81]. On distingue en phylogénie deux sortes d'approches. La première, dite classique, utilise comme

[OP96] D. Ocone, E. Pardoux, « Asymptotic stability of the optimal filter with respect to its initial condition », *SIAM Journal on Control and Optimization 34*, 1, 1996, p. 226–243.

[Fel81] J. Felsenstein, « Evolutionnary trees from DNA sequences: a maximum likelihood approach », *Journal of Molecular Evolution 17*, 1981, p. 368–376.

données des caractéristiques morphologiques, physiologiques, comportementales... La seconde, dite moléculaire, utilise des séquences de gènes (ADN ou protéines).

Nous nous plaçons dans la seconde approche. Les données que nous utilisons sont des séquences de protéines.

Nous nous intéressons donc, à la modélisation du processus d'évolution des séquences protéiques. Cette modélisation peut être séparée en deux parties. La première consiste à décrire l'évolution au niveau d'un site, i.e d'une position, dans la séquence. Pour cela, on utilise des processus de Markov. La deuxième consiste à représenter l'évolution sur l'ensemble de la séquence. Cet aspect est souvent occulté, en considérant, que tous les sites d'une même séquence suivent le même processus d'évolution et ce, de manière indépendante.

Les modèles que nous avons étudiés considèrent que tous les sites d'une même séquence ne suivent pas le même processus d'évolution, et que des sites voisins dans la séquence n'évoluent pas de manière indépendante et ce, via des modèles de Markov cachés.

Ces modèles de Markov cachés permettent de tenir compte de la structure des protéines, chaque état caché correspondant à une catégorie de structure secondaire. L'intérêt de ces modèles réside dans le fait que la fonction d'une protéine est en étroite relation avec la structure de la molécule, structure qui est beaucoup plus conservée que sa séquence[TGJ96] [GTJ98]. Nous nous sommes ensuite penchés sur la comparaison des modèles, afin de voir si le fait de tenir compte de la structure secondaire pouvait améliorer, ou non, l'adéquation aux données.

5.4.2 Analyse de séquences

Mots clés : Protéine, phylogénie, chaîne de Markov, alignement multiple.

Participant : Bruno Torrésani.

Partant d'un alignement multiple de séquences de protéines, on s'intéresse au comportement de familles de matrices de transition obtenues en comparant les séquences deux à deux. L'objectif est de tester la qualité de la description donnée de l'alignement multiple par un modèle Markovien sur un arbre, et de cibler et décrire les écarts à de tels modèles.

Il s'agit d'une collaboration avec : A. Grossmann, C. Devauchelle, A. Hénaut, J.L. Risler (Génome et Informatique, Versailles), M. Holschneider (Géosciences Rennes), M. Monnerot (CGM Gif sur Yvette).

La comparaison de séquences génétiques est souvent effectuée à partir de séquences « alignées » (c'est à dire dans lesquelles on a mis face à face les états — les nucléotides dans le cas des ADN et ARN, et les acides aminés dans le cas des protéines — censés se correspondre). Nous avons développé une méthode de comparaison systématique de séquences protéiques dans un contexte d'évolution. En supposant les sites indépendants et identiquement distribués (hypothèses probablement simplistes mais classiques, et relativement « honnêtes » si le jeu de données étudiées est correctement choisi), on fait généralement l'hypothèse que les mutations ponctuelles sont gouvernées par une évolution Markovienne réversible à temps continu. Le

[TGJ96] J. L. Thorne, N. Goldman, D. T. Jones, «Combining protein evolution and secondary structure», *Molecular Biology and Evolution 13*, 5, 1996, p. 666–673.

[GTJ98] N. Goldman, J. L. Thorne, D. T. Jones, « Assessing the impact of secondary structure and solvent accessibility on protein evolution», *Genetics 149*, 1998, p. 445–458.

problème est généralement d'estimer les paramètres du modèle, et de tester l'adéquation de ce dernier aux données considérées

Partant d'un alignement multiple, on considère tous les sous-alignements deux à deux de séquences satisfaisant une condition (de nature algébrique) de proximité. A chacun de ces alignements deux à deux est associé une matrice de transition, dont on considère le logarithme (dont l'existence est assurée par la condition de proximité). Le point central est la comparaison des logarithmes de toutes les matrices ainsi obtenues, et leur confrontation à la prédiction du modèle. Cette comparaison est effectuée en utilisant des techniques classiques d'algèbre linéaire et de traitement des données. Dans le cas où le modèle donne une description acceptable des données, on obtient alors des estimations des « âges » des alignements, c'est à dire du temps écoulé depuis que deux séquences considérées ont divergé. Les « âges » ainsi obtenus sont généralement très proches de distances d'arbres (c'est à dire, satisfont de façon approximative la condition de Bünemann), de sorte qu'ils permettent la construction d'arbres phylogénétiques.

Les techniques développées ont été testées sur plusieurs familles de séquences (globines, génome mitochondrial...). Dans le cas des génomes mitochondriaux, elles permettent en particulier de montrer que différentes familles d'espèces (par exemple, vertébrés et mollusques) ne peuvent pas être décrites par le même modèle évolutif.

5.4.3 Analyse du transcriptome

Mots clés : Puce à ADN, analyse multifactorielle.

Participants : Bruno Torrésani, Marie–Christine Roubaud.

Les techniques basées sur des « puces à ADN » permettent une mesure indirecte de l'expression de gènes sélectionnés, dans des conditions données. L'analyse statistique des résultats expérimentaux reste toutefois très spéculative compte tenu de la grande variabilité des techniques existantes, des conditions expérimentales et des familles de gènes explorées. On s'intéresse à l'analyse de données obtenues sur des supports de type « membrane nylon » avec des marqueurs radioactifs. L'utilisation de techniques classiques d'analyse de données (analyse multifactorielle, classification) permet entre autres d'obtenir des renseignements sur les covariations de certaines familles de gènes. On espère ainsi obtenir de façon indirecte des informations sur les interactions entre gènes et la régulation.

Il s'agit d'une collaboration avec : P. Chiappetta (CPT Marseille), R. Houlgatte (CIML Marseille).

Les techniques d'analyse du transcriptome permettent de mesurer indirectement l'expression des gènes, c'est à dire très grossièrement l'efficacité de la transcription `gène` → `protéine` dans des conditions données. On fonde de grands espoirs sur l'analyse des résultats expérimentaux obtenus sur les puces à ADN, qui pourraient permettre d'inférer les processus d'interactions entre gènes.

Nous travaillons actuellement sur des jeux de données fournies par le Centre d'Immunologie de Marseille, portant sur l'expression d'environ 160 gènes répertoriés, dans des conditions données (cellules cancéreuses, avec des pathologies comparables). Les données sont des données « statiques », de sorte qu'aucune étude dynamique n'est a priori envisageable.

Les premiers résultats permettent de mettre en évidence des co-variations significatives de certaines familles de gènes. On observe en particulier des sous-familles de gènes dont les niveaux d'expression sont liés par une relation manifestement linéaire (ce qui est mis en évidence par des techniques standard de corrélation). D'autres sous familles sont en revanche significativement

liées par des relations manifestement non-linéaires, qui ont dans un (petit) nombre de cas été mises en relation avec des processus biologiques.

Les familles de gènes mises en évidence par ces approches semblent être confirmées par l'utilisation de méthodes de classification (arbres et amas) ; ces résultats sont toutefois très préliminaires.

5.4.4 Complexité de Kolmogorov

Participant : Enrico Formenti.

Collaborateurs extérieurs : Bruno Durand (LMI/Université de Provence) et Brigitte Mossé (IML/Université de Provence).

La classification fonctionnelle du code génétique est la continuation naturelle de l'énorme effort de séquençage du génome humain qui vient de s'achever. Pour cela, les chercheurs ont introduit plusieurs méthodes afin d'automatiser le travail des annotateurs (chaînes de Markov, réseaux de neurones etc). Mais leurs faibles performances laissent encore beaucoup de marge à de meilleures solutions.

En collaboration avec B. Durand, je cherche à donner à ce domaine une vision algorithmique du problème à l'aide de la complexité de Kolmogorov.

En effet, cette dernière est un outil très puissant pour l'analyse de suites de mots infinis et formalise exactement la notion de suite aléatoire. Il est clair qu'elle se prête bien à améliorer les algorithmes probabilistes existants, car elle permet de mieux prendre en compte, lors de la reconnaissance de portions d'ADN codantes, tous les critères algorithmiques possibles.

Un stage a permis de mettre « en pratique » ces idées en réalisant un programme de classification des parties codantes qui utilise une approche mixte « chaînes de Markov-complexité de Kolmogorov ». Les premiers résultats sont encourageants et permettent d'avoir un plus petit nombre d'états cachés par rapport aux approches classiques utilisant uniquement les chaînes de Markov (donc un gain en temps de calcul). De plus, les taux de succès sont parfaitement comparables à ceux des programmes utilisés couramment.

Le fait que, pour l'instant seulement un petit nombre de critères algorithmiques a été pris en compte, laisse espérer de meilleurs résultats.

Un aspect de la bio-informatique auquel je me suis intéressé est l'évolution. Les organismes évoluent lentement, génération après génération. Leur évolution laisse des traces dans leur code génétique. En collaboration avec Brigitte Mossé, mes recherches essayent de formaliser ce processus à l'aide des systèmes dynamiques symboliques. Dans ce cadre les questions qui viennent à l'esprit sont :

- existe-t-il un ensemble attracteur ? Si oui, est il commun à plusieurs individus ou espèces différents ?
- Y a-t-il des macro-structures qui gouvernent la dynamique de l'évolution ?

La première étape consiste en la définition d'un espace topologique bien adapté dans lequel les mutations sont vues comme des systèmes dynamiques.

Une partie des propriétés topologiques de cet espace a été étudiée par G. Varouchas lors d'un stage de MIM1 (Magistère d'Informatique et Modélisation) de l'ENS Lyon que j'ai encadré.

A présent, nous sommes en train d'étudier les propriétés des mutations, mais nous ne sommes pas encore en mesure d'en tirer des résultats généraux.

6 Actions régionales, nationales et internationales

6.1 Actions nationales

6.1.1 Groupe de travail info-bio-math-phy

Groupe de travail organisé par B. Torrésani. Séminaire bi-hebdomadaire, centré sur la génomique et la bio-informatique, organisé à Luminy.

6.1.2 Marseille-Génopole

Bruno Torrésani est coordinateur du groupe « bio-informatique ».

6.1.3 Bio–informatique : CNRS–INRA–INRIA–INSERM

SYSDYS a proposé un projet qui a été retenu dans le cadre de l'appel d'offre conjointe CNRS–INRA–INRIA–INSERM en bio–informatique (appel d'offre 2000). Le thème de ce contrat est exposé à la Section 5.4.3.

6.2 Actions internationales

6.2.1 École du CIMPA

Participants : Etienne Pardoux, Antoine Lejay, François Delarue.

On a organisé une École du CIMPA sur les liens entre équations aux dérivées partielles et processus stochastiques, avec 5 conférenciers : S. Kuznetsov (Boulder, Colorado), S. Méléard (Paris 10 et Paris 6), Y. Ouknine (Marrakech), E. Pardoux et B. Roynette (Nancy, projet Omega), à Marrakech, Maroc.

L'assistance comprenait de fortes délégations d'Algérie, Tunisie, Maroc, Sénégal et de France, avec environ 50 participants.

Antoine Lejay et François Delarue ont assistés Étienne Pardoux.

6.2.2 École d'été « Analyse In Silico de Séquences Génomiques »

Participant : Bruno Torrésani.

Ecole organisée par P. Chiappetta et B. Torrésani du 17 au 28 Juillet 2000, à Marseille-Luminy.

Il s'agissait d'une école pluridisciplinaire visant à apporter à des théoriciens (mathématiciens, physiciens, informaticiens) intéressés par une réorientation en génomique (bio-informatique et bio-mathématique), une formation de base solide en génomique, ainsi que sur les approches classiques de la bio-informatique.

6.2.3 Contrat PICS

Ce contrat s'est poursuivi cette année. Il a permis différents échanges avec des équipes de recherche de Moscou : notamment autour des problèmes d'homogénéisation (voir Section 5.2.1).

6.3 Visites, et invitations de chercheurs

Andrey Piatnitski (Lebedev Physical Institute, Moscou) a été invité par le projet SYSDYS en juillet et décembre. Thème : homogénéisation d'opérateurs aux dérivées partielles à coefficients aléatoires.

Elisabeth Remy (Université de São Paulo) a été invitée par le projet SYSDYS en septembre. Thème : marches aléatoires en milieux aléatoires.

Antoine Lejay (Université de Standford) a été invité par le projet SYSDYS en septembre et décembre. Thème : méthode de Monte Carlo pour les milieux aléatoires.

7 Diffusion de résultats

7.1 Animation de la Communauté scientifique

Marie–Christine Roubaud est membre suppléant de la commission de spécialistes de l'UJF en 26^e section.

7.2 Enseignement

Marie–Christine Roubaud • DEUG 1ère année mention SMb (Sciences de la matière, Physico–Chimie-Biologie): Cours (16h) et Cours–travaux dirigés (40h) de Mathématiques • DEUG 2ème année Mention MIAS (Mathématiques, Informatique et Applications aux sciences) et SMa (Sciences de la matière, physique–chimie): Cours (18h) et travaux dirigés (72h) de Mathématiques.

Bruno Torrésani • Licence de physique (99h) • Maîtrise d'ingénierie mathématiques (65h) • DESS de Probabilités, Statistique et Informatique (36h).

Fabien Campillo • DEA d'Informatique filières info–bio–maths (15h).

8 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] E. Pardoux, S. Peng, « Backward stochastic differential equations and quasi-linear parabolic partial differential equations », *in : Stochastic partial differential equations and their applications*, B. Rzosvskii, R. Sowers (éditeurs), *Lect. Notes in Control & Info. Sci*, *176*, Springer, p. 200–217, Berlin, Heidelberg, New York, 1992.

[2] E. Pardoux, D. Talay., « Discretization and simulation of stochastic differential equations », *Acta Applicandae Mathematicae 3*, 23, 1985.

Thèses et habilitations à diriger des recherches

[3] A. Lejay, *Méthodes probabilistes pour l'homogénéisation des opérateurs sous forme divergence : cas linéaires et semi-linéaires*, thèse de doctorat, Université de Provence, Janvier 2000.

Articles et chapitres de livre

[4] F. Campillo, F. Le Gland, Y. Kutoyants, « Small noise asymptotics of the GLR test for off-line change detection in misspecified diffusion processes », *Stochastics and Stochastic Reports 70*, 2000, p. 109–129.

[5] C. Devauchelle, A. Grossmann, A. Hénaut, M. Holschneider, M. Monnerot, J. Risler, B. Torrésani, « Rate matrices for the analysis of large families of protein sequences », *J. Comp. Biol.*, 2001, à paraître.

[6] M. Kleptsyna, A. Le Breton, M. Roubaud, « General approach to filtering with fractional Brownian noises », *Stochastics and Stochastics Reports*, à paraître.

[7] M. Kleptsyna, A. Le Breton, M. Roubaud, « Parameter estimation and optimal filtering for fractional type stochastic systems », *Statistical Inference for Stochastic Processes*, à paraître.

[8] A. Le Breton, M. Roubaud, « Asymptotic optimality of approximate filters in stochastic systems with colored noises », *siam-co*, à paraître.

[9] A. Lejay, « BSDE driven by Dirichlet process and semi–linear parabolic PDE. Application to homogenization », *Stochastic Processes and their Applications*, 2000, soumis.

[10] E. Pardoux, A. Rascanu, « Backward stochastic variational inequalities », *Stochastics 67*, 1999, p. 159–167.

[11] E. Pardoux, S. Tang, « Forward-Backward stochastic differential equations and quasilinear parabolic PDE's », *Probability Theory and Related Fields 114*, 1999, p. 123–150.

[12] E. Pardoux, « Homogenization of linear and semi-linear second order parabolic PDE's with periodic coefficients: a probabilistic approach », *Journal of Functional Analysis 167*, 1999, p. 498–520.

[13] A. Y. Veretennikov, E. Pardoux, « On the smoothness of an invariant measure of a Markov chain with respect to a parameter », *Dokl. Akad. Nauk 370*, 2, 2000, p. 158–160, en Russe.

Communications à des congrès, colloques, etc.

[14] A. L. Breton, M. Roubaud, « Approximate filters in stochastic systems with colored noises », *in : 14th International Conference on Mathematical Theory of Networks and Systems MTNS2000*, p. 19–23, June 2000.

[15] C. Devauchelle, A. Hénaut, B. Torrésani, M. Holschneider, J. Risler, A. Grossmann, « Une méthode d'exploitation des données d'alignement », *in : Actes de la conférence JOBIM, Montpellier*, G. Caraux, O. Gascuel, M. Sagot (éditeurs), p. 119–128, 2000.

[16] M. Kleptsyna, A. Le Breton, M. Roubaud, «An elementary approach to filtering in systems with fractional Brownian observation noise», *in : Probability Theory and Mathematical Statistics*, B. G. et al. (éditeur), p. 373–392, 1999.

[17] A. Lejay, « Weak solution of semi-linear PDE, BSDE and homogenization», *in : Monte Carlo 2000*, D. Talay (éditeur), July 2000. À paraître dans un numéro spécial de la revue *Monte Carlo Methods and Applications*.

[18] E. Pardoux, A. Piatnitski, « Averaging of nonlinear random parabolic operators », *in : Optimal Control and PDE's - Innovations and Applications. In honor of Alain Bensoussan's 60th anniversary*, J.-L. Menaldi, E. Rofman, A. Sulem (éditeurs), IOS Press, Amsterdam, 2000.

[19] E. Pardoux, « BSDE's, weak convergence and homogenization of semilinear PDE's », *in : Nonlinear analysis, Differential Equations and Control*, F. Clarke, R. Stern (éditeurs), Kluwer Acad. Pub., p. 503–549, 1999.

Rapports de recherche et publications internes

[20] F. Campillo, A. Lejay, « A Monte Carlo method without grid to compute the exchange coefficient in the double porosity model. Part I: From the matrix to the fissures », *rapport de recherche n° RR–4048*, INRIA, Novembre 2000.

Divers

[21] L. Dudoignon, «Phylogénie et structure secondaire de protéines», Mémoire de DEA, Université Rennes I, 2000.

[22] A. Lejay, « A probabilistic representation of the solution of some quasi-linear PDE with a divergence-form operator », soumis.