

# *Projet SYSDYS*

*Systèmes Dynamiques Stochastiques*

*Sophia Antipolis*

THÈME 4B



*R*apport  
*d'Activité*

2001



## Table des matières

<b>1</b>	<b>Composition de l'équipe</b>	<b>2</b>
<b>2</b>	<b>Présentation et objectifs généraux</b>	<b>2</b>
<b>3</b>	<b>Fondements scientifiques</b>	<b>3</b>
3.1	Analyse stochastique . . . . .	3
3.2	Homogénéisation et milieux aléatoires . . . . .	3
3.3	Probabilité, statistique et traitement du signal pour la génomique . . . . .	3
<b>4</b>	<b>Domaines d'applications</b>	<b>4</b>
4.1	Inférence statistique pour l'évolution des protéines . . . . .	4
4.2	Analyse du transcriptome . . . . .	4
<b>5</b>	<b>Résultats nouveaux</b>	<b>5</b>
5.1	Analyse stochastique . . . . .	5
5.1.1	Interprétation probabiliste d'une classe d'EDP semi-linéaires . . . . .	5
5.1.2	Équations différentielles stochastiques rétrogrades avec coefficient localement monotone . . . . .	5
5.2	Homogénéisation et milieux aléatoires . . . . .	6
5.2.1	Équation de Poisson et approximation-diffusion . . . . .	6
5.2.2	Homogénéisation et EDP stochastiques . . . . .	6
5.3	Probabilité, statistique et traitement du signal pour la génomique . . . . .	6
5.3.1	Inférence statistique pour l'évolution des protéines . . . . .	6
5.3.2	Analyse du transcriptome . . . . .	7
<b>6</b>	<b>Actions régionales, nationales et internationales</b>	<b>10</b>
6.1	Actions nationales . . . . .	10
6.1.1	Groupe de travail info-bio-math-phy . . . . .	10
6.1.2	Marseille-Génopole . . . . .	10
6.2	Actions internationales . . . . .	10
6.2.1	Contrat PICS . . . . .	10
6.3	Visites, et invitations de chercheurs . . . . .	10
<b>7</b>	<b>Diffusion de résultats</b>	<b>10</b>
7.1	Animation de la Communauté scientifique . . . . .	10
7.2	Enseignement . . . . .	10
<b>8</b>	<b>Bibliographie</b>	<b>11</b>

*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 Torrèsani

### Personnel Université Joseph Fourier

Marie-Christine Roubaud

### Chercheurs doctorants

François Delarue [bourse AMN, soutenance 7 janvier 2002]

Lorie Dudoignon [bourse Génopole]

## 2 Présentation et objectifs généraux

En modélisation, analyse ou simulation numérique « l'aléa » transparait 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 thème initial du projet SYSDYS, en plus de rentabiliser les compétences du projet Mefisto dans le domaine de l'identification et du filtrage non linéaire numériques (en collaboration avec François LeGland du projet Sigma2), était l'étude des milieux aléatoires et notamment des propriétés d'homogénéisation de tels milieux. Depuis quelques temps un nouvel axe s'est développé au sein du projet. Il concerne les outils probabilistes pour la « modélisation du génome ». Etant très différent du thème initial, il a été décidé de clore le projet SYSDYS afin d'étudier les possibilités d'un nouveau projet sur ce sujet. Concernant le thème initial, les principaux apports du groupe concernent l'étude des propriétés d'homogénéisation des milieux aléatoires, pour des modèles continus aussi bien que discrets. L'approche probabiliste, notamment menée par Etienne Pardoux, a conduit à l'obtention de résultats fins concernant

les équations différentielles stochastiques progressives/rétrogrades (EDSPR). Ces équations, très utilisées dans les modèles financiers et en contrôle stochastique, permettent en effet une interprétation probabiliste des équations aux dérivées partielles quasi-linéaires. Certaines interprétations probabilistes fournissent également des outils de type Monte Carlo très adaptés au calcul des caractéristiques effectives des milieux aléatoires fortement hétérogènes.

### 3 Fondements scientifiques

#### 3.1 Analyse stochastique

**Participants** : François Delarue, 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, Etienne Pardoux.

**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.

#### 3.3 Probabilité, statistique et traitement du signal pour la génomique

**Participants** : Fabien Campillo, Lorie Dudoignon, Marie-Christien Roubaud, Bruno

Torrésani.

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

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

## 4 Domaines d'applications

### 4.1 Inférence statistique pour l'évolution des protéines

**Participants** : Fabien Campillo, Lorie Dudoignon.

**Mots clés** : modèles de Markov, biologie, protéine, évolution des protéines.

Les relations entre la structure d'une protéine et sa fonction sont étroites. La cristallographie permet depuis de nombreuses années d'affiner les connaissances ces structures. Ainsi l'étude de l'évolution des protéines se doit de prendre en compte (et/ou de rendre compte de) ces structures<sup>[Ris96]</sup>. Les approches probabilistes<sup>[TG01]</sup> offrent une certaine souplesse d'analyse.

### 4.2 Analyse du transcriptome

**Participants** : Marie-Christine Roubaud, Bruno Torrésani.

**Mots clés** : biologie, puce à ADN, analyse du transcriptome.

Une application majeure des puces à ADN<sup>[Jor99]</sup> est l'analyse du transcriptome (ensemble des transcrits présents dans un tissu ou dans des conditions données). Les techniques d'hybridation sur puces à ADN permettent une analyse des niveaux d'expression pour des milliers de gènes simultanément et dans des conditions multiples à partir de la mesure des taux d'ARN messagers. Ces méthodes sont utilisées pour répondre à plusieurs objectifs dont les principaux sont la découverte des gènes candidats impliqués dans un processus physiologique ou pathologique spécifique, la compréhension de leurs mécanismes d'interaction et de régulation, et la caractérisation d'un phénomène ou d'un échantillon biologique par son profil d'expression pour un grand nombre de gènes connus.

Récemment ces techniques commencent à être appliquées à l'analyse des cancers humains et à l'étude de l'hétérogénéité moléculaire de certaines maladies comme le cancer du sein. C'est

---

[Ris96] J. L. RISLER, « Evolution des protéines », *Le dictionnaire du darwinisme*, 1996.

[TG01] J. THORNE, N. GOLDMAN, « Probabilistic models for the study of protein evolution », *in: Handbook of Statistical Genetics*, D. Balding, M. Bishop, C. Cannings (éditeurs), John Wiley & Sons, p. 67–82, New York, 2001.

[Jor99] B. JORDAN, « Mesure à grande échelle du niveau d'expression de gènes: techniques et applications », *in: Eurocancer99*, Boiron, Marty (éditeurs), p. 141, 1999.

dans ce cadre que ce situe cette étude fondée sur deux problématiques principales. La première est la comparaison des conditions dans le but de classifier les différentes pathologies en fonction de l'expression des gènes. La seconde est la recherche de covariations de l'expression des gènes afin d'essayer d'en extraire des informations sur leurs interactions.

## 5 Résultats nouveaux

### 5.1 Analyse stochastique

#### 5.1.1 Interprétation probabiliste d'une classe d'EDP semi-linéaires

**Mots clés** : Processus de Dirichlet, opérateur sous forme divergence, équations différentielles stochastiques rétrogrades.

**Participant** : Etienne Pardoux.

*Il s'agit d'une collaboration avec Vlad Bally (Université Paris 6) et Lucreiu Stoica (Institut de Mathématiques, Bucarest), qui fait suite au travail initialisé en 2000.*

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 une interprétation probabiliste de l'EDP semi-linéaire

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

avec une condition initiale en  $t = 0$ . Ici  $A$  est un opérateur elliptique du second ordre sous forme divergence, avec des coefficients peu réguliers, et surtout une matrice de diffusion qui peut dégénérer. Par ailleurs,  $A$  est supposé fermé. La formule probabiliste pour la solution 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  $A$ . Le travail de cette année traite le cas d'une fonction  $f$  satisfaisant une hypothèse de monotonie, et par ailleurs à croissance assez générale en  $u$ .

#### 5.1.2 Équations différentielles stochastiques rétrogrades avec coefficient localement monotone

**Participant** : Etienne Pardoux.

*Il s'agit d'une collaboration avec K. Bahlali, de l'Université de Toulon.*

Il est bien connu que les hypothèses de Lipschitz, ou bien de Lipschitz-Monotonie ne peuvent pas être localisées aussi simplement pour les EDS rétrogrades, que pour les EDS usuelles. Cependant, Khaled Bahlali a récemment montré comment on pouvait localiser l'hypothèse de monotonie, à condition de faire une hypothèse de croissance strictement sous-linéaire du coefficient, et de supposer que la constante de monotonie locale croît comme le logarithme du rayon de la boule.

Le but du travail en commun est d'adapter ce type d'idées au cas localement monotone.

## 5.2 Homogénéisation et milieux aléatoires

### 5.2.1 Équation de Poisson et approximation–diffusion

**Participant** : Etienne Pardoux.

*Il s'agit d'une collaboration avec A. Veretennikov université de Leeds (UK).*

Pardoux et Veretennikov ont poursuivi leur collaboration sur l'étude de l'équation de Poisson dans l'espace euclidien. Suite à deux articles acceptés pour publication aux *Annals of Probability*, les auteurs traitent maintenant le cas d'une diffusion dégénérée, l'hypothèse assurant l'irréductibilité étant assez faible. On donne alors un sens faible à l'équation de Poisson, c'est à dire que,  $\{X_t, t \geq 0\}$  étant une diffusion  $d$ -dimensionnelle possédant une unique probabilité invariante  $\mu$ , et  $f : R^d \rightarrow R$  telle que  $\int f(x)\mu(dx) = 0$ ,  $u(x) = \int_0^\infty E_x f(X_t)dt$  est l'unique solution de l'équation de Poisson  $Lu + f = 0$ , où  $L$  désigne le générateur infinitésimal de la diffusion  $\{X_t\}$ , au sens où  $u(X_t) - u(x) - \int_0^t f(X_s)ds$  est une martingale. Malgré le peu de régularité de la solution  $u$ , ce résultat nous permet d'étudier la limite en loi du processus  $\{Y_t^\varepsilon\}$  lorsque  $\varepsilon \rightarrow 0$ , où  $\{Y_t^\varepsilon\}$  est la solution de l'EDO

$$\frac{dY_t^\varepsilon}{dt} = F(X_{t/\varepsilon^2}, Y_t^\varepsilon) + \varepsilon^{-1}G(X_{t/\varepsilon^2}, Y_t^\varepsilon),$$

sous l'hypothèse  $\int G(x, y)\mu(dx) = 0, \forall y$ . Notons que, contrairement au cas non-dégénéré étudié précédemment, l'équation pour  $Y_t^\varepsilon$  n'est pas une EDS (elle ne contient pas d'intégrale stochastique). Un article est en cours de rédaction.

### 5.2.2 Homogénéisation et EDP stochastiques

**Participant** : Etienne Pardoux.

*Il s'agit d'une collaboration avec A. Piatnitski, Laboratoire Lebedev, Moscou.*

Il s'agit d'une extension du travail de Campillo, Kleptsina, Piatnitski, au cas d'EDP semi-linéaires. On étudie l'homogénéisation d'une EDP à coefficients périodiques, et perturbés par le processus stochastique  $\xi_t$ , de la forme

$$\frac{\partial u^\varepsilon}{\partial t}(t, x) = \frac{\partial}{\partial x_i} a_{ij}\left(\frac{x}{\varepsilon}, \xi_{t/\varepsilon^2}\right) \frac{\partial u^\varepsilon}{\partial x_j}(t, x) + \frac{1}{\varepsilon} g\left(\frac{x}{\varepsilon}, \xi_{t/\varepsilon^2}, u^\varepsilon(t, x)\right).$$

La limite est une EDP stochastique non linéaire, qui possède une unique solution. Un article a été publié aux actes du Colloque en l'honneur d'Alain Bensoussan. Un article plus complet est soumis pour publication. Le travail de cette année a consisté à mettre au point l'ensemble des démonstrations sous des hypothèses minimales, et à établir l'unicité de la solution de l'équation limite.

## 5.3 Probabilité, statistique et traitement du signal pour la génomique

### 5.3.1 Inférence statistique pour l'évolution des protéines

**Participants** : Fabien Campillo, Lorie Dudoignon.

**Mots clés** : modèles de Markov, biologie, protéine, évolution, structure secondaire des

protéines.

*Il s'agit d'une collaboration avec Jean-Loup Risler (Laboratoire Génome et Informatique, Évry).*

Dans les années 70, Dayhoff et ses collègues<sup>[DEP72,DSO78]</sup> proposent un modèle probabiliste d'évolution de protéines : ils ont déterminé une famille de matrices de taux de mutation empiriques des acides aminés (AA) — matrices PAM — identifiées à partir de plusieurs dizaines de familles de protéines. Il s'agit d'un modèle où l'évolution est indépendante de site à site et suit la même loi markovienne (réversible).

Il est bien connu des biologistes que le taux d'évolution n'est pas homogène le long de la séquence : pour des raisons tenant à la localisation du site dans la séquence ainsi qu'à la nature même de l'AA qui s'y trouve, certains sites évoluent plus vite que d'autres. Il est en revanche délicat de proposer des modèles utilisables en pratique permettant de prendre en compte ce phénomène.

On propose un modèle fondé sur des matrices de type Dayhoff mais autorisant une hétérogénéité de la loi de site en site au travers d'un paramètre qui rend compte du taux d'évolution en chaque site [8].

Une protéine présente différents niveaux de structure. La structure primaire est la séquence des AA de la chaîne peptidique. La structure secondaire rend compte des configurations locales régulières que l'on retrouve dans toutes les protéines (hélices- $\alpha$ , feuillets- $\beta$  etc.) La structure tertiaire rend compte de la forme spatiale de la protéine.

Dans [8] nous constatons que les zones structurées (i.e. correspondant à des hélices- $\alpha$  ou des feuillets- $\beta$ ) évoluent en moyenne moins vite que les zones moins structurées. Ceci rejoint le fait que la fonction de la protéine est déterminée par sa forme et que, comme c'est celle-ci qui prime, la structure semble plus conservée que la séquence.

### 5.3.2 Analyse du transcriptome

**Mots clés :** puce à ADN, analyse factorielle, classification, cartes de Kohonen, séparation de sources en aveugle.

**Participants :** Marie-Christine Roubaud, Bruno Torrèsani.

*Il s'agit d'une collaboration avec Pierre Chiappetta (Centre de Physique Théorique, Marseille-Luminy) et Rémi Houlgatte (Centre d'Immunologie de Marseille, équipe TAGC, Marseille-Luminy)*

Nous travaillons sur des jeux de données statiques (membranes nylon) fournis par le Centre d'Immunologie de Marseille (équipe TAGC). Le premier jeu de données était constitué des niveaux d'expression de 166 gènes répertoriés, dans 36 conditions données (cellules cancéreuses). Actuellement, nous étudions un jeu d'environ 1000 gènes pour 66 conditions dont 12 correspondent à des lignées cultivées. Les objectifs de cette étude s'inscrivent principalement dans deux problématiques. La première est la comparaison des conditions dans le but de classifier

---

[DEP72] M. O. DAYHOFF, R. V. ECK, C. M. PARK, « A model of Evolutionary Change in Proteins », *Atlas of Protein Sequence and Structure* 5, 1972, p. 89-99.

[DSO78] M. DAYHOFF, R. SCHWARTZ, B. ORCUTT, « A model of evolutionary change in proteins », *in* : *Atlas Of Protein Sequence and Structure*, M. Dayhoff (éditeur), 5, *Suppl. 3*, National Biomedical Foundation, Washington DC, 1978, p. 345-352.

les différentes pathologies à partir de l'expression des gènes. La seconde est la recherche de covariations de l'expression des gènes afin d'essayer d'en extraire des informations sur leurs interactions.

**Conditionnement des données** Tout indique que la variable naturelle à considérer pour ce problème est le logarithme du niveau d'expression mesuré (des résultats allant dans ce sens ont été publiés récemment, notamment par M. Vingron). Ceci pose le problème du traitement des valeurs nulles ou faibles. La procédure choisie est de remplacer ces valeurs par des valeurs aléatoires non nulles. Plusieurs variantes ont été testées, notamment l'algorithme SEM (algorithme EM stochastique).

Après correction logarithmique, l'ensemble des données présente une distribution proche de la loi gaussienne.

**Classification sur les conditions** Diverses approches ont été testées pour la comparaison entre conditions (notamment la classification hiérarchique, utilisant diverses distances et diverses procédures d'aggrégation). Globalement, les classes obtenues se corrélaient de façon satisfaisante aux données cliniques. Nos résultats vont dans la même direction que les résultats déjà obtenus par les biologistes de TAGC.

**Classification sur les gènes** Après avoir effectué une analyse exploratoire classique des données (ACP, classification automatique, ...), une procédure du type classification mixte a été élaborée afin de partitionner l'ensemble gènes en classes homogènes. Dans une première étape, l'algorithme de Kohonen (SOM)<sup>[Koh97]</sup> a été appliqué dans le but de représenter les logarithmes des gènes sur un réseau de dimension 2 et en donner une première partition en respectant la topologie locale de l'espace des observations, i.e deux observations voisines dans l'espace d'entrée appartiennent à la même classe ou à des classes voisines à la sortie.

Les centres des classes issus de cet algorithme sont ensuite utilisés comme initialisation de l'algorithme dit *K-means* et permettent d'atteindre un « bon » minimum de la variance intra-classes. Cependant ce type d'algorithme converge vers un optimum local. Pour remédier partiellement à cet inconvénient, une méthode de *groupements stables* suggérée principalement par E. Diday a été mise en œuvre. Cette technique consiste à effectuer plusieurs fois la même procédure de classification et à ne retenir comme groupements stables les ensembles d'individus qui ont toujours été affectés à une même classe dans chacune des partitions obtenues.

Comme le nombre de classes est fixé a priori assez grand, il est ensuite utile de procéder à un regroupement. La seconde étape consiste à effectuer une classification ascendante hiérarchique selon le critère de Ward où les éléments à classer sont les centres des groupements stables retenus et pondérés par leurs effectifs. Ceci permet de définir une nouvelle partition avec un nombre réduit de classes, partition que l'on améliore par réaffectations en utilisant de nouveau un algorithme de *K-means*. On obtient ainsi une partition finale constituée d'un petit nombre de *super-classes*.

Les classes obtenues sont cohérentes. On retrouve en particulier dans certaines des familles de gènes connues pour appartenir à une voie de régulation (notamment une voie faisant intervenir le récepteur œstrogène, et une voie faisant intervenir le récepteur prolactine).

---

[Koh97] T. KOHONEN, *Self Organized Maps*, Springer Verlag, 1997, 2nd edition.

Un des principaux inconvénients de la version des cartes de Kohonen telles qu'elles ont été utilisées ici est la forme de la carte qui doit être fixée a priori. Un développement ultérieur possible de ce travail consisterait à introduire une liberté supplémentaire à ce niveau. Une piste possible est l'analyse en composantes curvilignes, qui permet de transformer la base de données multidimensionnelles dans un sous espace à 2 dimensions de manière non linéaire, susceptible de mettre en évidence une structuration en groupes interprétable suivant les appartenances aux catégories a priori connues par les biologistes.

**Séparation de sources** Parallèlement à la classification, nous avons également développé une approche potentiellement intéressante pour l'identification de voies de régulation. Il s'agit de modéliser le tableau de données (plus précisément, les logarithmes des données) comme mélange linéaire de sources statistiquement indépendantes. L'idée sous-jacente est de rechercher des « processus intrinsèques » indépendants, liés à la régulation, les observations étant une combinaison linéaire inconnue de ces sources. Le problème revient alors à identifier la matrice de mélange et les sources, il s'agit d'un problème classique d'identification de sources qui a reçu une grande attention récemment, notamment dans la littérature sur les réseaux de neurones<sup>[Car98,AH99]</sup>.

Compte tenu du volume de données disponibles, l'estimation des sources basée sur des critères d'indépendance (négentropie, information mutuelle) souffre d'un manque de robustesse certain. On leur préfère alors des critères de « non-gaussianité »<sup>[AH99]</sup>. L'estimation basée sur la kurtosis par exemple conduit à des résultats relativement stables, qui font notamment apparaître des « sources indépendantes » possédant une signification biologique (d'après les biologistes de TAGC). Dans la plupart des cas, la méthode détecte des écarts à la gaussianité provoqués par des valeurs extrêmes de l'expression de certains gènes. On y retrouve notamment les voies de régulation mentionnées plus haut (celles du récepteur prolactive, du récepteur œstrogène,...), et apparaissent également dans ces sources indépendantes des familles de gènes significativement corégulées négativement, inaccessibles si on se limite à une approche purement « classification ».

Les prolongements naturels de ce programme concernent la validation de la méthode par une comparaison systématique des résultats obtenus en utilisant des critères « d'indépendance » différents, et une étude plus poussée de l'estimation automatique du nombre de sources à sélectionner. Reste également à évaluer le potentiel de ce type d'approche pour le problème « dual », à savoir le problème de la caractérisation de conditions : si tant est qu'une « source indépendante » est réellement significative, la corrélation des conditions avec celle-ci possède-t-elle une valeur en termes de prédiction ?

---

[Car98] J. CARDOSO, « Blind signal separation: statistical principles », *Proc. of the IEEE* 9, 10, 1998, p. 2009–2025.

[AH99] E. O. A. HYVÄRINEN, « Independent Component Analysis: a tutorial », 1999, <http://www.cis.hut.fi/projects/ica>.

## 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. Torrèsani. Séminaire bi-hebdomadaire, centré sur la génomique et la bio-informatique, organisé à Luminy.

#### 6.1.2 Marseille-Génopole

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

### 6.2 Actions internationales

#### 6.2.1 Contrat PICS

Le PICS « Milieux aléatoires et Homogénéisation » a continué en 2001. Il a permis en particulier d'inviter Andrey Piatnitski, du Laboratoire Lebedev de l'Académie des Sciences de Russie à Moscou, et Alexander Veretennikov, de l'Université de Leeds (Royaume Uni).

### 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.

## 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<sup>e</sup> section.

### 7.2 Enseignement

MARIE-CHRISTINE ROUBAUD • IUP Modélisation et Simulation des Systèmes Industriels (centre Joseph Fourier de Valence) Cours et travaux dirigés d'Algèbre linéaire (36h) en 1ère année, cours et travaux dirigés de probabilités (10h) en 2ème année • 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) : Cours (18h), travaux dirigés (12h) de Mathématiques, Travaux Pratiques en Maple (24h).

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

FABIEN CAMPILLO • DEA d'Informatique filière Info-Bio-Maths (15h).

ETIENNE PARDOUX à l'université de Provence • DESS de Génie Informatique et Statistique, cours sur les algorithmes stochastiques (24h) • DEA de Mathématiques Appliquées, cours sur l'homogénéisation (24h) • DEA de Bioinformatique, Biologie Structurale et Génomique, cours de Probabilités–Statistique (14h) • Ecole Supérieure de Mécanique de Marseille, cours sur les chaînes de Markov (14h).

## 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. Rozosvskii, 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] F. DELARUE, *Équation différentielles stochastiques progressives–rétrogrades. Application à l'homogénéisation des EDP quasi-linéaires*, thèse de doctorat, Université de Provence, Janvier 2002.

### Articles et chapitres de livre

- [4] G. GAUDRON, E. PARDOUX, « EDSR, convergence en loi et homogénéisation d'EDP paraboliques semi-linéaires », *Ann. Inst. IHP, Probabilités–Statistique* 37, 2001, p. 1–42.
- [5] E. PARDOUX, A. VERETENNIKOV, « On Poisson equation and diffusion approximation », *Ann. Probab.* 1, 2001.

### Communications à des congrès, colloques, etc.

- [6] L. DUDOIGNON, E. REMY, J.-L. RISLER, F. CAMPILLO, « Variation du taux d'évolution le long de séquences de protéines », *in* : *JOBIM*, Toulouse, INRA – IMPG, p. 35–36, 2001.
- [7] E. PARDOUX, A. PIATNITSKI, « Averaging of random parabolic operators », *in* : *Optimal Control And Partial Differential Equations*, J. Menaldi, et al (éditeurs), IOS Press, p. 268–276, 2001.

### Rapports de recherche et publications internes

- [8] F. CAMPILLO, L. DUDOIGNON, « Un modèle probabiliste d'évolution de protéines », *Rapport de recherche n°4332*, INRIA, 2001, <http://www.inria.fr/rrrt/rr-4332.html>.