

SOMMAIRE DU N° 130

SMF	
Mot du Président	3
MATHÉMATIQUES	
Équation de Vlasov-Poisson et amortissement Landau, <i>C. Mouhot</i>	7
Sur la route des ondelettes, <i>A. Cohen</i>	19
MATHÉMATIQUES ET PHYSIQUE	
Physique Combinatoire I : Groupes à un paramètre, <i>G.H.E. Duchamp, K.A. Penson, C. Tollu</i>	37
STATISTIQUES ET GÉNOME	
En guise d'introduction, <i>F. Picard</i>	51
Statistiques de motifs, <i>S. Schbath</i>	60
Segmentation pour l'analyse de puces CGH, <i>E. Lebarbier</i>	65
Tests multiples en génomique, <i>P. Neuvial</i>	71
Réseaux biologiques, <i>J. Chiquet</i>	76
MATHÉMATIQUES ET ART	
Les beaux arts au service des mathématiques, <i>C.P. Bruter</i>	83
INFORMATIONS	
Bilan du concours chercheurs 2011, <i>P. Dehornoy</i>	91
Journée parité en mathématiques du 6 juin 2011, <i>M. Ribot, B. Schapira</i>	93
Table ronde « Égalité des chances : mythe ou réalité ? »	98
LIVRES	105

Éditorial

Les mathématiques ne sont certes pas une science isolée, coupée du monde réel. La physique, la biologie, l'informatique, et même les arts se nourrissent de mathématiques tout en les enrichissant.

Ce numéro de la Gazette explore ainsi les liens profonds entre la mécanique quantique et la combinatoire, l'utilisation des fameuses ondelettes de Meyer pour coder les images, et l'analyse mathématique au secours du phénomène d'amortissement des plasmas de Laudau. Sans oublier les arts, avec le compte-rendu de la création en 2010 de la Société Européenne pour les Mathématiques et les Arts. Enfin, et c'est le cœur de ce numéro, nous découvrons l'analyse statistique liée à la génétique.

En effet, il y a 10 ans et quelques mois, les revues Science et Nature publiaient les premières cartes du génome humain. En un modeste écho, la Gazette propose dans ce numéro un dossier spécial sur les travaux de statistique mathématique suscités par les progrès de la génomique, et qui rejoignent de plus en plus souvent d'autres domaines mathématiques ou informatiques : structures de données complexes, théorie des graphes, etc. Un grand merci à Patricia Raynaud-Bouret pour avoir monté ce dossier.

La diversité des mathématiques est une grande richesse et un atout important de notre discipline. Elle se nourrit également de notre diversité à nous, mathématiciennes et mathématiciens de toutes origines. Le problème de la parité a déjà été évoqué dans cette Gazette, mais les réponses ne sont pas faciles à trouver. Ce numéro fait part de deux rencontres récentes sur ce sujet, à l'IHP et au CIRM.

Enfin, puisque je prends désormais la suite de Frédéric Patras aux commandes de cette Gazette, j'en profite pour le féliciter pour le travail qu'il a accompli (jusqu'au dossier « Statistique et Génome » de ce numéro !) et pour m'avoir laissé une équipe extrêmement compétente et efficace. Je ferai de mon mieux pour continuer à faire de la Gazette une publication que tous les mathématiciens francophones lisent avec plaisir et intérêt.

— San Vĩ Ngoc

SMF

Mot du Président

Quand vous lirez ces lignes, deux grands événements où la SMF s'est fortement investie auront eu lieu : la fête des 30 ans du CIRM et la célébration du bicentenaire de la naissance d'Évariste Galois autour de nombreuses manifestations (colloque scientifique, exposition, conférences grand public...). N'hésitez pas à visiter à la bibliothèque de l'IHP l'exposition consacrée à Galois, qui a été préparée par Caroline Ehrhardt. J'espère que ces deux événements auront un réel impact sur le grand public.

Il y a aussi beaucoup d'autres annonces en cette rentrée.

Mercredi 21 septembre ont eu lieu la conférence de presse de CARMIN et la réunion du comité de pilotage de ce labex. Vous avez sans doute en mémoire les difficultés de sa conception et les nombreuses craintes exprimées. La présence face à la presse des quatre directeurs J.-P. Bourguignon (IHÉS), C. Cibils (CIMPA), P. Foulon (CIRM) et C. Villani (IHP) présentant les projets communs aux quatre institutions et leur approbation à l'unanimité par le comité de pilotage laissent augurer que CARMIN va permettre effectivement une meilleure coordination entre nos grands outils nationaux. Cette conférence de presse a été aussi l'occasion d'annoncer pour l'IHP la création d'une chaire Poincaré¹ et pour le CIRM la création d'une chaire Jean Morlet². Les conseils d'administration du CIRM et de la SMF avaient conditionné la création de cette chaire et le bouclage financier de la rénovation de la maison du jardinier au CIRM à la présentation d'un projet scientifique global où, sans exclure un soutien de fonds privés, l'AMU (la nouvelle université rassemblant les 3 universités marseillaises) serait fortement impliquée. Les discussions récentes que P. Foulon a menées avec les 3 vice-présidents recherche des universités constituantes de l'AMU ont permis une clarification du financement de cette chaire par l'AMU et dissiperont quelques inquiétudes soulevées en juillet par des membres de l'Institut Mathématique de Luminy.

Avec la SFdS et la SMAI, la SMF a soutenu dans une lettre commune le nouveau programme PRIAM présenté par le Réseau National des Bibliothèques de Mathématiques dans le cadre de la seconde vague du programme equipex du

¹ Financée par les fonds initialement prévus pour le prix Clay décerné à Perelman pour sa résolution de la conjecture de Poincaré.

² Du nom de l'ingénieur chez Elf à l'origine de la création des ondelettes.

grand emprunt. Le lecteur se rappelle que nous avons déploré le sort réservé au programme « PurMath » lors de la première vague. Nous espérons que le jury constitué par l'ANRU³ comprendra mieux dans la deuxième vague le rôle clef de la documentation dans notre discipline.

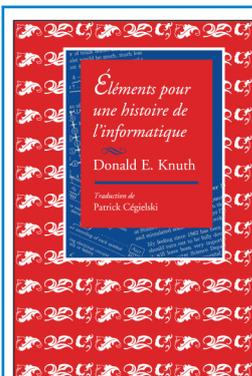
Sous la coordination d'Animath, la SMF et de nombreuses associations impliquées dans l'animation mathématique vers les collégiens et les lycéens ont présenté un projet de consortium « Cap'Maths » dans le cadre du volet « Culture scientifique et technique et égalité des chances » du grand emprunt. Ce projet vient d'être accepté et donne des moyens financiers importants pour une période de 5 ans. Plusieurs opérations que nous menons avec Animath et d'autres partenaires, comme le cycle « Un texte, un mathématicien » et le programme « Promenades Mathématiques », sont susceptibles de rentrer dans le champ des appels d'offre de Cap'Maths. Dans le cadre de ce volet, en plus de Cap'Maths, d'autres programmes où les mathématiciens sont présents seront également financés comme « Des Maisons pour la science et la technologie au service des Professeurs », porté par l'Académie des sciences, « Maths au CP » et « Lecture et petite enfance » proposé par l'association Agir pour l'école et enfin le CREDEC⁴. Ainsi contrairement à ce que nous avons observé dans d'autres appels à projet (Labex, Idex), nous sommes heureux de constater que le dynamisme des mathématiciens dans ce secteur a été largement reconnu.

Enfin, c'est aussi un plaisir d'annoncer que le projet de carte des masters soutenu par les 3 sociétés savantes avec le soutien de l'INSMI, conçu pour informer un large public, est maintenant finalisé et d'un accès très facile par notre site.

Le 1^{er} octobre 2011
Bernard Helffer

³ Agence Nationale pour la Rénovation Urbaine.

⁴ Centre de Recherche de l'Ecole de la Deuxième Chance de Marseille.



Éléments pour une histoire de l'informatique

Donald E. Knuth

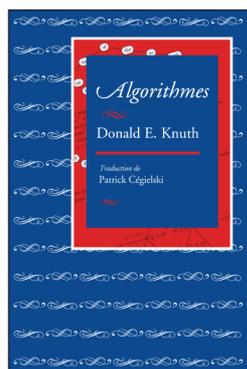
(traduction de Patrick Cégielski)

prix public* : 28 € - prix membre* : 19 €

* frais de port non compris

ISBN : 978-1-57586-622-2

Voici enfin la traduction en français d'une série de quinze articles classiques de Donald Knuth portant sur des moments clés de l'histoire de l'informatique, plus exactement l'algorithmique, que ce soit à l'époque babylonienne ancienne, au IX^e siècle lors de l'âge d'or des sciences arabes ou après l'invention des ordinateurs : étude des premiers programmes, des premiers langages de programmation et des premiers compilateurs. On y retrouve également l'historique des notions chères à Knuth : découpage d'un paragraphe en ligne, grammaires à attributs et vocabulaire de l'analyse des algorithmes.



Algorithmes

Donald E. Knuth

(traduction de Patrick Cégielski)

prix public* : 28 € - prix membre* : 19 €

* frais de port non compris

ISBN : 978-1-57586-620-8

Donald Knuth, professeur émérite à l'université Stanford, est l'un des pionniers de l'algorithmique. Il est l'auteur d'une centaine d'articles et d'une dizaine de livres sur l'algorithmique et les mathématiques discrètes. Les quatre premiers volumes de «The Art of Computer Programming» demeurent un ouvrage de référence. Afin d'avoir une bonne qualité de mise en page pour la deuxième édition de cette série d'ouvrages, Knuth a créé deux logiciels libres, par la suite largement utilisés en typographie professionnelle et en mathématique, TeX et METAFONT. Son intérêt pour la typographie l'a également poussé à créer la police Computer Modern, police par défaut TeX.

**Société
Mathématique
de France**



Institut Henri Poincaré
11 rue Pierre et Marie Curie
F - 75231 PARIS CEDEX 05

<http://smf.emath.fr>



Revue d'histoire des mathématiques

Tome 17, Fascicule 1

Sommaire

Michèle Audin, Roland Brasseur - Addendum à «Publier sous l'Occupation I»,
Revue d'histoire des mathématiques, 15 (2009), 5-57.

Olivier Bruneau - Le *De Linearum* de MacLaurin : entre Newton et Poncelet.

Jenny Boucard -Louis Poinot et la théorie de l'ordre : un chaînon manquant
entre Gauss et Galois ?

Reinhard Siegmund-Schultze

TEXTES ET DOCUMENTS : Opposition to the Boycott of German Mathematics in
the Early 1920s : Letters by Edmund Landau (1877-1938) and Edwin Bidwell Wilson
(1879-1964).

prix public* : 38 € - prix membre* : 27 €

* frais de port non compris

Catalogue disponible à l'adresse :

<http://smf.emath.fr>

Information

Maison de la SMF

avenue de Luminy, BP 67, F-13274 Marseille cedex 9, France

Tél : 33 4 91 83 30 25 • Fax : 33 4 91 41 17 51 • email : smf@smf.univ-mrs.fr



Institut Henri Poincaré
11 rue Pierre et Marie Curie
F - 75231 PARIS CEDEX 05

<http://smf.emath.fr>

MATHÉMATIQUES

Stabilité non-linéaire pour l'équation de Vlasov-Poisson et amortissement Landau

Clément Mouhot¹

Nous présentons l'un des travaux de C. Villani cité pour la médaille Fields qui lui a été attribuée l'été dernier.

1. Un détour par la physique : la découverte de Landau

Lev Davidovitch Landau est l'un des plus grands physiciens du XX^e siècle. Il a révolutionné un grand nombre de domaines. En ce qui concerne la physique des plasmas, il publie deux articles retentissants. En 1936, il explique comment modifier la célèbre équation de Ludwig Boltzmann (décrivant les collisions dans un gaz) au cas des collisions dans un plasma, *i.e.* un gaz d'électrons et d'ions qui interagissent à travers le champ électro-magnétique. Cependant en 1938, un autre physicien russe collaborateur de Landau, Anatoly Vlasov, montre qu'en général ces équations de collision ne décrivent les plasmas que sur des temps très longs mais pas sur les temps d'observation usuels. Il propose alors une nouvelle équation dans laquelle les électrons du plasma interagissent à travers le champ électrique moyen qu'ils créent collectivement, sans collision. Dans le cas où l'on ne considère que le champ électrique et l'on néglige le champ magnétique, c'est l'équation dite de *Vlasov-Poisson*.

En 1946, Landau met alors en évidence un phénomène étonnant : il suggère au moyen d'un calcul astucieux sur l'équation de Vlasov-Poisson que les plasmas « s'homogénéisent » en se réorganisant au cours du temps. Autrement dit la répartition des électrons s'égalise et donc le champ électrique créé par le plasma tend vers zéro au fur et à mesure que la neutralité électrique est rétablie. Landau propose même une estimation de la vitesse (exponentielle en temps) à laquelle se produit cette « homogénéisation ». C'est ce phénomène que l'on appelle aujourd'hui amortissement Landau. Pourquoi cette prédiction est-elle si surprenante ? Parce que contrairement aux gaz où les particules entrent en collision, le plasma décrit par l'équation de Vlasov-Poisson est « réversible », de la même façon qu'un système dynamique décrit par les équations de Newton.²

¹ DPMMS, Université de Cambridge.

² En fait, l'équation de Vlasov-Poisson est très exactement la *limite de champ moyen* des équations de Newton pour un très grand nombre de particules, et hérite de leur structure hamiltonienne. La preuve mathématique rigoureuse de cette limite reste un important problème ouvert.

Cet article est le point de départ d'un demi-siècle de débats chez les physiciens. La découverte de Landau fut jugée « stupéfiante » par ses contemporains ; c'est maintenant l'une des bases théoriques de la physique classique des plasmas (parmi un très grand nombre de références citons [1, 16]). Des effets d'amortissement similaires ont été prédits en astrophysique [17], mais aussi dans d'autres domaines de la physique.

Plusieurs difficultés se posent pour confirmer et surtout expliquer l'intuition originale de Landau. Premièrement, il semble y avoir une contradiction entre une équation mathématique *réversible* et un phénomène *irréversible* (le champ électrique tend vers zéro). Deuxièmement, le calcul de Landau simplifie l'équation en négligeant l'interaction du plasma sur lui-même (la non-linéarité) et il n'y a aucune garantie *a priori* que l'étude *linéarisée* de cette équation hamiltonienne gouverne correctement le comportement non-linéaire en temps long, et qu'il n'y ait pas de développement d'instabilité. Enfin et troisièmement, il se produit un phénomène bien connu des physiciens dans les plasmas, la *filamentation*, *i.e.* le développement de filaments de plus en plus rapprochés dans l'espace des phases (position et vitesse), qui semble contraire à l'idée d'homogénéisation.

2. Le problème mathématique

Le « modèle standard » de la physique des plasmas (en mécanique classique) est l'équation de Vlasov–Poisson–Landau [13, 25], ici écrite avec des conditions aux limites périodiques et en unités adimensionnées :

$$(1) \quad \frac{\partial f}{\partial t} + v \cdot \nabla_x f + F[f] \cdot \nabla_v f = \frac{\log \Lambda}{2\pi\Lambda} Q_L(f, f),$$

où $f = f(t, x, v) \geq 0$ est la fonction de distribution des électrons ($t \geq 0$, $v \in \mathbb{R}^3$, $x \in \mathbb{T}^3 = \mathbb{R}^3/\mathbb{Z}^3$), et

$$(2) \quad F[f](t, x) = - \int_{\mathbb{R}^3} \int_{\mathbb{T}^3} \nabla W(x - y) f(t, y, w) dw dy$$

est la force auto-induite dite de « champ moyen », soit le champ électrique pour un plasma. Dans ces formules, $W(x) = 1/|x|$ est le potentiel d'interaction coulombien, et Q_L est l'opérateur bilinéaire de collision de Landau, décrit par exemple dans [16] ou [23]. Le paramètre Λ est très grand, variant typiquement entre 10^2 et 10^{30} .

Sur de très grandes échelles de temps (disons $O(\Lambda/\log \Lambda)$), les phénomènes dissipatifs jouent un rôle non négligeable, et l'augmentation de l'entropie est supposée forcer la convergence (lente) vers un équilibre homogène en espace et gaussien en vitesse. Grâce aux progrès récents sur l'hypocoercivité (dans lesquels C. Villani a joué un rôle majeur), ce mécanisme est maintenant assez bien compris mathématiquement parlant, dès que l'on dispose d'estimations de régularité globale, voir [24] et les références incluses.

On peut alors formuler la prédiction de Landau [13] de la façon suivante : au voisinage d'un équilibre homogène vérifiant certaines conditions – par exemple une fonction radiale régulière et monotone de la vitesse v , mais pas forcément gaussienne – il y a « stabilité » au sens où le champ électrique converge vers zéro et la solution s'homogénéise (en la variable d'espace) sur des échelles de temps

beaucoup plus courtes (disons $O(1)$), par le jeu de mécanismes purement conservatifs. Ce phénomène est en fait une propriété de l'équation (non collisionnelle) de Vlasov-Poisson, obtenue en posant $\Lambda = \infty$ dans (1).

Cette même équation (sans opérateur de collision) est également l'un des principaux modèles statistiques pour les « gaz d'étoiles » que sont les galaxies, lorsqu'on considère le potentiel d'interaction *attractif* newtonien $W(x) = -1/|x|$ des forces gravitationnelles, et une prédiction d'amortissement inspirée de celle de Landau en physique des plasmas fut émise par le physicien britannique Donald Lynden-Bell dans les années 1960.

En dehors de la littérature physique et de quelques rares articles de simulation numérique [26, 11], l'état de l'art mathématique se divise principalement en trois classes de résultats. La première concerne l'étude linéarisée, dans la lignée du calcul original de Landau, et dans ce cas le phénomène d'amortissement est compris depuis longtemps [2, 5, 9, 10, 19, 22]. Nous allons y revenir plus en détails.

La deuxième classe de résultats, qui est aussi la plus développée mathématiquement et fait l'objet d'une théorie autonome intéressante en tant que telle, concerne la question de la « stabilité orbitale » dans des espaces L^p , voir par exemple parmi de nombreuses références [8, 15, 14]. Le terme « orbital » se réfère au fait que l'invariance (galliléenne) par translation de l'équation induit un paramètre libre dans l'étude de stabilité dans ces espaces fonctionnels (qui sont eux-mêmes invariants par translation); la stabilité est uniforme pour tous les temps, mais au prix de ce paramètre libre. Le but de cette théorie est de façon générale de montrer qu'un équilibre vérifiant la condition de stabilité linéarisée possède cette propriété de stabilité non-linéaire, *i.e.* qu'il existe un voisinage L^p de piège autour de cet équilibre, aux translations près. Mentionnons par exemple un résultat obtenu récemment par M. Lemou, F. Méhats et P. Raphaël dans le cas gravitationnel : ils démontrent qu'un équilibre continu à support compact et « sphérique » au sens où il est fonction uniquement de l'énergie microscopique $e = [|v|^2/2 + (W * \rho)]$ est orbitalement stable, ce qui résout mathématiquement une conjecture classique en dynamique des galaxies. Un résultat similaire était déjà connu dans le cas (bien plus simple) des interactions répulsives coulombiennes [3]. La force de ces résultats est d'utiliser des espaces de fonctions très grands, et donc de s'appliquer à une classe très large de solutions. C'est aussi leur faiblesse, car l'invariance par translation de ces espaces (et de façon générale le fait que ces derniers ne mesurent pas la régularité) ne permet pas de déduire des propriétés de convergence asymptotique plus précises sur les solutions.

Enfin la troisième classe de résultats concerne la preuve de l'*existence* de solutions amorties [4, 12] par des méthodes de point fixe sur les conditions en temps infini. Bien qu'ils montrent la possibilité mathématique de l'amortissement pour l'équation non-linéaire, ces résultats ne disent rien sur les données initiales qui pourraient y mener, et donc si cet amortissement se produit effectivement ou pas.

Le théorème démontré par C. Villani et l'auteur de cette notice [20, 21] établit pour la première fois cet effet d'amortissement dans un cadre non-linéaire, en temps infini et avec le taux exponentiel prédit par Landau.

3. Le résultat d'amortissement non-linéaire

Théorème (en simplifiant un petit peu...). Soit $W : \mathbb{T}^3 \rightarrow \mathbb{R}$ un potentiel d'interaction tel que sa transformée de Fourier vérifie

$$(3) \quad \forall k \in \mathbb{Z}^3 \setminus \{0\}, \quad |\hat{W}(k)| \leq \frac{C_W}{|k|^2}$$

et soit $f^0 = f^0(v) \geq 0$ un équilibre spatialement homogène avec régularité analytique et qui vérifie une condition de stabilité linéarisée correspondant à ce potentiel d'interaction (voir plus loin).

Alors si $f_i \geq 0$ est telle que

$$(4) \quad \delta := \|f_i - f^0\|_{\mathcal{A}(\lambda)} \quad \text{est suffisamment petit}$$

pour une norme $\mathcal{A}(\lambda)$ contrôlant la régularité analytique en x et v et la décroissance exponentielle en la variable v (l'exposant $\lambda > 0$ se réfère à la régularité analytique en v), l'unique solution de l'équation de Vlasov non-linéaire

$$(5) \quad \frac{\partial f}{\partial t} + v \cdot \nabla_x f + F[f] \cdot \nabla_v f = 0, \quad F[f](t, x) = - \int_{\mathbb{T}^3} \int_{\mathbb{R}^3} \nabla W(x-y) f(t, y, w) dw dy,$$

définie pour tous les temps et de condition initiale $f(0, \cdot) = f_i$, vérifie (pour tout $0 \leq \lambda' < \lambda$ et des constantes associées)

$$(6) \quad \|\rho(t, \cdot) - \rho_\infty\|_{C^r(\mathbb{T}^3)} \leq C \delta e^{-2\pi\lambda'|t|} \quad \text{et} \quad \|F[f](t, \cdot)\|_{C^r(\mathbb{T}^3)} \leq C \delta e^{-2\pi\lambda'|t|}$$

où $\rho(t, x) = \int f(t, x, v) dv$, $\rho_\infty = \iint f_i(x, v) dv dx$. En outre il existe des profils analytiques $f_{+\infty}(v)$, $f_{-\infty}(v)$ tels que $f(t, \cdot) \xrightarrow{t \rightarrow \pm\infty} f_{\pm\infty}$ au sens faible (par exemple dans des espaces de Sobolev négatifs).

Ce théorème est entièrement constructif. On peut l'interpréter de différentes façons, l'une d'entre elles étant certainement de dire que pour cette équation hamiltonienne, la stabilité linéarisée implique la stabilité non-linéaire, mais pour des normes de régularité « glissante » (voir plus loin), ce qui rappelle donc la stabilité orbitale. D'un point de vue de système dynamique, un voisinage entier – en topologie analytique – de f^0 (équilibre linéairement stable) est rempli par des trajectoires homoclines ou hétéroclines. C'est la dimension infinie qui permet ce comportement remarquable.

En terme d'optimalité des hypothèses, les conditions aux limites périodiques sont bien sûr contestables; cependant, au vu des contre-exemples de Glassey et Schaeffer [6, 7], un mécanisme de confinement – ou tout au moins de limitation des longueurs d'onde spatiales – est indispensable. L'analyticité du profil f^0 et de la perturbation est nécessaire pour pouvoir observer une convergence exponentielle, mais la preuve peut être étendue au cas d'une régularité Gevrey. La condition de stabilité linéaire sur laquelle nous allons revenir est essentiellement optimale et couvre les cas physiquement intéressants : l'interaction attractive de Newton (gravitation) aux longueurs d'onde inférieures à la longueur d'instabilité de Jeans, et l'interaction répulsive de Coulomb (force électrique) pour toutes les longueurs d'ondes autour de profils f^0 radialement symétriques en dimension supérieure ou égale à 3. La condition (3) sur le potentiel d'interaction en revanche n'intervient

que dans la stabilité non-linéaire, elle inclut à nouveau les interactions de Newton ou de Coulomb. Quant à la condition (4), son caractère perturbatif est naturel au vu des spéculations théoriques et des études numériques sur le sujet.

Discutons maintenant plus précisément de la signification du résultat. La convergence en temps grand repose sur un mécanisme réversible, purement déterministe, sans fonctionnelle de Lyapunov ni interprétation variationnelle. Ce n'est pas non plus un mécanisme de dispersion, puisque nous sommes dans une configuration de confinement en espace et la distribution reste uniformément localisée en vitesse au cours du temps. Les profils asymptotiques $f_{\pm\infty}$ gardent d'ailleurs la mémoire de la donnée initiale et de l'interaction. Remarquons enfin que la convergence en temps grand de la fonction de distribution n'a lieu qu'au sens faible ; les normes des dérivées en vitesse croissent très vite en temps, ce qui traduit la filamentation dans l'espace des phases, et un transfert d'énergie (ou d'information) des basses vers les hautes fréquences (« turbulence faible »). C'est ce mécanisme de transfert d'information aux petites échelles qui permet de réconcilier la réversibilité de l'équation de Vlasov–Poisson avec l'apparente irréversibilité de l'amortissement Landau.

4. La théorie linéarisée

La stabilité linéaire est la première étape de notre étude et se révèle fondamentale pour comprendre la structure que nous chercherons à propager à travers le schéma itératif de construction de solutions pour l'équation non-linéaire.

Commençons par le cas très simple de l'équation de transport libre dans le tore

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = 0, \quad x \in \mathbb{T}^3, \quad v \in \mathbb{R}^3$$

qui décrit simplement le mouvement en ligne droite des particules dans notre boîte périodique. On résout cette équation par la méthode des caractéristiques qui donne la formule $f(t, x, v) = f_i(x - tv, v)$ pour une donnée initiale f_i . Si on note, pour $k \in \mathbb{Z}^d$, $\eta \in \mathbb{R}^d$, les transformées de Fourier

$$\hat{f}(k, v) = \int_{\mathbb{T}^d} f(x, v) e^{-2i\pi k \cdot x} dx, \quad \tilde{f}(k, \eta) = \iint_{\mathbb{T}^d \times \mathbb{R}^d} f(x, v) e^{-2i\pi k \cdot x} e^{-2i\pi \eta \cdot v} dv dx,$$

on obtient alors en variables de Fourier : $\tilde{f}(t, k, \eta) = \tilde{f}_i(k, \eta + tk)$. On observe alors que toute décroissance de \tilde{f}_i en η (et uniformément par rapport à k)

$$\left| \tilde{f}_i(k, \eta) \right| \leq \varphi(\eta) \xrightarrow{|\eta| \rightarrow \infty} 0$$

se traduit par le contrôle sur la solution $|\tilde{f}(t, k, \eta)| \leq \varphi(\eta + tk)$, c'est-à-dire une estimation de localisation de l'énergie dans l'espace de Fourier. Cette hypothèse sur f_i correspond à une régularité en v uniforme par rapport à x . En particulier, si cette dernière régularité est analytique, la fonction φ décroît exponentiellement et on obtient

$$(7) \quad \left| \tilde{f}(t, k, \eta) \right| \leq C e^{-\lambda|\eta + tk|}.$$

Pour $\eta = 0$ cela implique alors en particulier que la transformée de Fourier de la densité ρ vérifie $|\hat{\rho}(t, k)| \leq C e^{-\lambda|k|t}$, soit la décroissance exponentielle de tous les modes non nuls. Il est également très facile de voir que (7) implique la convergence

faible de f_t en $t \rightarrow \pm\infty$ vers un profil homogène. L'amortissement Landau est ainsi démontré. L'idée clef est alors de se concentrer maintenant sur l'estimation (7) qui correspond à une structure de *mélange de phase* et de montrer qu'elle reste vraie pour l'équation linéarisée (sous une condition de stabilité) puis pour l'équation non-linéaire au niveau perturbatif.

Considérons l'équation de Vlasov linéarisée autour de $f^0(v)$:

$$(8) \quad \frac{\partial h}{\partial t} + v \cdot \nabla_x h - (\nabla W * \rho) \cdot \nabla_v f^0 = 0, \quad \rho = \int_{\mathbb{R}^3} h \, dv.$$

Cette équation se découple en une infinité d'équations indépendantes régissant l'évolution des modes de ρ : pour tout $k \in \mathbb{Z}^d$ et tout $t \geq 0$,

$$(9) \quad \hat{\rho}(t, k) - \int_0^t K^0(t - \tau, k) \hat{\rho}(\tau, k) \, d\tau = \tilde{h}_i(k, kt),$$

où h_i est la donnée initiale, et K^0 un noyau intégral qui dépend de f^0 :

$$(10) \quad K^0(t, k) = -4\pi^2 \hat{W}(k) \tilde{f}^0(kt) |k|^2 t.$$

On déduit alors de résultats classiques sur les équations de Volterra que pour tout $k \neq 0$ la décroissance de $\hat{\rho}(t, k)$ quand $t \rightarrow +\infty$ est essentiellement contrôlée par le pire de deux taux de convergence suivants :

- (a) le taux de convergence du membre de droite de (9), qui ne dépend que de la régularité de la donnée initiale dans la variable de vitesse ;
- (b) $e^{-\lambda t}$, où λ est le plus grand réel positif tel que la transformée de Fourier-Laplace de K^0 ne s'approche pas de 1 dans la bande complexe $\{0 \leq \Re z \leq \lambda\}$.

Une fois la décroissance sur ρ démontrée, il est ensuite facile de remonter à la solution complète h , pour obtenir sur celle-ci l'estimation (7) de mélange de phase.

5. Quelques idées sur la preuve de l'amortissement non-linéaire

L'idée centrale est de montrer que la structure mathématique de mélange de phase (du type (7)) décrite plus haut survit à la perturbation non-linéaire du champ moyen.

Afin de pouvoir suivre ces estimations de mélange de phase au niveau non-linéaire, nous devons tout d'abord les reformuler en introduisant des normes analytiques « hybrides » (reposant sur la taille des dérivées successives dans la variable de vitesse, et sur la taille des coefficients de Fourier dans la variable de position) et « glissantes » (la norme utilisée changera au cours du temps pour tenir compte des transferts dans l'espace des phases). Cinq indices permettent d'obtenir toute la souplesse nécessaire :

$$(11) \quad \|f\|_{\mathcal{Z}_\tau^{\lambda, (\mu, \gamma); p}} = \sum_{k \in \mathbb{Z}^d} \sum_{n \in \mathbb{N}^d} e^{2\pi\mu|k|} (1+|k|)^\gamma \frac{\lambda^n}{n!} \left\| (\nabla_v + 2i\pi\tau k)^n \hat{f}(k, v) \right\|_{L^p(dv)}$$

et on notera cette norme juste $\mathcal{Z}_\tau^{\lambda, \mu; p}$ dans le cas $\gamma = 0$.

Les normes \mathcal{Z} possèdent des propriétés remarquables vis-à-vis de la composition et du produit. Le paramètre τ permet de compenser, dans une certaine mesure, la filamentation, et de suivre le mélange de phase. Enfin le caractère hybride est

bien adapté à la géométrie du problème. Si f ne dépend que de x , la norme (11) coïncide avec la norme d'algèbre $\mathcal{F}^{\lambda\tau+\mu,\gamma}$ définie par

$$(12) \quad \|f\|_{\mathcal{F}^{\lambda\tau+\mu,\gamma}} = \sum_{k \in \mathbb{Z}^d} |\hat{f}(k)| e^{2\pi(\lambda\tau+\mu)|k|} (1+|k|)^\gamma.$$

(Nous utilisons également la version « homogène » $\dot{\mathcal{F}}^{\lambda\tau+\mu,\gamma}$ lorsque l'on n'inclut pas le mode $k = 0$ dans la somme.)

L'équation de Vlasov est résolue par un schéma itératif de Newton

$$f^n = f^0 + h^1 + \dots + h^n,$$

dont la première étape est la solution du linéarisé autour de l'équilibre homogène f^0 en partant de la fluctuation initiale

$$\begin{cases} \partial_t h^1 + v \cdot \nabla_x h^1 + F[h^1] \cdot \nabla_v f^0 = 0 \\ h^1(0, \cdot) = f_i - f^0, \end{cases}$$

et les étapes suivantes correspondent à des problèmes linéarisés autour de la solution approchée f^n , sans donnée initiale mais avec un terme source provenant de l'erreur quadratique en l'incrément h^n de l'étape précédente

$$(13) \quad \begin{cases} \partial_t h^{n+1} + v \cdot \nabla_x h^{n+1} + F[f^n] \cdot \nabla_v h^{n+1} + F[h^{n+1}] \cdot \nabla_v f^n = -F[h^n] \cdot \nabla_v h^n \\ h^{n+1}(0, \cdot) = 0. \end{cases}$$

Dans une première étape préliminaire, on établit la régularité analytique en temps petit des $h^n(\tau, \cdot)$ en norme $\mathcal{Z}_\tau^{\lambda,(\mu,\gamma);1}$ et pour un intervalle de temps indépendant de n ; cette étape, dans l'esprit d'un théorème de Cauchy–Kowalevskaya, repose sur l'identité

$$(14) \quad \left. \frac{d}{dt} \right|_{t=\tau} \|f\|_{\mathcal{Z}_\tau^{\lambda(t),\mu(t);p}} \leq -\frac{K}{1+\tau} \|\nabla f\|_{\mathcal{Z}_\tau^{\lambda(\tau),\mu(\tau);p}},$$

où $\lambda(t) = \lambda - Kt$, $\mu(t) = \mu - Kt$.

Ensuite il s'agit d'établir des estimations uniformes en temps sur chaque $h^{n+1}(\tau, \cdot)$ solution de (13). Le prix à payer pour l'utilisation d'un schéma de Newton à la place d'un point fixe de Picard est bien sûr le fait que la linéarisation soit faite autour d'une solution $f^n = f^n(t, x, v)$ non stationnaire. Si l'on compare ainsi (13) au problème linéarisé (8) qu'étudiait Landau, on voit principalement deux nouvelles difficultés :

(1) La présence d'un terme de champ de force $F[f^n] \cdot \nabla_v h^{n+1}$ qui courbe les trajectoires des particules, avec une intensité qui ne tend *pas* vers zéro lorsque n tend vers l'infini (car on attend une convergence de f^n vers une solution non stationnaire et qui n'est pas homogène pour tout temps); malgré tout on pourra exploiter que ce terme décroît très rapidement *en temps* à condition d'avoir démontré la structure de mélange de phase sur la solution approchée f^n au rang n .

(2) Le fait que dans le terme de « réaction » $F[h^{n+1}] \cdot \nabla_v f^n$ la fonction f^n soit non stationnaire (*i.e.* non homogène en espace) implique une croissance en $O(t)$ due au gradient en vitesse (c'est l'effet de la filamentation).

On propage alors un certain nombre d'estimations à travers le schéma, les plus importantes étant (en simplifiant légèrement)

$$(15) \quad \sup_{\tau \geq 0} \left\| \int_{\mathbb{R}^d} h^n(\tau, \cdot, v) dv \right\|_{\mathcal{F}^{\lambda_n \tau + \mu_n}} \leq \delta_n,$$

$$(16) \quad \sup_{t \geq \tau \geq 0} \left\| h^n(\tau, \Omega_{t,\tau}^n) \right\|_{\mathcal{Z}^{\lambda_n(1+b), \mu_n; 1}} \leq \delta_n, \quad b = b(t) = \frac{B}{1+t},$$

$$(17) \quad \left\| \Omega_{t,\tau}^n - \text{Id} \right\|_{\mathcal{Z}^{\lambda_n(1+b), (\mu_n, \gamma); \infty}} \leq C \left(\sum_{k=1}^n \frac{\delta_k e^{-2\pi(\lambda_k - \lambda_{n+1})\tau}}{2\pi(\lambda_k - \lambda_{n+1})^2} \right) \min\{t - \tau; 1\}.$$

Pour résoudre la première difficulté on utilise une méthode partiellement eulérienne et partiellement lagrangienne, en intégrant l'équation le long des caractéristiques $(X_{\tau,t}^n, V_{\tau,t}^n)$ créées par la force $F[f^n]$. (Ici τ est le temps initial, t le temps courant, (x, v) les conditions initiales, (X^n, V^n) l'état courant.) La régularité de ces caractéristiques est exprimée par des contrôles en norme hybride sur les opérateurs $\Omega_{t,\tau}^n(x, v) = (X_{t,\tau}^n, V_{t,\tau}^n)(x + v(t - \tau), v)$, qui comparent la dynamique perturbée à la dynamique libre, informellement appelés *opérateurs de déflexion*. Ainsi les estimations (17) sont obtenues par des applications répétées de théorèmes de point fixe en normes analytiques.

On note, dans (15), l'augmentation linéaire de la régularité de la densité spatiale, qui fait contreponds à la détérioration de régularité dans la variable de vitesse. Dans (16), le léger décalage des indices par la fonction $b(t)$ est crucial pour absorber les termes d'erreur provenant de la composition avec les opérateurs de déflexion; la constante B est elle-même choisie en fonction des estimations en temps petit réalisées précédemment. Enfin, dans (17), remarquons le contrôle uniforme en t , et l'amélioration des estimées dans les deux régimes asymptotiques $t \rightarrow \tau$ et $\tau \rightarrow \infty$; ceci également est important pour la gestion de ces termes d'erreur. En effet, en propageant ces décalages d'indices tout au long de nos estimations, nous pourrions finalement les absorber grâce à un mécanisme de transfert de régularité analytique sur les moyennes en vitesse dont voici une version simplifiée. étant données deux fonctions de distribution h et f , dépendant de t, x, v , définissons

$$\sigma(t, x) = \int_0^t \int_{\mathbb{R}^d} (F[h] \cdot \nabla_v f)(\tau, x - v(t - \tau), v) dv d\tau.$$

L'interprétation de σ est comme suit : si les particules distribuées selon h exercent une force sur des particules distribuées selon f , alors σ est la variation de densité $\rho[h] = \int h dv$ causée par la réaction de f sur h . Ce terme σ correspond au terme de réaction de f^n sur h^{n+1} dans le schéma (13), sans courbure des trajectoires. Alors on démontre

$$(18) \quad \left\| \sigma(t, \cdot) \right\|_{\mathcal{F}^{\lambda t + \mu}} \leq \int_0^t \left\| F[h(\tau, \cdot)] \right\|_{\mathcal{F}^{\lambda[\tau - b(t - \tau)] + \mu, \gamma}} \left\| \nabla f(\tau, \cdot) \right\|_{\mathcal{Z}^{\lambda(1+b), (\mu, 0); 1}} d\tau.$$

On voit dans (18) que la régularité de σ est meilleure que celle de $F[h]$, avec un gain qui dégénère en temps grand et lorsque $\tau \rightarrow t$. Ce mécanisme est réminiscent

des « lemmes de moyenne » pour les équations cinétiques. Il est remarquable qu'il produise un gain de régularité dépendant du temps qui permet de compenser *exactement* les termes d'erreur ci-dessus dus à la courbure des trajectoires.

Les constantes λ_n et μ_n décroissent à chaque étape du schéma, convergeant – pas trop rapidement – vers des limites $\lambda_\infty, \mu_\infty$ positives ; dans le même temps, les constantes δ_n convergent extrêmement vite vers 0 (essentiellement avec un taux en « double exponentiel » caractéristique du schéma de Newton), ce qui garantit « par rétroaction » l'uniformité des constantes du membre de droite de (17).³

Pour résoudre la deuxième difficulté et ainsi pouvoir passer de l'étape n à l'étape $n+1$ avec des bornes *uniformes en temps*, il nous faut à nouveau contrôler le terme de réaction mais maintenant en temps long. Oublions pour le moment le problème de la courbure des caractéristiques. Nous démontrons tout d'abord un mécanisme d'*extorsion de régularité*, dont voici une version simplifiée :

$$(19) \quad \|\sigma(t, \cdot)\|_{\dot{X}^{\lambda t + \mu}} \leq \int_0^t K(t, \tau) \|F[f(\tau, \cdot)]\|_{\mathcal{F}^{\lambda\tau + \mu, \gamma}} d\tau,$$

où

$$K(t, \tau) = \left[\sup_{0 \leq s \leq t} \left(\frac{\|\nabla_v \bar{f}(s, \cdot)\|_{\dot{X}_s^{\bar{\lambda}, \bar{\mu}; 1}}}{1 + s} \right) \right] \\ \times (1 + \tau) \sup_{k \neq 0, \ell \neq 0} \frac{e^{-2\pi(\bar{\lambda} - \lambda)|k(t - \tau) + \ell\tau|} e^{-2\pi(\bar{\mu} - \mu)|\ell|}}{1 + |k - \ell|^\gamma}$$

et $\lambda < \bar{\lambda}$, $\mu < \bar{\mu}$. Intuitivement une partie de la régularité analytique en x et v s'est convertie en décroissance temporelle, ce qui nous permet de ramener l'intégrale du noyau $K(t, \tau)$ d'un $O(t^2)$ à un $O(t)$.

Néanmoins, on est alors confronté au fait que le noyau $K(t, \tau)$ est encore « trop grand » : il est d'intégrale $O(t)$ quand $t \rightarrow \infty$, ce qui laisse craindre une croissance exponentielle indépendante de n que le schéma ne pourra compenser, et donc une instabilité violente. Heuristiquement si on pose $y(t) = \|\rho(t, \cdot)\|_{\dot{X}^{\lambda t + \mu}}$ nous sommes proches d'une inégalité différentielle de la forme

$$(20) \quad y'(t) \leq A + \int_0^t y(\tau) K(t, \tau) d\tau,$$

et si $K(t, \tau) = B$ est constant cela conduit à une estimation $y(t) \leq C e^{Bt}$.

La solution vient alors du fait qu'au fur et à mesure que t augmente, le noyau $K(t, \tau)$ est en fait *de plus en plus concentré* sur des temps discrets $\tau = kt/(k - \ell)$ reliés à des fréquences d'espace k et ℓ ; c'est la manifestation des *échos plasmas*, découverts et observés expérimentalement dans les années soixante [18]. Le rôle stabilisant du phénomène d'écho, en relation avec l'amortissement Landau, est mis à jour dans notre étude. Expliquons cet effet de stabilisation sur notre modèle différentiel heuristique (20). L'estimation de croissance exponentielle correspond au cas d'un noyau $K(t, \tau)$ d'intégrale $O(t)$ mais *uniformément réparti* sur les temps $\tau \in [0, t]$. Considérons alors le cas exactement

³ On le voit dans ce terme (17) mais c'est un principe général tout au long de la preuve : il est crucial de garder trace des différentes échelles au cours du schéma en « stratifiant » les estimations.

opposé d'un noyau concentré (en une masse de Dirac) sur un seul temps (un seul « écho ») $K(t, \tau) = t \delta_{\theta t}(\tau)$ pour $\theta \in [0, 1]$. Ce qui conduit à l'inégalité différentielle $y'(t) \leq A + B t y(\theta t)$. Pour $\theta = 0$, on obtient une estimation excellente de croissance au plus quadratique en t . Le cas $\theta = 1$ au contraire ne donne trivialement *aucun* contrôle en temps. Le cas intermédiaire $\theta = 1/2$ donne enfin un contrôle du type $y(t) \leq \exp(C\sqrt{t})$ pour une constante $C > 0$ en temps grand. On peut alors espérer qu'une telle croissance sous-exponentielle soit compensée par l'amortissement exponentiel grâce aux légères pertes sur les indices de régularité λ_n et μ_n au cours du schéma.

On analyse alors la réponse non-linéaire due aux échos et notre heuristique est sérieusement compliquée par plusieurs difficultés : (a) il y a un nombre dénombrable d'échos θ_k , $k \in \mathbb{N}$; (b) ces derniers s'approchent asymptotiquement du point critique $\theta = 1$; (c) ces échos ne sont en fait pas exactement des masses de Dirac mais en deviennent proches en temps grand. Pour des potentiels d'interaction moins singuliers que Coulomb et Newton, on peut finalement montrer que cette réponse est sous-exponentielle, de sorte que l'on peut la contrôler par une très légère perte de régularité glissante, au prix d'une constante gigantesque, qui sera plus tard absorbée par la convergence ultra-rapide du schéma de Newton. Dans le cas critique des interactions de Coulomb et Newton, des estimations encore plus fines sont nécessaires. Pour pouvoir traiter ce cas, nous étudions la réponse non-linéaire mode par mode, c'est-à-dire en estimant $\hat{\rho}(t, k)$ pour tous les k , *via* un système infini d'inégalités. Cela permet de tirer avantage du fait que les échos qui se produisent à des fréquences différentes sont asymptotiquement plutôt bien séparés.

6. Quelques mots de conclusion entre mathématiques et physique

Au cours de la preuve nous avons dû résoudre les difficultés expliquées en introduction : réversibilité de l'équation, filamentation, non-linéarité. Les réponses ont été l'occasion de surprises et de mises à jour de liens avec d'autres résultats en mathématiques et en physique.

Tout d'abord, il est bien sûr possible de montrer l'irréversibilité d'une partie d'un système alors même que le système complet est réversible. Et dans notre cas la répartition des électrons dans le domaine suit une évolution irréversible, mais la distribution de leurs positions et vitesses suit une évolution réversible. Cela correspond mathématiquement à la notion de convergence faible. D'un point de vue de la physique cela rejoint l'idée d'irréversibilité macroscopique de Boltzmann : lorsque l'on peut déterminer l'évolution de certaines quantités moyennées d'un système de particules (un gaz par exemple) uniquement à partir d'elles-mêmes, alors on observe une irréversibilité de ces quantités quand bien même les particules, elles, suivent des lois réversibles. Cela est résumé dans sa formule de l'entropie : cette dernière est reliée au nombre d'états « microscopiques » associés à un état moyenné « macroscopique » donné, et cette quantité croît au cours du temps tant que l'équilibre n'est pas atteint.

La surprise dans notre cas est de pouvoir observer ce phénomène de retour à l'équilibre pour une quantité moyennée (le champ électrique) dont l'évolution ne dépend pas seulement de ses états précédents, mais aussi de toute la trajectoire microscopique du système. Autrement dit une équilibration sans augmentation de l'entropie. Cela est lié à la solution que nous donnons du deuxième problème,

celui de la filamentation. En fait, cette filamentation correspond exactement à la traduction visible du mélange de phase dont nous avons vu que la quantification est le cœur de la preuve. Les filaments se ramifient infiniment mais ce faisant, ils se moyennisent dans l'espace... Ils nous aident donc au lieu d'être un problème.

Pourquoi alors cette structure survit-elle à la perturbation non-linéaire? Nous pouvons ici faire un parallèle avec le théorème KAM (Kolmogorov, Arnold, Moser) en théorie des systèmes dynamiques. Dit de manière très simplifiée, ce théorème montre que si l'on perturbe (de façon régulière) un système dynamique complètement intégrable (*i.e.* dont on peut construire des variables qui le réduisent à un produit d'oscillateurs) et vérifiant une condition de non-dégénérescence, alors l'essentiel de l'espace des phases reste recouvert par des « tores invariants », soit des zones où le système évolue indéfiniment à travers un mouvement quasi-périodique. Notre théorème propose en quelque sorte un analogue de ce résultat dans le cas d'un continuum de particules. La « structure persistante » qui remplace les tores invariants est la structure de mélange de phase, et le système « complètement intégrable » autour duquel nous perturbons est l'équation de Vlasov-Poisson linéarisée. Les changements de variables successifs du schéma itératif sont ici les Ω^n et nous arrivons à montrer la convergence de ce schéma en temps fini grâce au mécanisme de transfert de régularité analytique sur les moyennes en vitesse. En ce qui concerne le temps long, les problèmes de résonances de la théorie KAM ne se posent plus au niveau de l'équation de Vlasov-Poisson car ces trajectoires sont de mesure nulle. Mais un nouveau phénomène de résonance apparaît, en lien avec la non-linéarité de l'équation aux dérivées partielles elle-même, à travers les rétroactions du système sur lui-même *via* le champ électrique, ce qui est quantifié dans le noyau K dont nous avons discuté. Et le fait que, pour des petites perturbations analytiques, ces rétroactions du plasma sur lui-même ne déstabilisent pas la structure de mélange de phase est lié au fait qu'elles fonctionnent en échos, *i.e.* avec un délai et de manière très localisée en temps.

Le problème de l'amortissement Landau mélange de façon fascinante la physique et les mathématiques. On voit ici que les mathématiques peuvent se révéler utiles pour démêler des débats en physique mais aussi, et c'est ce qui en fait la beauté, pour faire ressortir les structures sous-jacentes à différentes situations apparemment sans lien entre elles.

7. Références

- [1] AKHIEZER, A., AKHIEZER, I., POLOVIN, R., SITENKO, A., ET STEPANOV, K. *Plasma electrodynamics. Vol. I : Linear theory, Vol. II : Non-linear theory and fluctuations.* Pergamon Press, 1975 (English Edition). Translated by D. ter Haar.
- [2] BACKUS, G. Linearized plasma oscillations in arbitrary electron distributions. *J. Mathematical Phys.* 1, 3 (1960), 178–191, 559.
- [3] BATT, J., ET REIN, G. A rigorous stability result for the Vlasov-Poisson system in three dimensions. *Ann. Mat. Pura Appl. (4)* 164 (1993), 133–154.
- [4] CAGLIOTI, E., ET MAFFEI, C. Time asymptotics for solutions of Vlasov–Poisson equation in a circle. *J. Statist. Phys.* 92, 1-2 (1998), 301–323.
- [5] DEGOND, P. Spectral theory of the linearized Vlasov–Poisson equation. *Trans. Amer. Math. Soc.* 294, 2 (1986), 435–453.
- [6] GLASSEY, R., ET SCHAEFFER, J. Time decay for solutions to the linearized Vlasov equation. *Transport Theory Statist. Phys.* 23, 4 (1994), 411–453.

- [7] GLASSEY, R., ET SCHAEFFER, J. On time decay rates in Landau damping. *Comm. Partial Differential Equations* 20, 3-4 (1995), 647–676.
- [8] GUO, Y., ET REIN, G. A non-variational approach to nonlinear stability in stellar dynamics applied to the King model. *Comm. Math. Phys.* 271, 2 (2007), 489–509.
- [9] HAYES, J. N. Damping of plasma oscillations in the linear theory. *Phys. Fluids* 4 (1961), 1387–1392.
- [10] HAYES, J. N. On non-Landau damped solutions to the linearized Vlasov equation. *Nuovo Cimento (10)* 30 (1963), 1048–1063.
- [11] HEATH, R., GAMBA, I., MORRISON, P., ET MICHLER, C. A discontinuous Galerkin method for the Vlasov–Poisson system. À paraître dans *Jour. Comp. Phys.*, 2011.
- [12] HWANG, H. J., ET VELÁZQUEZ, J. J. L. On the existence of exponentially decreasing solutions of the nonlinear Landau damping problem. *Indiana Univ. Math. J.* 58, 6 (2009), 2623–2660.
- [13] LANDAU, L. On the vibration of the electronic plasma. *J. Phys. USSR* 10 (1946), 25. English translation in *JETP* 16, 574. Reproduced in *Collected papers of L.D. Landau*, edited and with an introduction by D. ter Haar, Pergamon Press, 1965, pp. 445–460; and in *Men of Physics : L.D. Landau*, Vol. 2, Pergamon Press, D. ter Haar, ed. (1965).
- [14] LEMOU, M., MÉHATS, F., ET RAPHAËL, P. Orbital stability of spherical systems. arXiv :1007.4095.
- [15] LEMOU, M., MÉHATS, F., ET RAPHAËL, P. A new variational approach to the stability of gravitational systems. *Comm. Math. Phys.* 302, 1 (2011), 161–224.
- [16] LIFSHITZ, E. M., ET PITAEVSKIÏ, L. P. *Course of theoretical physics (“Landau–Lifshits”)*. Vol. 10. Pergamon Press, Oxford, 1981. Translated from the Russian by J. B. Sykes et R. N. Franklin.
- [17] LYNDEN-BELL, D. The stability and vibrations of a gas of stars. *Mon. Not. R. astr. Soc.* 124, 4 (1962), 279–296.
- [18] MALMBERG, J., WHARTON, C., GOULD, R., ET O’NEIL, T. Plasma wave echo experiment. *Phys. Rev. Letters* 20, 3 (1968), 95–97.
- [19] MASLOV, V. P., ET FEDORYUK, M. V. The linear theory of Landau damping. *Mat. Sb. (N.S.)* 127(169), 4 (1985), 445–475, 559.
- [20] MOUHOT, C., ET VILLANI, C. Landau damping. *J. Mathematical Phys.* 51, 1 (2010), 015204, 7.
- [21] MOUHOT, C., ET VILLANI, C. On Landau damping. À paraître dans *Acta Mathematica*. Disponible en ligne <http://hal.archives-ouvertes.fr/hal-00376547/>.
- [22] SÁENZ, A. W. Long-time behavior of the electric potential and stability in the linearized Vlasov theory. *J. Mathematical Phys.* 6 (1965), 859–875.
- [23] VILLANI, C. A review of mathematical topics in collisional kinetic theory. In *Handbook of mathematical fluid dynamics*, Vol. 1. North-Holland, Amsterdam, 2002, pp. 71–305.
- [24] VILLANI, C. *Hypocoercivity*, vol. 202 of *Mem. Amer. Math. Soc.* 2009.
- [25] VLASOV, A. A. On the oscillation properties of an electron gas. *Zh. Èksper. Teoret. Fiz.*, 8 (1938), 291–318.
- [26] ZHOU, T., GUO, Y., ET SHU, C.-W. Numerical study on Landau damping. *Physica D* 157 (2001), 322–333.

Sur la route des ondelettes

Albert Cohen¹

Depuis près de trois décennies, les ondelettes se sont imposées comme un outil puissant en analyse mathématique, et dans des domaines appliqués tels que le traitement du signal et de l'image, les statistiques ou encore la simulation numérique. La construction de cet outil est en quasi-totalité effectuée durant la décennie 1980-1990, tout comme l'identification de ses principales propriétés. L'attribution du prix Gauss à Yves Meyer nous donne l'occasion de nous repencher sur cette floraison scientifique, à l'interface entre l'analyse harmonique et le calcul numérique, dont il fut à la fois l'artisan, le chef d'orchestre, et le critique avisé. Si certaines des promesses formulées dans l'enthousiasme des premières découvertes n'ont pas toujours été suivies des succès escomptés, d'autres développements actuels inattendus et spectaculaires doivent beaucoup à ce moment fécond.

1. Représenter les fonctions : de Fourier à Gabor

Analyser, reconstruire et représenter des fonctions quelconques à l'aide de fonctions élémentaires, parfois surnommées *atomes*, est une démarche scientifique à l'origine de nombreuses avancées fondamentales et appliquées depuis plusieurs siècles. La possibilité offerte plus récemment par l'ordinateur d'implémenter une telle démarche au moyen d'algorithmes performants lui confère un intérêt supplémentaire pour le calcul numérique.

Dans l'exemple fondateur de la *transformée de Fourier*, on effectue l'analyse en fréquence d'une fonction f , définie sur \mathbb{R} par la formule

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t)e^{-i\omega t} dt.$$

Sous des hypothèses convenables sur f , la fonction \hat{f} est bien définie et permet la *synthèse* de f par la formule d'inversion de Fourier

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega)e^{i\omega t} d\omega.$$

Dans cet exemple, les atomes sont les oscillations

$$e_{\omega}(t) = e^{i\omega t},$$

aux différentes fréquences $\omega \in \mathbb{R}$, et la combinaison de ces fonctions affectées des poids $\hat{f}(\omega)$ permet de reconstruire f . Les atomes e_{ω} servent aussi à l'analyse, puisque l'on peut formellement écrire $\hat{f}(\omega)$ comme un *produit scalaire*

$$\hat{f}(\omega) = \langle f, e_{\omega} \rangle,$$

avec la notation $\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t)\overline{g(t)} dt$. La fonction \hat{f} est une *représentation* de f au sens où la connaissance de celle-ci est équivalente à celle de f .

¹ Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris.

Les mêmes remarques s'appliquent aux *séries de Fourier* pour l'analyse des fonctions définies sur un intervalle borné : si f est définie sur $[a, b]$, on étudie sa décomposition sous la forme de la série de fonctions de période $T = b - a$,

$$f(t) = \sum_{n \in \mathbb{Z}} c_n e^{i \frac{2\pi n t}{T}}.$$

Les coefficients de Fourier c_n peuvent être vus comme les *coordonnées* de f dans la base orthogonale de l'espace de Hilbert $L^2([a, b])$ constituée des fonctions $e_n(t) = e^{i 2\pi n t / T}$, et sont donnés par les produits scalaires

$$c_n = c_n(f) = \frac{1}{T} \int_a^b f(t) e^{-i \frac{2\pi n t}{T}} dt.$$

Il est bien connu que le *calcul approché* était l'une des motivations à l'origine des séries de Fourier, l'approximation de f par la somme partielle $\sum_{n=-N}^N c_n e^{i 2\pi n t / T}$ pouvant servir à l'évaluation de solutions d'équations différentielles ou aux dérivées partielles. Les applications des représentations de Fourier se sont intensifiées avec l'apparition des ordinateurs et le développement du calcul numérique, ainsi que l'invention de l'algorithme de transformée de Fourier rapide, permettant de calculer en $\mathcal{O}(N \log N)$ opérations la version discrète de la transformée de Fourier, où N représente la taille du vecteur représentant la fonction discrétisée.

Du point de vue du calcul, la représentation d'une fonction f est d'autant plus efficace qu'elle *concentre* sur un petit nombre de valeurs numériques l'essentiel de l'information permettant de reconstruire cette fonction avec précision. Les représentations de Fourier sont ainsi efficaces pour des combinaisons linéaires d'oscillations de la forme $\sum_{n=1}^N a_n e_{\omega_n}(t)$, puisque leur transformée de Fourier est une combinaison de masses de Dirac aux fréquences ω_n correspondantes, ainsi que pour les fonctions uniformément régulières : l'application itérée de la formule

$$\widehat{f'}(\omega) = (-i\omega) \widehat{f}(\omega)$$

montre que la régularité de f se traduit par la décroissance à l'infini de \widehat{f} .

Ces propriétés de concentration disparaissent lorsque f est régulière sauf en un point isolé de discontinuité, une telle fonction étant perçue comme globalement peu régulière par l'analyse en fréquence. Les transformées de Fourier sont aussi mises en échec lorsque l'on cherche à identifier des fréquences qui apparaissent uniquement dans certains intervalles de temps, telles des notes dans une partition de musique. Ces limitations traduisent le caractère global des transformées de Fourier : toutes les valeurs de f influencent toutes les valeurs de \widehat{f} et réciproquement.

Une première idée pour *localiser* l'analyse en fréquence est de multiplier la fonction f par une fonction g régulière et bien localisée, par exemple la gaussienne $g(t) = e^{-t^2}$ ainsi que le propose Dennis Gabor en 1945, et par ses translatées $g(t - \tau)$, avant d'appliquer la transformée de Fourier. On obtient ainsi la *transformée de Fourier locale* (ou à *fenêtre glissante*)

$$Gf(\omega, \tau) = \int_{-\infty}^{+\infty} f(t) g(t - \tau) e^{-i\omega t} dt,$$

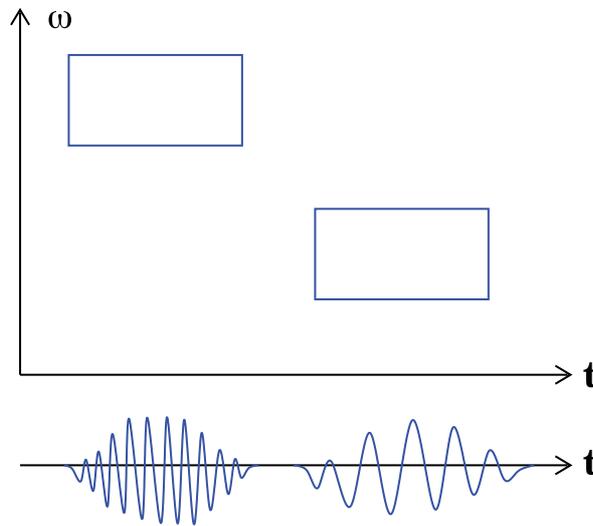


FIG. 1. Atomes temps-fréquence

qui mesure l'intensité de la fréquence ω dans un voisinage de l'instant τ . Elle peut aussi s'écrire $Gf(\omega, \tau) = \langle f, g_{\omega, \tau} \rangle$, où les atomes d'analyse sont donnés par

$$g_{\omega, \tau}(t) = g(t - \tau)e^{-i\omega t}, \quad \omega, \tau \in \mathbb{R}.$$

Remarquons que l'on a $\hat{g}_{\omega, \tau}(\xi) = e^{-i\tau\xi}\hat{g}(\xi - \omega)$, et que \hat{g} est aussi une gaussienne : les atomes sont à la fois bien localisées dans le temps autour de l'instant t et dans le domaine fréquentiel autour de la fréquence ω . On peut les visualiser symboliquement comme des rectangles de formes fixes, localisés à divers emplacements du plan temps-fréquence, repéré par les axes (ω, t) (Figure 1), les dimensions horizontales et verticales du rectangle correspondant aux variances de g et \hat{g} . Le principe d'incertitude limite cette double localisation par une aire minimale imposée à ces rectangles. On peut reconstruire f à partir de sa transformée Gf grâce à la formule

$$f(t) = C \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Gf(\omega, \tau) e^{i\omega t} d\omega d\tau \quad \text{avec} \quad C = \left(\int_{-\infty}^{+\infty} g(t) dt \right)^{-1},$$

qui est une conséquence immédiate de la formule d'inversion de Fourier.

2. Transformées en ondelettes et frames

La résolution temporelle de la transformée de Fourier locale est limitée par l'échelle du support de g : la détection d'une oscillation localisée sur un intervalle de temps significativement plus petit se heurte inévitablement au même problème que dans le cas de la transformée de Fourier globale.

Au début des années 1980, Jean Morlet propose une solution différente : partant d'une fonction ψ bien localisée dans le plan temps-fréquence et oscillante au sens

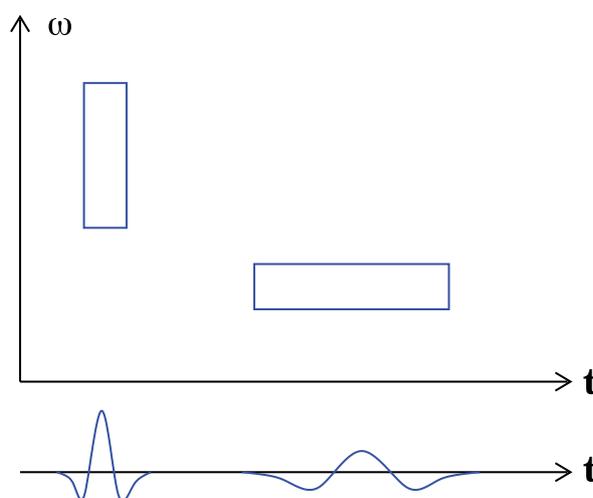


FIG. 2. Ondelettes

où $\int_{-\infty}^{+\infty} \psi(t) dt = 0$, il construit une famille de fonctions analysantes $\psi_{a,b}$, dites *ondelettes*, au moyen de translations et dilatations

$$\psi_{a,b}(t) = a^{-1/2} \psi\left(\frac{t-b}{a}\right), \quad a > 0, \quad b \in \mathbb{R}.$$

Le facteur $a^{-1/2}$ normalise cette famille, au sens où $\|\psi_{a,b}\|_{L^2}$ est indépendante de a et b . La transformée en ondelettes de f est alors définie par

$$Wf(a, b) := \langle f, \psi_{a,b} \rangle = a^{-1/2} \int_{-\infty}^{+\infty} f(t) \psi\left(\frac{t-b}{a}\right) dt.$$

La formule de reconstruction

$$f(t) = C \int_0^{+\infty} \int_{-\infty}^{+\infty} Wf(a, b) \psi_{a,b}(t) \frac{db da}{a^2} \quad \text{avec} \quad C = \left(\int_{-\infty}^{+\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega \right)^{-1},$$

conjecturée par Jean Morlet est démontrée en 1982 par Alexandre Grossman, qui l'identifie comme un cas particulier d'une formule de résolution de l'identité en théorie des représentations des groupes.

Par contraste avec les transformées précédentes, le paramètre d'échelle a ouvre l'accès à l'analyse de phénomènes d'oscillation ou de régularité arbitrairement localisés en temps au prix d'une perte de localisation en fréquence : quand a tend vers 0, les ondelettes $\psi_{a,b}$ sont visualisées par des rectangles fins en temps de l'ordre de l'échelle a et longs en fréquences de l'ordre de $1/a$ (Figure 2).

Dans le cadre du calcul numérique, on est naturellement amené à *échantillonner* les transformées Gf et Wf . Dans le cas de la transformée de Fourier locale, il est assez clair que ceci peut se faire sans perte d'information sur f : si g est à support dans un intervalle $[a, b]$, positive et ne s'annule pas sur un intervalle fermé de longueur t_0 contenu dans $[a, b]$, on peut caractériser f par la donnée de toutes les fonctions $f_m(t) = f(t)g(t - mt_0)$ pour $m \in \mathbb{Z}$, puisque $f(t) =$

$(\sum_{m \in \mathbb{Z}} f_m(t)) / (\sum_{m \in \mathbb{Z}} g(t - mt_0))$, où le dénominateur est strictement positif. Les fonctions f_m sont elles mêmes caractérisées par la donnée de leurs coefficients de Fourier

$$c_n(f_m) = Gf(mt_0, n\omega_0) = \langle f, g_{n\omega_0, mt_0} \rangle, \quad n \in \mathbb{Z},$$

avec $\omega_0 = 2\pi/(b - a)$. L'échantillonnage naturel de Gf est ainsi donné par le réseau $(mt_0, n\omega_0)_{m, n \in \mathbb{Z}}$, ce qui correspond à effectuer un recouvrement du plan temps-fréquence par les rectangles associés aux atomes

$$g_{n, m} := g_{n\omega_0, mt_0}, \quad n, m \in \mathbb{Z}.$$

Dans le cas des ondelettes, un tel recouvrement est plus naturellement associé à un réseau du type $(a_0^n, mb_0 a_0^{-n})_{n, m \in \mathbb{Z}}$, avec $a_0 > 1$ et $b_0 > 0$ fixés, qui prend en compte le changement des formes des rectangles en fonction de a . Pour ψ , a_0 et b_0 fixés, on définit ainsi la famille

$$\psi_{n, m} := \psi_{a_0^n, mb_0 a_0^{-n}}, \quad n, m \in \mathbb{Z},$$

et l'on cherche à comprendre si l'échantillonnage $Wf(a_0^{-n}, mb_0 a_0^{-n}) = \langle f, \psi_{n, m} \rangle$ permet de caractériser f et de la reconstruire.

La réponse à ce problème est liée à la notion suivante : une suite $(f_n)_{n \geq 0}$ dans un espace de Hilbert H est appelée *frame*, si et seulement si il existe deux constantes $0 < A \leq B$ telles que pour tout $f \in H$,

$$A\|f\|^2 \leq \sum_{n \geq 0} |\langle f, f_n \rangle|^2 \leq B\|f\|^2.$$

Un tel encadrement montre que la suite des produits scalaires $c_n = \langle f, f_n \rangle$ caractérise f de façon stable au sens où la norme de f dans H est équivalente à la norme hilbertienne discrète de la suite $(c_n)_{n \geq 0}$. Il signifie aussi que l'opérateur F qui associe à x la suite (c_n) est continu et inversible à gauche, ce qui permet de construire une deuxième suite $\tilde{f}_n = (F^*F)^{-1}f_n$, dite *frame dual*, qui conduit à la formule de reconstruction

$$f = \sum_{n \geq 0} \langle f, f_n \rangle \tilde{f}_n,$$

que l'on peut implémenter numériquement. Dans le cas des ondelettes, on cherche donc à savoir si la famille $(\psi_{n, m})_{n, m \in \mathbb{Z}}$ est un frame pour l'espace de Hilbert $L^2(\mathbb{R})$. En 1984, Ingrid Daubechies montre que ceci est vrai dès que le réseau d'échantillonnage est suffisamment fin, c'est-à-dire si $a_0 b_0 < K$ où K est une constante qui dépend de ψ .

3. Les bases d'ondelettes d'Yves Meyer

La décomposition d'une fonction f dans un frame est généralement *redondante*, au sens où on n'a pas imposé l'indépendance linéaire des atomes f_n , par contraste avec une base de fonctions. Ceci constitue un défaut lorsque l'on souhaite obtenir une représentation la plus économique possible, ce qui est souvent le cas dans les applications numériques. La construction des bases orthonormées d'ondelettes proposée par Yves Meyer en 1985 est un moment clé qui fait le lien entre l'évolution décrite précédemment et d'autres programmes scientifiques.

Il faut d'abord rappeler que l'existence d'une telle base était connue depuis le début du 20^e siècle : il s'agit du système introduit par Alfred Haar dans sa thèse et qui a la forme

$$\psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k), \quad j, k \in \mathbb{Z},$$

où l'ondelette génératrice

$$\psi = \chi_{[0,1/2[} - \chi_{[1/2,1[},$$

est la fonction constante par morceau qui vaut 1 sur $[0, 1/2[$ et -1 sur $[1/2, 1[$. Il est immédiat de constater que ces fonctions sont orthogonales entre elles. Le fait qu'elles constituent une base de $L^2(\mathbb{R})$ est aussi facile à vérifier, en introduisant pour $f \in L^2(\mathbb{R})$ ses approximations constantes par morceaux

$$P_j f(t) = 2^j \int_{I_{j,k}} f(s) ds, \quad t \in I_{j,k} := [2^{-j}k, 2^{-j}(k+1)[, \quad k \in \mathbb{Z}.$$

On remarque que la projection sur les ondelettes de résolution 2^{-j} coïncide avec la différence de deux approximations successives

$$Q_j f = P_{j+1} f - P_j f = \sum_{k \in \mathbb{Z}} \langle f, \psi_{j,k} \rangle \psi_{j,k},$$

et on conclut par le fait que $\lim_{j \rightarrow +\infty} P_j f = f$ et $\lim_{j \rightarrow -\infty} P_j f = 0$. Le défaut principal du système de Haar est son manque de régularité puisque ψ n'est pas continue, ce qui équivaut à un manque de localisation de $\hat{\psi}$.

Un autre exemple élémentaire de base d'ondelettes $(\psi_{j,k})_{j,k \in \mathbb{Z}}$ est obtenu en définissant la fonction ψ par sa transformée de Fourier

$$\hat{\psi}(\omega) = \chi_{]-2\pi, -\pi]}(\omega) + \chi_{[\pi, 2\pi[}(\omega)$$

qui vaut 1 sur les intervalles $]-2\pi, -\pi]$ et $[\pi, 2\pi[$, et 0 ailleurs. Comme pour le système de Haar, l'orthogonalité des fonctions $\psi_{j,k}$ est facile à établir, et le fait qu'elles constituent une base se montre de façon analogue mais en utilisant des approximations successives définies par troncature en fréquence

$$\widehat{P_j f}(\omega) = \hat{f}(\omega) \chi_{[-2^j \pi, 2^j \pi]}(\omega).$$

La fonction ψ est parfois appelée ondelette de *Shannon* par référence au théorème de Shannon-Nyquist qui caractérise le pas d'échantillonnage permettant de reconstruire les fonctions à support limité en fréquence au moyen d'une base orthonormée. L'ondelette de Shannon est, à l'inverse de celle de Haar, parfaitement localisée en fréquence mais mal localisée en temps.

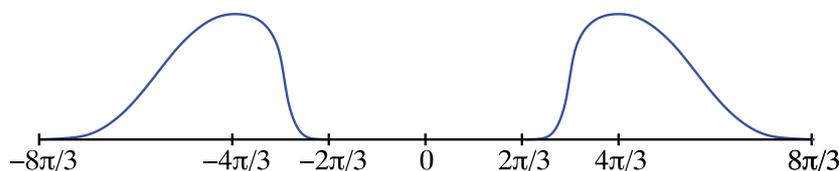


FIG. 3. L'ondelette d'Yves Meyer (graphe de $|\hat{\psi}|$)

L'existence d'une base orthonormée d'ondelettes engendrée par une fonction ψ à la fois bien localisée en temps et en fréquence semble cependant loin d'être acquise, car une obstruction existe pour le même problème dans le cas des atomes temps-fréquence $g_{m,n}$ issus de la discrétisation de la transformée de Gabor. Un théorème dû à Roger Balian et Francis Low affirme en effet que ces atomes ne peuvent constituer une base orthonormée que si l'un des moments quadratiques $\int_{-\infty}^{+\infty} |t|^2 |g(t)|^2 dt$ ou $\int_{-\infty}^{+\infty} |\omega|^2 |\hat{g}(\omega)|^2 d\omega$ est divergent. C'est en se posant la question d'un résultat similaire pour les ondelettes qu'Yves Meyer parviendra finalement à construire une base avec une fonction ψ dont les moments de tous les ordres convergent ainsi que ceux de $\hat{\psi}$. L'ondelette proposée à la forme d'une version régularisée en fréquence de l'ondelette de Shannon (Figure 3) : la fonction $\hat{\psi}$ est infiniment régulière, à support dans $[-8\pi/3, -2\pi/3] \cup [2\pi/3, 8\pi/3]$, son module est une fonction paire vérifiant

$$\sum_{n \in \mathbb{Z}} |\hat{\psi}(\omega + 2n\pi)|^2 = 1 \text{ et } \sum_{j \in \mathbb{Z}} |\hat{\psi}(2^j \omega)|^2 = 1.$$

On vérifie aisément par la formule de Poisson que la première relation assure l'orthogonalité des $\psi_{j,k}$ pour un niveau d'échelle j fixé. Un choix astucieux de la phase suivant $e^{i\omega/2}$ pour définir $\hat{\psi}(\omega)$, assure l'orthogonalité entre les échelles.

Une autre base orthonormée d'ondelettes bien localisée en temps et en fréquence avait déjà été construite en 1981 par Jan Olov Stromberg, en suivant une approche tout à fait différente, plus proche de celle qui est discutée dans la section suivante.

Dans sa conception, l'ondelette d'Yves Meyer s'inspire d'une *décomposition dyadique*

$$f = \sum_{j \in \mathbb{Z}} \Delta_j f$$

proposée par John Littlewood et Raymond Paley dans les années 1930 pour l'analyse harmonique des espaces L^p . Les blocs dyadiques de la décomposition de Littlewood-Paley sont définis par filtrage de f aux fréquences de l'ordre de 2^j

$$\widehat{\Delta_j f}(\omega) = \hat{\Psi}(2^{-j}\omega) \hat{f}(\omega),$$

où la fonction $\hat{\Psi}$ est du même type que $|\hat{\psi}|$. Dans la décomposition en ondelettes, les blocs $\Delta_j f$ sont remplacés par les projections

$$Q_j f := \sum_{k \in \mathbb{Z}} \langle f, \psi_{j,k} \rangle \psi_{j,k},$$

l'indice k apportant ainsi un niveau supplémentaire de discrétisation.

L'analyse de Littlewood-Paley permet de caractériser avec précision les propriétés de régularité de fonctions, et Yves Meyer montre qu'il en est de même des décompositions en ondelettes. Par exemple, si α est un nombre positif non-entier et $f \in L^\infty(\mathbb{R})$, on peut montrer que les trois propriétés suivantes sont équivalentes :

(1) La fonction f est uniformément Hölderienne d'exposant α sur \mathbb{R} , c'est-à-dire que pour tout $t \in \mathbb{R}$ il existe un polynôme P_t de degré $m < \alpha$ tel que

$$|f(s) - P_t(s)| \leq A|s - t|^\alpha,$$

où $A > 0$ est indépendant de s et t .

(2) Les blocs de Littlewood-Paley de f vérifient

$$\|\Delta_j f\|_{L^\infty} \leq B 2^{-\alpha j},$$

où $B > 0$ est indépendante de j .

(3) Les coefficients d'ondelettes de f vérifient

$$|\langle f, \psi_{j,k} \rangle| \leq C 2^{-(\alpha + \frac{1}{2})j},$$

où $C > 0$ est indépendante de j et k .

On retrouve ainsi avec plus de précision l'idée, déjà présente dans l'analyse de Fourier, que la régularité se traduit par de la décroissance dans la représentation. La présence de l'indice k dans la décomposition en ondelettes permet en outre d'apporter une information supplémentaire sur la régularité *locale* de f . Nous reviendrons un peu plus loin sur ce point important.

La possibilité de caractériser l'appartenance d'une fonction à la quasi-totalité des espaces fonctionnels classiques (Hölder, Lebesgue, Sobolev, Besov) par des conditions portant uniquement sur les *modules* des coefficients de cette fonction dans une base d'ondelettes traduit le fait que celle-ci a la propriété de *base inconditionnelle* pour ces espaces. Par définition, une base $(f_n)_{n \geq 0}$ d'un espace de Banach X est inconditionnelle si et seulement si il existe une constante $C > 0$ telle que pour tout $N > 0$, on ait la propriété

$$|c_n| \leq |d_n|, \quad n = 0, \dots, N \Rightarrow \left\| \sum_{n=0}^N c_n f_n \right\|_X \leq C \left\| \sum_{n=0}^N d_n f_n \right\|_X.$$

Cette propriété est fondamentale pour les applications numériques car elle assure la stabilité dans la norme $\|\cdot\|_X$ d'opérations telles que la mise à zéro de certains coefficients ou leur approximation à une certaine précision par l'ordinateur.

4. Analyse multirésolution et filtres numériques

Dans le domaine du traitement numérique du signal et de l'image, les ingénieurs utilisent depuis les années 1970 des décompositions hiérarchiques présentant des similarités avec l'analyse de Littlewood-Paley. Ainsi, partant d'une image discrétisée sur une grille de pixels, on peut considérer ses approximations obtenues en moyennant successivement l'intensité lumineuse sur des carrés de 2×2 , 4×4 , 8×8 pixels etc. De telles approximations sont *hiérarchisées* au sens où l'image approchée à une certaine résolution se déduit de celle approchée à la résolution deux fois plus fine, et on cherche alors à caractériser l'information manquante entre chaque niveau.

En 1986, Stéphane Mallat et Yves Meyer remarquent que ce point de vue conduit naturellement aux bases d'ondelettes, à travers la notion fondamentale d'*analyse multirésolution*. Cette notion désigne une suite emboîtée d'espaces

$$\{0\} \subset \dots \subset V_{j-1} \subset V_j \subset V_{j+1} \subset \dots \subset L^2(\mathbb{R}),$$

l'entier j étant associé à la résolution 2^{-j} au sens où V_j est engendré par une base du type

$$\varphi_{j,k}(t) = 2^{j/2} \varphi(2^j t - k), \quad k \in \mathbb{Z}.$$

La fonction φ est appelée *fonction d'échelle*. Dans le cas des espaces de fonctions constantes par morceaux qui mène au système de Haar, on a tout simplement $\varphi = \chi_{[0,1]}$, c'est-à-dire $\varphi(t) = 1$ sur $[0, 1]$ et 0 ailleurs. On désigne par P_j la projection orthogonale sur V_j et on exige que pour tout $f \in L^2(\mathbb{R})$,

$$\lim_{j \rightarrow -\infty} \|P_j f\|_{L^2} = 0 \text{ et } \lim_{j \rightarrow +\infty} \|f - P_j f\|_{L^2} = 0.$$

La deuxième propriété signifie que les espaces V_j permettent d'approcher n'importe quelle fonction lorsque l'on fait tendre la résolution 2^{-j} vers 0.

En pratique, on s'intéresse souvent à d'autres types de convergence de $P_j f$ vers f (par exemple la convergence uniforme), et on souhaite de plus avoir des informations sur la vitesse de cette convergence. Dans le cas des fonctions constantes par morceaux, il est facile de vérifier que si f est suffisamment régulière au sens où $f' \in L^p(\mathbb{R})$, on a

$$\|f - P_j f\|_{L^p} \leq 2^{-j} \|f'\|_{L^p},$$

et qu'une telle estimation ne peut être améliorée même si f est beaucoup plus régulière. Ceci reflète le fait que la précision de l'approximation par des fonctions constantes est d'ordre 1 en fonction de la résolution. Afin d'augmenter l'ordre d'approximation, une idée naturelle est de remplacer les constantes par des polynômes de degré plus élevé, ce qui permet aussi d'augmenter la régularité des fonctions de V_j . On peut ainsi définir V_j comme le sous-espace de $L^2(\mathbb{R})$ constitué des fonctions polynômiales de degré m sur les intervalles $I_{j,k}$ définis plus haut, et globalement $m - 1$ fois continûment dérivables : ce sont des *fonctions splines* de degré m , bien connues des ingénieurs pour la modélisation numérique des courbes et surfaces. Dans ce cas, on peut prendre pour fonction d'échelle φ la *B-spline* obtenue en effectuant $m + 1$ produits de convolution de $\chi_{[0,1]}$ par elle-même

$$\varphi = \chi_{[0,1]} * \dots * \chi_{[0,1]} = (*)^{m+1} \chi_{[0,1]}.$$

On note qu'en dehors du cas $m = 0$, la base $(\varphi_{j,k})_{k \in \mathbb{Z}}$ n'est pas orthonormée.

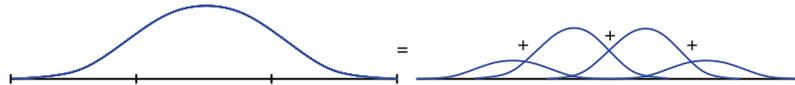


FIG. 4. B-spline quadratique
 $\varphi(t) = \frac{3}{4}(\varphi(2t - 1) + \varphi(2t - 2)) + \frac{1}{4}(\varphi(2t) + \varphi(2t - 3))$

Partant d'une analyse multirésolution, on cherche à présent à construire une ondelette ψ permettant de caractériser l'information manquante $P_{j+1} f - P_j f$ entre deux niveaux d'approximation, au moyen des fonctions $(\psi_{j,k})_{k \in \mathbb{Z}}$. Une remarque essentielle pour cette construction est que la fonction d'échelle vérifie une équation d'autosimilarité

$$\varphi(t) = \sum_{n \in \mathbb{Z}} h_n \varphi(2t - n),$$

qui exprime l'emboîtement des espaces V_j . On peut vérifier, sous des hypothèses minimales, que le support de la suite h_n est de même taille que celui de la fonction φ . Par exemple, dans le cas de la B-spline de degré m , on a $\text{Supp}(\varphi) = [0, m + 1]$ et

$h_n = 2^{-m} \binom{m+1}{n}$ pour $n = 0, \dots, m+1$ et $h_n = 0$ sinon. La Figure 4 illustre ainsi cette équation dans le cas de la B-spline quadratique ($m = 2$).

Yves Meyer démontre que si la base $(\varphi_{j,k})_{k \in \mathbb{Z}}$ de V_j est orthonormée, on peut définir l'ondelette ψ par la relation

$$\psi(t) = \sum_{n \in \mathbb{Z}} g_n \varphi(2t - n),$$

avec $g_n = (-1)^n h_{1-n}$, les fonctions $(\psi_{j,k})_{k \in \mathbb{Z}}$ obtenues par changement d'échelle constituant alors une base du supplémentaire orthogonal de V_j dans V_{j+1} . Par le même raisonnement que pour le système de Haar et l'ondelette de Shannon on aboutit ainsi à une base orthonormée $(\psi_{j,k})_{j,k \in \mathbb{Z}}$ de $L^2(\mathbb{R})$.

Le problème qui subsiste est la construction de fonctions d'échelles φ telles que $(\varphi_{j,k})_{k \in \mathbb{Z}}$ constitue une base orthonormée, et ayant des propriétés intéressantes pour les applications : support compact, régularité, ordre élevé d'approximation pour les espaces V_j correspondants. On peut chercher à utiliser les splines en construisant une nouvelle fonction d'échelle dont les translatés sont orthogonaux, mais le défaut de cette construction est que la fonction φ obtenue n'est plus à support compact, et par conséquent plus difficile à manipuler numériquement que la B-spline de départ.

Ce problème est finalement entièrement résolu par Ingrid Daubechies en 1988 dans un travail qui fait l'unanimité pour sa pertinence et sa profondeur. L'approche suivie consiste à définir directement la fonction φ comme une solution de l'équation d'échelle, et de mettre au point les coefficients h_n dans cette équation de façon à assurer les propriétés requises sur la solution φ . Le tour de force accompli par Ingrid Daubechies est à la fois d'identifier la nature de ces propriétés sous forme d'équations algébriques portant sur les h_n et de résoudre ces équations. Elle construit pour tout m une suite de coefficients $(h_n)_{n=0, \dots, 2m-1}$, telle que la solution φ de l'équation d'échelle, à support compact dans $[0, 2m-1]$, engendre une base orthonormée $(\varphi_{j,k})_{k \in \mathbb{Z}}$ d'espaces d'approximation d'ordre m , et montre en outre que la régularité de la fonction φ tend vers $+\infty$ avec m . Le cas $m = 1$ correspond simplement au système de Haar. Pour $m > 1$, les fonctions φ et ψ n'ont pas d'expression mathématique explicite (contrairement aux splines), et sont déterminées implicitement par la donnée des coefficients h_n . Cette construction est ensuite étendue par Albert Cohen, Ingrid Daubechies et Jean-Christophe Feauveau en 1992, au cadre plus général des bases *biorthogonales*, qui est le plus utilisé en pratique en raison de sa flexibilité (on peut ainsi en particulier obtenir des ondelettes splines à support compact).

L'introduction de l'analyse multirésolution donne non seulement un éclairage nouveau sur la construction des bases d'ondelettes, mais elle permet aussi par sa flexibilité d'adapter cette construction à de nombreuses situations géométriques : on sait ainsi construire des analyses multirésolutions et des ondelettes sur des domaines multidimensionnels (par exemple en partant d'espaces d'éléments finis emboîtés), ou plus généralement sur certains types de variétés.

Le cadre des approximations hiérarchiques joue en outre un rôle essentiel pour la mise en œuvre d'algorithmes permettant le calcul des coefficients d'ondelettes en partant de données discrétisées à un certain niveau de résolution 2^{-j} . Ces données sont assimilées à une fonction d'un espace V_j et on calcule de proche en proche les

coefficients aux échelles $j' \leq j$ en allant vers la résolution plus grossière. Ces algorithmes utilisent les suites h_n et g_n intervenant dans les équations d'échelles puisque ce sont celles-ci qui relient les ondelettes et fonctions d'échelles aux différentes résolutions. Ces suites jouent le rôle de filtres numériques, et les algorithmes de décomposition obtenus sont du même type que ceux développés par les ingénieurs dans le domaine du traitement du signal et de l'image. De même, les algorithmes de reconstruction, qui remontent vers la résolution plus fine, se révèlent être du même type que les *schémas de subdivision* proposés en modélisation géométrique pour générer des courbes et surfaces à partir d'un petit nombre de points de contrôle par des raffinement itératifs.

5. Du traitement du signal à la simulation numérique

En à peine dix ans, la théorie des ondelettes est ainsi forgée dans toute sa variété, grâce à des apports provenant de nombreux domaines de recherche : analyse harmonique, traitement du signal et de l'image, théorie de l'approximation, modélisation des courbes et surfaces. Le premier grand congrès dédié aux ondelettes (Luminy 1989) témoigne de cette diversité thématique et de l'enthousiasme suscité par l'émergence d'un domaine permettant de telles rencontres, ainsi que par la multiplicité des applications envisagées.

En premier lieu, c'est la capacité des ondelettes à analyser les phénomènes localisés, à la manière d'un microscope mathématique, qui intéresse pour l'étude de signaux provenant de phénomènes acoustiques, électromagnétiques, économiques ou biologiques. Un effort particulier est porté sur l'analyse de la turbulence développée, autour de spécialistes tels que Uriel Frisch et Marie Farge, avec l'espoir de comprendre plus finement les lois de transfert d'énergie à travers les échelles.

Revenons en particulier sur la caractérisation de la régularité Hölderienne d'une fonction par la décroissance des coefficients d'ondelettes en fonction du niveau d'échelle suivant l'estimation

$$|\langle f, \psi_{j,k} \rangle| \leq C 2^{-(\alpha + \frac{1}{2})j}.$$

Cette estimation est vérifiée uniformément sur j et k si la fonction f est uniformément Hölderienne d'exposant α sur tout \mathbb{R} , et en supposant que l'ondelette vérifie les relations de moments nuls

$$\int_{-\infty}^{+\infty} t^n \psi(t) dt = 0,$$

pour tout entier $n < \alpha$. Si l'ondelette ψ est à support compact, il est facile de montrer que si f est Hölderienne d'exposant α seulement en un point t , alors l'estimation ci-dessus sera assurée pour les indices j et k tels que $|2^j t - k| \leq M$ où M est une constante arbitraire (mais dont la taille influence la constante C), c'est-à-dire pour les ondelettes dont les supports se concentrent vers le point t lorsque la résolution j augmente. En 1991, Stéphane Jaffard montre une réciproque à ce résultat sous une hypothèse de régularité uniforme minimale. Il est ainsi possible de caractériser la *régularité locale* d'une fonction par la taille de ses coefficients d'ondelettes.

Cette propriété ouvre des perspectives nouvelles en *analyse multifractale*, domaine dans lequel on cherche à quantifier avec précision la taille des ensembles de

points où une fonction f donnée présente une certaine régularité. On s'intéresse en particulier aux propriétés du *spectre de singularité* de f qui est la fonction $\alpha \mapsto d(\alpha)$ mesurant la dimension de Hausdorff de l'ensemble des points t où l'exposant de Hölder α est atteint. L'observation des signaux turbulents amène Uriel Frisch et Giorgio Parisi à proposer une conjecture appelée *formalisme multifractal*. Cette conjecture relie le spectre de singularité avec la transformée de Legendre de la fonction $s \mapsto s(p)$ qui mesure l'ordre de dérivabilité de f dans l'espace L^p (c'est-à-dire la valeur maximale de s telle que f admet une dérivée fractionnaire d'ordre s dans L^p). L'utilisation des bases d'ondelettes permet à Stéphane Jaffard de montrer en toute rigueur que le formalisme multifractal est mis en défaut par certaines fonctions, le spectre de singularité pouvant avoir une forme quelconque, mais que sa validité est néanmoins générique au sens de Baire sur les espaces fonctionnels classiques.

Si les ondelettes permettent d'identifier un point régulier dans un environnement moins régulier, c'est surtout la propriété inverse qui intéresse les numériciens : si f est une fonction régulière par morceaux présentant des singularités en des points isolés, alors ses coefficients d'ondelettes décroissent rapidement en fonction du niveau de résolution j , à l'exception d'une minorité d'entre eux pour lesquels le support de $\psi_{j,k}$ rencontre les points de singularité. La Figure 6, illustre cette propriété dans le cas où la fonction décomposée est une image : les coefficients d'ondelettes, représentés par l'image du centre, sont presque tous nuls ou très proche de 0, à l'exception de ceux qui sont localisés au voisinage des contours de l'image. Ces contours correspondent à des discontinuités de l'intensité lumineuse localisées sur des courbes isolées.

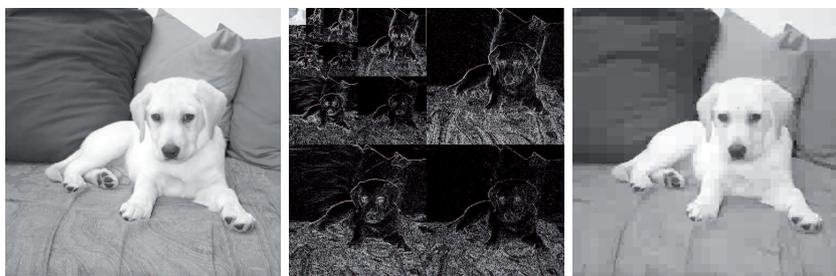


FIG. 5. Image 512×512 pixels, coefficients d'ondelettes, compression Haar 1/100

Cette propriété de concentration de l'information sur un petit nombre de coefficients – souvent appelée *parcimonie* ou *sparsity* – signifie qu'il est possible de reconstruire avec précision l'image en ne conservant qu'un petit nombre de coefficients obtenus par une procédure de *seuillage*. L'image de droite sur la Figure 5 montre ainsi la reconstruction obtenue à partir des 2500 plus grands coefficients dans la base de Haar (adaptée aux images bidimensionnelles), soit environ 1/100 du nombre total de coefficients d'ondelettes qui est égal à la dimension 512×512 de l'image de départ. On constate des artefacts visuels, liés à l'utilisation du système de Haar et qui disparaissent avec des ondelettes plus régulières, mais qui permettent

ici d'illustrer le fait que la résolution est *localement adaptée* et se raffine au voisinage des contours puisque les coefficients aux résolutions fines y ont été conservés par la procédure de seuillage.

On voit ainsi apparaître un lien entre les procédures de seuillage des coefficients d'ondelettes et le raffinement de maillage. Ces procédures sont par essence *adaptatives* et *non-linéaires*, l'ensemble des coefficients ou le maillage retenu variant selon la fonction que l'on approche. La *théorie de l'approximation non-linéaire*, développée autour de Ronald DeVore à partir des années 1980, offre un cadre général pour analyser leurs performances. Un résultat élémentaire de cette théorie propose une mesure mathématique du degré de parcimonie d'une représentation dans une base hilbertienne : si $\sigma_N(f)$ désigne l'erreur en norme L^2 obtenue si l'on ne conserve que les N plus grands coefficients de f , et si $(d_n)_{n>0}$ est la suite des modules des coefficients réarrangés par ordre décroissant, on a l'équivalence

$$\sup_{N>0} N^s \sigma_N(f) < \infty \Leftrightarrow \sup_{n>0} n d_n^p < \infty \quad \text{avec} \quad \frac{1}{p} = \frac{1}{2} + s.$$

La première estimation signifie que $\sigma_N(f)$ décroît comme N^{-s} , et la deuxième affirme que la suite des coefficients est *faiblement* ℓ^p sommable. L'espace $w\ell^p$ des suites faiblement ℓ^p sommables est proche mais plus grand que l'espace ℓ^p (par exemple, la suite $1/n$ appartient à $w\ell^1(\mathbb{N})$ mais sa série diverge logarithmiquement). Le degré de parcimonie est ainsi mesuré de manière équivalente par vitesse de convergence de l'approximation non-linéaire à N termes et par l'appartenance de la suite des coefficients aux espace $w\ell^p$, et ce degré augmente lorsque $p \rightarrow 0$ ou $s \rightarrow +\infty$. Dans le cas des bases d'ondelettes, il est également possible de relier ces propriétés à l'appartenance de la fonction à certains espaces fonctionnels mesurant la régularité (en particulier les espaces de Besov). Nous évoquons dans la section suivante un exemple important d'un tel résultat.

Une première application de la parcimonie est la compression des données. La représentation adaptative de l'image par ses N plus grands coefficients d'ondelettes n'est pas à proprement parler un algorithme de compression d'image puisque le codage de ces coefficients sans erreur d'arrondi exige a priori un nombre infini de bits. Cette représentation suggère néanmoins une stratégie de compression naturelle, consistant à ne coder avec précision que les plus grands coefficients d'ondelettes, et à seuiller les plus petits. Une telle stratégie sous-entend que l'on code aussi les indices correspondant aux coefficients numériquement significatifs. Ces idées simples sont à la base des algorithmes de compression les plus efficaces à l'heure actuelle, et qui ont été intégrés à la fin des années 1990 dans le standard JPEG 2000. Puisque les ingénieurs disposaient déjà des outils de décompositions multiéchelle des images au moyen de filtres numériques, il est légitime de se demander si le détour par les bases d'ondelettes proposé par les mathématiciens a joué un rôle fondamental dans le développement de nouvelles stratégies de compression. Il semble que l'on peut affirmer sans exagération que les travaux sur l'approximation non-linéaire par les bases d'ondelettes ont fait évoluer l'intuition dans ce domaine d'application : la propriété essentielle recherchée dans une représentation du signal ou de l'image en vue de la compression n'est plus la décorrélation (qui menait à la recherche de composantes principales) mais la parcimonie.

Les ondelettes reçoivent aussi un fort écho en statistiques à partir de 1990, à travers les travaux de David Donoho, Iain Johnstone, Dominique Picard et Gérard Kerkycharian sur les méthodes d'estimation par seuillage des coefficients. Ces méthodes consistent à ajuster le seuil au niveau du bruit de mesure, afin de capturer uniquement les coefficients numériquement significatifs qui émergent au-dessus de ce niveau. Leur succès est par conséquent conditionné aux propriétés de parcimonie des décompositions en ondelettes. Il s'agit là aussi de procédures non-linéaires, par contraste avec les techniques de filtrage plus anciennes qui présentent le défaut de lisser les discontinuités. Dans le domaine du traitement d'image, ces méthodes sont à comparer avec celles utilisant les EDP de diffusion non-linéaires ou avec celles plus récentes reposant sur des moyennages non-locaux, proposées par Jean-Michel Morel et qui constituent l'état de l'art actuel.

Depuis le début des années 1990, les bases d'ondelettes suscitent enfin un intérêt croissant dans le domaine de la simulation numérique des équations aux dérivées partielles. Si elles ne diagonalisent pas les opérateurs différentiels intervenant dans les EDP, les coefficients des matrices obtenues se concentrent néanmoins autour de la diagonale, et, comme le remarque Yves Meyer, c'est aussi le cas pour les opérateurs intégraux de type Calderon-Zygmund. Ces remarques sont à l'origine d'un algorithme proposé en 1990 par Gregory Beylkin, Ronald Coifman et Vladimir Rokhlin, pour le calcul rapide de produits matrices-vecteurs.

On peut d'autre part chercher à utiliser les représentations dans les bases d'ondelettes pour calculer sous une forme approchée la solution d'une EDP de la façon la plus économique possible. Nous avons observé que le seuillage des coefficients revient à effectuer un raffinement local de l'approximation aux voisinages des singularités d'un signal ou d'une image. En simulation numérique de tels raffinements sont nécessaires aux voisinages des singularités des solutions : ondes de chocs, ruptures, singularités. Contrairement à un signal que l'on comprime, la solution de l'EDP n'est pas une donnée explicite (elle est donnée sous forme implicite par l'équation), et on ne connaît pas par avance l'emplacement de ses coefficients d'ondelettes les plus significatifs. Les *stratégies adaptatives* visent à acquérir ces informations au fil du calcul. En 1998, Albert Cohen, Wolfgang Dahmen et Ronald DeVore montrent, dans le cadre des EDP elliptiques, qu'il existe des stratégies *optimales* en un sens suggéré par la théorie de l'approximation non-linéaire : la précision $\varepsilon(N)$ obtenue en fonction du nombre de coefficients d'ondelettes N calculés pour décrire la solution de l'EDP est comparable à la précision optimale que l'on obtiendrait si cette solution était une donnée connue que l'on approche en conservant ses N plus grands coefficients d'ondelettes. Il serait aujourd'hui inexact d'affirmer que les méthodes ondelettes ont surpassé les méthodes d'éléments finis adaptatifs qui sont leur concurrents naturels, mais il est intéressant de constater que le point de vue offert par l'approximation non-linéaire a souvent été repris par les numériciens pour mieux comprendre les performances des méthodes de raffinement de maillage.

6. À la recherche de la parcimonie

Afin d'évaluer plus précisément les performances des méthodes fondées sur les ondelettes pour le traitement d'image, il est nécessaire de disposer d'un modèle pour décrire l'ensemble des images plausibles. L'espace BV des *fonctions à variation bornée* est fréquemment utilisé, car il semble bien adapté pour décrire la structure

régulière par morceaux des images peu texturées. Rappelons qu'une fonction f est dans l'espace $BV(\Omega)$ pour Ω domaine de \mathbb{R}^d si et seulement si son gradient ∇f est une mesure de Radon finie. La variation totale de f est la masse de cette mesure, que l'on peut définir par dualité suivant

$$TV(f) := \sup \left\{ \int_{\Omega} f \operatorname{div}(g) : g \in \mathcal{D}(\Omega), \|g\|_{L^\infty} \leq 1 \right\},$$

où $\mathcal{D}(\Omega)$ est l'ensemble des fonctions C^∞ à support compact, ou par l'intégrale $\int_{\Omega} |\nabla f|$ lorsque $\nabla f \in L^1(\Omega)$. En particulier, une fonction de deux variables f régulière par morceaux présentant une discontinuité sur un ensemble Γ de courbes rectifiables de longueur finie, a une variation totale bornée et donnée par

$$TV(f) = \int_{\Omega \setminus \Gamma} |\nabla f| + \int_{\Gamma} |[f]|,$$

où $[f]$ désigne le saut de f . L'espace BV est aussi le cadre utilisé pour décrire la régularité des solutions faibles d'EDP hyperboliques nonlinéaires qui développent des profils de chocs en temps fini.

Cet espace n'est pas séparable, et n'admet pas de base inconditionnelle, mais il est néanmoins presque possible de le caractériser au moyen des bases d'ondelettes : en 1998, Albert Cohen, Ronald DeVore, Pencho Petrushev et Hong Xu démontrent que si $(d_n)_{n>0}$ est la suite des modules des coefficients d'ondelettes d'une image f réarrangés par ordre décroissant, on a l'encadrement

$$A \sup_{n>0} n d_n \leq TV(f) \leq B \sum_{n>0} d_n,$$

où $0 < A < B$ sont indépendantes de f , ce qui signifie que l'espace des coefficients d'ondelettes des fonctions BV en dimension 2 contient ℓ^1 et est contenu dans $w\ell^1$.

Il est d'autre part possible de montrer que le comportement en $1/n$ de la suite d_n est optimal et ne peut pas être amélioré, même pour des images très simples telles que l'indicatrice d'un polygone ou d'un cercle. D'après les remarques de la section précédente, ceci signifie aussi que la vitesse d'approximation par seuillage des N plus grands coefficients se comporte en $N^{-1/2}$ et pas mieux pour de telles fonctions. Ce résultat cerne ainsi les limites des ondelettes et du modèle BV en traitement d'image.

D'un point de vue intuitif, la parcimonie des représentations en ondelettes d'objets bidimensionnels présentant des contours est limitée par l'*isotropie* du changement d'échelle : le nombre d'ondelettes interceptant un contour à la résolution 2^{-j} est de l'ordre de 2^j . Afin de briser cette limite, il est nécessaire de mettre au point d'autres représentations dans lesquelles les changements d'échelles sont *anisotropes*, et qui peuvent en particulier tirer parti de la *régularité géométrique* des contours. Ainsi Emmanuel Candes et David Donoho proposent en 2000 la famille des *curvelets* qui ont la forme générale

$$\psi_{j,k,l}(x) = 2^{3j/2} \psi(D_j R_l x - k), \quad j \in \mathbb{Z}, \quad k \in \mathbb{Z}^2, \quad l = 0, \dots, 2^j - 1,$$

où ψ est une fonction de deux variables localisée en espace et en fréquence, et d'intégrale nulle dans la direction verticale, R_l est la matrice de rotation d'angle $2^{-j} l \pi$, et D_j est une matrice de dilatation diagonale anisotrope de facteur 2^j et 2^{2j}

dans les directions horizontale et verticale. Ils montrent que de telles familles constituent des frames de $L^2(\mathbb{R}^2)$ et que les coefficients d'une image régulière par morceaux avec des contours géométriquement réguliers appartiennent à l'espace $w\ell^p$ pour tout $p > 2/3$, améliorant ainsi le degré de parcimonie obtenu avec les ondelettes.

Dans la foulée, de nombreuses autres familles apparaissent avec les mêmes objectifs, à quelques variantes près qui peuvent se révéler importantes dans la pratique du calcul numérique. Citons en particulier la construction en 2005 des *bandelettes* par Erwan Le Pennec et Stéphane Mallat qui permettent d'améliorer ces résultats lorsque la régularité des contours augmente. Un concurrent naturel à ces approches est le raffinement de maillage anisotrope dans le cadre des méthodes d'éléments finis.

L'anisotropie joue aussi un rôle important lorsque l'on cherche à approcher numériquement des fonctions d'un nombre de variables $d \gg 1$. Les méthodes de discrétisation classiques, telles que l'approximation à partir des valeurs échantillonnées sur un réseau carré, se heurtent à une difficulté de complexité : par exemple, deux points de discrétisation par variable en dimension $d = 40$, donnent un nombre total de 2^{40} points, ce qui n'est pas gérable par l'ordinateur. L'idée des *sparse grids* souvent utilisée par les numériciens confrontés aux problèmes de grande dimension, est directement héritée des bases d'ondelettes : on considère la base anisotrope obtenue par tensorisation

$$\Psi_{j,k}(x) = \psi_{j_1,k_1}(x_1) \cdots \psi_{j_d,k_d}(x_d), \quad j = (j_1, \dots, j_d), \quad k = (k_1, \dots, k_d)$$

et on retient uniquement les indices tels que $j_1 + \dots + j_d \leq J$, où J pilote la précision recherchée, ce qui signifie qu'une résolution fine en une variable est compensée par une résolution plus grossière pour les autres variables.

La recherche de représentations parcimonieuses des signaux et des images conduit aussi à remplacer une base orthonormée fixée, par une famille de plusieurs bases, ou plus généralement par un *dictionnaire* redondant. En augmentant la taille du dictionnaire, l'objectif est de mieux capturer les fonctions d'intérêt avec très peu de termes. Les exemples caricaturaux tels qu'un dictionnaire contenant toutes les fonctions, et dans lequel n'importe quelle fonction est capturée par un seul terme, sont rapidement disqualifiés par des considérations de théorie de l'information et de calcul : la taille du dictionnaire entre en jeu lorsque l'on souhaite pouvoir coder ses éléments par un nombre raisonnable de bits, ou effectuer la recherche d'un élément en un temps limité. En pratique, on considère par exemple des dictionnaires \mathcal{G} contenant des atomes temps-fréquences et temps-échelles, ou d'autres fonctions élémentaires. Puisqu'il ne s'agit plus de bases, la représentation d'une fonction f suivant

$$f = \sum_{g \in \mathcal{G}} c_g g,$$

n'est pas unique. Le problème qui se pose est alors le suivant : si f admet une représentation exacte ou approchée utilisant un petit nombre d'éléments de \mathcal{G} , comment identifier ceux-ci et leurs coefficients ? L'approche classique des moindres carrés, qui consiste à minimiser $\sum_{g \in \mathcal{G}} |c_g|^2$ parmi toutes les suites (c_g) telles que $\sum_{g \in \mathcal{G}} c_g g = f$, conduit à une représentation qui n'est pas parcimonieuse. Ce

problème se relie à une question souvent considérée en statistique : identifier un petit nombre de variables *explicatives* du phénomène observé.

En 1998, David Donoho propose l'approche du *basis pursuit*, qui consiste à remplacer le critère des moindres carrés par la somme des modules $\sum_{g \in \mathcal{G}} |c_g|$, c'est-à-dire minimiser la norme ℓ^1 sous la contrainte $\sum_{g \in \mathcal{G}} c_g g = f$. Il constate, et démontre dans certains cas, que la solution de norme ℓ^1 minimale est aussi la plus parcimonieuse, c'est-à-dire celle qui contient le plus petit nombre de termes. Ces idées aboutissent à partir de 2005 aux travaux d'Emmanuel Candes et Terence Tao qui ouvrent la voie au domaine actuellement très actif du *compressed sensing*. L'un des résultats marquant du *compressed sensing* affirme que si x est un vecteur de taille $N \gg 1$ qui ne contient que k composantes non-nulles en des emplacements inconnus, il est possible de reconstruire exactement ce vecteur à partir d'un vecteur y de m mesures linéaires

$$y = \Phi x,$$

où Φ est une matrice $m \times n$ et m est de l'ordre de $k \log(N/k)$. L'algorithme utilisé est précisément le *basis pursuit*, c'est-à-dire la minimisation $\|z\|_{\ell^1}$ sous la contrainte $\Phi z = y$. Le point clé réside dans le choix de la matrice de mesure Φ , les seules constructions optimales étant actuellement obtenues par des techniques probabilistes. Ces résultats sont fortement liés à ceux obtenus depuis les années 1970 en géométrie des espaces de Banach sur les épaisseurs des convexes, tout en partant d'une motivation concrète en traitement du signal et de l'image issue de la recherche de représentations parcimonieuses. Signalons aussi le lien récemment établi par Yves Meyer et Basarab Matei entre le *compressed sensing* et la *théorie des modèles*, introduite à la fin des années 1960 par Yves Meyer et qui fournit la description mathématique des quasi-cristaux.

Tant par leurs fascinantes propriétés que par leurs limitations intrinsèques, les ondelettes ont ainsi stimulé l'activité scientifique dans de nombreuses directions.

7. Un peu de lecture

Deux références fondamentales sur la théorie des ondelettes :

- [1] Y. MEYER, *Wavelets and operators*, vol. 1 and 2, revised version of 3, Cambridge university press, 1992-1997.
- [2] I. DAUBECHIES, *Ten lectures on wavelets*, SIAM, Philadelphia, 1992.

Un ouvrage orienté vers les applications des ondelettes :

- [3] S. JAFFARD, Y. MEYER and R. RYAN, *Wavelets, tools for science and technology*, SIAM, Philadelphia, 2001.

Autour des applications au traitement du signal et de l'image, et de la notion de parcimonie :

[4] S. MALLAT, *A wavelet tour of image processing. Third edition : the sparse way*, Academic Press, 2008.

Autour des applications à la simulation numérique :

[5] W. DAHMEN, *Wavelets and operator equations*, Acta Numerica, 1997.

[6] A. COHEN, *Numerical analysis of wavelet methods*, Elsevier, 2003.

Une référence sur l'approximation non-linéaire :

[7] R. DEVORE, *Nonlinear approximation*, Acta Numerica, 1998.

Autour de l'analyse de l'espace BV et des phénomènes oscillants :

[8] Y. MEYER, *Oscillating patterns in image processing and in some nonlinear evolution equations*, Jacqueline Lewis Lectures, AMS, 2001.

MATHÉMATIQUES ET PHYSIQUE

Physique Combinatoire I : Groupes à un paramètre

Gérard H. E. Duchamp¹, Karol A. Penson² et Christophe Tollu³

1. Introduction

En 1925, en l'espace d'un an, après plus de 25 ans de tâtonnements et de discussions scientifiques entre des physiciens et mathématiciens de premier plan, ce n'est pas un ni deux, mais bien trois modèles de la Mécanique Quantique qui ont été proposés (les articles paraissent entre juin 1925 et juin 1926).

Ces modèles sont ceux de Dirac, de Schrödinger, fondé sur une équation aux dérivées partielles, et d'Heisenberg, fondé sur la relation

$$(1) \quad AB - BA = Id .$$

Cette relation apparaît clairement lorsque l'on examine quelques propriétés de l'oscillateur harmonique quantique dont l'Hamiltonien, en une dimension, s'écrit

$$(2) \quad H = \frac{p^2}{2m} + \frac{\gamma}{2}q^2$$

où p (resp. q) désigne l'opérateur moment (resp. l'opérateur position) qui agit sur la fonction d'onde $\varphi(x)$ comme $-i\hbar d/dx$ (resp. comme la multiplication par x). Le terme $p^2/2m$ représente l'énergie cinétique et le terme $(\gamma/2)q^2$ l'énergie potentielle, γ désignant le coefficient de rappel. Les opérateurs p et q satisfont à la relation de commutation

$$(3) \quad [q, p] = qp - pq = i\hbar .$$

À partir de p et q , on peut construire deux nouveaux opérateurs,

$$(4) \quad P = (\gamma m)^{-1/4} p \quad \text{et} \quad Q = (\gamma m)^{1/4} q ,$$

qui satisfont à la même relation de commutation

$$(5) \quad [Q, P] = i\hbar .$$

Avec ces nouveaux opérateurs, l'Hamiltonien se réécrit :

$$(6) \quad H = \frac{1}{2}\omega(P^2 + Q^2), \quad \text{où} \quad \omega = \left(\frac{\gamma}{m}\right)^{1/2} .$$

¹ Université Paris XIII, Laboratoire d'Informatique de Paris Nord. CNRS UMR 7030.

² Université Pierre et Marie Curie, Laboratoire de Physique Théorique de la Matière Condensée. CNRS UMR 7600.

³ Université Paris XIII, Laboratoire d'Informatique de Paris Nord. CNRS UMR 7030.

On peut maintenant définir deux opérateurs qui sont mutuellement conjugués :

$$(7) \quad a^\dagger = \frac{1}{\sqrt{2\hbar}}(Q - iP)$$

et

$$(8) \quad a = \frac{1}{\sqrt{2\hbar}}(Q + iP) .$$

Ils satisfont la relation de commutation

$$(9) \quad [a, a^\dagger] = 1 ,$$

et l'Hamiltonien se réécrit avec déplacement

$$(10) \quad H = \hbar\omega \left(a^\dagger a + \frac{1}{2} \right) .$$

Les couples d'opérateurs (A, B) tels que

$$(11) \quad AB - BA = I$$

où I est l'élément identique d'une algèbre associative, se rencontrent maintenant dans de nombreux domaines, évidemment en Physique Quantique [12, 13, 27] et plus récemment en Combinatoire [26, 37] et en Physique Combinatoire [8, 9].

2. Algèbre de Heisenberg-Weyl (un mode)

À partir de maintenant toutes les algèbres sont associatives avec élément unité non nul.

2.1. Définition formelle

L'apparition de la relation (11) en 1925 a immédiatement contraint Born, Heisenberg et Jordan à considérer des matrices infinies. En effet, un simple calcul de traces montre immédiatement que la relation (11) ne peut, en caractéristique zéro⁴, se représenter par des matrices finies (non vides). La représentation par des opérateurs continus n'est pas non plus possible dans un espace de Banach. En effet, avec $k = \mathbb{R}$ ou \mathbb{C} , on peut considérer le groupe à un paramètre $(e^{tB})_{t \in \mathbb{R}}$ et faire le calcul suivant :

$$(12) \quad ad_A(e^{tB}) = [A, \sum_{n \geq 0} \frac{t^n}{n!} B^n] = \sum_{n \geq 0} \frac{t^n}{n!} [A, B^n] = \sum_{n \geq 1} \frac{t^n}{n!} [A, B^n] = \sum_{n \geq 1} \frac{t^n}{n!} n B^{n-1} = t e^{tB} ,$$

ce qui est contraire au fait que le rayon spectral de ad_A (qui est continu) soit borné.

Les premières possibilités de représentation fidèle de (11) sont au nombre de deux (ce sont d'ailleurs les deux représentations utilisées traditionnellement).

La première se fait par des opérateurs (fermables et densément définis, mais nécessairement non bornés, comme le montre le calcul ci-dessus) dans un espace

⁴ En caractéristique finie, on peut trouver de tels couples en dimension finie, cette possibilité commence à être utilisée en physique [39].

de Hilbert (espace de Fock traditionnel), la deuxième par des opérateurs continus dans un espace de Fréchet (Bargmann-Fock).

On peut définir formellement l'algèbre de Heisenberg-Weyl [29] par

$$(13) \quad HW_{\mathbb{C}} = \mathbb{C}\langle A, B \rangle / \mathcal{J}_{HW}$$

où $\mathbb{C}\langle A, B \rangle = \mathbb{C}[\{A, B\}^*]$ est l'algèbre du monoïde libre $\{A, B\}^*$ [3, 4, 30]⁵ i.e., l'algèbre des polynômes non commutatifs et \mathcal{J}_{HW} est l'idéal bilatère engendré par le polynôme $AB - BA - 1$.

On notera que cette définition, donnée avec la flèche

$$(14) \quad s : \mathbb{C}\langle A, B \rangle \rightarrow HW_{\mathbb{C}},$$

permet de clarifier toutes les ambiguïtés concernant les formes normales qui sont traditionnelles en Physique Quantique : image commutative (opération double-dot) et ordre normal [33]. À partir de maintenant, on posera $a = s(A)$ et $a^\dagger = s(B)$; l'opérateur a (resp. a^\dagger) est appelé opérateur d'annihilation (resp. de création)⁶.

En général, par *ordre normal* [7] d'une expression $F(a^\dagger, a)$, on entend une écriture $\mathcal{N}(F(a^\dagger, a))$ obtenue en déplaçant (à l'aide de la réécriture $aa^\dagger \rightarrow a^\dagger a + 1$, dérivée de Eq. (11)) tous les opérateurs d'annihilation sur la droite. Cette procédure conduit à un opérateur dont l'action est équivalente à celle de l'opérateur original, c'est-à-dire que $\mathcal{N}(F(a^\dagger, a)) \equiv F(a^\dagger, a)$ en tant qu'opérateurs bien que leurs écritures (dont les composantes homogènes vivent dans $\mathbb{C}\langle A, B \rangle$) puissent être complètement différentes. En utilisant (14), c'est un exercice facile que de montrer que $((a^\dagger)^i a^j)_{i,j \in \mathbb{N}}$ est une base linéaire de $HW_{\mathbb{C}}$ (base des formes normales).

Les constantes de structure de $HW_{\mathbb{C}}$, calculées dans cette base, sont données dans [8] et s'obtiennent par la formule suivante⁷

$$(15) \quad (a^\dagger)^{i_1} a^{j_1} (a^\dagger)^{i_2} a^{j_2} = \sum_{k=0}^{m=\min(j_1, i_2)} \binom{m}{k} k! (a^\dagger)^{i_1+i_2-k} a^{j_1+j_2-k}.$$

2.2. Graduation de l'algèbre de Heisenberg-Weyl

Avec la regraduation de $\mathbb{C}\langle A, B \rangle$ donnée par

$$(16) \quad \deg(B) = -\deg(A) = 1,$$

on voit que l'idéal \mathcal{J}_{HW} est gradué. Cette graduation quotient est donc, de façon pratique, déterminée, pour $E \in \mathbb{Z}$, par

$$(17) \quad HW_{\mathbb{C}}^{(E)} = \text{Vect}_{\mathbb{C}}((a^\dagger)^i a^j)_{i-j=E}$$

Elle fait de $HW_{\mathbb{C}}$ une algèbre \mathbb{Z} -graduée.

La graduation précédente apparaît d'ailleurs comme naturelle quand on représente l'algèbre $HW_{\mathbb{C}}$ par les deux opérateurs sur $\mathbb{C}[z]$

⁵ Ce monoïde est noté $\text{Mo}(\{A, B\})$ dans [15], et l'algèbre correspondante $\text{Libas}_{\mathbb{C}}(\{A, B\})$.

⁶ Que ce soit dans cette algèbre ou dans ses représentations.

⁷ Formule qui peut être facilement obtenue à partir de l'approche (équivalente à celle du théorème de Wick) par les « nombres de placements de tours » (« rook numbers », en anglais) [32, 35].

$$(18) \quad \begin{cases} \rho_{BF}(a) = \frac{d}{dz} \\ \rho_{BF}(a^\dagger) = (S \mapsto zS) . \end{cases}$$

Mais on peut varier les espaces. Si l'on dispose, sur une algèbre associative \mathcal{A} de
 a) une dérivation $\partial \in \mathfrak{Der}(\mathcal{A})$
 b) un élément x tel que $\partial(x) = 1$,
 on obtient automatiquement une représentation de $HW_{\mathbb{C}}$ en posant pour $\rho(B)$ la multiplication par x ($S \mapsto xS$), et $\rho(A) = \partial$. Cela donne lieu à toute une famille de représentations fidèles⁸, en particulier sur $\mathbb{C}[z], \mathbb{C}[[z]]$ et $\mathbb{C}\langle X \rangle, \mathbb{C}\langle\langle X \rangle\rangle, \mathbb{C}[X], \mathbb{C}[[X]]$ avec $x = \sum_{l \in X} l$, mais aussi sur les espaces de fonctions C^∞ , les espaces de distributions. Cette famille de représentations est connue sous le nom de « représentations de Bargmann-Fock ».

En général, et ceci de façon très concrète, on peut associer beaucoup d'opérateurs importants en Physique Quantique avec des éléments de $HW_{\mathbb{C}}$ ou d'espaces de fonctions (qui sont souvent des complétés dans lesquels l'image d'une orbite de $HW_{\mathbb{C}}$ est dense). En particulier, si on se donne un élément $\Omega \in HW_{\mathbb{C}}$, il est intéressant d'étudier le groupe à un paramètre (groupe d'évolution)[16]

$$(19) \quad \left(e^{\lambda \rho(\Omega)} \right)_{\lambda \in \mathbb{R}} .$$

De tels groupes sont importants en dynamique quantique (ici, le paramètre λ est le temps) ou en mécanique statistique quantique (où λ est l'opposé de l'inverse de la température).

Les questions que l'on se pose sont :

- (1) Est-ce que le groupe (19) est bien défini ? À travers quelle représentation ρ ? Quel est le domaine du paramètre (dans le cas où c'est un groupe local) ?
- (2) Quelles sont les méthodes combinatoires qui peuvent être extraites de la connaissance de ce(s) groupe(s) ?

Notre premier travail va donc être d'étudier la combinatoire des formes normales des puissances Ω^n grâce aux matrices infinies.

⁸ Toutes ces représentations sont fidèles car on vérifie aisément qu'un tel x n'a pas de polynôme minimal et que le $HW_{\mathbb{C}}$ -module engendré par x est isomorphe à celui de la représentation de Bargmann-Fock sur $\mathbb{C}[x]$.

3. Combinatoire des matrices infinies

3.1. Opérateurs homogènes et nombres de Stirling généralisés

Avant de définir des représentations (ou des réalisations) du groupe à un paramètre $(e^{\lambda\Omega})_{\lambda \in \mathbb{R}}$, on peut (et on doit) considérer le problème qui consiste à mettre les puissances de Ω en ordre normal.

$$(20) \quad \mathcal{N}(\Omega^n) = \sum_{i,j \in \mathbb{N}} \alpha(n,i,j) (a^\dagger)^i a^j .$$

C'est en général un problème à trois paramètres, mais en tirant parti de la graduation précédente, on peut le ramener à deux pour les opérateurs homogènes. Un tel opérateur de degré (ou excès) E s'écrit

$$(21) \quad \Omega = \sum_{i-j=E} \alpha(i,j) (a^\dagger)^i a^j .$$

On a alors

$$(22) \quad \mathcal{N}(\Omega^n) = \begin{cases} (a^\dagger)^{nE} \sum_{k=0}^{\infty} S_{\Omega}(n,k) (a^\dagger)^k a^k & \text{si } E \geq 0 \\ \left(\sum_{k=0}^{\infty} S_{\Omega}(n,k) (a^\dagger)^k a^k \right) a^{n|E|} & \text{si } E < 0 . \end{cases}$$

C'est la définition des « Nombres de Stirling Généralisés » telle qu'elle a été introduite et utilisée dans [5, 6] pour les monômes (Ω est un produit de facteurs) et généralisée aux opérateurs homogènes dans [21] (voir aussi [31]). Ces nombres ont récemment attiré l'attention de combinatoristes [25] qui ont considéré qu'ils constituaient une généralisation non triviale de quantités combinatoires connues depuis 200 ans [2].

La justification du nom *Stirling Numbers* se trouve dans l'exemple ci-dessous, à la suite duquel on donne deux autres exemples.

Avec $\Omega = a^\dagger a$, on obtient la matrice des nombres de Stirling de seconde espèce.

$$(23) \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 3 & 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 7 & 6 & 1 & 0 & 0 & \dots \\ 0 & 1 & 15 & 25 & 10 & 1 & 0 & \dots \\ 0 & 1 & 31 & 90 & 65 & 15 & 1 & \dots \\ \vdots & \ddots \end{pmatrix}$$

Cette matrice est bien le tableau des nombres $(S(n,k))_{n,k \in \mathbb{N}}$ tels que

$$(24) \quad \sum_{n,k \geq 0} S(n,k) \frac{x^n}{n!} y^k = e^{y(e^x-1)} .$$

Ces nombres ont une interprétation combinatoire très simple : $S(n, k)$ est le nombre de relations d'équivalence sur un ensemble à n éléments qui ont k classes ou, de façon équivalente, le nombre de partitions de $\{1, \dots, n\}$ en k sous-ensembles non vides (appelés blocs en combinatoire⁹).

Pour $\Omega = a^\dagger aa^\dagger$, on a

$$(25) \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 2 & 4 & 1 & 0 & 0 & 0 & 0 & \dots \\ 6 & 18 & 9 & 1 & 0 & 0 & 0 & \dots \\ 24 & 96 & 72 & 16 & 1 & 0 & 0 & \dots \\ 120 & 600 & 600 & 200 & 25 & 1 & 0 & \dots \\ 720 & 4320 & 5400 & 2400 & 450 & 36 & 1 & \dots \\ \vdots & \ddots \end{pmatrix}$$

et pour $\Omega = a^\dagger aaa^\dagger a^\dagger$,

$$(26) \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 2 & 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 12 & 60 & 54 & 14 & 1 & 0 & 0 & 0 & 0 & \dots \\ 144 & 1296 & 2232 & 1296 & 306 & 30 & 1 & 0 & 0 & \dots \\ 2880 & 40320 & 109440 & 105120 & 45000 & 9504 & 1016 & 52 & 1 & \dots \\ \vdots & \ddots \end{pmatrix}$$

Pour tout mot $w(a, a^\dagger)$ (et plus généralement, pour tout opérateur homogène), toutes les lignes $(S_w(n, k))_{k \in \mathbb{N}}$ de la matrice de Stirling sont à support fini¹⁰. Nous appelons ces matrices « row-finite » [21]. On peut montrer qu'elles codent certaines classes d'opérateurs (par exemple les opérateurs continus de l'espace de Fréchet $\mathbb{C}[[z]]$ muni de la topologie de Treves [38]¹¹).

L'espace des matrices « row-finite » est une algèbre (notée, ci-dessous, $\text{RF}(\mathbb{N}; \mathbb{C})$).

⁹ On remarquera ainsi que $S(0, 0) = 1$ (équivalence au graphe vide), que $S(0, k) = 0$ pour $k \geq 1$, que $S(n, 0) = 0$ si $n \geq 1$ et, plus généralement que $S(n, k) = 0$ dès que $n < k$, que $S(n, n) = 1$ (singletons), que $S(n, 1) = 1$ dès que $n \geq 1$ (une seule classe, l'ensemble entier) et enfin, la récurrence (qui définit le tableau dès que l'on a fixé la première ligne)

$$S(n+1, k) = S(n, k-1) + kS(n, k)$$

qui se voit directement au niveau des partitions.

¹⁰ On peut montrer facilement que la matrice est en général en escalier et que la « marche » a pour profondeur le nombre d'annihilations, ce qui explique la forme de la matrice (26).

¹¹ C'est la topologie définie par les semi-normes

$$\rho_n\left(\sum_{m \geq 0} a_m z^m\right) = |a_n|.$$

3.2. Transformations de suites

À chaque matrice « row-finite » $(M[n, k])_{n, k \in \mathbb{N}}$, on peut associer un opérateur $\Phi_M \in \text{End}(\mathbb{C}[[x]])$ tel que l'image de $f = \sum_{k \in \mathbb{N}} a_k \frac{x^k}{k!} \in \mathbb{C}[[x]]$ soit définie par

$$(27) \quad \Phi_M(f) = \sum_{n \in \mathbb{N}} b_n \frac{x^n}{n!} \text{ avec } b_n = \sum_{k \in \mathbb{N}} M[n, k] a_k .$$

On peut remarquer que si on munit $\mathbb{C}[[x]]$ de la structure d'espace de Fréchet de la convergence simple sur les coefficients, chaque Φ_M est continu. La proposition suivante, dont la preuve est laissée au lecteur, montre qu'il n'y a pas d'autre cas.

Proposition 1. *La correspondance $\varphi : M \rightarrow \Phi_M$ entre $\text{RF}(\mathbb{N}; \mathbb{C})$ et $\mathcal{L}(\mathbb{C}[[x]])$ (endomorphismes continus) est un isomorphisme d'espaces vectoriels. C'est aussi un isomorphisme d'algèbres car*

$$(28) \quad \begin{aligned} \Phi_{MN} &= \Phi_M \circ \Phi_N \\ \Phi_I &= \text{Id}_{\mathbb{C}[[x]]} \end{aligned}$$

Cette proposition a une application immédiate pour les groupes à un paramètre $e^{\lambda\Omega}$ engendrés par des opérateurs homogènes Ω , parce que, à travers la représentation de Bargmann-Fock ρ_{BF} , la matrice $\varphi^{-1}(\rho_{BF}(\Omega))$ est

- strictement triangulaire inférieure quand $E < 0$
- diagonale quand $E = 0$
- strictement triangulaire supérieure quand $E > 0$.

Et donc $e^{\lambda\Omega}$ admet toujours une représentation comme groupe d'opérateurs sur un espace approprié. Pour $E < 0$, c'est chacun des espaces de polynômes $\mathbb{C}_{\leq n}[x]$ ¹², pour $E \geq 0$ c'est $\mathbb{C}[[x]]$ muni de la topologie de Treves [38].

3.3. Groupes à un paramètre et matrices de Stirling

Dans ce paragraphe, nous nous consacrons à la combinatoire des monômes qui contiennent au plus un opérateur d'annihilation (dans la représentation de Bargmann-Fock, ce sont des opérateurs différentiels du premier ordre au plus). Leur image par ρ_{BF} est de la forme

$$(29) \quad q(x) \frac{d}{dx} + v(x)$$

(la somme d'un champ scalaire et d'un champ de vecteurs sur la droite). Les groupes à un paramètre engendrés par ces opérateurs peuvent, bien sûr, être intégrés en utilisant des techniques d'équations aux dérivées partielles [16] mais, ici, on va privilégier le point de vue géométrique et montrer qu'un champ du type (29) est conjugué du champ de vecteurs pur $q(x) \frac{d}{dx}$.

Aussi, pour calculer $e^{t(q(x) \frac{d}{dx} + v(x))} [f]$ (f est une fonction d'un espace approprié et t est suffisamment petit), on peut utiliser la procédure suivante (q et v sont supposés au moins continus, mais q ne satisfait pas nécessairement à une condition de Cauchy-Lipschitz). Considérons d'abord le cas $v \equiv 0$ (cas du champ pur) :

¹² Chaque $\mathbb{C}_{\leq n}[x]$, constitué des polynômes de degré moindre que n , est laissé stable par cette action.

- si $q \equiv 0$, $e^{t\rho_{BF}(\Omega)}[f] = f$ (action triviale);
- si $q \not\equiv 0$, on choisit un intervalle ouvert $I \neq \emptyset$ dans lequel q ne s'annule pas et $x_0 \in I$;

– pour tout $x \in I$, on pose

$$(30) \quad F(x) = \int_{x_0}^x \frac{dt}{q(t)} .$$

Soit $J = F(I)$ (intervalle ouvert); $F : I \rightarrow J$ est un difféomorphisme (car F est strictement monotone);

– on pose

$$(31) \quad s_t(x) = F^{-1}(F(x) + t)$$

pour les couples (x, t) pour lesquels cette formule a un sens, soit l'ouvert ci-dessous (où I est choisi maximal)

$$(32) \quad \mathcal{O}_{I,q} = \{(x, t) \in \mathbb{R}^2 \mid x \in I, F(x) + t \in J\} .$$

Il est clair que s_t est une déformation de l'identité au sens où $(x, t) \mapsto s_t(x)$ est continue (et même de classe C^1) sur son domaine et que $s_0(x) = x$;

– on a alors que $e^{t(q(x)\frac{d}{dx})}$ coïncide avec la substitution $f \mapsto f \circ s_t$.

Note 1. On prendra garde au fait que le domaine de validité de la formule $e^{t(q(x)\frac{d}{dx})}[f(x)] = f(s_t(x))$ peut être restreint par la nature même de f .

Maintenant, nous sommes en position d'intégrer le groupe à un paramètre $e^{t(q(x)\frac{d}{dx} + v(x))}$ pour un champ scalaire général v , (I , F , s_t sont comme ci-dessus).

– Sur I , posons

$$(33) \quad u(x) = e^{\int_{x_0}^x \frac{v(t)}{q(t)} dt} ;$$

– on vérifie facilement la formule de conjugaison suivante

$$(34) \quad \rho_{BF}(\Omega) = (q(x)\frac{d}{dx} + v(x)) = \frac{1}{u} (q(x)\frac{d}{dx}) u ,$$

qui signifie que, sur chaque fonction, $\rho_{BF}(\Omega)$ opère comme la composition de

- la multiplication de f par u (qui est définie, ainsi que son inverse, sur I)
- l'action du champ de vecteurs $(q(x)\frac{d}{dx})$ (maintenant sur uf)
- la division par u ;

– alors, en utilisant le fait que l'exponentiation commute avec la conjugaison, l'exponentielle devient

$$(35) \quad e^{t(q(x)\frac{d}{dx} + v(x))} = u^{-1} e^{t(q(x)\frac{d}{dx})} u .$$

En utilisant les calculs précédents, on a le groupe à un paramètre sous la forme

$$(36) \quad U_t[f](x) = e^{t(q(x)\frac{d}{dx} + v(x))}[f](x) = \frac{u(s_t(x))}{u(x)} f(s_t(x)) .$$

On peut se convaincre *a posteriori* de la validité de cette procédure en utilisant une technique de vecteur tangent comme suit :

– tester que, pour des petites valeurs λ, θ du paramètre, on a

$$(37) \quad U_\lambda \circ U_\theta = U_{\lambda+\theta} ;$$

(groupe à un paramètre local)

– tester que

$$(38) \quad \left. \frac{d}{dt} \right|_{t=0} U_t[f](x) = (q(x) \frac{d}{dx} + v(x))f(x) .$$

Les transformations du type

$$(39) \quad f \rightarrow g \cdot (f \circ s)$$

s'appellent, en Physique Combinatoire, des *substitutions avec préfonction* [21].

Exemple 1. Donnons maintenant un exemple d'intégration du groupe à un paramètre $e^{\lambda \rho_{BF}(\Omega)}$ pour

$$(40) \quad \Omega = (a^\dagger)^2 a a^\dagger + a^\dagger a (a^\dagger)^2 .$$

On a la forme conjuguée

$$(41) \quad \rho_{BF}(\Omega) = x^2 \frac{d}{dx} x + x \frac{d}{dx} x^2 = x^{-\frac{3}{2}} (2x^3 \frac{d}{dx}) x^{\frac{3}{2}} .$$

En utilisant la procédure ci-dessus, on obtient le groupe à un paramètre de transformations U_λ

$$(42) \quad U_\lambda[f](x) = \sqrt[4]{\frac{1}{(1-4\lambda x^2)^3}} \times f\left(\sqrt{\frac{x^2}{1-4\lambda x^2}}\right) .$$

Le lecteur est invité à vérifier par un calcul direct que que l'on a bien une formule d'addition pour les petites valeurs du paramètre i.e.,

$$(43) \quad |\lambda| + |\theta| < \frac{1}{4x^2} \implies U_\lambda \circ U_\theta = U_{\lambda+\theta} .$$

Une fois intégré, le groupe à un paramètre U_λ permet d'obtenir la matrice de Stirling comme le montre le résultat suivant dont des éléments de preuve peuvent être trouvés dans [19, 21].

Proposition 2. Avec les définitions déjà introduites, U_λ désignant le groupe à un paramètre $\exp(\lambda \rho_{BF}(\Omega))$ et $e \geq 0$, les conditions suivantes sont équivalentes

i)

$$(44) \quad \sum_{n,k \geq 0} S_{\Omega}(n, k) \frac{x^n}{n!} y^k = g(x) e^{y\varphi(x)}$$

ii)

$$(45) \quad U_\lambda[f](x) = g(\lambda x^e) f(x(1 + \varphi(\lambda x^e))) .$$

Exemple 1. (suite)

Avec $\Omega = (a^\dagger)^2 a a^\dagger + a^\dagger a (a^\dagger)^2$, on a le groupe à un paramètre

$$(46) \quad U_\lambda[f](x) = \sqrt[4]{\frac{1}{(1-4\lambda x^2)^3}} \times f\left(\sqrt{\frac{x^2}{1-4\lambda x^2}}\right).$$

En appliquant la correspondance précédente, on obtient

$$(47) \quad \sum_{n,k \geq 0} S_\Omega(n,k) \frac{x^n}{n!} y^k = \sqrt[4]{\frac{1}{(1-4x)^3}} e^{y(\sqrt{\frac{1}{1-4x}}-1)} = \sqrt[4]{\frac{1}{(1-4x)^3}} e^{y(\sum_{n \geq 1} c_n x^n)}$$

où $c_n = \binom{2n}{n}$ désigne la famille des coefficients binomiaux centraux.

3.4. Deux exponentielles ou comment les diagrammes apparaissent

Les premiers développements de cette théorie sont expliqués en détail dans [10, 11, 28, 32, 33, 34], et la structure précise a été établie dans [21, 24].

$$(48) \quad F(z) = \sum_{n \geq 0} a_n \frac{z^n}{n!}, \quad G(z) = \sum_{n \geq 0} b_n \frac{z^n}{n!}, \quad \mathcal{H}(F, G) := \sum_{n \geq 0} a_n b_n \frac{z^n}{n!},$$

on vérifie facilement que

$$(49) \quad \mathcal{H}(F, G) = F\left(z \frac{d}{dx}\right) G(x) \Big|_{x=0}.$$

Lorsque les termes constants de F et G sont non nuls on peut, puisque \mathcal{H} est bilinéaire, normaliser les fonctions de façon que $F(0) = G(0) = 1$, le but étant de disposer de formules universelles, on écrit les fonctions comme des exponentielles libres

$$(50) \quad F(z) = \exp\left(\sum_{n=1}^{\infty} L_n \frac{z^n}{n!}\right), \quad G(z) = \exp\left(\sum_{n=1}^{\infty} V_n \frac{z^n}{n!}\right).$$

Les expressions (50) vivent dans $\mathbb{C}[[\mathbb{L} \cup \mathbb{V}]]$ (où $\mathbb{L} = \{L_1, L_2, \dots, L_k, \dots\}$ et $\mathbb{V} = \{V_1, V_2, \dots, V_k, \dots\}$, voir [21, 24] pour les détails) et la sommabilité est assurée par la topologie de la convergence stationnaire des coefficients.

C'est un classique de la Combinatoire que de déduire aussitôt de (50) que

$$(51) \quad F(z) = \sum_{n \geq 0} \frac{z^n}{n!} \sum_{\alpha \in \mathbb{N}(\mathbb{L})} \text{numpart}(\alpha) \mathbb{L}^\alpha, \quad G(z) = \sum_{n \geq 0} \frac{z^n}{n!} \sum_{\beta \in \mathbb{N}(\mathbb{V})} \text{numpart}(\beta) \mathbb{V}^\beta,$$

où, pour un multiindice $\alpha \in \mathbb{N}(\mathbb{L})$ (resp. $\beta \in \mathbb{N}(\mathbb{V})$), $\text{numpart}(\alpha)$ (resp. $\text{numpart}(\beta)$) est le nombre de partitions de l'ensemble $\{1, \dots, n\}$ de type α (resp. de type β) (voir ci-dessous). Avec $n = \sum_{i \geq 1} i \alpha(L_i) = \sum_{j \geq 1} j \beta(V_j)$, on obtient

$$(52) \quad \mathcal{H}(F, G) = \sum_{n \geq 0} \frac{z^n}{n!} \sum_{P_1, P_2 \in UP_n} \mathbb{L}^{Type(P_1) \nabla Type(P_2)}$$

où UP_n désigne l'ensemble des partitions [14] de $\{1 \dots n\}$ en ensembles non vides (usuellement appelés blocs).

Le type d'une partition $P \in UP_n$ (noté $Type(P)$) est le multiindice $\alpha \in \mathbb{N}^{(\mathbb{L})}$ (resp. $\beta \in \mathbb{N}^{(\mathbb{V})}$) tel que $\alpha(L_k)$ (resp. $\beta(V_k)$) soit le nombre de k -blocs (c'est-à-dire de parties de cardinal k).

C'est ici que la formule s'« intrique » et que les diagrammes de la théorie émergent. On doit remarquer que

- le calcul du monôme $\mathbb{L}^{Type(P_1) \nabla Type(P_2)}$ requiert bien moins d'informations que ce qui est contenu dans le couple de partitions (P_1, P_2) (par exemple on peut réétiqueter les éléments)
- à deux partitions (P_1, P_2) du même ensemble, on peut associer la matrice d'incidence

$$(\text{card}(Y \cap Z))_{(Y,Z) \in P_1 \times P_2}$$

- cette matrice peut être représentée par un graphe bicoloré comme suit
 - (1) tout bloc de P_1 (resp. de P_2) étiquette un point noir (resp. blanc)
 - (2) on trace des lignes entre le point noir $Y \in P_1$ et le point blanc $Z \in P_2$ en nombre $\text{card}(Y \cap Z)$.

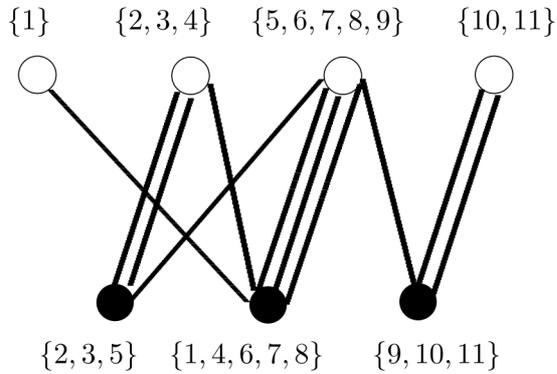


FIG. 1. Diagramme engendré par un couple de partitions P_1, P_2 de $\{1, \dots, 11\}$. $P_1 = \{\{2, 3, 5\}, \{1, 4, 6, 7, 8\}, \{9, 10, 11\}\}$ et $P_2 = \{\{1\}, \{2, 3, 4\}, \{5, 6, 7, 8, 9\}, \{10, 11\}\}$ (étiquetant respectivement les points noirs et les points blancs).

On obtient ainsi un graphe avec p ($= \text{card}(P_1)$) points noirs et q ($= \text{card}(P_2)$) points blancs, aucun sommet isolé et des multiplicités entières. Nous notons l'ensemble de ces diagrammes **diag**.

La formule du produit devient

$$(53) \quad \mathcal{H}(F, G) = \sum_{n \geq 0} \frac{z^n}{n!} \sum_{\substack{d \in \text{diag} \\ |d|=n}} \text{mult}(d) \mathbb{L}^{\alpha(d)} \mathbb{V}^{\beta(d)}$$

où $\alpha(d)$ (resp. $\beta(d)$) est le « type blanc » (resp. « type noir »), i.e. le multiindice α (resp. β) tel que $\alpha(L_i)$ (resp. $\beta(V_i)$) soit le nombre de sommets blancs (resp. de sommets noirs) de degré i et $\text{mult}(d)$ est le nombre de couples de partitions de $\{1 \dots |d|\}$ (ici $|d| = |\alpha(d)| = |\beta(d)|$) est le nombre d'arêtes de d).

Nous avons donc une flèche

$$(54) \quad d \mapsto \mathbb{L}^{\alpha(d)} \mathbb{V}^{\beta(d)}$$

qui se prolonge par linéarité à $k^{(\text{diag})}$.

La question principale est la suivante :

(Q) existe-t-il une structure d'algèbre de Hopf sur $k^{(\text{diag})}$ qui relève les opérations d'algèbre de Hopf de $k[\mathbb{L} \cup \mathbb{V}]$?

La réponse est positive en ce qui concerne le produit et l'on peut relever le coproduit d'un sous-alphabet. L'algèbre obtenue admet un relèvement non commutatif qui se déforme [24] en des algèbres de Hopf combinatoires classiques [17].

4. Conclusion

Les groupes à un paramètre qui sont si naturellement indispensables pour la théorie de Lie, mais *a priori* de nature continue peuvent conduire à des identités non triviales dans le monde combinatoire et discret. Les classes de graphes justiciables de la formule exponentielle [36] fournissent, quand on en fait la statistique par nombre de sommets et de composantes connexes, des matrices infinies unipotentes du type (23), c'est-à-dire correspondant à des substitutions sans préfonction (cas du champ de vecteurs pur). Le dictionnaire

champs de vecteurs \longleftrightarrow matrices infinies

est biunivoque, mais aucune théorie descriptive satisfaisante ne semble s'être constituée.

Dans un prochain article nous montrerons, à travers la notion de fonction représentative sur un semi-groupe et la combinatoire classique du monoïde libre que, dans l'espace de Fock (voir la section 2.1, ci-dessus), le calcul des « paquets de changement de niveau » (ensemble de produits d'opérateurs montants et descendants) conduit à des questions simples et profondes de représentation symbolique des fractions continues non commutatives. Ces représentations sont suffisamment expressives pour permettre de résoudre des questions de rationalité [23] en géométrie non commutative et le problème mentionné (les paquets de changement de niveau dans l'espace de Fock). De plus, la structure d'orbite finie des éléments ainsi représentés permet une bonne manipulation des éléments du dual de Sweedler de l'algèbre de Hopf des polynômes non commutatifs [22, 23].

5. Références

- [1] C. M. BENDER, D. C. BRODY, AND B. K. MEISTER, Quantum field theory of partitions, *J. Math. Phys.* **40** (1999) 3239–3245.
- [2] M. BERNSTEIN AND N.J.A. SLOANE, Some canonical sequences of integers. *Linear Algebra Appl.* **57** (1995) 226–228.
- [3] J. BERSTEL AND D. PERRIN, *Theory of codes*, Academic Press, 1985.
- [4] J. BERSTEL AND C. REUTENAUER, *Rational series and their languages*, Springer (EATCS Monographs on Theoretical Computer Science), 1988.
- [5] P. BLASIAK, K.A.PENSON AND A. I. SOLOMON, The general boson normal ordering problem, *Phys. Lett. A* **309** (2003) 198–205.
- [6] P. BLASIAK, K.A.PENSON AND A. I. SOLOMON, The boson normal ordering problem and generalized Bell numbers, *Annals of Combinatorics* **7** (2003) 127–139.
- [7] P. BLASIAK, A. HORZELA, K.A. PENSON, A.I. SOLOMON AND G.H.E. DUCHAMP, Combinatorics and Boson normal ordering : A gentle introduction, *Am. J. Phys.* **75** (2008) 639–646. arXiv : 0704.3116 [quant-ph]
- [8] P. BLASIAK, G. H. E. DUCHAMP, A. HORZELA, K. A. PENSON, AND A. I. SOLOMON, Heisenberg-Weyl algebra revisited : combinatorics of words and paths, *J. Phys. A : Math. and Theoret.* **41** (2008) 415204.
- [9] P. BLASIAK AND P. FLAJOLET, Combinatorial models of creation-annihilation, arXiv : 1010.0354v2. Séminaire Lotharingien de Combinatoire, **65** (2011).
- [10] P. BLASIAK, A. HORZELA, K. A. PENSON, G. H. E. DUCHAMP, AND A.I. SOLOMON, Boson normal ordering via substitutions and Sheffer-Type Polynomials, *Phys. Lett. A* **338** (2005) 108.
- [11] P. BLASIAK, K. A. PENSON, A.I. SOLOMON, A. HORZELA, AND G. H. E. DUCHAMP, Some useful combinatorial formulas for bosonic operators, *J. Math. Phys.* **46** (2005) 052110-052114.
- [12] M. BORN, W. HEISENBERG UND P. JORDAN, Zur Quantenmechanik II, *Z. Phys.* **35** (1926) 557–615.
- [13] M. BORN UND P. JORDAN, Zur Quantenmechanik, *Z. Phys.* **34** (1925) 858–888.
- [14] N. BOURBAKI, *Theory of sets*, Springer, 2004.
- [15] N. BOURBAKI, *Algebra, chapter I to III*, Springer, 2006.
- [16] G. DATTOLI, P.L. OTTAVIANI, A. TORRE AND L. VÁSQUEZ, Evolution operator equations : integration with algebraic and finite difference methods, *La Rivista del Nuovo Cimento* **20**(1) (1997) 3–133.
- [17] G.H.E. DUCHAMP, F. HIVERT, AND J.-Y. THIBON, Noncommutative symmetric functions VI : Free quasi-symmetric functions and related algebras, *Int. J. Alg. Comput.* **12**(5) (2002) 671–717.
- [18] G.H.E. DUCHAMP, J.-G. LUQUE, J.-C. NOVELLI, C. TOLLU, AND F. TOUMAZET, Hopf algebras of diagrams, *Int. J. Alg. Comput.* (to appear), DOI No : 10.1142/S0218196711006418.
- [19] G.H.E. DUCHAMP, L. POINSOT, A.I. SOLOMON, K.A. PENSON, P. BLASIAK, AND A. HORZELA, Ladder Operators and Endomorphisms in Combinatorial Physics, *Discrete Mathematics & Theoretical Computer Science*, **12**(2) (2010).
- [20] G.H.E. DUCHAMP, C. TOLLU, K. A. PENSON AND G. A. KOSHEVOY, Deformations of Algebras : Twisting and Perturbations, *Séminaire Lotharingien de Combinatoire* **B62e** (2010) 14 p., <http://www.emis.de/journals/SLC/>.
- [21] G.H.E. DUCHAMP, K.A. PENSON, A.I. SOLOMON, A. HORZELA AND P. BLASIAK, One-Parameter Groups and Combinatorial Physics, In J. Govaerts, M. N. Hounkonnou and A. Z. Msezane (eds), *Proc. of the Third International Workshop on Contemporary Problems in Mathematical Physics (COPROMAPH3)*, Porto-Novo (Benin), November 2003, p. 436, World Scientific Publishing, Singapore, 2004. arXiv : quant-ph/0401126
- [22] G.H.E. DUCHAMP AND C. TOLLU, Sweedler's duals and Schützenberger's calculus, In K. Ebrahimi-Fard, M. Marcolli and W. van Suijlekom (eds), *Combinatorics and Physics*, p. 67–78, Amer. Math. Soc. (Contemporary Mathematics, vol. 539), 2011.
- [23] G.H.E. DUCHAMP, C. REUTENAUER, *Un critère de rationalité provenant de la géométrie noncommutative* *Invent. Math.* **128** 613-622. (1997).

- [24] G.H.E. DUCHAMP, A. I. SOLOMON, P. BLASIAK, A. HORZELA AND K. A. PENSON, A multipurpose Hopf deformation of the algebra of Feynman-like diagrams, In J.L. Birman, S. Catto and B. Nicolescu (eds), *Proc. of the 26th International Colloquium on Group Theoretical Methods in Physics*, New York, Canopus Publishing, 2009.
- [25] P. FLAJOLET AND R. SEDGEWICK, *Analytic Combinatorics*. Cambridge University Press, 2009. [Version préliminaire, <http://algo.inria.fr/flajolet/Publications/books.html>]
- [26] S. FOMIN Duality of graded graphs, *J. Alg. Comb.*, **3**(4) (1994) 357–404.
- [27] W. HEISENBERG, Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen, *Z. Phys.* **33** (1925) 879–893.
- [28] A. HORZELA, P. BLASIAK, G.H.E. DUCHAMP, K. A. PENSON AND A.I. SOLOMON, A product formula and combinatorial field theory, In C. Burdik, O. Navratil, and S. Posta (eds), *Proc. of the XI International Conference on Symmetry Methods in Physics (SYMPHYS-11)* Prague, Czech Republic, June 2004, JINR Publishers, Dubna, 2004. arXiv : quant-ph/0409152
- [29] KATRIEL J., DUCHAMP G., *Ordering relations for q-boson operators, continued fractions techniques, and the q-CBH enigma*. Journal of Physics A **28** 7209-7225 (1995).
- [30] M. LOTHFAIRE, *Combinatorics on Words*, Cambridge University Press (Cambridge Mathematical Library), 1997.
- [31] M. MENDEZ, P. BLASIAK AND K. A. PENSON, Combinatorial approach to generalized Bell and Stirling numbers, *J. Math. Phys.* **46** (2005) 083511-083518.
- [32] A.I. SOLOMON, G.H.E. DUCHAMP, P. BLASIAK, A. HORZELA AND K.A. PENSON, Normal Order : Combinatorial Graphs, In P.C. Argyres, T.J. Hodges, F. Mansouri, J.J. Scanio, P. Suranyi, L.C.R. Wijewardhana (eds), 'Quantum Theory and Symmetries', *Proc. of the 3rd International Symposium on Quantum Theory and Symmetries*, p. 398–406, World Scientific Publishing, 2004. arXiv : quant-ph/0402082
- [33] A.I. SOLOMON, P. BLASIAK, G.H.E. DUCHAMP, A. HORZELA AND K.A. PENSON, Combinatorial Physics, Normal Order and Model Feynman Graphs, In B. Gruber, G. Marmo and N. Yoshinaga (eds.), *Proc. of the Symposium 'Symmetries in Science XIII', Bregenz, Austria, 2003*, p. 527, Kluwer Academic Publishers, 2004. arXiv : quant-ph/0310174
- [34] A.I. SOLOMON, G.H.E. DUCHAMP, P. BLASIAK, A. HORZELA AND K. A. PENSON, Partition functions and graphs : A combinatorial approach, In C. Burdik, O. Navratil, and S. Posta (eds), *Proc. of the XI International Conference on Symmetry Methods in Physics (SYMPHYS-11)* Prague, Czech Republic, June 2004, JINR Publishers, Dubna, 2004. arXiv : quant-ph/0409082
- [35] A. VARVAK, Rook numbers and the normal ordering problem, *J. Comb. Theory A*, **112**(2) (2005) 292–307.
- [36] R. STANLEY, *Enumerative Combinatorics, Vol I and II*, Cambridge University Press, 1993 and 1999.
- [37] R.P. STANLEY, Differential posets. *J. Amer. Math. Soc.*, **1**(4) (1988) 919–961.
- [38] F. TREVES, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, 1967.
- [39] A. VOURDAS, Quantum systems with positions and momenta on a Galois field, *J. Phys. : Conf. Ser.* **104** (2008) 012014.

Les auteurs, tous participants du projet « Physique Combinatoire » Programme PHYSCOMB No. ANR-08-BLAN-0243-2 de l'Agence Nationale de la Recherche, tiennent à remercier l'agence pour son soutien financier ainsi que celui du projet PICS PAN/CNRS N° 4339 (2008-2010).

STATISTIQUES ET GÉNOME

En guise d'introduction

Franck Picard ¹

1. L'introduction de l'introduction

Alors que nous fêtons les dix ans du séquençage du génome humain, c'est avec plaisir que nous nous regroupons pour partager avec la *Gazette des mathématiciens* notre expérience dans le domaine des statistiques appliquées à la génomique. En guise d'introduction, nous avons envie de retracer l'historique des échanges entre ces deux disciplines. C'est un aperçu partiel, biaisé, succinct, qui fera sûrement hurler bon nombre de spécialistes de tel ou tel sujet, mais nous espérons que son fil conducteur aidera les lecteurs à comprendre l'articulation des problèmes statistiques rencontrés sous la lumière des avancements et des grandes orientations de la biologie moléculaire de ces trente dernières années. Plus modestement, j'espère que cette introduction permettra de comprendre les liens entre les différents articles de ce dossier !

2. Comment faisait-on avant l'ADN ?

On pourrait considérer la génomique comme une branche récente de la génétique. En effet, les deux disciplines ont pour objectif central d'élucider les fondements biologiques de l'hérédité, et pour principe commun d'étudier les liens entre variations de caractères observés (ou phénotype), et variations de facteurs héréditaires (génotype). La distinction des deux branches est avant tout historique. Jusqu'à la découverte de la molécule d'ADN dans les années cinquante, le concept de gène n'était fondé que sur des observations de transmissions de caractères et n'avait pas de réalité physique. Les travaux fondateurs de Mendel et Morgan montrent que les facteurs héréditaires sont transmis par des entités distinctes qui sont organisées linéairement (mais on ne sait pas sur quel support). Morgan montre aussi que la distance entre les gènes détermine leur transmission d'une génération à une autre. À cette époque, la génétique est plutôt une branche de la statistique, avec notamment les méthodes de cartographie génétique s'appuyant sur les modèles linéaires et sur les modèles probabilistes de génétique des populations dans la lignée des premiers développements de Fisher. Le terme de génome est employé pour la première fois en 1932 pour désigner l'intégralité de l'information génétique d'un organisme. Alors que la génétique est profondément ancrée dans l'étude des données populationnelles, la génétique moléculaire et la génomique

¹ LBBE, UMR CNRS 5558 Université Lyon 1, F-69622, Villeurbanne, France.

s'intéresseront plus particulièrement à la molécule d'ADN et à ses propriétés. La génomique en tant que science connaît un essor considérable à partir des années 90 avec la disponibilité des génomes complets.

3. Seulement 4 lettres ! (oui mais enroulées en double hélice)

La transition de la génétique vers la génomique commence par l'exploration des bases moléculaires de la génétique. C'est grâce à deux étapes fondamentales que l'on commence à matérialiser la notion de gène. Beadle et Tatum montrent d'abord que le gène constitue l'information à la source de molécules individuelles dans une voie de régulation biologique, et posent le postulat « un gène une enzyme » [6]. Ensuite on démontre que l'hérédité possède un support physique d'une nature chimique différente des protéines [12, 5]. C'est grâce à la résolution de la structure 3D [33] de la molécule d'ADN que l'on obtient la clé pour expliquer par quels mécanismes l'ADN peut être la molécule vecteur de l'hérédité. L'ADN – Acide Désoxyribo Nucléique – est une molécule double brin dont chaque brin est un polymère orienté composé de 4 molécules élémentaires (nucléotide), Adénine, Thymine, Cytosine, Guanine, qui sont reliées de manière covalente (forte). Chaque nucléotide a la propriété de s'apparier de manière non covalente (faible) avec un autre nucléotide, et ce de manière spécifique, A-T, G-C, c'est le principe d'appariement des bases qui permet de déduire le deuxième brin du premier. Cette propriété de complémentarité explique à elle seule comment un brin peut servir de modèle pour la synthèse d'un autre brin, ouvrant ainsi la voie aux mécanismes de la réplication, qui est à la source de tout mécanisme de transmission. De plus, l'existence de deux brins complémentaires et séparables permet d'expliquer comment des erreurs de réplication peuvent conduire à des mutations transmises à la descendance par l'intermédiaire de l'une des deux molécules filles. La découverte de la structure de la molécule d'ADN offre donc les bases moléculaires des théories de l'évolution. Elle explique également les liens précoces de la génomique et de l'algorithmique du texte. En 1975 Frederick Sanger annonce qu'il a mis au point une technique permettant de déterminer la succession des nucléotides le long de la molécule d'ADN [22]. C'est le début du séquençage des gènes et de leur modélisation par un texte écrit dans un alphabet de 4 lettres : A, T, G, C. Le séquençage des génomes entiers commencera dans les années quatre-vingt-dix.

4. Le développement des bases de données

Dès la fin des années soixante-dix se mettent en place des initiatives nationales et internationales pour stocker et cataloguer systématiquement les informations de séquences afin de les rendre disponibles à la communauté scientifique, à travers des initiatives comme GeneBank et l'EMBL [26, 8]. C'est le début d'un « naturalisme moderne » qui consiste à collecter les génomes comme on a collecté les spécimens aux XVIII^e et XIX^e siècles [26]. L'explosion de la micro-informatique et la popularisation des micro-ordinateurs ont alimenté l'attrait pour l'analyse automatique des séquences d'ADN, parce que d'une part les séquences sont illisibles sans modèle, et d'autre part les méthodes mathématiques et informatiques étaient de toute évidence la seule manière de faire face au déluge de séquences produites par les projets de séquençage. Les progrès toujours constants des techniques de séquençage posent

désormais d'importants défis informatiques aux banques de données publiques, à tel point que la survie même du modèle de base de données centralisée est remise en question à l'heure actuelle. Alors qu'il a fallu dix ans pour produire une première ébauche du génome humain (3,5 milliards de nucléotides), le projet 1000 Génomes propose par exemple de cataloguer les variations génétiques de 1000 individus sur la totalité de leur génome. Aujourd'hui, séquencer un génome humain peut se faire en quelques mois par un seul laboratoire pour quelques milliers d'euros. En avril 2011, GenBank contenait 126 551 501 141 nucléotides dans 135 440 924 séquences. De nouvelles initiatives voient donc le jour pour faire face à des jeux de données qui atteignent désormais le peta-byte (10^{15}) [8].

5. Les débuts de la bioinformatique

Le concept de bioinformatique émerge dans les années soixante-dix, avec l'idée sous-jacente que les sciences de l'information pouvaient aider à la compréhension des systèmes biologiques. Un axe historique des développements mathématiques pour l'étude du génome est la génomique comparative. L'idée centrale est que les séquences biologiques (ADN et protéines) contiennent des traces de l'évolution qui sont étudiées par comparaison. L'hypothèse de base est que les dissimilarités entre séquences proviennent de mécanismes évolutifs comme les insertions de nucléotides, les délétions, et les substitutions. Se développent alors des modèles mathématiques d'évolution moléculaire [14, 16] permettant la reconstruction d'une histoire évolutive en utilisant les processus stochastiques et la combinatoire des arbres phylogénétiques.

Le calcul de similarité entre séquences s'effectue par des techniques d'alignement pour déterminer si deux chaînes de caractères se ressemblent. Ce sont les travaux (entre autres) de Needleman et Wunsch [19], Sankoff [22], et Smith et Waterman [32] dans les années soixante-dix qui donnent des solutions algorithmiques à ce problème d'alignement de séquences, à l'aide de la programmation dynamique qui permet de calculer des scores optimaux d'alignements. Le score le plus simple serait par exemple de noter 1 si deux lettres correspondent et 0 sinon, puis ensuite de chercher les segments de séquences de score élevé. En 1990 est publié BLAST (Basic Local Alignment Search Tool, [3]), qui propose à la communauté un outil de recherche de séquences dans les banques. C'est une des publications les plus citées dans la littérature scientifique des années quatre-vingt-dix. Mais très vite se pose une question cruciale : quel sens donner au score d'alignement de deux séquences ? Ce score est-il significatif ? Le point de vue statistique mettra l'accent sur l'aspect stochastique de l'évolution des séquences en proposant un formalisme probabiliste aux méthodes d'alignements avec notamment les travaux de Samuel Karlin et Volker Brendel [15]. Le principe de l'approche statistique de l'alignement est de tester si le score observé est plus élevé qu'attendu sous un modèle aléatoire. Les modèles de Markov permettent alors de définir un modèle de référence pour les séquences biologiques, et donnent alors accès à la loi du score d'alignement. La notion d'exceptionnalité du score est quantifiée en calculant la probabilité que le long d'une séquence, le score du segment de meilleur alignement dépasse un seuil. Ces développements font appel à la théorie des maxima de sommes partielles de processus. La notion d'exceptionnalité s'applique également

à la sur/sous-représentation de certains mots dans les séquences [21], avec l'idée sous-jacente qu'une chaîne de caractères (motif) exceptionnellement présente dans un texte pourrait avoir un sens biologique particulier. Ces développements feront l'objet de l'article de S. Schbath.

Les premières contributions des statistiques à la génomique s'inscrivent également dans le courant de l'analyse automatique du langage, qui se développe dans les années soixante-dix, avec la mise au point et l'utilisation des modèles de chaînes de Markov cachées qui permettent de modéliser et de localiser des changements dans les fréquences d'apparition des nucléotides le long de la séquence. Ces modèles probabilistes permettent alors l'annotation automatique des génomes qui constitue une dynamique majeure des années quatre-vingts. Cet effort considérable de cartographie physique permet de découvrir, de positionner les gènes le long de l'ADN et de les munir d'une signature de séquence. Il contribue à la vision des années soixante/soixante-dix du génome : un code présent dans l'ADN qui donne lieu à des produits fonctionnels.

6. Après la séquence ?

Alors que les projets de séquençage permettent d'établir des modèles de fonctionnement et d'évolution des génomes de plus en plus précis, la notion de gène se complexifie de plus en plus. Un gène devient un segment d'ADN qui contribue au phénotype ou à une fonction biologique. Les techniques d'annotations montrent que la structure des gènes est extrêmement complexe. Les premières structures élucidées étaient les plus simples, mais il apparaît rapidement que les gènes ont des structures éclatées en plusieurs morceaux (plusieurs exons). Si un gène est constitué de trois exons A, B, C, l'agencement de ces exons peut produire des transcrits ayant des fonctions différentes. Ce phénomène s'appelle l'épissage alternatif : alors que la quantité de gènes annotés n'augmente quasiment plus aujourd'hui chez l'homme, c'est le nombre de formes différentes par gène qui explose. Aujourd'hui les annotations contenues dans GENCODE [13] contiennent une moyenne de 5, 4 formes différentes par gène (transcrits alternatifs), et plus de la moitié des gènes a un site d'initiation de la transcription alternatif. Alors que le texte contenu dans la séquence d'ADN est déterminé, les produits de l'expression de ce texte apparaissent extrêmement divers. Par conséquent, une des conclusions majeures des projets de séquençage est que l'information de séquence ne constitue qu'une part (certes fondatrice) de la complexité des bases moléculaires des phénomènes biologiques étudiés. Alors que dans les années 2000 les génomes de plus de 800 organismes ont été séquencés, le déséquilibre est criant entre le nombre de séquences stockées et la connaissance que l'on a de leurs fonctions biologiques. L'élan qui succède aux projets de séquençage a pour objectif de s'intéresser à grande échelle, en plus du génome, aux produits provenant de l'expression des séquences génomiques : les ARN et les protéines.

Le dogme central de la biologie moléculaire est posé par Francis Crick dans les années soixante [27] et propose un modèle de passage de l'information de la séquence d'ADN aux protéines par les ARN (Acide Ribo Nucléique). Dans ce modèle, l'ADN constitue la matrice qui contient l'information qui est transcrite dans

un autre alphabet de 4 lettres (la Thymine -T- est remplacée par l'Uracile -U) dont le support sont les « transcrits » composés d'ARN. Les protéines constituent les molécules « exécutantes » des fonctions biologiques, et l'ARN la molécule de transition (le messenger) entre les acides nucléiques et les protéines. Ce sont d'ailleurs les ARN messagers qui sont « traduits » en protéines à l'aide du code génétique. Après avoir étudié le « texte » que constituent les molécules d'ADN, une extension naturelle de la génomique était l'étude des produits des gènes. Pour des raisons techniques, il est plus facile d'étudier les ARN messagers que les protéines, et les techniques classiques ne permettaient d'étudier la présence que de quelques dizaines d'ARN différents simultanément. C'est l'émergence de la technologie des puces à ADN dans les années quatre-vingt-dix qui permet l'essor de la biologie dite « à haut débit ».

7. La biologie à haut débit et « petits » problèmes statistiques associés

La technologie des puces à ADN (DNA microarrays) repose sur la propriété de complémentarité des bases de la molécule d'ADN. Considérons la portion d'ADN correspondant à la séquence d'un gène. Cette molécule comporte deux brins complémentaires pouvant être séparés. On appelle hybridation spécifique la capacité d'un simple brin à « retrouver » son brin complémentaire. Simplifiée à l'extrême, la technologie des puces à ADN consiste à fixer un catalogue de séquences connues simple brin sur une lame de verre, et à déposer sur cette lame un extrait des ARN d'une cellule (pour des raisons techniques ces ARN sont convertis en ADN). Si sur la lame sont déposées les séquences correspondant aux transcrits des gènes A,B,C, et que les produits des gènes A,C,D,E sont présents dans l'échantillon, alors la puce à ADN permettra de quantifier la présence des produits des gènes A et C, les produits des gènes D et E n'étant pas détectés sur la lame (car non prévus dans le catalogue). Historiquement, la technologie des puces à ADN a été mise au point grâce aux transferts de technologie provenant des circuits imprimés des cartes à puce [18], et les projets de séquençage ont permis de créer des catalogues extrêmement larges. Les deux technologies qui s'imposent alors sont mises au point par Pat Brown à Stanford [24], et par la société Affymetrix [10]. Elles reposent sur la fixation sur des supports fixes de milliers de séquences connues, et sur la mesure de la quantité de séquences hybridées sur la lame par des signaux de fluorescence.

C'est l'essor du domaine de la biologie à haut débit et des sous-disciplines en « -omique » : on caractérise les cellules non seulement grâce à leur séquence d'ADN (génomique), mais aussi grâce à l'ensemble de leurs transcrits et de leurs protéines (transcriptomique, protéomique). La contribution des statistiques à ce domaine est alors centrale. En effet la technologie des microarrays produit des signaux de fluorescence (continus) dont l'analyse relève naturellement des statistiques : normalisation des données, analyse des différences d'expression entre conditions (tests), planification expérimentale, classification des gènes et/ou des individus, prédiction et apprentissage statistique. On considère désormais l'ensemble des produits d'expression des gènes d'un organisme pour identifier des signatures moléculaires. Les deux études emblématiques du début de la transcriptomique concernent la prédiction de deux types de leucémies sur la base de l'expression des gènes des patients [11], et

la distinction de sous-types de cancers sur des critères moléculaires [2], ouvrant la voie au diagnostic moléculaire des maladies, et en particulier des cancers. La technologie des microarrays est aussi étendue à d'autres problématiques, comme la cartographie des anomalies chromosomiques ou des interactions entre l'ADN et certaines protéines.

C'est l'époque du lancement du projet ENCODE [7] qui, à l'image des grands projets de séquençage, a pour objectif de fournir une encyclopédie du fonctionnement du génome humain (ENCyclopedia Of DNA Elements). Cette époque voit également se populariser l'utilisation de méthodes statistiques jusqu'à présent absentes des laboratoires de biologie moléculaire. C'est même l'ère de l'utilisation quasi-obligatoire des ordinateurs et des logiciels d'analyse (autres qu'Excel!) afin de pouvoir analyser des données comportant plusieurs dizaines de milliers d'enregistrements. Du côté des statisticiens, c'est le développement considérable du logiciel libre R² qui, grâce à l'essor d'internet permet aux statisticiens de diffuser leurs méthodes sous forme de logiciels (packages).

Cependant, même si les tâches statistiques sont bien identifiées, les techniques classiques reposent sur le postulat que le nombre d'observations n était beaucoup plus grand que le nombre de variables p . Cette tendance s'inverse dans les années 1990-2000 car il devient beaucoup moins coûteux de mesurer plusieurs milliers de variables sur peu d'individus par les techniques à haut débit, que de trouver un échantillon représentatif sur lequel fonder des prédictions solides. Les fondements asymptotiques des statistiques classiques ne sont donc plus valides, et les tâches autrefois simples deviennent un enjeu méthodologique majeur. Les microarrays ne sont qu'un exemple (avec le nombre de gènes étudiés correspondant au nombre de variables) et l'analyse d'images ou de courbes rencontre le même problème. Ces problématiques statistiques s'inscrivent dans une dynamique très populaire dans les années 1990-2000, celle de l'apprentissage statistique, qui regroupe des chercheurs en informatique, mathématique et statistique sous l'impulsion (entre autres) des travaux de Vapnik et Chervonenkis dans les années soixante-dix [29]. L'objectif est de construire un classifieur (ou règle de décision) prenant en compte les descripteurs pour prédire une étiquette d'appartenance à une classe ou label. La construction du classifieur se fait sur un échantillon pour lequel les labels sont connus (échantillon d'apprentissage). En génomique, on cherchera à prédire le statut biologique d'un patient (son label) à l'aide de ses caractéristiques génomiques (une variable quantifie la quantité d'expression d'un gène). Un élément crucial de la construction d'un classifieur performant est bien entendu la sélection des variables les plus informatives/prédictives. Le thème de la sélection de variables est également un domaine extrêmement actif de ces dernières années, avec la mise au point de techniques dites de « régression creuse » (sparse regression) sous l'impulsion des travaux de R. Tibshirani [28]. Au vu de l'importance des développements méthodologiques suscités par l'essor de l'étude des données de puces à ADN, les articles de E. Lebarbier et P. Neuvial de ce dossier y seront consacrés.

² <http://www.r-project.org/>

8. Une médecine personnalisée ?

On assiste alors au développement de la « translational research » qui consiste à transférer le plus rapidement possible les avancées de la recherche fondamentale à la recherche clinique pour une meilleure prise en charge du patient. Alors que des budgets considérables sont mis en œuvre (notamment pour la recherche contre le cancer), et après une succession d'annonces sur les promesses de la médecine moléculaire personnalisée et sur l'identification « de biomarqueurs » (« Array of Hope » [17]), les statisticiens ont souvent eu un rôle de modérateur, voire de rabat-joie. En effet, malgré les progrès des méthodes statistiques pour la prédiction, si l'échantillon d'apprentissage est trop petit, les règles de prédictions montreront des performances modestes sur de nouveaux individus. De plus les études moléculaires ont mis en lumière une incroyable variabilité inter-individuelle concernant les maladies complexes. À ce titre, les années 2000 ont vu progresser le nombre de publications rapportant des résultats non reproductibles, avec des titres évocateurs comme « The cancer biomarker problem » [23], ou « Valid Concerns » [4] ! Malgré ces limitations, la technologie des microarrays a cependant eu un impact considérable dans l'appréhension des phénomènes moléculaires à l'échelle du génome entier avec des protocoles utilisables par chaque laboratoire. La technologie des microarrays est maintenant utilisée en routine et permet d'étudier des phénomènes moléculaires sans *a priori* sur le gène d'intérêt. Au travers d'une utilisation plus rigoureuse des statistiques par l'ensemble de la communauté des « génomiciens » nous avons également pu faire passer l'idée qu'il était crucial de distinguer la variabilité technique (qui peut être contrôlée par des procédés expérimentaux plus fiables) de la variabilité biologique, dont la quantification requiert un nombre minimum d'individus. Le sur-optimisme des études prédictives est sans aucun doute lié à la séparation des communautés de la génétique quantitative et de la génomique, les uns ayant peu l'habitude des données à grande dimension, les autres ne connaissant pas forcément les principes fondamentaux des études génétiques. On voit maintenant émerger le concept de génomique-génétique ou de génomique des populations : le retour des problématiques (classiques ?) de génétique étudiées avec les outils et au regard des connaissances apportées par la génomique. On peut désormais superposer différentes complexités, moléculaires et populationnelles.

9. Après les transcrits ? C'est complexe !

De même que la progression des projets de séquençage a mis en lumière le besoin de comprendre le monde des ARN, après dix années de projets « post-génomiques », un changement de mode de pensée était nécessaire pour appréhender l'ensemble des données collectées. Alors que les grands projets de génomique cataloguent les bases moléculaires des génomes et de leur expression, progressivement les nouveaux projets ont adopté un nouvel angle d'attaque en se focalisant directement sur les fonctions biologiques dans leur ensemble. Or ces fonctions sont généralement le résultat d'interactions entre protéines, entre voies de signalisation et de régulation. Plutôt que d'envisager les strates d'information une par une, l'enjeu réside désormais dans la compréhension des interactions entre ces différentes « échelles de complexité ».

La complexité! Les réseaux! Ce sont certainement les deux mots clés de ces dernières années. Selon l'ISI Web of Knowledge le champ thématique des « complex networks » était la principale thématique de recherche en mathématiques en 2008. Cette dynamique provient d'un élan commun à plusieurs disciplines : l'explosion des réseaux sociaux et l'intérêt grandissant pour l'étude de leurs structures, la disponibilité des données en physique des particules, et la quantité considérable d'informations collectées en biologie moléculaire. La distinction individus/variables est désormais trop restrictive : les données sont constituées d'agents dont les interactions permettent le fonctionnement d'un système. Les fondements de la biologie des systèmes sont donc ancrés dans l'intégration des informations récoltées par les projets de génomique, afin de modéliser les processus biologiques, et de concevoir aussi de nouvelles expériences. On assiste à une implication croissante de la physique dans cette discipline, qui voit la cellule comme un système contrôlable, et qui ouvre la voie à une véritable biologie in-silico. Les graphes constituent l'objet mathématique central de l'analyse des réseaux biologiques, ce qui explique l'implication extrêmement forte des mathématiques discrètes. La statistique quant à elle permet l'inférence des mécanismes de régulation (graphes de régulation), dans un contexte de ultra-haute dimension (la taille du problème explose potentiellement avec le nombre de gènes [30]). Les phénomènes de régulation sont aussi modélisés par des réseaux booléens en grande dimension. Les statisticiens se sont également intéressés à la détection de structures particulières dans les réseaux biologiques. On parle de « modules » par exemple, qui regroupent des gènes présentant un fonctionnement homogène du point de vue d'une fonction donnée. On peut également rechercher des structures de connectivité au sens large [9, 1]. La notion de motif, axe phare de l'analyse des séquences, a également été généralisée aux réseaux biologiques [25, 20], avec l'idée sous-jacente que certaines sous-structures pouvaient être à l'origine du fonctionnement de l'ensemble du réseau. Le thème de l'analyse statistique des réseaux biologiques fera l'objet du dernier article de ce dossier.

10. Et après ?

Décidément, nous assistons à une révolution tous les 5 à 10 ans dans le domaine de la génomique! Aujourd'hui, c'est le grand retour des technologies de séquençage qui crée un nouvel engouement. Et pour cause : après les microarrays, la nouvelle transition technologique permet de séquencer des génomes d'organismes entiers en quelques semaines pour « quelques » milliers d'euros! Des consortiums internationaux nous ont déjà permis d'accéder aux génomes de ~250 eucaryotes (organisme possédant un noyau cellulaire), ~4000 bactéries et virus, avec des génomes complets pour l'homme, la souris, le rat, le chien, le chimpanzé, la vache, mais aussi pour un marsupial, l'ornithorynque et pour un oiseau! La nouvelle révolution est que ce séquençage de génomes complets est maintenant réalisable par des laboratoires de taille modeste, et pas uniquement pour des organismes modèles. Cet accès grandissant aux technologies à haut (très haut) débit permet de sortir du modèle de « génome de référence » : les projets internationaux cherchent désormais à étudier la diversité génétique à l'échelle du génome entier et de populations. C'est le cas du projet 1000 génomes par exemple [31].

En tant que statisticiens, nous savons d'expérience qu'il est très difficile de prédire les grandes tendances des prochaines années! Sans trop prendre de risques,

on peut néanmoins supposer que la thématique de la « grande dimension » sera à l'honneur, avec des quantités de données à modéliser qui rivalisent avec les quantités de la physique des particules. Peut être un rapprochement entre l'informatique, les mathématiques et la statistique ? L'accès, le stockage, la manipulation est déjà problématique, et les statisticiens sont souvent moins bien armés que d'autres pour modéliser des données à structures exotiques (par rapport à la théorie des graphes par exemple). Une bonne occasion pour encourager nos collègues mathématiciens qui souhaiteraient nous rejoindre dans l'aventure ! Et aussi pour vous souhaiter bonne lecture de ce dossier !

11. Références

- [1] E. M. AIROLDI, D. M. BLEI, S. E. FIENBERG, and E. P. XING. Mixed Membership Stochastic Blockmodels. *J Mach Learn Res*, 9 :1981–2014, Sep 2008.
- [2] A. ALIZADEH and AL. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403 :503–511, Feb 2000.
- [3] S. F. ALTSCHUL, W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. Basic local alignment search tool. *J. Mol. Biol.*, 215 :403–410, Oct 1990.
- [4] No authors listed. Valid concerns. *Nature*, 463 :401–402, Jan 2010.
- [5] O. T. AVERY, C. M. MACLEOD, and M. MCCARTY. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79 :137–158, Feb 1944.
- [6] G. W. BEADLE and E. L. TATUM. Genetic Control of Biochemical Reactions in Neurospora. *Proc. Natl. Acad. Sci. U.S.A.*, 27 :499–506, Nov 1941.
- [7] E. BIRNEY and AL. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447 :799–816, Jun 2007.
- [8] G. COCHRANE and AL. Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, 37 :19–25, Jan 2009.
- [9] J.J. DAUDIN, F. PICARD, and S. ROBIN. A mixture model for random graph. *Statistics and computing*, 18(2) :1–36, 2008.
- [10] S. P. FODOR, J. L. READ, M. C. PIRRUNG, L. STRYER, A. T. LU, and D. SOLAS. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251 :767–773, Feb 1991.
- [11] T. R. GOLUB and AL. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286 :531–537, Oct 1999.
- [12] F. GRIFFITH. The Significance of Pneumococcal Types. *J Hyg (Lond)*, 27 :113–159, Jan 1928.
- [13] J. HARROW, F. DENOEUDE, A. FRANKISH, A. REYMOND, C. K. CHEN, J. CHRAST, J. LAGARDE, J. G. GILBERT, R. STOREY, D. SWARBRECK, C. ROSSIER, C. UCLA, T. HUBBARD, S. E. ANTONARAKIS, and R. GUIGO. GENCODE : producing a reference annotation for ENCODE. *Genome Biol.*, 7 Suppl 1 :1–9, 2006.
- [14] T.H. JUKES and C.R. CANTOR. *Evolution of protein molecules*. Mammalian protein metabolism. New York : Academic Press., 1969.
- [15] S. KARLIN and V. BRENDEL. Chance and statistical significance in protein and DNA sequence analysis. *Science*, 257 :39–49, Jul 1992.
- [16] M. KIMURA. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.
- [17] E. S. LANDER. Array of hope. *Nat. Genet.*, 21 :3–4, Jan 1999.
- [18] T. LENOIR and E. GIANNELLA. The emergence and diffusion of DNA microarray technology. *J Biomed Discov Collab*, 1 :11, 2006.
- [19] S. B. NEEDLEMAN and C. D. WUNSCH. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48 :443–453, Mar 1970.
- [20] F. PICARD, J. J. DAUDIN, M. KOSKAS, S. SCHBATH, and S. ROBIN. Assessing the exceptionality of network motifs. *J. Comput. Biol.*, 15 :1–20, 2008.

- [21] G. REINERT, S. SCHBATH, and M. S. WATERMAN. Probabilistic and statistical properties of words : an overview. *J. Comput. Biol.*, 7 :1–46, 2000.
- [22] F. SANGER. The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proc. R. Soc. Lond., B, Biol. Sci.*, 191 :317–333, Dec 1975.
- [23] C. L. SAWYERS. The cancer biomarker problem. *Nature*, 452 :548–552, Apr 2008.
- [24] M. SCHENA, D. SHALON, R. W. DAVIS, and P. O. BROWN. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 :467–470, Oct 1995.
- [25] S. S. SHEN-ORR, R. MILO, S. MANGAN, and U. ALON. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, 31 :64–68, May 2002.
- [26] B. J. STRASSER. Genetics. GenBank–Natural history in the 21st Century ? *Science*, 322 :537–538, Oct 2008.
- [27] B. J. STRASSER, L. PAULING, and F. CRICK. A world in one dimension : Linus Pauling, Francis Crick and the central dogma of molecular biology. *Hist Philos Life Sci*, 28 :491–512, 2006.
- [28] R. TIBSHIRANI. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 85(1) :267–288, 1996.
- [29] V. N. VAPNIK. *The Nature of Statistical Learning Theory*. Springer, 2000.
- [30] N. VERZELEN. Minimax risks for sparse regressions : Ultra-high-dimensional phenomenons. Technical report, Arxiv, 2010.
- [31] M. VIA, C. GIGNOUX, and E. G. BURCHARD. The 1000 Genomes Project : new opportunities for research and social challenges. *Genome Med*, 2 :3, 2010.
- [32] M.S. WATERMAN, T.F. SMITH, and W.A. BEYER. Some biological sequence metrics. *Adv. Math.*, 20 :367–387, 1976.
- [33] J.D. WATSON and F.H.C. CRICK. A structure of deoxyribonucleic acid. *Nature*, 171 :964–967, 1953.

Statistiques de motifs

Sophie Schbath ¹

Cet article traite de l'identification de motifs d'ADN fonctionnels le long des génomes par des approches statistiques. Par motif d'ADN, on entend une courte suite de lettres (généralement pas plus d'une quinzaine) dans l'alphabet des nucléotides $\mathcal{A} = \{a, c, g, t\}$, et par motif fonctionnel, on entend un motif dont les occurrences sur le génome seront reconnues par une protéine qui se fixera alors sur l'ADN pour entrer en action.

L'identification de motifs d'ADN fonctionnels reste un problème biologique encore loin d'être résolu pour plusieurs raisons : (i) leur longueur est variable selon la nature de la protéine, (ii) la reconnaissance de chacune des lettres par la protéine peut être imprécise, autrement dit il n'y a pas nécessairement unicité d'une lettre à chaque position, (iii) l'activité protéique peut dépendre de la présence à proximité d'autres protéines elles-mêmes reconnaissant d'autres motifs ADN, (iv) les motifs fonctionnels ne sont généralement pas conservés d'un organisme à l'autre.

Bien souvent, ces motifs fonctionnels (appelés simplement motifs par la suite) se caractérisent par des répartitions bien particulières le long du génome, d'où la construction de critères ou scores dont on cherchera à mesurer la significativité : on cherche en effet à repérer des événements (ici des occurrences de motifs) qui

¹ INRA, Unité Mathématique, Informatique et Génome, Jouy-en-Josas, France.

auraient très peu de chance de se produire au hasard. Le « hasard » sera déterminé par des séquences aléatoires $X_1 X_2 \cdots X_\ell$ dont les lettres X_i seront tirées dans l'alphabet \mathcal{A} selon un modèle probabiliste plus ou moins sophistiqué.

Typiquement, on utilise des modèles de chaîne de Markov d'ordre m : la probabilité de générer la lettre $b \in \mathcal{A}$ à la position i dépend des m lettres² $a_1 a_2 \cdots a_m$ qui précèdent, c'est-à-dire aux positions $i - m, \dots, i - 1$. Ces probabilités, dites *de transition*, des lettres $a_1 a_2 \cdots a_m$ vers la lettre b sont estimées à partir de la composition en mots de taille $m + 1$ de la séquence d'ADN analysée. Ce modèle a l'énorme avantage de comparer ce que l'on observe dans la séquence d'ADN étudiée avec ce que l'on pourrait attendre dans des séquences aléatoires ayant en moyenne la même composition en lettres mais aussi en mots de tailles 2, 3, ..., $(m + 1)$.

Par exemple, certains motifs se caractérisent par une fréquence anormalement élevée³ sur le génome entier, ou seulement sur une partie du génome. Pour découvrir d'autres motifs potentiels ayant la même propriété statistique, on est amené à étudier la loi de probabilité du comptage d'un mot dans une chaîne de Markov (cf. section 1) et regarder ceux ayant un comptage significativement élevé (probabilité critique proche de zéro).

D'autres motifs se caractérisent par leur présence dans certaines régions caractéristiques du génome, par exemple en amont des gènes⁴. Si l'on considère alors toutes⁵ les sous-séquences de quelques centaines de lettres situées en amont des gènes, cela se traduit par un nombre anormalement élevé de ces sous-séquences contenant au moins une occurrence du motif en question. Dans ce cas, on est amené à évaluer la probabilité qu'un motif donné soit présent (peu importe son nombre d'occurrences) dans une chaîne de Markov (cf. section 2).

D'autres types de questions statistiques relatives aux occurrences de motifs existent mais ne seront pas traitées dans ce dossier. On pourra cependant noter que la formalisation d'un certain nombre d'entre elles utilise non plus un modèle de séquences aléatoires, mais des processus ponctuels pour modéliser les occurrences elles-mêmes. On citera par exemple l'utilisation de processus de Poisson, pour détecter des régions anormalement riches ou pauvres en certains motifs ou pour tester si deux séquences sont aussi riches en un motif donné. Ou encore l'utilisation de processus de Hawkes pour détecter si les occurrences d'un ou plusieurs motifs présentent des distances favorisées ou évitées.

1. Comptage attendu ou anormal ?

Le problème est le suivant : on observe par exemple 762 occurrences du motif de longueur 8 `gctggtgg` dans une séquence d'ADN de longueur $\ell = 4\,638\,858$ (en fait le génome complet de *E. coli*) et on se demande si ce comptage ne serait

² La valeur de m est choisie par le modélisateur et permet de fixer en moyenne la composition des séquences aléatoires jusqu'aux mots de taille de $m + 1$.

³ C'est le cas du motif `gctggtgg` avec 762 occurrences le long du génome de la bactérie *Escherichia coli* long de $4.6 \cdot 10^6$ lettres, ou du motif `aagtgcggg` avec 740 occurrences le long du génome de la bactérie *Haemophilus influenzae* de longueur $1.8 \cdot 10^6$.

⁴ C'est le cas des sites de fixation des facteurs de transcription indispensables à la transcription des gènes en ARN.

⁵ Plusieurs milliers, autant que le nombre de gènes par organisme.

pas significativement élevé. Intuitivement, en effet, si les 4^8 mots possibles de taille 8 avaient la même fréquence dans la séquence, on s'attendrait à les observer chacun 72 fois. Pour savoir si l'écart entre 762 (l'observé) et 72 (l'attendu) est significatif, il faut calculer la probabilité de l'événement $\{N \geq 762\}$ sous le modèle choisi, appelée *probabilité critique* (ou *p-value* en anglais). On se placera ici dans le modèle de chaîne de Markov d'ordre m ($0 \leq m \leq h - 2$, où h est la longueur du mot étudié) qui permet de s'ajuster sur la composition de la séquence d'ADN en mots de taille 1 à $m + 1$.

Le calcul de cette probabilité serait trivial si on connaissait la distribution du comptage N , mais cette dernière est complexe à obtenir du fait que l'on compte des occurrences qui potentiellement peuvent se chevaucher⁶ dans la séquence. En effet, le comptage N est une somme de variables aléatoires de Bernoulli Y_i ($Y_i = 1$ si le motif est présent à la position i , et 0 sinon) non indépendantes. Sa loi n'est donc pas une loi binomiale comme l'on aurait pu être tenté de dire.

Plusieurs approches ont été proposées soit pour calculer exactement la probabilité critique soit pour l'approcher. L'une des approches exactes consiste à calculer la distribution du temps d'attente T_n de la n -ième occurrence du mot, pour tout $1 \leq n \leq 762$, puis d'utiliser le principe de dualité suivant : $\mathbb{P}(T_{762} \leq \ell) = \mathbb{P}(N \geq 762)$. La distribution exacte du temps d'attente s'obtient par récurrence [4] ou via sa fonction génératrice [10] qu'il convient ensuite de développer en série de Taylor. Néanmoins, la valeur exacte de la probabilité critique n'est réellement calculable numériquement que pour des séquences de quelques dizaines de milliers de lettres et pour des ordres de modèles très faibles, 0 ou 1, ce qui en limite grandement l'usage pour l'analyse de génomes entiers.

Des approximations de la probabilité critique sont donc plutôt utilisées en pratique. Deux types d'approches ont été poursuivies : d'une part celles visant à approcher la distribution du comptage par des lois paramétriques explicites; d'autre part celles qui approchent directement la queue de la distribution par des approches de grandes déviations [1]. Ces dernières sont en effet pertinentes lorsque les motifs sont très exceptionnels. Parmi les distributions approchées, on peut noter la loi gaussienne pour les motifs plutôt fréquents [2], des lois de Poisson composées pour les motifs plutôt rares [8], voire la loi binomiale pour des motifs non périodiques (dont les occurrences ne peuvent jamais se chevaucher).

Si l'on reprend le problème mentionné au départ de ce paragraphe, il s'avère que la probabilité critique d'observer au moins 762 occurrences du motif gctggtgg dans une chaîne de Markov de longueur $\ell = 4\,638\,858$ et ayant en moyenne la même composition en mots de taille 1 à 7 que le génome de *E. coli*, vaut environ $8.7 \cdot 10^{-27}$ (en utilisant l'approximation gaussienne). Autrement dit, ce motif est significativement fréquent le long de ce génome (seul deux autres motifs de taille 8 ont une probabilité critique encore plus petite). D'un point de vue biologique, ce motif est vital à la bactérie puisqu'il intervient dans la réparation du génome en cas de cassure des brins d'ADN.

⁶ Par exemple, le motif périodique atgatga a deux occurrences chevauchantes dans la séquence gatgatgatgattc : l'une à la position 3 (soulignée), l'autre à la position 6 (en italique).

2. Présence attendue ou anormale ?

Ici ce qui nous intéresse est de savoir si le fait qu'un motif soit présent dans une séquence (plutôt courte) soit exceptionnel, c'est-à-dire non dû au hasard, ou pas. Pour cela, on peut bien sûr calculer ou approcher la probabilité $\mathbb{P}(N \geq 1) = \mathbb{P}(T_1 \leq \ell)$ en utilisant les approches décrites dans le paragraphe précédent.

Dans le cadre de la recherche de sites de fixation de facteurs de transcription, on est très vite amené à considérer l'occurrence de 2 motifs \mathbf{m}_1 et \mathbf{m}_2 (voire plus) à une certaine distance d plus ou moins variable dans une séquence ($d_1 \leq d \leq d_2$). Par exemple, ttgactt suivi de ataataa 16 à 18 lettres après. On s'intéresse donc à l'occurrence du motif composite résultant, noté $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$. On pourrait considérer ce motif composite comme un ensemble de longs motifs « simples » (de taille 30 à 32 dans l'exemple) et leur appliquer les résultats précédents, mais la distance est généralement trop grande, produisant un ensemble de motifs « simples » beaucoup trop grand (de l'ordre de 10^{18} dans l'exemple). Il faut donc recourir à des méthodes ad-hoc pour traiter ces motifs composites.

Voici trois approches qui ont été proposées pour évaluer la probabilité qu'un motif composite donné $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$ soit présent dans une chaîne de Markov.

– On peut approcher la probabilité $\mathbb{P}(N = 0) = 1 - \mathbb{P}(N \geq 1)$ par le produit $(\mathbb{P}(Y_i = 0))^{\ell-h+1}$ où Y_i vaut 1 si le motif composite (i.e. la première lettre du motif) apparaît à la position i dans la séquence, ou 0 sinon, et h est la taille du motif composite [13] ce qui revient à faire comme si les variables Y_i étaient indépendantes. On peut sensiblement améliorer l'approximation en considérant une dépendance d'ordre 1 entre les variables Y_i ce qui ramène à calculer le produit $\mathbb{P}(Y_1 = 0)(\mathbb{P}(Y_i = 0 | Y_{i-1} = 0))^{\ell-h}$ [6]⁷. Le calcul de $\mathbb{P}(Y_i = 0)$ pour un motif composite est moins trivial que pour un motif « simple » mais s'obtient à partir de la distribution du temps d'attente entre deux motifs « simples » (problème lié au paragraphe précédent). En effet, le motif composite $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$ apparaît en position i si et seulement si le premier motif \mathbf{m}_1 apparaît en position i et le 2^e motif \mathbf{m}_2 apparaît à une distance comprise entre d_1 et d_2 après le 1^{er} motif. Le calcul de $\mathbb{P}(Y_i = 0 | Y_{i-1} = 0)$ est plus technique mais faisable pour un motif composite composé de 2 motifs simples (le travail reste à faire pour 3 motifs ou plus).

– On peut décomposer le temps d'attente avant la première occurrence du motif composite comme une somme aléatoire de temps d'attente indépendants entre motifs simples [11]. Il faut néanmoins faire l'hypothèse que chacun des motifs simples ne peut apparaître plus d'une fois dans le motif composite. Dans le cas du motif composite $\mathbf{m}_1(d_1 : d_2)\mathbf{m}_2$, on obtient la décomposition suivante (cf. figure 1) : il faut attendre le premier motif \mathbf{m}_1 puis (***) attendre le prochain motif parmi $(\mathbf{m}_1, \mathbf{m}_2)$; s'il s'agit de \mathbf{m}_2 , on vérifie si la distance qui le sépare de \mathbf{m}_1 est bonne (entre d_1 et d_2) auquel cas on a trouvé le motif composite; si la distance n'est pas bonne, il faut rechercher le prochain motif \mathbf{m}_1 et repartir de (**); s'il s'agissait du motif \mathbf{m}_1 alors on repart de (**). Connaissant les lois du temps d'attente avant la prochaine occurrence de \mathbf{m}_1 et du temps d'attente avant l'occurrence d'un des motifs $(\mathbf{m}_1, \mathbf{m}_2)$ [5], [10], on en déduit la loi du temps d'attente du motif composite. Cette approche n'est cependant valable que pour des motifs composites à 2 motifs.

⁷ $P(A|B)$ désigne la probabilité de l'événement A sachant l'événement B .

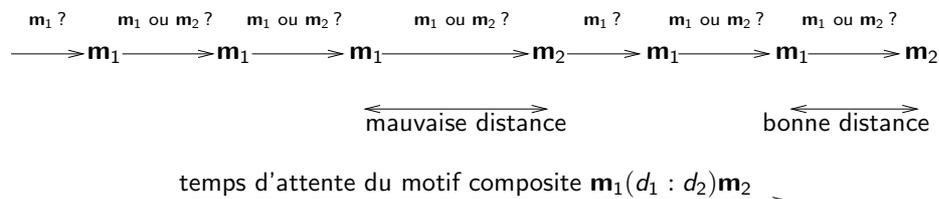


FIG. 1. Décomposition du temps d'attente avant l'occurrence du motif composite $m_1(d_1 : d_2)m_2$ comme la somme de temps d'attente indépendants entre les occurrences de m_1 et m_2 .

– L'approche précédente peut se reformuler en termes de construction d'un processus semi-markovien dans lequel les états du processus correspondent aux différentes étapes nécessaires pour obtenir une occurrence du motif composite [12]. Pour 2 motifs il n'y a que 3 états : (*état 1*) le motif m_1 est atteint (l'état suivant peut être l'état 1 ou l'état 2), (*état 2*) le motif m_2 est atteint mais à la mauvaise distance de m_1 (dans ce cas, l'état suivant sera l'état 1) et enfin (*état 3*) le motif m_2 est atteint à la bonne distance de m_1 (cet état est un état absorbant du processus semi-markovien). Si on reprend l'exemple de la figure 1, la succession des états serait la suivante : $\rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 3$. Le temps d'attente du motif composite est donc égal au temps d'absorbance du processus semi-markovien⁸ dont on sait calculer les probabilités de transition et les lois des temps de séjour dans chaque état grâce aux lois des temps d'attente entre motifs simples. Cette approche est facilement généralisable à un nombre quelconque de motifs simples.

3. Conclusion

Le lecteur intéressé par plus de détails mathématiques pourra se reporter au livre [7] ou aux chapitres d'ouvrages [3] et [9].

4. Références

- [1] NUEL, G. (2004). LD-SPatt : Large Deviations Statistics for Patterns on Markov Chains. *Journal of Computational Biology*, **11**, 1023–1033.
- [2] PRUM, B., RODOLPHE, F. and TURCKHEIM, Å. (1995). Finding words with unexpected frequencies in DNA sequences, *Journal of the Royal Statistical Society series B*, **57**, 205–220.
- [3] REINERT, G., SCHBATH, S. and WATERMAN, M. (2005). *Applied Combinatorics on Words*. volume 105 of *Encyclopedia of Mathematics and its Applications*, chapter Statistics on Words with Applications to Biological Sequences. Cambridge University Press.
- [4] ROBIN, S. and DAUDIN, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters, *J. Appl. Prob.* **36**, 179–193.
- [5] ROBIN, S. and DAUDIN, J.-J. (2001). Exact distribution of the distances between any occurrences of a set of words, *Ann. Inst. Statist. Math.* **36**, 895–905.

⁸ Le processus est semi-markovien dans le sens où les lois des temps de séjour dans chaque état ne sont pas géométriques, contrairement au cas markovien.

- [6] ROBIN, S., DAUDIN, J.-J., RICHARD, H., SAGOT, M.-F. and SCHBATH, S. (2002). Occurrence probability of structured motifs in random sequences, *Journal of Computational Biology*, **9**, 761–773.
- [7] ROBIN, S., RODOLPHE, F. and SCHBATH, S. (2003). *ADN, mots et modèles*. BELIN.
- [8] SCHBATH, S. (1995). Compound Poisson approximation of word counts in DNA sequences, *ESAIM : Probability and Statistics*, **1**, 1–16.
- [9] SCHBATH, S. and ROBIN, R. (2009). *Scan Statistics – Methods and Applications*. (J. Glaz, I. Pozdnyakov, and S. Wallenstein, ed.), chapter How can pattern statistics be useful for DNA motif discovery? Statistics for Industry and Technology. Birkhauser.
- [10] STEFANOV, V.T. (2003). The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models : an algorithmic approach, *J. Appl. Prob.* **40**, 881–892.
- [11] STEFANOV, V.T., ROBIN, S., and SCHBATH, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Appl. Math.* **155**, 868–880.
- [12] STEFANOV, V., ROBIN, S. and SCHBATH, S. (2011). Occurrence of structured motifs in random sequences : Arbitrary number of boxes. *Discrete Appl. Math.* **159**, 826–831.
- [13] SANDVE, G.K. and ABUL, O. and DRABLOS, F. (2008). Compo : composite motif discovery using discrete models. *BMC Bioinformatics*, **9** :527.

Segmentation pour l'analyse de puces CGH

Emilie Lebarbier ¹ et Franck Picard ²

Chaque espèce possède un nombre caractéristique de copies des chromosomes : chez la grande majorité des êtres vivants, comme chez l'homme, chaque chromosome est présent en deux copies (organisme diploïde) mais d'autres organismes peuvent être polyploïdes (ayant plus de 2 copies de chaque chromosome), comme par exemple la pomme de terre ou l'huître avec 4 copies. Une déviation de ce nombre de copies par rapport au nombre normal pour l'espèce (surnuméraire ou manquante) entraîne de ce fait un déséquilibre du nombre de copies des gènes, pouvant être à l'origine de maladies majeures. Un exemple classique chez l'humain est la trisomie 21, qui se caractérise, comme son nom l'indique, par la présence de 3 copies du chromosome 21. La détection de ces défauts chromosomiques est donc un élément majeur dans l'établissement du diagnostic et/ou du traitement de la pathologie. Si la détection se place à l'échelle du chromosome, elle est rendue possible grâce au traditionnel caryotype. Cependant, la perte ou le gain peut ne toucher qu'une portion du chromosome. Et lorsque cette anomalie chromosomique est de très petite taille, elle peut passer inaperçue. C'est en 1992 que l'étude de ces « petites » anomalies connaît un essor considérable grâce à la mise au point d'une nouvelle technique, l'hybridation génomique comparative ou CGH. La résolution a été nettement améliorée par l'utilisation de la technologie des microarrays (microarrays CGH ou puces CGH) permettant la détection d'aberrations d'environ 50kb. Après avoir été exclusivement appliquées à l'étude de la génomique du cancer, les microarrays CGH sont aujourd'hui utilisées dans d'autres études de génétique humaine.

¹ UMR Agroparistech/INRA MIA 518.

² LBBE, UMR CNRS 5558 Université Lyon 1 Villeurbanne, France.

L'objectif des expériences de microarrays CGH est donc de détecter et de cartographier des aberrations chromosomiques à l'échelle du génome, en une seule expérience. Son principe consiste à compter les variations du nombre de copies de séquences d'ADN entre deux échantillons d'ADN génomiques (un échantillon test et un échantillon de référence). Ce nombre n'étant pas mesurable directement, la technologie des microarrays est utilisée (une description de cette technologie est donnée page 5 du chapitre d'introduction) : les deux échantillons sont marqués par deux molécules de couleur différente. Le mélange est ensuite déposé sur une lame sur laquelle sont fixées des séquences d'ADN du génome de l'organisme étudiée (appelées séquences cibles). Les séquences ADN colorées vont alors « s'apparier » avec leur séquence complémentaire déposée sur cette lame et autant de fois que la séquence est présente dans le mélange. Ensuite de l'excitation des molécules de couleurs, sont récupérées deux intensités (une par couleur). Ainsi la valeur de ces intensités va refléter l'abondance des séquences cibles dans chacun des échantillons. Afin d'en avoir une représentation visuelle, le log-ratio de l'intensité de l'échantillon test par rapport à celle de référence est calculé pour chaque séquence cible puis ordonné selon la position physique de ces séquences sur le génome (position connue). On obtient alors ce que l'on appelle un profil CGH.

Pour l'étude de syndrômes humains, l'échantillon de l'individu sain est pris comme référence et celui de l'individu malade comme test. L'homme étant diploïde, le nombre de copies de chaque séquence dans l'échantillon de référence est toujours 2. Dans la mesure où la quantité d'intérêt est un nombre relatif de copies de séquences d'ADN, les log-ratios devraient appartenir à un ensemble de valeurs prédéfinies : par exemple la perte d'une séquence dans l'échantillon test (une délétion) devrait montrer un rapport de $\log(1/2)$ alors que le gain (une amplification) devrait montrer un rapport de $\log(3/2)$. C'est ce que montre la figure 1 (Gauche). Notons que l'amplification peut être multiple (2, 3 voire 10 copies) et la délétion peut apparaître sur les deux chromosomes. Cependant, le signal obtenu en pratique est loin d'être discret (cf. figure 1 (Droite)). Cette variabilité, induite non seulement par la variabilité naturelle (nature des génomes étudiés) mais aussi par l'expérience, rend très compliquée une analyse « à la main ». La statistique s'avère alors essentielle pour pouvoir tenir compte de cette variabilité et extraire l'information biologique sous-jacente. Comme on peut le voir sur la figure 1 (Droite), la spécificité des données de microarrays CGH est qu'elles sont spatialement ordonnées. En effet, elles se présentent le long du génome comme une succession de segments représentant des régions sur le génome dont les séquences partagent en moyenne le même nombre de copies. La détection et la localisation de ces régions permettent alors une interprétation biologique du signal : des segments de moyenne positive s'interpréteront comme des régions amplifiées sur le génome de l'échantillon test et des segments de moyenne négative comme des régions délétées.

Les outils statistiques naturels pour cette problématique sont les méthodes de segmentation dont l'objectif général est de détecter les changements abrupts qui apparaissent dans certaines caractéristiques d'un signal. Dans le cadre particulier des données de microarrays CGH, il s'agira de déterminer : (i) quelles sont les caractéristiques soumises aux changements (c'est l'étape de modélisation) et (ii) combien il y a de changements et où ils sont (c'est l'étape d'inférence statistique). La segmentation est un sujet important en statistique et fait l'objet d'une recherche

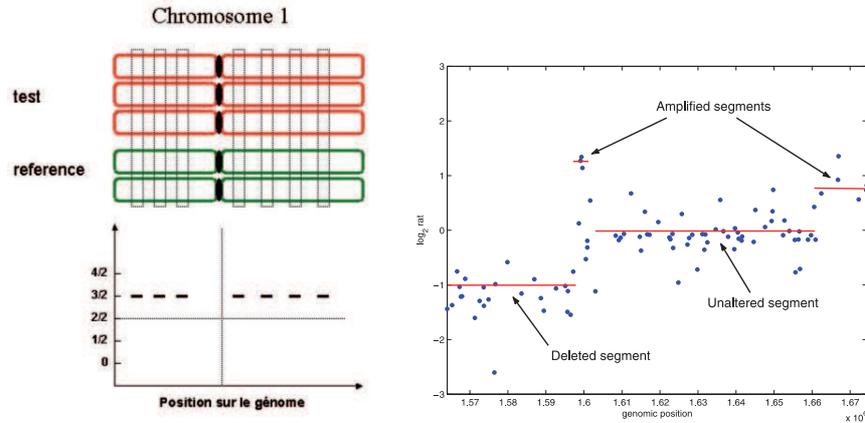


FIG. 1. Gauche : profil CGH théorique. Droite : profil CGH réel et interprétation. Ces figures sont extraites de [7].

intensive depuis plusieurs années. Cet intérêt est largement motivé par les nombreuses applications pratiques dans des domaines aussi variés que la biologie, la finance, la climatologie,... De nombreuses méthodes de segmentation ont été proposées pour l'analyse des données de microarrays CGH. Dans ce chapitre, nous en présentons une en particulier [7], qui s'est révélée être l'une des plus efficaces pour l'analyse de ce type de données [4].

1. Quel modèle pour les microarrays CGH ?

Il s'agit de déterminer un modèle décrivant aussi bien que possible le processus biologique. Ce processus peut être décrit par une fonction constante par morceaux : il existe une partition du génome notée I_1, \dots, I_K telle que les niveaux, valant μ_t à la position x_t et représentant le vrai log-ratio, sont constants à l'intérieur d'un intervalle et différents d'un segment à l'autre. On note μ_k la valeur du niveau au sein du segment I_k . Ces intervalles sont délimités par des instants, notés abusivement $t_1 < t_2 < \dots < t_{K-1}$, appelés instants de rupture. Le signal observé n'est alors qu'une version bruitée de ce vrai signal : le log-ratio observé à la position x_t est défini par une variable aléatoire Y_t (pour $t = 1, \dots, n$) telle que

$$Y_t = \mu_k + E_t \text{ si } t \in I_k.$$

La variabilité induite par l'expérience est prise en compte au travers de la variable aléatoire E_t . Ces variables sont supposées indépendantes et être issues d'une distribution gaussienne de moyenne nulle et de variance σ^2 . Cela revient à supposer que la distribution des Y_t est une gaussienne de moyenne μ_k et de variance σ^2 ($\mathcal{N}(\mu_k, \sigma^2)$) si t appartient au segment I_k . Ce modèle est noté \mathcal{M}_m . Notons cependant qu'un autre modèle pourrait être envisagé : à la fois la moyenne et la variance peuvent être sujettes aux changements. Dans ce cas, le modèle noté \mathcal{M}_{mv} s'écrit : $Y_t \sim \mathcal{N}(\mu_k, \sigma_k^2)$.

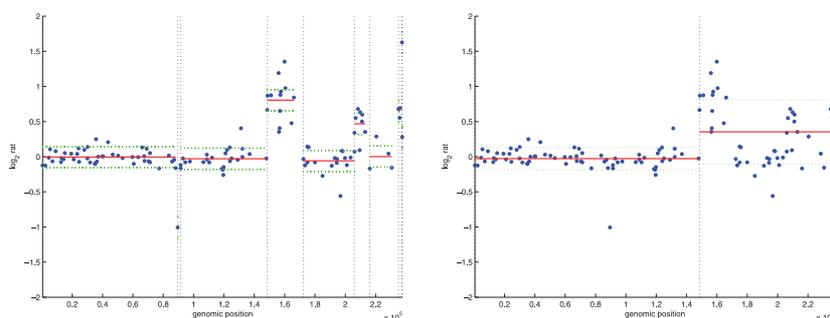


FIG. 2. Segmentations pour le modèle \mathcal{M}_m (Gauche), pour le modèle \mathcal{M}_{mv} (Droite)

Pour savoir quel modèle est le plus adapté à l'analyse des données CGH, il est non seulement important d'avoir une bonne connaissance des données mais aussi de bien connaître le comportement des deux modèles. Par définition, le modèle \mathcal{M}_{mv} autorise des variances différentes selon les segments. Ainsi avec ce modèle de petites régions successivement de niveaux différents pourront être regroupées en une longue région de grande variance, alors qu'elles seront chacune considérées comme un segment avec le modèle \mathcal{M}_m , comme l'illustre la figure 2. Cet argument milite en faveur de ce dernier modèle puisqu'il permet de mettre en exergue plus de régions, certes des régions pouvant être issues d'artefacts techniques mais aussi des régions d'intérêt biologique. On peut aussi noter une récente procédure de segmentation proposée par [1] dans le cadre de changements dans la moyenne mais sans contrainte sur la variance.

2. Comment obtenir la meilleure segmentation ?

Pour un nombre K de segments, une segmentation est décrite par les instants de rupture t_k et les moyennes μ_k . Ainsi rechercher la « meilleure » segmentation en K segments se réduit à déterminer les « meilleures » valeurs de ces paramètres. À cet effet, il est nécessaire de définir un critère de qualité des segmentations. Un critère classique en statistique est le critère des moindres carrés qui permet de mesurer l'écart existant entre la segmentation proposée et le signal observé. Il s'écrit ici :

$$(1) \quad J_K = \sum_{k=1}^K \sum_{t \in I_k} (Y_t - \mu_k)^2.$$

Trouver les valeurs des μ_k et t_k qui minimisent ce critère revient donc à chercher la segmentation en K segments qui s'ajuste le mieux aux données. La particularité des paramètres des instants de rupture t_k est qu'ils sont discrets (les moyennes étant continues). Ainsi la minimisation du critère portant sur ces paramètres nécessite, dans le cas d'une recherche exhaustive, l'exploration de C_{n-1}^{K-1} configurations, rendant impossible l'utilisation d'un algorithme naïf. Grâce à la propriété d'additivité de J_K en K , la segmentation optimale peut être obtenue en un temps raisonnable par un algorithme de Programmation Dynamique [2], emprunté à la théorie des graphes.

Cet algorithme repose sur le principe d'optimalité énoncé par le mathématicien Richard Bellman : « tout chemin optimal est formé de sous-chemins optimaux ». Cet algorithme a été récemment amélioré [10] permettant l'analyse de signaux de très grande taille, comme c'est le cas des nouvelles données de microarrays CGH qui comptent au moins un million de points.

Deux questions se posent alors. (i) Comment choisir le nombre K de segments ? (ii) Quelle confiance peut-on accorder à la segmentation obtenue ? Les réponses à ces questions ne sont pas si simples. Là encore le problème principal est la nature discrète des instants de rupture, qui ne permet pas d'utiliser les outils classiques de statistique.

Choix du nombre de segments K .

On dispose donc de la meilleure segmentation des données en K segments. Cependant, en pratique ce nombre n'est pas connu. Et même si l'on dispose d'informations a priori sur les données, il est difficile de se le fixer à l'avance, il s'agit donc de le déterminer. La qualité de cette meilleure segmentation est naturellement mesurée par la valeur du minimum de J_K en μ_k et t_k à K fixé, et noté \hat{J}_K . \hat{J}_K traduit l'ajustement du modèle aux données. Ce terme diminue avec le nombre de segments K : plus il y a de segments, plus la segmentation résultante sera proche des données. On est alors tenté de choisir le nombre maximal de segments pour avoir le meilleur ajustement. Dans ce cas, la segmentation résultante coïncide complètement avec les données, ce qui est sans intérêt. On a plutôt envie de choisir une segmentation suffisamment bien ajustée aux données sans être trop complexe, c'est-à-dire avec un nombre raisonnable de segments, afin d'obtenir une représentation des données suffisamment informative sans qu'elle soit difficilement interprétable. La traduction mathématique de ce compromis s'obtient en ajoutant au critère des moindres carrés, une fonction notée *pen* et appelée pénalité :

$$\hat{J}_K + \text{pen}(K).$$

Cette pénalité doit donc être une fonction qui reflète la complexité de la segmentation et augmenter avec elle. Le nombre de segments est ensuite choisi en minimisant ce critère, appelé critère pénalisé. Bien sûr toute la difficulté réside dans le choix d'une fonction de pénalité judicieuse. Bien que depuis ces dernières années de nombreux travaux aient été menés sur ce sujet ([5], [3], [6], [12], ...), il n'existe pas de critère universellement meilleur. La sélection du nombre de segments reste un problème ouvert.

Qualité de la segmentation obtenue.

La méthode de segmentation précédente propose au final une segmentation particulière. Mais à quel point peut-on avoir confiance en cette segmentation ? À quel point est-on sûr de la rupture localisée en t ? Cette dernière question se traduit statistiquement par « quelle est la probabilité qu'une rupture soit localisée en t ? ». Plus la probabilité est proche de 1, plus forte sera la certitude sur son existence. Une approche possible est l'approche bayésienne. Dans ce cadre, les paramètres (les μ_k et t_k) sont vus comme des variables aléatoires, donnant alors un sens à la probabilité qu'une rupture soit localisée en t ou comprise dans un intervalle. Le calcul exact de ces probabilités nécessite l'exploration de l'espace des segmentations (plus ou moins restreint selon la probabilité d'intérêt), qui, comme évoqué précédemment, pose d'importantes difficultés algorithmiques. Ces difficultés

peuvent être contournées pour certains modèles de segmentations [11], comme par exemple pour le modèle \mathcal{M}_{mv} .

3. Vers une nouvelle dimension

Aujourd'hui l'avancée importante de la technologie des expériences de microarrays offre la possibilité d'analyser simultanément plusieurs échantillons biologiques. L'arrivée de ces nouvelles données a généré de nouvelles questions, comme l'analyse jointe d'aberrations chromosomiques sur un ensemble de profils. Une telle analyse offre d'une part l'opportunité de corriger des artefacts techniques, qui sont partagés par tous les signaux issus d'une même lame, et d'autre part d'intégrer des informations sur chaque échantillon, comme des informations cliniques dans l'étude des cancers. Le passage au niveau multi-profils pose évidemment de nouvelles questions de modélisation statistique et amène de nouveaux défis algorithmiques [8].

4. Références

- [1] S. Arlot and A. Celisse, *Segmentation of the mean of heteroscedastic data via cross-validation*, Statistics and Copmputing, in prints (technical report arXiv :0902.3977v2, 2009) (2010).
- [2] R.E. Bellman and S.E. Dreyfus, *Applied dynamic programming*, Princeton University Press, 1962.
- [3] L. Birgé and P. Massart, *Minimal penalties for gaussian model selection*, Probability Th. and Related Fields **138** (2007), 33–73.
- [4] W.R. Lai, M.D. Johnson, R. Kucherlapati, and P. J. Park, *Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data*, Bioinformatics **0** (2005), n° 0, 1–8.
- [5] M. Lavielle, *Using penalized contrasts for the change-point problem*, Signal Processing **85** (2005), n° 8, 1501–1510.
- [6] E. Lebarbier, *Detecting multiple change-points in the mean of gaussian process by model selection*, Signal Processing **85** (2005), 717–736.
- [7] F. Picard, *Process segmentation/clustering. application to the analysis of array CGH data*, Ph.D. thesis, 2005.
- [8] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaiill, B. Thiam, and Robin, *Joint segmentation, calling and normalization of multiple CGH profiles*, Biostatistics **12** (2011), n° 3, 413–428.
- [9] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, *A statistical approach for array CGH data analysis*, BMC Bioinformatics **6** (2005), n° 27, 1, www.biomedcentral.com/1471-2105/6/27.
- [10] G. Rigaiill, *Pruned dynamic programming for optimal multiple change-point detection*, Tech. report, arXiv :1004.0887v1, 2010.
- [11] G. Rigaiill, E. Lebarbier, and S. Robin, *Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problems*, Tech. report, arXiv :1004.4347, 2011.
- [12] N. R. Zhang and D. O. Siegmund, *A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data*, Biometrics **63** (2007), n° 1, 22–32.

Tests multiples en génomique

Pierre Neuvial¹

1. Recherche de gènes différentiellement exprimés

L'émergence récente de problèmes de grande dimension, c'est-à-dire pour lesquels le nombre p de variables excède le nombre n d'observations – et ce parfois de plusieurs ordres de grandeur, nécessite des développements majeurs en statistique. En effet, les outils mathématiques classiques pour traiter des questions usuelles comme les problèmes de tests d'hypothèses ou de classification (supervisée ou non supervisée) ont été développés dans des contextes où $p < n$, et ne sont donc pas applicables tels quels en grande dimension.

Nous nous intéressons dans ce chapitre à la question des *tests multiples en génomique*, que nous introduisons grâce à l'exemple de la recherche de gènes dont le niveau d'expression (c'est-à-dire le niveau d'activité) diffère entre deux groupes de patients à partir de données de puces à ADN. On parle de *gènes différentiellement exprimés*. Notons qu'il existe des applications de la théorie des tests multiples dans tous les domaines faisant intervenir des données de grande dimension, non seulement en génomique mais aussi par exemple en imagerie médicale ou en astronomie.

En guise d'illustration, nous utilisons un des premiers jeux de données publiés : les données de Golub [2]. Il s'agit d'une collection de $n = 38$ expériences de puces à ADN effectuées sur des prélèvements sanguins de deux groupes de 11 et 27 patients atteints de deux types de leucémies (cancers du sang). Chacune de ces expériences fournit une mesure du niveau d'expression des mêmes $p = 3051$ gènes chez un des 38 patients. Un des objectifs de l'étude est l'identification de gènes différentiellement exprimés entre les deux types de leucémies. Dans les jeux de données plus récents, p est généralement de l'ordre de plusieurs dizaines de milliers, tandis que n dépasse rarement une centaine d'observations.

Ces données de puces à ADN sont représentées dans la figure 1. L'asymétrie de la matrice (a) est typique des données de grande dimension : ses lignes correspondent aux 3051 gènes et ses colonnes aux 38 patients. La matrice (b) est une extraction des 367 lignes de la matrice (a) dont le niveau d'expression diffère le plus fortement (en un sens précisé à la Section 2) entre les deux groupes, c'est-à-dire des gènes les plus différentiellement exprimés. L'objectif de ce chapitre est d'expliquer, en s'appuyant sur cet exemple, les enjeux statistiques sous-tendus par l'extraction d'un tel sous-ensemble d'hypothèses.

¹ Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 – USC INRA.

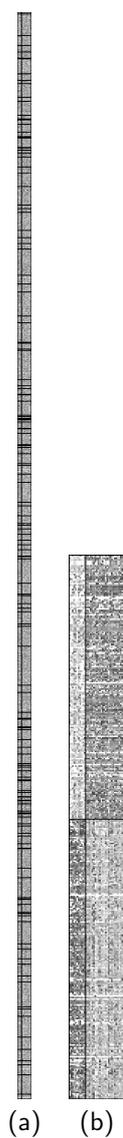


FIG. 1. Recherche de gènes différentiellement exprimés en grande dimension : matrices d'expression, avec les individus (patients) en colonne, et les variables (gènes) en ligne. La matrice des données originales (a) est de dimension 3051×38 ; la matrice (b), de dimension 367×38 , est un exemple de résultat : elle correspond à l'extraction des 367 lignes de (a) marquées en noir. Le niveau d'expression d'un gène est représenté par une couleur allant du noir (faible expression) au blanc (forte expression).

Si la théorie des tests statistiques permet, pour un gène donné, de quantifier la différence entre les niveaux d'expression des deux groupes, et même d'associer un degré de confiance à cette quantification, elle ne permet pas en revanche d'associer un tel degré de confiance à un ensemble de gènes testés simultanément. C'est l'objet essentiel de la théorie des tests multiples, qui inclut d'une part la définition de mesures de risques adaptées à un ensemble de tests, et d'autre part la recherche d'algorithmes permettant de contrôler ces risques ainsi que de conditions réalistes garantissant l'applicabilité de ces procédures en pratique. Pour une introduction plus complète à la théorie des tests multiples, nous renvoyons à un article récent du *Journal de la Société Française de Statistique* [4].

2. Tests

Un test statistique est une procédure permettant de valider ou d'invalider une hypothèse de travail. On s'intéresse ici au cas particulier où on cherche à identifier une différence entre les niveaux observés d'une variable pour deux groupes d'individus. Dans notre exemple, considérons un gène (parmi p). La variable observée est le niveau d'expression de ce gène, et les individus sont des patients atteints de deux types de leucémies. Elle correspond à une ligne donnée dans les matrices de la figure 1. Les ingrédients classiques pour la mise en place d'un test sont les suivants.

(1) Un modèle probabiliste, qui formalise les hypothèses faites sur la distribution de la variable observée. Par exemple, on pourra supposer que les niveaux d'expression du gène suivent une loi gaussienne d'espérance (c'est-à-dire de moyenne) μ_1 pour le groupe 1 et μ_2 pour le groupe 2.

(2) Une hypothèse de travail, qu'on cherche à valider ou invalider, qu'on appelle « hypothèse nulle » et qu'on note \mathcal{H}_0 . Dans notre exemple, il s'agit d'une définition mathématique de l'absence de différences réelles, comme l'égalité des moyennes μ_1 et μ_2 .

(3) Une règle de décision, qui associe à un jeu de données (ici, 11 observations issues du premier groupe et 27 issues du second) le rejet ou l'acceptation de l'hypothèse nulle. Par construction, elle garantit que la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie ne dépasse pas une valeur pré-spécifiée. Cette valeur est appelée *niveau* du test, et sera notée α . On appelle *probabilité critique* associée à un jeu de données le plus petit niveau de test permettant de rejeter l'hypothèse nulle.

La contribution principale de la statistique se situe à l'étape de construction d'une règle de décision. Dans notre exemple, l'approche la plus classique est le test de Student [3], dont la mesure est le ratio entre une estimation de la différence des moyennes des deux groupes, et une estimation de la variabilité. Sous l'hypothèse nulle, c'est-à-dire si $\mu_1 = \mu_2$, ce ratio suit une loi de probabilité connue, ce qui induit une règle de décision. Les 367 gènes de la figure 1(b) correspondent aux variables de plus grands ratios parmi les 3051 gènes initiaux.

3. Tests multiples

La « recette » de la section précédente permet de tester *une* hypothèse à un niveau α donné. Pour un gène donné, on sait donc décider s'il existe une

différence d'expression réelle entre deux groupes d'échantillons, avec un risque d'erreur contrôlé. Considérons maintenant la situation où l'on effectue simultanément le test de p hypothèses au niveau α . Chaque test a 4 issues possibles, selon que l'hypothèse nulle \mathcal{H}_0 est rejetée ou non, et selon qu'elle est vraie ou fausse (Tableau 1).

	\mathcal{H}_0 acceptée (négatifs)	\mathcal{H}_0 rejetée (positifs)	Ensemble
\mathcal{H}_0 vraie	VN	FP	p_0
\mathcal{H}_0 fausse	FN	VP	$p - p_0$
Ensemble	$p - R$	R	p

TAB. 1. Ventilation de p hypothèses testées en termes de nombres de vrais négatifs (VN), faux négatifs (FN), faux positifs (FP), vrais positifs (VP). R est le nombre total d'hypothèses rejetées, et p_0 le nombre (inconnu) d'hypothèses nulles vraies.

Notons p_0 le nombre total d'hypothèses nulles vraies (gènes non différentiellement exprimés). Le nombre FP de faux positifs, c'est-à-dire d'hypothèses rejetées à tort, est en moyenne $p_0\alpha$. Par conséquent, en supposant qu'aucun gène n'est réellement différentiellement exprimé (c'est-à-dire que $p_0 = p$), on attend environ 150 faux positifs si l'on effectue le test de $p = 3051$ gènes au seuil $\alpha = 5\%$.

3.1. Taux d'erreurs par famille : FWER

Pour limiter cette inflation du nombre de faux positifs, il est naturel de diminuer le niveau de chacun des tests individuels, dans le but de contrôler un risque global, associé aux p tests. Ainsi, si chaque test est effectué au niveau α' , il est possible de contrôler la probabilité qu'il y ait au moins un faux positif, que l'on appelle taux d'erreur par famille, ou FWER pour Family-Wise Error Rate. En effet, on a par sous-additivité de la mesure de probabilité

$$\text{FWER} \leq \sum_{\{i/\mathcal{H}_0 \text{ est vraie pour } i\}} \mathbb{1}_{\{\mathcal{H}_0 \text{ est rejetée pour } i\}}$$

Chaque rejet étant un événement de probabilité α' par définition, on obtient $\text{FWER} \leq p_0\alpha'$. Ainsi, effectuer chaque test au niveau $\alpha' = \alpha/p$ assure que la probabilité qu'il y ait au moins un faux positif est majorée par α (car $p_0 \leq p$). Cette procédure de test multiples, dite de Bonferroni, est peu utilisée dans les applications génomiques. En effet, le grand nombre p d'hypothèses testées la rend souvent trop stringente le seuil α/p étant fréquemment si petit qu'aucune hypothèse n'est déclarée significative. Elle fonctionne cependant bien dans l'exemple des données de Golub, où l'on obtient 98 rejets au seuil $\alpha = 5\%$.

3.2. Taux de fausses découvertes : FDR

La mesure de risque la plus couramment utilisée aujourd'hui dans les problématiques de tests multiples est le False Discovery Rate (FDR) [1]. Le FDR est la proportion attendue de faux positifs parmi les hypothèses rejetées, qui correspond donc à la proportion de gènes considérés à tort comme différentiellement exprimés parmi les gènes sélectionnés, soit, avec les notations du Tableau 4.1 :

$$\text{FDR} = \mathbb{E} \left[\frac{\text{FP}}{R \vee 1} \right].$$

Benjamini et Hochberg ont prouvé qu'une procédure (due à Simes [5]) permet de contrôler le FDR dans le cas où les hypothèses nulles vraies sont indépendantes. Soient $(X_i)_{1 \leq i \leq p}$ les probabilités critiques associées à p tests d'hypothèses. On note

$$\hat{k} = \max\{1 \leq i \leq p : X_{(i)} \leq \alpha i / m\},$$

où $(X_{(i)})_{1 \leq i \leq p}$ est un réarrangement croissant de $(X_i)_{1 \leq i \leq p}$. La procédure de Simes rejette toutes les hypothèses dont les probabilités critiques sont en deçà de $X_{(\hat{k})}$. Avec les données de Golub, on obtient 367 rejets pour un niveau FDR cible de 5% : ce sont les gènes de la figure 1(b).

Le risque FDR et la procédure de Simes sont rapidement devenus des standards pour les problèmes de tests multiples. Parallèlement, cette mesure de risque a engendré de nombreux développements statistiques récents en théorie des tests multiples. Nous en évoquons trois très brièvement, et renvoyons à [4] pour plus de détails.

- L'amélioration des procédures de contrôle du FDR via l'estimation de la proportion $\pi_0 = p_0/p$ d'hypothèses nulles vraies. En effet, la procédure de Simes appliquée à un niveau α garantit en fait un contrôle du FDR au niveau $\pi_0\alpha$, donc strictement inférieur au niveau prescrit dès que $\pi_0 < 1$.

- L'étude de l'impact de la dépendance entre les hypothèses testées sur le contrôle du FDR, afin de rendre les hypothèses faites sur cette dépendance plus proches des structures de dépendance observées en pratique, notamment dans les données génomiques.

- La définition d'autres mesures de risque inspirées par des limitations du contrôle du FDR. En particulier, le FDR est la valeur moyenne de la proportion (aléatoire) de faux positifs parmi les hypothèses rejetées, qui est notée FDP pour False Discovery Proportion. Par conséquent, contrôler le FDR ne renseigne pas sur les fluctuations de cette proportion ; il peut donc être intéressant de contrôler d'autres caractéristiques de la *distribution* du FDP, notamment ses queues de distribution.

4. De nouveaux enjeux méthodologiques

Outre les développements mathématiques que nous venons d'évoquer, la théorie des tests multiples doit faire face à de nouveaux enjeux méthodologiques, motivés par les possibilités d'exploiter de nouveaux types de données génomiques. Un premier axe est l'utilisation d'informations *a priori* sur les hypothèses à tester, notamment sur leur structure de dépendance. Dans le cas de la recherche de gènes différentiellement exprimés, les biologistes disposent souvent, outre des données

de puces à ADN, d'informations sur les réseaux d'interactions régissant les relations fonctionnelles entre les gènes (voir le chapitre sur les réseaux). Plutôt que de procéder en deux étapes : analyse différentielle pour sélectionner une liste de gènes, puis interprétation des résultats à la lumière des connaissances fonctionnelles, il paraît naturel de tester directement l'expression différentielle de réseaux de gènes (connus). Ceci implique d'une part la définition de procédures de tests appropriées pour un réseau donné, et d'autre part le développement de procédures de tests multiples adaptées.

Un autre axe important est l'arrivée récente de données de séquençage à haut débit, notamment pour la quantification des niveaux d'expression des gènes (RNA-seq). Ces données sont d'encore plus grande dimension, puisqu'on lit désormais des *millions* de séquences en une seule expérience, et leur nature diffère fondamentalement des données de puces à ADN car ce sont des données de comptage.

5. Références

- [1] Y. BENJAMINI and Y. HOCHBERG. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1) : 289–300, 1995.
- [2] T.R. GOLUB, D.K. SLONIM, P. TAMAYO, C. HUARD, M. GAASENBEEK, J.P. MESIROV, H. COLLIER, M.L. LOH, J.R. DOWNING, M.A. CALIGIURI, C.D. BLOOMFIELD, and E.S. LANDER. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286(5439) :531–537, October 1999.
- [3] W.S. GOSSET. The probable error of a mean. *Biometrika*, 6(1) :1–25, 1908.
- [4] E. ROQUAIN. Type I error rate control for testing many hypotheses : a survey with proofs. *J. Soc. Fr. Stat.*, To appear.
- [5] R.J. SIMES. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3) :751–754, 1986.

Réseaux biologiques

Julien Chiquet¹

L'essor d'internet et des réseaux sociaux a popularisé les graphes comme outil de représentation de relations entre variables dans tous les champs d'application. En biologie des systèmes, on les retrouve à divers niveaux de description des mécanismes, ainsi i) les réseaux métaboliques (graphes entre réactions et produits de réaction); ii) les réseaux protéines/protéines (graphes de proximité des séquences protéiques); iii) les réseaux de régulation (graphes des interactions du produit d'un gène sur la transcription d'autres gènes); ou iv) les réseaux de co-expression (graphes des corrélations entre niveau d'expressions de gènes).

Si l'analyse des réseaux biologiques a beaucoup emprunté aux télécommunications et aux sciences sociales, domaines où la théorie des graphes est solidement ancrée, les systèmes biologiques ont généré de nouvelles problématiques dues à l'extrême complexité des mécanismes en jeu, à leur instabilité et à la nature des sources de données associées (typiquement un très grand nombre de variables

¹ Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 – USC INRA.

pour peu d'observations). Dès lors les statistiques se sont imposées pour gérer les incertitudes liées aux phénomènes et à l'afflux massif de données de grande dimension, rendant souvent prohibitive l'utilisation à grande échelle de modèles déterministes.

Cet article se veut une brève introduction à quelques outils de statistique pour l'étude des réseaux biologiques. Nous présentons un cadre de modélisation largement répandu chez les statisticiens et nous survolons deux visions du réseau suscitant de nombreux travaux en statistiques mathématiques et appliquées : la première considère le réseau biologique comme étant l'observation ; la seconde s'appuie sur les données omiques pour reconstruire les arêtes du réseau. Nous tâcherons d'insister sur les spécificités de ces outils liées aux contraintes d'ordre biologique et à l'échantillonnage.

1. Modélisation

Notations. Pour fixer les idées, on considère dans la suite uniquement les graphes non dirigés composés d'un ensemble fixe $\mathcal{P} = \{1, \dots, p\}$ de nœuds, correspondant aux molécules d'intérêt (gènes ou protéines, par exemple). La présence d'une arête entre deux nœuds i et j de \mathcal{P} est définie par la variable aléatoire indicatrice $\Theta_{ij} = \mathbb{1}_{\{i \leftrightarrow j\}}$, ainsi $\Theta = (\Theta_{ij})_{i,j \in \mathcal{P}^2}$ décrit la matrice (aléatoire) d'adjacence du graphe. Par convention, les nœuds ne sont pas connectés à eux-mêmes, c'est-à-dire que $\Theta_{ii} = 0$ pour tout $i \in \mathcal{P}$.

Partant d'une collection de réseaux connus, on souhaite naturellement proposer un modèle suffisamment souple pour reproduire les attributs statistiques dominants de cette collection, comme la distribution des degrés des nœuds : typiquement, les réseaux biologiques se caractérisent par la présence d'un grand nombre de nœuds à faible degré et d'un petit nombre au degré très élevé, appelés « hubs ». Une telle distribution peut être décrite par une loi de puissance, popularisée par la physique statistique (voir [2]). Elle s'avère cependant trop souvent restrictive pour une description fine d'autres caractéristiques des réseaux biologiques, telles leur forte hétérogénéité ou le faible nombre effectif d'arêtes parmi le nombre possible : le nombre d'arêtes est typiquement de l'ordre du nombre de nœuds.

Le modèle à « blocs stochastiques »² s'intègre bien à ces problématiques : il s'agit d'un modèle de graphe aléatoire issu des sciences sociales réintroduit indépendamment par de nombreux auteurs. On trouvera dans [4] une présentation complète et de récentes extensions. Ce modèle propose de distribuer les nœuds \mathcal{P} en un ensemble $\mathcal{Q} = \{1, \dots, Q\}$ de classes latentes³ de probabilités *a priori* $\alpha = (\alpha_q)_{q \in \mathcal{Q}}$. Par la suite, on suppose pour simplifier Q connu et fixé. L'appartenance du nœud i à une classe est commodément décrite par le vecteur $\mathbf{Z}_i = (Z_{iq})_{q \in \mathcal{Q}}$ qui suit une loi multinomiale $\mathcal{M}(1, \alpha)$. Chacune des classes permet de représenter un groupe de variables dont le comportement dans le graphe est homogène, aussi bien entre elles que vis-à-vis des variables extérieures au groupe. On pense notamment à un ensemble de molécules impliquées dans les mêmes voies métaboliques. Cette propension à la connexion interne ou externe des nœuds d'un

² *Stochastic Bloc Model* en anglais.

³ C'est-à-dire non observées.

groupe est décrite par une série de paramètres $\pi = (\pi_{ql})_{q,l \in \mathcal{Q}}$, indiquant la probabilité pour qu'un nœud de la classe q se connecte avec un nœud de la classe l . La probabilité de présence d'une arête entre les nœuds i et j est définie *conditionnellement*⁴ à leurs classes respectives selon une loi de Bernoulli :

$$(1) \quad \Theta_{ij} \mid \{Z_{iq}Z_{jl} = 1\} \sim \mathcal{B}(\pi_{ql}), \quad i \neq j.$$

Ce modèle s'avère particulièrement souple, permettant la génération d'une grande variété de topologies de graphes composés de sous-graphes bipartites, en communauté ou en étoile par exemple.

2. Analyse des réseaux aléatoires

Une première classe de problèmes consiste à considérer l'objet réseau Θ comme étant l'observation. Ajuster un modèle à un graphe biologique est très utile à la découverte d'éléments structurants qui n'étaient pas directement visibles : les classes de nœuds correspondant à des classes de gènes peuvent permettre aux biologistes de formuler des hypothèses recoupant les connaissances disponibles souvent parcellaires. Par exemple, si dans un réseau de régulation un gène se trouve impliqué dans une structure en étoile, il est possible qu'il contrôle la traduction d'une protéine régulant la transcription d'autres gènes ; de même, l'existence d'une communauté de gènes fortement connectés suggère une implication des protéines sous-jacentes dans des voies métaboliques communes, parfois inconnues. Ainsi l'analyse de réseau permet la formulation d'hypothèses de travail pour les biologistes, voire leur validation dans certains cas, en vue d'isoler des structures ou motifs caractérisant une espèce ou une maladie.

Du point de vue statistique, l'ajustement du modèle à blocs stochastiques correspond à déterminer le paramètre β regroupant les proportions de classe et les probabilités de connexion, soit $\beta = \{\alpha, \pi\}$. La méthode d'estimation du maximum de vraisemblance est la plus naturelle en statistique : elle consiste à déterminer les valeurs des paramètres les plus vraisemblables vis-à-vis des données, en maximisant la probabilité d'occurrence des données sous le modèle choisi. On note $\mathbb{P}_\beta(\Theta)$ cette probabilité, correspondant à la loi du graphe observé sous les paramètres β , inconnus. La log vraisemblance associée, notée ℓ , « renverse » le problème en considérant les observations comme étant un paramètre fixé et β la variable. Ainsi, on cherche à résoudre le problème

$$\arg \max_{\beta} \ell_{\Theta}(\beta) = \arg \max_{\beta} \mathbb{P}_{\beta}(\Theta).$$

La spécificité du problème courant tient au fait qu'une partie de l'information disponible dans les données est manquante : on observe le graphe Θ alors que le modèle (1) requiert l'appartenance aux classes des nœuds via $\mathbf{Z} = \{Z_i\}_{i \in \mathcal{P}}$. La log vraisemblance du modèle ne peut être écrite qu'en marginalisant sur \mathbf{Z} , la distribution des arêtes Θ_{ij} n'étant connue qu'à classes fixées :

$$\ell_{\Theta}(\beta) = \mathbb{P}_{\beta}(\Theta) = \log \sum_{\mathbf{Z} \in \mathcal{Z}} \mathbb{P}_{\beta}(\Theta, \mathbf{Z}) = \log \sum_{\mathbf{Z} \in \mathcal{Z}} \mathbb{P}_{\beta}(\Theta | \mathbf{Z}) \mathbb{P}_{\beta}(\mathbf{Z}).$$

⁴ La loi conditionnelle est symbolisée par le symbole \mid .

Maximiser $\ell(\beta)$ s'avère impossible directement numériquement de par la présence de $Q^{|\mathcal{P}|}$ termes dans la sommation, correspondant à toutes les configurations possibles de \mathbf{Z} . Cette famille de problèmes, à savoir la maximisation de la vraisemblance d'un modèle à classes latentes, est bien connue en statistique depuis les années 1970 et se contourne à l'aide de la stratégie EM. Celle-ci repose sur la décomposition suivante de la log vraisemblance, utilisant la vraisemblance des données *complétées* par les variables latentes :

$$(2) \quad \ell_{\Theta}(\beta) = \log \mathbb{P}_{\beta}(\Theta, \mathbf{Z}) - \log \mathbb{P}_{\beta}(\mathbf{Z}|\Theta).$$

L'algorithme EM consiste à alterner une étape approchant la distribution conditionnelle de la classification \mathbf{Z} pour une valeur fixe des paramètres β (i.e., le deuxième terme dans (2)) et une étape déterminant les valeurs des paramètres β maximisant la vraisemblance complétée (le premier terme dans (2)) calculée sous l'estimation courante de la distribution conditionnelle de \mathbf{Z} . La suite ainsi construite pour les paramètres β conduit à un maximum local de ℓ_{Θ} .

Malheureusement, la stratégie EM échoue elle aussi dans le cas du modèle à blocs stochastiques car elle requiert le calcul explicite de $\mathbb{P}(\mathbf{Z}|\Theta)$, ce qui s'avère impossible de par la complexité des structures de dépendances entre variables dans le cas des modèles de graphes aléatoires. Pour résoudre ce problème, les statisticiens se sont tournés notamment vers les approches variationnelles : elles consistent à maximiser une borne inférieure de la vraisemblance, obtenue en approchant la distribution $\mathbb{P}(\mathbf{Z}|\Theta)$ par une forme factorisable, rendant possible les calculs (voir par exemple [3]).

Diverses extensions du modèle à blocs stochastiques et des méthodes d'estimation associées ont mené à de nombreuses publications, souvent dues au soucis de cohérence avec la biologie. Nous pensons par exemple à la possibilité pour un nœud d'appartenir à plusieurs classes, c'est-à-dire, pour une protéine, d'être au carrefour de plusieurs voies métaboliques : on parle de classe *chevauchante*.

À titre d'exemple, nous proposons à la figure 1 une vue du modèle à blocs stochastiques à 5 classes ajusté au réseau de régulation d'*Escherichia coli*, organisme modèle en biologie. Les nœuds sont les gènes de l'organisme et les arêtes décrivent des relations de régulations (activation ou inhibition) induites par les protéines codées par chacun des gènes. On distingue une structure très forte, en particulier la présence de « hubs » et de communautés correspondant respectivement à des gènes régulateurs, appelés facteurs de transcription, et à des groupes formés de gènes tous régulés par un même autre.

3. Reconstruction

Le problème de la reconstruction de réseau, parfois abusivement appelé « inférence de réseau », est radicalement différent. Il consiste, à partir d'une source de données relatives aux variables de \mathcal{P} , à déterminer les arêtes du graphe, c'est-à-dire à déterminer les éléments non nuls et éventuellement leur intensité dans la matrice Θ : le graphe considéré est alors *valué*.

À cet effet, il convient de définir une mesure qui permette de rendre compte des interactions effectives entre les variables d'étude. Cette mesure dépend directement de la nature des données biologiques en jeu. Dans le cadre de l'analyse du transcriptome, un échantillon correspond typiquement au niveau d'expression

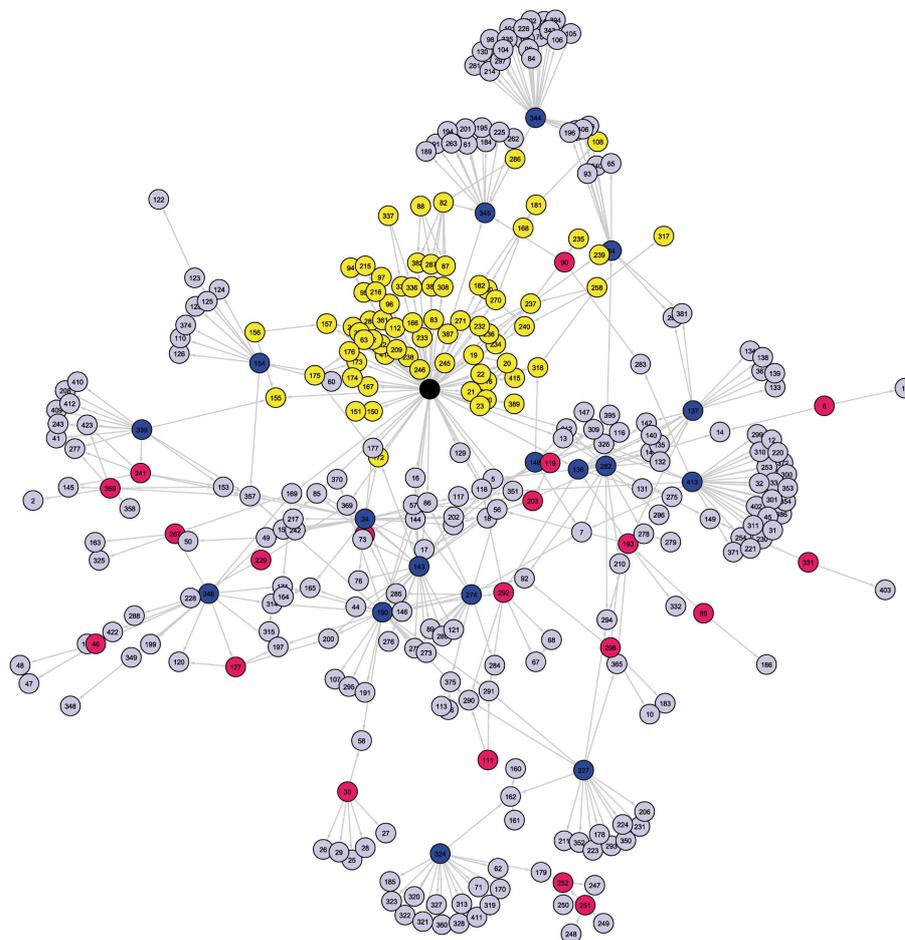


FIG. 1. Ajustement du modèle à blocs stochastiques avec $Q = 5$ classes au réseau de régulation d'*Escherichia coli* (source : [5]). Chaque niveau de gris correspond à une classe de nœuds.

des gènes d'un individu, soit de l'ordre de 22 000 variables pour l'être humain. Le nombre d'échantillons disponibles dépasse rarement la centaine. Les méthodes d'analyse différentielle (décrites précédemment) permettent de sélectionner les gènes prédisant au mieux une réponse d'intérêt, telle l'appartenance à un sous-ensemble de l'échantillon (individu sain/malade). L'inférence de réseau se situe naturellement dans la continuité de l'analyse différentielle, tâchant de déterminer parmi les gènes les plus explicatifs lesquels interagissent et régulent éventuellement les autres.

Un modèle couramment adopté en statistique et qui connaît un large succès dans les applications à l'inférence de réseau est celui des modèles graphiques gaussiens [6] : on décrit naturellement une expérience de puce à ADN à l'aide d'un vecteur aléatoire gaussien X de taille p dont la structure de corrélation est expliquée par la matrice $\Sigma = (\Sigma_{ij})_{i,j \in \mathcal{P}}$ de variance-covariance. Chaque entrée du

vecteur X correspond à un gène. À partir de n puces notées $\mathbf{X} = (X^1, \dots, X^n)$, l'estimation de la matrice de covariance doit permettre d'inférer la structure de dépendance entre les gènes, pour ainsi reconstruire un réseau. Une première idée pour reconstruire les interactions du réseau serait d'utiliser directement la covariance empirique entre les variables, se fondant sur le fait que⁵ $X_i \perp\!\!\!\perp X_j \Leftrightarrow \Sigma_{ij} = 0$ dans le cas gaussien. Pourtant, de par la complexité des systèmes biologiques en jeu, les réseaux reconstruits uniquement sur cette base seraient très peu informatifs : les fortes corrélations entre gènes sont trop nombreuses et les effets de cascades aboutiraient à la détermination de réseaux « pleins », chaque gène étant lié aux autres. On utilise plutôt comme mesure d'interaction entre variables la notion de corrélation partielle, qui consiste à ne s'intéresser qu'aux interactions *directes* entre variables. On s'appuie alors sur le fait que les (in)dépendances *conditionnelles* sont décrites par l'inverse de la matrice de covariance empirique dans les modèles graphiques gaussiens⁶ :

$$X_i \perp\!\!\!\perp X_j \mid \{X_{\mathcal{P} \setminus \{i,j\}}\} \Leftrightarrow \Sigma_{ij}^{-1} = 0.$$

Ainsi, la matrice $\Theta = \Sigma^{-1}$ décrit exactement le graphe de corrélations partielles ou de relations directes entre variables, et par là même entre les gènes. La log vraisemblance s'écrit en fonction de la covariance empirique $\mathbf{S}_n = \mathbf{X}'\mathbf{X}/n$ en vue de la détermination du paramètre d'intérêt, ici la matrice Θ :

$$(3) \quad \ell_{\mathbf{X}}(\Theta) = \frac{n}{2} \log \det(\Theta) - \frac{n}{2} \text{Trace}(\mathbf{S}_n \Theta) - \frac{n}{2} \log(2\pi).$$

Néanmoins, la maximisation de (3) pose problème : lorsque l'on dispose de suffisamment de données ($n > p$), on montre aisément que le maximum existe et est défini par \mathbf{S}_n^{-1} . Cependant, le graphe correspondant n'est que peu informatif puisque qu'aucune entrée de \mathbf{S}_n^{-1} n'est nulle et toutes les arêtes du graphes sont actives. Par ailleurs, lorsque $n < p$, l'inversion n'est pas possible. Une approche extrêmement populaire issue de problèmes de régression consiste à *pénaliser* le problème pour contraindre Θ à la forme voulue : le terme de pénalité vise non seulement i) à régulariser la solution ; ii) à induire de la parcimonie, c'est-à-dire des éléments exactement nuls dans le réseau reconstruit, afin de sélectionner les interactions les plus significatives. On peut même envisager d'introduire via cette pénalité un *a priori* sur la structure latente \mathbf{Z} du réseau à reconstruire, issue de techniques présentées dans la section précédente ou de connaissance biologique. L'estimateur régularisé peut donc s'écrire comme le problème d'optimisation

$$\arg \min_{\Theta} -\ell_{\mathbf{X}}(\Theta) + \lambda \cdot \text{pen}(\Theta, \mathbf{Z}),$$

où le paramètre $\lambda > 0$ permet de doser le compromis en le terme de vraisemblance aux données et la forme attendue du réseau.

À titre d'exemple nous avons réalisé à l'aide des outils décrits dans [1] une étude sur un lot de puces à ADN concernant des patientes atteintes de cancer du sein. Certaines d'entre elles avaient une très bonne réponse à la chimiothérapie, d'autres

⁵ Le symbole $\perp\!\!\!\perp$ est utilisé pour dénoter l'indépendance de deux variables aléatoires.

⁶ La notation $X_i \perp\!\!\!\perp X_j \mid \cdot$ est utilisée pour l'indépendance conditionnelle, et l'ensemble $\mathcal{P} \setminus \{i,j\}$ correspond à toutes les variables de \mathcal{P} sauf i et j . Ainsi $X_i \perp\!\!\!\perp X_j \mid \{X_{\mathcal{P} \setminus \{i,j\}}\}$ signifie que l'expression des gènes i et j est indépendante lorsque l'expression de tous les autres gènes est fixée.

ayant encore des résidus tumoraux après traitement. L'inférence de réseau, réalisée sur les puces correspondant à ces deux sous-ensembles, a permis d'identifier des interactions qui différaient entre les deux types de patientes (cf figure 2). Une étude poussée auprès des biologistes pourrait idéalement permettre de mettre à jour la défaillance de certains mécanismes de transcription chez les malades.

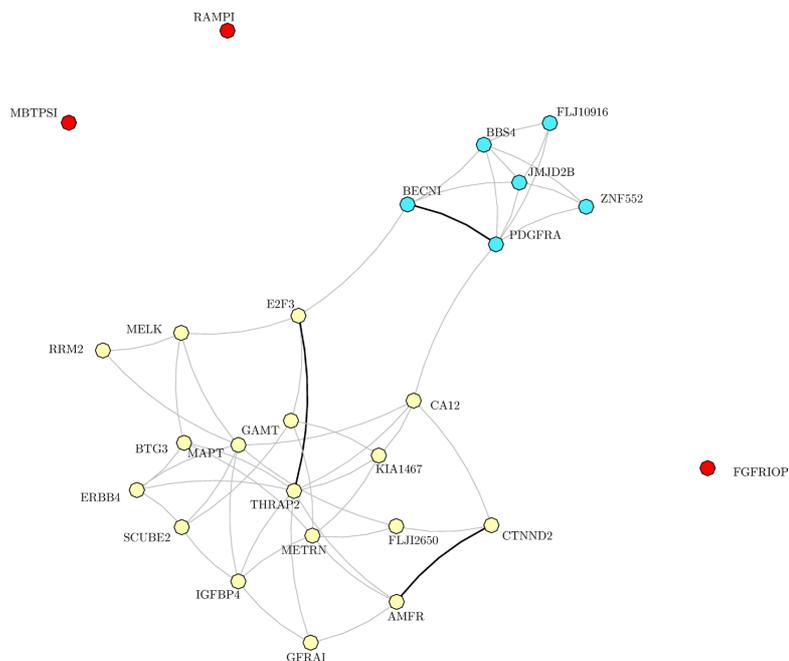


FIG. 2. Comparaison des réseaux reconstruits pour les données cancer du sein selon la réponse au traitement. Les traits épais indiquent les arêtes différenciant d'une série d'une patientes à l'autre.

4. Références

- [1] C. AMBROISE; J. CHIQUET; C. MATIAS. Inferring sparse Gaussian graphical models with latent structure, *Electronic Journal of Statistics*, 3 : 205–238, 2009.
- [2] A. BARABÁSI; R. ALBERT. Emergence of scaling in random networks *Science* 286 (5439) : 509-512, 1999.
- [3] J.-J. DAUDIN, F. PICARD, S. ROBIN, Mixture model for random graphs, *Statistics and Computing*, 2008.
- [4] P. LATOCHE, *Modèles de graphes aléatoires à structure cachée pour l'analyse des réseaux*, mémoire de thèse, 2011.
- [5] Groupe de travail SSB – Statistics for Systems Biology, *page web du logiciel MixNet*, <http://stat.genopole.cnrs.fr/logiciels/mixnet>.
- [6] J. WHITTAKER, *Graphical Models in applied multivariate Statistics*, 1996.

Le comité de rédaction remercie chaleureusement Patricia Reynaud-Bouret qui a coordonné ce dossier pour la *Gazette*.

MATHÉMATIQUES ET ART

Les beaux arts au service des mathématiques

Claude P. Bruter¹

Cet article reprend, en partie et dans sa version en langue française, celui paru dans la *Newsletter de l'European Mathematical Society* en septembre 2010, sous le titre équivalent de *Fine arts to the service of Mathematics*. Il s'enrichit ici d'un rapport sur une expérience pédagogique très positive menée à l'occasion d'une exposition récente à Aime (Savoie).

Il fut une période pendant laquelle les mathématiques ont eu une mauvaise presse². Il y eut comme un rejet général de la part de nos sociétés. N'ai-je pas entendu dire, l'année dernière encore, que « les enfants ont peur des mathématiques » ! Reste présent dans l'esprit de beaucoup une mésintelligence de ce que sont les mathématiques, mésintelligence qui a fortement contribué à ce rejet.

Les raisons de ce rejet tiennent sans doute aussi et en partie au comportement des mathématiciens eux-mêmes, aux caractères propres de leur milieu culturel. Nombre d'entre eux vivent presque en vase clos, travaillant avec bonheur, sans relâche, sur leurs sujets de prédilection, également soumis à une pression constante de la part de leurs pairs et du système administratif d'avancement et de reconnaissance dans lequel ils sont plongés. La mise en forme et la communication de leurs travaux obéit à des règles pleines de vertu professionnelle, mais leur caractère rigide est évident et modèle une part de leur attitude psychologique. Si ces règles de communication sont adaptées à leur communauté d'adultes professionnels, dont les habitudes mentales ont été façonnées par des dizaines d'années d'exercice ininterrompu, il est loin d'être certain qu'elles soient adaptées à tous les autres publics auxquels ils s'adressent ou pourraient s'adresser. Nombreux ont été les échecs

¹ ESMA, IHP, 11 rue Pierre et Marie Curie, 75231 Paris cedex 05.

² San Vu Ngoc m'a posé la question pertinente : « serait-il possible de préciser un peu la période pendant laquelle les maths ont eu une mauvaise presse ? » C'est en fait un véritable petit essai que l'on pourrait peut-être écrire à ce propos, relatif à l'image et à la place des mathématiques dans les diverses strates des sociétés humaines depuis les temps les plus reculés, à l'histoire de leur enseignement. La contestation a pris une forme particulière en France, liée à des considérations d'ordre sociologique et politique. Les réformes dans l'enseignement introduites par la Commission Lichnerowicz (1966-1972) furent à l'origine de cette contestation. On ne peut en quelques lignes décrire la manière dont elle s'est développée, les formes qu'elle a prises. Les lettres de lecteurs parues dans *Le Monde* du 6 juillet 1971 témoignent de l'introduction dans le public des débats entre mathématiciens. Il est possible que l'absence révélatrice, dans le monde politique et pour différentes raisons, de mathématiciens d'une grande ouverture d'esprit, ait contribué à maintenir une forme de crainte à l'encontre de la discipline.

pédagogiques, pour ne citer que les propos tenus par une part du public d'anciens élèves et étudiants.

Ces échecs de communication vers les jeunes générations se prolongent en échecs de communication et de compréhension entre mathématiciens et le public en général, qui fut jeune, et qui s'est presque totalement détourné du monde mathématique. Ce public fort large comprend maintenant et entre autres, celui de la majorité, si ce n'est la totalité, des décideurs économiques et politiques.

Ce sont tous ces publics, jeunes et moins jeunes, avec qui il serait humainement avisé et sage de reprendre contact dans la bonne humeur et l'intelligence réciproque.

Par quels moyens ?

Il en est sans doute plusieurs, mais l'un d'eux semble avoir montré son efficacité au cours des millénaires passés. Qu'il soit architectural, musical, pictural, poétique, sculptural, l'outil auquel je fais allusion est l'outil artistique. C'est à travers des œuvres de cette nature que les hommes se sont exprimés avec le plus de force de conviction, et finalement de persuasion, car une belle œuvre, par son seul pouvoir esthétique qui est d'abord un pouvoir attractif, se laisse contempler, admirer, entendre, écouter, générations après générations. Aucun effort intellectuel n'est a priori demandé. Le message diffuse pourtant ainsi subtilement auprès du public le plus large, qui finit par en assimiler plus ou moins complètement le contenu ou la signification.

Liés au pouvoir pénétrant de la lumière, les arts plastiques ont de tout temps occupé une place prépondérante dans les techniques de communication. Pourquoi ne pourrait-il pas en être de même dans les échanges entre le public et les mathématiciens, et cela avec d'autant plus de pertinence qu'une des parts fondamentales des mathématiques est la représentation et la description de l'espace, ce à quoi s'emploient également bien des artistes ?

C'est dans cet esprit que la collaboration entre mathématiciens et artistes peut conduire à affaiblir les barrières psychologiques qui séparent la communauté des mathématiciens du public en général.

En s'inspirant des formes si diverses que présentent les objets mathématiques, en les incluant de manière évidente ou plus cachée dans leurs œuvres, en réalisant une manière d'incarnation des mathématiques, les artistes les font connaître auprès des néophytes, familiarisent le public avec elles. Leur caractère abstrait s'estompe, elles acquièrent une matérialité, une évidence concrète, palpable, directement assimilable par les sens, et par induction, assimilable par la pensée. Le rejet s'affaiblit, s'évanouit. Les obstacles à la conversation, à une première explication sont levés.

Pour coordonner et aider les efforts en ce sens faits ici et là en Europe par divers de nos collègues, une Société Européenne pour les Mathématiques et les Arts³ a été officiellement créée en janvier 2010. Qu'il n'y ait point de méprise : nos amis d'autres pays et notamment américains y ont pleinement leur place. On peut espérer qu'avec vos encouragements, votre participation, votre soutien ne serait-ce que moral, elle parviendra à atteindre pleinement les buts qu'elle s'est assignée.

Un mot bref sur les activités récentes de l'ESMA : son premier colloque à Paris (juillet 2010 à l'IHP, volume des actes publié par Springer à paraître), accompagné d'une première exposition⁴, sera suivi d'un second colloque l'an prochain à Berlin.

³ <http://www.mathart.eu>

⁴ <http://www.mathart.eu/ihp10/index.html>

Dans le cadre du Festival Agora organisé par l'IRCAM en juin 2011, en même temps qu'une conférence internationale sur le thème « mathématique et musique »⁵, a été mise en place, de juillet à septembre au Palais de la Découverte qui en a assuré la médiation, une toute petite exposition ESMA. Une exposition de taille plus respectable a été présente du 7 juin au 23 juillet dans la ville savoyarde d'Aime, située au pied de la station de ski de La Plagne.

Cette exposition a été accompagnée d'exposés devant des élèves et des adultes. Leur succès fortifie la philosophie et la position de l'ESMA. Voici d'abord le type de propos d'adultes recueillis au moment de cette exposition, comme d'ailleurs lors des précédentes autrefois organisées par l'ARPAM, par exemple dans la belle galerie du Centre Culturel Christiane Peugeot à Paris⁶ : « Vous savez, je suis de formation littéraire, je n'ai jamais rien compris aux mathématiques, mais vraiment, je le regrette, j'aurais dû m'y intéresser davantage. » D'abord sensibles aux qualités esthétiques des œuvres présentées, ces personnes diffusent autour d'elles un nouveau rapport aux mathématiques. Les exposés à Aime devant des scolaires provenant du collège de Bourg-St-Maurice et du lycée de Moutiers ont été faits par Patrice Jeener (devant des élèves de quatrième) et par Jos Leys (à l'intention d'élèves de troisième, puis d'un ensemble comprenant des terminales L ayant choisi une option mathématiques et des premières S d'une classe européenne maîtrisant l'anglais - la date tardive d'ouverture de l'exposition n'a pas permis la présence de terminales S). D'une manière générale, tous les exposés ont été suivis dans le silence et avec attention. À la demande du professeur de mathématiques du Lycée de Moutiers qui accompagnait les élèves, et illustre certains de ces cours avec le DVD *Dimensions*, le second exposé de Jos Leys a été fait anglais. À la fin des exposés, un questionnaire en six points a été remis aux élèves. Voici donc quelques commentaires significatifs faits par des élèves ayant assisté à ce second exposé, on voudrait les citer tous :

1) *Votre première réaction après l'écoute de l'exposé est-elle :*

A- Positive B- Indifférente C- Négative ?

(merci de justifier votre réponse en quelques lignes)

Réponse « A » à l'unanimité.

« Jos Leys parlait très bien l'anglais, il était facile à comprendre. Ses explications étaient claires et précises. Il parlait avec passion, ce qui donnait envie de l'écouter ». « C'était facile à comprendre, et on ne parlait pas trop de mathématiques "pures" ». « C'était bien plus intéressant que ce que j'espérais ». « Cet exposé m'a permis d'envisager les mathématiques d'une façon différente, autre que des lignes de calcul ». « Exposé très clair, très accessible. Les nombreuses images ajoutent une dimension pédagogique, voire ludique. Donc très intéressant et bien mené ! »

2) *Ces exposés vous ont-ils permis de découvrir un nouveau visage des mathématiques ?*

A- Oui B- Non (merci de justifier votre réponse en quelques lignes)

⁵ <http://mcm2011.ircam.fr>

⁶ <http://hermay.org/ARPAM/>

Une exception intéressante commente sa réponse : « Non, au sens où je connaissais déjà cet aspect artistique, mais c'est toujours agréable de découvrir de nouvelles œuvres, ou des images/vidéos inédites ».

Sinon, un commentaire fréquent accompagne le Oui général :

« Jamais je n'aurais pensé qu'il était possible de créer des images des dessins grâce à des formules mathématiques ». « Oui, totalement, je ne pensais pas que les mathématiques pouvaient être artistiques ».

3) *Les images et objets vus et/ou commentés ont-ils éveillé votre curiosité ?*

A- Oui, sur quels plans ? (contenu mathématique, aspects techniques de fabrication...) B- Non, pour quelles raisons ?

Unanimité du « Oui » :

« Sur le plan technique, surtout avec les fractales et les nombres complexes (Mandelbröt) ».

Commentaire répété : « En les voyant, on se demande comment c'est possible de créer des images pareilles, comment les artistes font-ils pour conserver une logique dans le placement des formes ».

« Oui, car on peut se demander quel type de mathématiques peut se cacher derrière ces objets et comment on peut y parvenir ».

4) *Considérez-vous que les œuvres présentées sont des œuvres artistiques ?*

A- Oui, pourquoi et en quoi B- Non, pourquoi et en quoi

Unanimité du « Oui » :

« Car c'est beau ». « Tout est de l'art dans la vie ». « Oui car ils transforment des maths en image. C'est artistique quand les formes sont toutes emboîtées et se suivent ». « C'est un genre nouveau, abstrait, et ça incite les personnes à faire marcher leur imagination ».

5) *Après cet exposé, percevez-vous autrement leur dimension mathématique ?*

A- Oui B- Non

Unanimité du « Oui », aucun commentaire.

6) *Considérez-vous que la visite de cette exposition aura, à plus long terme, un effet sur :*

A- Votre manière d'envisager les mathématiques ? Oui/Non B- L'intérêt que vous portez aux arts ? Oui/Non C- Votre rapport aux sciences et à la connaissance en général ? Oui/Non

Alors que les premières S répondent en général « Non » aux questions A et B, les terminales littéraires y répondent plutôt par « Oui ».

Unanimité du « Oui » en réponse à la question C).

Devant son succès, les professeurs du Lycée de Moutiers ont souhaité le renouvellement de cette expérience et entendent agir en ce sens. Une des difficultés qu'ils rencontreront est bien sûr d'ordre financier (transport des œuvres, indemnités pour les conférenciers). Qu'on me permette ici de remercier la Présidente de

la commission des affaires culturelles du Canton d'Aime, également professeur de philosophie, d'avoir facilité la mise en place de l'exposition et des exposés qui l'ont accompagnée, également les personnes responsables de la Maison des Arts d'Aime, qui, par leur ouverture d'esprit, leur curiosité intellectuelle, ont pris l'initiative de provoquer cette exposition.

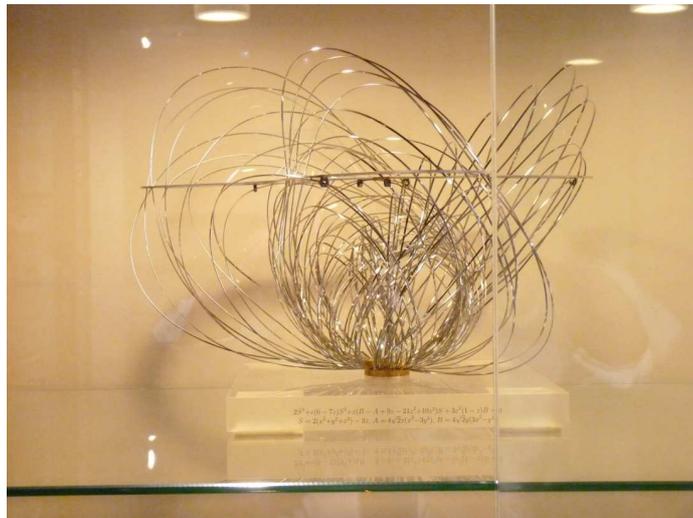
Une exposition du même genre que celle qui fut présente à Aime, avec donc exposés devant des élèves, se tiendra à Paris dans la première quinzaine de Février prochain, dans la grande salle Capitant de la Mairie du V^e, située face au Panthéon.

Souhaitons que d'autres établissements, d'autres villes accueillent des expositions de cette nature dont l'effet positif est à la fois réjouissant et porteur d'espoir.

Puis-je formuler un regret ? Nos ressources financières sont très maigres, elles ne nous permettent pas d'éditer un catalogue pédagogique quelconque, sans doute coûteux par le nombre d'images et d'œuvres que nous possédons, mais qui serait fort utile.

Les œuvres exposées relèvent non seulement de la peinture et de la gravure, mais aussi de la sculpture. À Aime, je fus un moment assez proche de la vitrine devant laquelle un groupe d'élèves s'étaient agglutinés, et où quelques objets de cette nature étaient présentés. J'ai eu la surprise d'entendre un garçon s'exclamer, il devait avoir dans les 16 ans : « Pour moi, c'est là ce qu'il y a de plus beau ! ».

Détail de la vitrine



Tablette haute : La surface de Boy en tant qu'immersion de l'espace projectif euclidien de dimension 2, ici feuilletée par des ellipses (François Apéry).



Seconde tablette : arêtes internes du dodécaèdre, au milieu un nœud de trèfle régulier (Philippe Rips)⁷.



Troisième tablette : deux polyèdres de George Hart (le polyèdre de gauche est lié à l'icosidodécaèdre et à l'icosaèdre, celui de droite au grand dodécaèdre), au milieu triangles d'arêtes internes à l'icosaèdre (Philippe Rips).

⁷ Ces configurations apparaissent également dans l'ouvrage de Alan Hoden, « Shapes, space and symmetry », remarque que je dois à l'obligeance de Gary Greenfield. Philippe Rips les a construites dans l'ignorance totale de cet ouvrage.



Quatrième tablette : au milieu, quatre nœuds de trèfle provenant des arêtes internes d'un tétraèdre et donnant une représentation topologique de la surface de Boy (Bruter-Rips), de part et d'autre deux nodus de Kozlov (nœuds cycliques) à courbure elliptique : toutes ces petites « sculptures » peuvent être aplaties et retrouver leur forme originale dès la contrainte d'aplatissement levée.



Tablette basse : deux surfaces minimales bordées par des nœuds (John Sullivan).

Je voudrais rappeler que les enfants sont, pour la plupart d'entre eux, encore très proches de la réalité concrète, et qu'ils peuvent entrer rapidement dans l'univers des objets mathématiques par d'autres voies que celle faisant appel à la manipulation d'entités abstraites avec lesquelles ils sont fort loin d'être spontanément familiers. Mais jusqu'à quel âge est-on enfant ? François Apéry, qui fait un remarquable travail d'archivage et de reconstitution des petites sculptures mathématiques, ne manque pas de rappeler que bien des modèles présents à l'IHP servaient d'outils pédagogiques efficaces à Darboux et aux géomètres de son époque.

Là où les salles dans lesquelles peuvent être présentées les œuvres jouent un rôle essentiel dans l'appréciation de leur qualité, de leur beauté, dans leur impact sur le public. Si l'IHP par exemple n'est pas encore l'endroit idéal où l'exposition peut être du plus bel et du plus efficace effet, la présence d'expositions en ce lieu, et quel lieu pour les mathématiciens, a une valeur plus que symbolique. En 2008 et 2009, deux expositions Mathématiques et Arts [4] ont été tenues à Paris, Avenue de la Grande Armée, dans les salles parfaitement adaptées du Centre Culturel Christiane Peugeot : elles furent alors fortement appréciées d'un public tout à fait étranger aux mathématiques. Ce fut une chance qu'elles puissent être montrées en ce lieu. Sans doute, nos grands musées pourraient-ils leur ouvrir leurs portes. Mais nous vivons encore dans une époque où le mot « Mathématiques » conserve des connotations négatives, auxquelles semblent très fortement sensibles les directeurs de musée. L'attitude à notre égard des artistes locaux du canton d'Aïme n'en est que davantage exemplaire.

Un dernier mot. Nos bâtiments scolaires et universitaires sont pour la plupart d'un misérabilisme esthétique aussi déprimant que peu digne du renom que nous souhaiterions leur voir attribué. Leur qualité architecturale n'est guère favorable à l'épanouissement intellectuel, n'incite guère l'étudiant au travail. Contrairement à ce qui frappe en général dans les pays étrangers, ils sont dans leur très grande majorité totalement dénués de halls d'accueil généreux et beaux, de salles de théâtre, de lieux d'exposition ouverts à l'enrichissement culturel, aucune peinture, aucune sculpture. Mais n'appartiendrait-il pas à chacun des membres de la communauté de peser pour offrir à tous un cadre de travail et de réflexion plus adapté aux conditions psychologiques subtiles qui favorisent la réussite ?

INFORMATIONS

Bilan du concours chercheurs 2011

Patrick Dehornoy¹

Un total de 28 postes de chercheur a été ouvert au recrutement en section 01 (mathématiques et applications), faisant l'objet de huit concours, soit

- un concours pour 9 postes de DR2 non profilés,
- un concours pour 2 postes de CR1 non profilés,
- un concours pour 9 postes de CR2 non profilés,
- un concours pour 4 postes de CR2 profilés « interactions des mathématiques en relation avec d'autres disciplines »,
- quatre concours pour chacun 1 poste d'échange destiné à être affecté dans un de leurs laboratoires respectifs par l'INP (physique), l'INS2I (informatique), l'INSIS (ingénierie), et l'INEE (écologie), ce dernier concours profilé « modélisation numérique du vivant et/ou de l'environnement (dynamique de la biodiversité) » avec affectation dans un laboratoire de Montpellier.

Par ailleurs, trois concours d'autres sections prévoyaient une affectation par l'INSMI dans un de ses laboratoires, soit

- un concours en section 02 (théories physiques, méthodes, modèles, et applications) pour 1 poste de CR2 profilé « interface physique-mathématiques »,
- un concours en section 04 (atomes et molécules, optique et lasers, plasmas chauds) pour 1 poste de CR2 profilé « thématique ITER »,
- un concours en section 07 (sciences et technologies de l'information) pour 1 poste de CR2 non profilé.

Sélection des candidats

Après avoir procédé, pour les concours CR, à une audition des candidats recevables, les sections du Comité National de la Recherche Scientifique (CoNRS) ont constitué pour chacun des concours une liste ordonnée de candidats déclarés admissibles. À cette étape, deux des concours ont été déclarés infructueux, à savoir le concours de la section 01 pour deux postes de CR1, et le concours de la section 04 pour un poste de CR2 profilé « thématique ITER ».

Un second jury, dit d'admission, a constitué la liste des candidats admis en se basant sur les listes d'admissibles. Les jurys d'admission ont tous suivi les classements du CoNRS concernant les mathématiques.

Tenant notamment compte des concours infructueux (mais sans que cela procède d'une démarche automatique), la direction du CNRS a accordé deux

¹ INSMI-CNRS.

postes supplémentaires de CR2 à l'INSMI, ce qui a permis de recruter les candidates classées premières de la liste complémentaire du concours CR2 section 01 non profilé et du concours CR2 section 02 profilé « interface physique-mathématiques ».

Affectations des chercheurs entrants

Globalement, 28 chercheurs ont été lauréats du concours 2011 avec vocation à être affectés dans un laboratoire de l'INSMI. Parmi ceux-ci, 21 sont des chercheurs nouveaux au CNRS (2 au niveau DR2, et 19 au niveau CR2), et 7 sont des chargés de recherche promus directeurs. Les deux DR2 recrutés à l'extérieur du CNRS sont issus, pour l'une, d'une université étrangère (USA), et pour l'autre, d'un poste de maître de conférence dans une université française.

Même si les résultats globaux restent éloignés de la parité, on pourra noter que, parmi les 28 lauréats du concours, 7 (soit 25%) sont des femmes, avec 3 sur 9 au niveau DR et 4 sur 19 au niveau CR.

Comme en 2010, l'affectation des chercheurs entrants a été préparée par une phase de concertation entre la direction de l'INSMI, les chercheurs, et les directeurs d'unité durant la période séparant la diffusion des listes d'admissibilité (mi-avril) et les jurys d'admission (mi-juin). Les propositions de la direction tiennent compte de trois facteurs principaux : vœux et projets individuels des candidats, politiques des laboratoires, et mission d'animation de l'INSMI vis-à-vis du réseau national des laboratoires, cette dernière impliquant d'une part une répartition géographique raisonnée avec un équilibre Paris-province et l'irrigation d'un grand nombre de laboratoires et, d'autre part, une mobilité au recrutement CR et au passage CR-DR. Les discussions avec les candidats ont montré que ces principes de mobilité sont désormais bien connus et acceptés dans la communauté mathématique, même si, au cas par cas, des difficultés liées aux contraintes individuelles peuvent subsister, notamment en ce qui concerne un éloignement de la région parisienne. La direction de l'INSMI remercie les chercheurs et les directeurs d'unité qui ont participé de façon constructive à ces discussions.

Au total, la répartition géographique des 28 lauréats 2011 est de 5 DR et 8 CR pour la région parisienne (dont 1 DR et 5 CR pour Paris intra muros) et de 4 DR et 9 CR pour la province – soit, globalement, un équilibre Paris-province. On peut noter que 25 laboratoires de l'INSMI se voient affecter au moins un chercheur en 2011, nombre passant à 34 si on cumule 2010 et 2011 – ce qui signifie qu'il n'existe que 11 laboratoires CNRS de mathématiques qui n'ont pas eu de chercheur affecté au cours des deux dernières années. Ces chiffres ne sont pas le signe d'une volonté de saupoudrage, mais une constatation a posteriori des besoins des laboratoires et des possibilités d'affectation.

Perspectives 2012

Les cadrages budgétaires pour 2012 ne sont pas encore faits. Les priorités de l'INSMI seront, au niveau CR2, le maintien d'un concours blanc suffisant, d'un concours « interactions » sur des thématiques pluridisciplinaires, et de postes d'échange avec d'autres sections du Comité National, et, aux niveaux CR1 et DR2, la possibilité de recrutements externes. Le maintien d'un flux important de CR passant professeurs dans des universités est un enjeu stratégique.

Concernant les affectations, en particulier les questions de mobilité, il est naturellement malaisé de convaincre les chercheurs d'infléchir leurs vœux a posteriori. Comme en 2011, nous demanderons à chaque candidat d'avoir réfléchi à **au moins deux affectations dont l'une au moins située hors de la région parisienne.**

Cas particulier du concours DR2

Nous souhaitons insister sur le fait que, à la différence des promotions CR2-CR1 ou DR2-DR1-DR0, l'accession au niveau DR2 est un concours de recrutement et qu'il doit être traité comme tel. De même qu'une candidature PR à l'université, une candidature DR au CNRS doit s'accompagner d'un projet de recherche et d'insertion dans un laboratoire qui ne soit pas seulement la poursuite de l'activité antérieure. Comme pour les PR, une mobilité est attendue pour les recrutements DR. Par ailleurs, la direction de l'INSMI rappelle que le concours DR2 est ouvert à tous et souhaite que des candidats externes au CNRS (candidats étrangers, mais aussi MC et PR des universités françaises) participent au concours.

Journée parité en mathématiques du 6 juin 2011

Magali Ribot, Barbara Schapira

Le lundi 6 juin s'est tenue à l'Institut Henri Poincaré la première journée sur la parité en mathématiques. Cette journée s'annonçait d'emblée comme un succès, au vu du grand nombre de participant-e-s (presque une centaine). Nous souhaitons prolonger ici cette journée en tenant compte des différents témoignages et des débats qui l'ont animée et vous proposer un rapide état des lieux de la parité en mathématiques, en nous concentrant sur les femmes déjà en poste. Le programme et une partie des transparents des intervenants sont disponibles à l'adresse : http://postes.smai.emath.fr/parite/journee/journee_parite.php.

Des statistiques parlantes

Le tableau de la situation a été brossé dès le début de la journée grâce à l'intervention de Laurence Broze, très riche en données chiffrées très parlantes. Tous ces chiffres, ainsi que des graphiques probants, se retrouvent sur le site d'Image des Maths : <http://images.math.cnrs.fr/Les-chiffres-cles-de-la-parite-en.html>.

Une des premières remarques est que les bonnes volontés des instances pour fournir des chiffres liés à la parité ne suffisent pas : les pourcentages fournis ne permettent pas d'avoir une interprétation fine et les indicateurs statistiques avancés ne sont pas toujours pertinents. Les chiffres présentés par L. Broze ont donc été collectés grâce à un travail extrêmement fastidieux.

À la figure 1, nous présentons le pourcentage de femmes à l'université, toutes disciplines confondues en haut et en mathématiques (sections 25-26) en bas. Avec seulement 27% (resp. 42% globalement) de femmes parmi les maîtres de conférences et 11% (resp. 23% globalement) parmi les professeurs,

les mathématiques sont dans une situation critique. La section 25, avec 14% de femmes, est même la section la moins féminisée de toutes, toutes disciplines confondues.

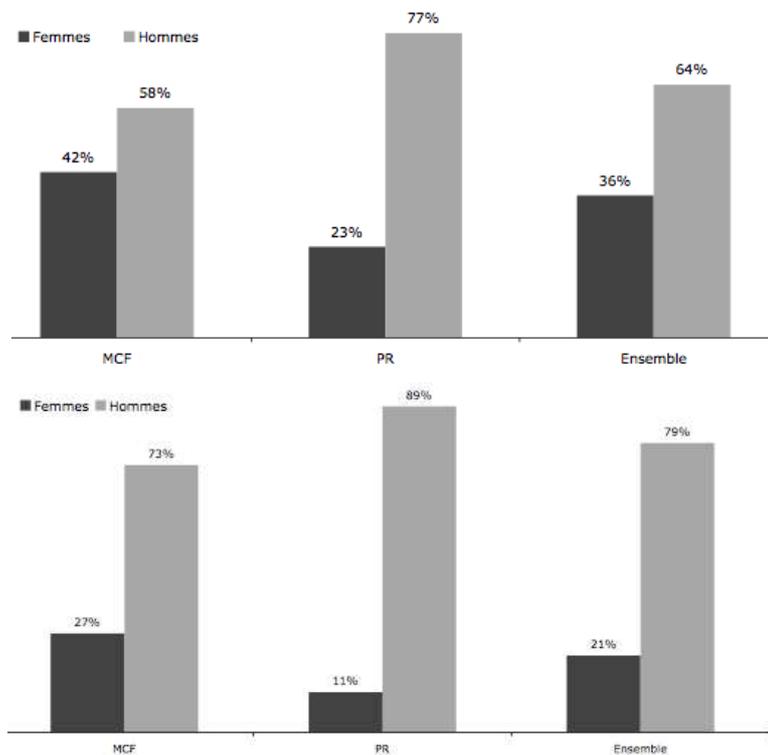


FIG. 1. La proportion de femmes toutes sections confondues (en haut) et en mathématiques (en bas).

On pourrait au moins espérer que la situation se soit améliorée ces dernières années. Or, il n'en est rien. Le pourcentage des femmes en mathématiques stagne depuis 15 ans : une belle suite constante ! Pire, en section 25, les périodes d'embellie de postes (1996-2005) n'ont pas profité aux femmes, les périodes de réduction (2006-2010) leur ont été très défavorables. Pour affiner ces données, il faut distinguer les sections 25 et 26 et séparer les maîtres de conférences (MCF) des professeurs (PR) : la proportion de femmes en 26 croît faiblement, alors qu'en 25, on observe une nette décroissance depuis 1996, que ce soit chez les MCF ou chez les PR. Le pourcentage de femmes professeurs en section 25 est même dramatiquement faible : 6,5%.

Et au CNRS ? Le nombre de femmes stagne lui aussi depuis 20 ans alors que le nombre d'hommes a augmenté régulièrement. On est donc passé de 19% de femmes en section 1 à 15% en 20 ans. Là aussi, la section 1 est la deuxième section la moins féminisée du CNRS.

Enfin, voici un dernier chiffre intéressant et très récent : lors des comités de sélection de 2011, un quart environ des comités ne comportait aucune femme.

Essayons maintenant de donner un caractère un peu plus humain à tous ces chiffres et d'extraire quelques facteurs explicatifs obtenus à travers les témoignages de cette journée et les exposés divers.

Quelques éléments d'explication

De l'avis de toutes, les femmes sont contentes d'être là ! Pas de sexisme apparent dans les laboratoires, une liberté identique à celle des hommes,... sont des facteurs favorisant l'accueil des femmes. Mais les problèmes persistent. Alors, pourquoi ?

Rappelons déjà le contexte général, défavorable aux femmes sur le plan professionnel : taux plus fort de chômage chez les femmes, plus de sous-emploi (travail à temps partiel), un salaire moindre à travail égal, un niveau de retraite nettement plus faible, un taux de pauvreté plus grand... De plus, la répartition des tâches ménagères reste très inégalitaire de nos jours. Des enquêtes à intervalles réguliers permettent de faire le point : ce sont toujours les femmes qui s'investissent le plus à la maison. À noter que lorsque le nombre d'enfants augmente, le taux de participation des hommes au travail ménager décroît et l'investissement dans leur activité professionnelle augmente. Finalement, la seule façon d'investir les pères dans la vie du foyer est d'exercer soi-même une activité valorisée.

Parlons aussi des reprises difficiles après les congés maternité. La fatigue d'un nouveau-né à la maison, les 4 mois (ou plus !) d'arrêt de travail, la nouvelle condition de maman, sont autant de facteurs qui pénalisent les femmes et retardent leurs carrières. Et puis des enfants à la maison rendent la mobilité très difficile, souvent bien plus difficile pour la mère que pour le père d'ailleurs...

Un autre point bien connu et très visible est l'auto-censure des femmes. On la retrouve à tous les niveaux. Hésiter à se lancer dans le bain de la prépa, ne pas « oser » faire de la recherche et préférer une carrière dans l'enseignement, tarder à passer l'HDR¹ et postuler PR : le phénomène classique d'auto-censure des femmes se retrouve à l'université. Les femmes ont aussi besoin de s'identifier pour pouvoir « oser » ; or, le très faible pourcentage de femmes PR rend difficile cette identification. À quoi bon passer l'HDR s'il y a si peu de femmes PR ? Franchement, le modèle du professeur en mathématiques « classique » est peu encourageant pour les femmes : il faut bien avouer qu'il s'agit souvent d'un homme plutôt jeune et plutôt issu des grandes écoles... Enfin, n'oublions pas les préjugés sociétaux. Oui, le préjugé de la femme nulle en mathématiques perdure et reste en sourdine dans tous les esprits. Une femme matheuse aura besoin de s'affirmer bien plus qu'un homme. Il est intéressant de noter qu'en Allemagne de l'Est, qui avait une politique favorable aux mères, la différence entre garçon et fille au niveau du lycée et des sciences était peu perceptible. Serait-ce un modèle à suivre ?

Ce qui existe

Faisons le point de quelques mesures en faveur de la parité.

Depuis 2009, la CPU² a signé une charte pour la parité. Un groupe de travail a été créé, qui réunit la MIPADI (Mission de la PARité et de la lutte contre les Discriminations du Ministère de l'Éducation Nationale et de la Recherche), les

¹ Habilitation à Diriger les Recherches.

² Conférence des Présidents d'université.

chargés de mission parité des universités et la CPU. Ce groupe s'intéresse aux carrières des femmes dans la recherche, à la mixité des populations étudiantes et au développement d'études sur le genre.

En 2001, est née la Mission pour la place des femmes au CNRS afin de promouvoir l'égalité homme-femme, un des objectifs du contrat 2009-2013 du CNRS. Cette mission permet entre autres d'avoir des données sexuées et d'élaborer des journées de formation, notamment pour les cadres du CNRS. À noter, le projet européen INTEGER (INstitutional Transformation for Effecting Gender Equality in Research) du 7e PCRD³ a débuté cette année et a pour mission « de créer un changement structurel durable au sein des institutions de recherche et des universités, au bénéfice de la progression des carrières des chercheuses en Science, Technology, Engineering and Mathematics (STEM) ». Ce projet étudiera plus particulièrement, au cours des quatre prochaines années, la situation des femmes en mathématiques et en physique. L'INSMI⁴ a également mis en place un groupe de travail sur la parité.

Enfin, au niveau de l'Europe, la Direction Générale de la Recherche a créé une unité Femmes et Sciences, avec rapports statistiques et appels à projets. Sans oublier les prises de position de l'EPWS (European Platform of Women Scientists) ou l'existence en France d'associations comme Femmes et Maths ou Femmes et Sciences.

Comment faire avancer les choses ?

On peut tout d'abord noter que les préjugés dénoncés lors de cette journée sont en grande partie inconscients. Il est vrai qu'en tant que femme, nous ne ressentons pas dans notre vie professionnelle quotidienne de sexisme particulier à notre encontre. En revanche, les chiffres montrent clairement que les femmes dans leur ensemble se trouvent défavorisées. Une des difficultés principales est donc de lutter contre un sexisme inconscient, mais bien présent. Une des premières choses à faire est de communiquer sur les problèmes que rencontrent les femmes !

Prenons l'exemple du CNU⁵. Lors des premières sessions, très peu de femmes ont été promues. La réaction a été rapide : le CNU fait maintenant attention à garder des femmes parmi les candidatures discutées. Il préserve ainsi l'équilibre homme-femme, tout en prenant aussi en compte l'équilibre géographique, thématique ou la question de l'âge. Mais cela est rendu difficile par l'auto-censure des femmes, les femmes postulent moins en moyenne. L'exemple du CNU pourrait être suivi pour les comités de sélection par exemple. À noter que ce ne sont pas aux femmes présentes au CNU d'être garantes des décisions relatives à la parité, elles sont là comme les hommes pour représenter l'ensemble de la communauté du fait de leur mérite scientifique auquel elles sont légitimement attachées.

L'augmentation du recrutement de femmes est donc un travail de longue haleine. Il semble indispensable d'accroître la représentation des femmes dans les diverses instances : comités de sélection, comités de rédaction des revues, conférenciers invités lors des colloques, comités scientifiques, jurys de thèse... Pourquoi déroger à la règle des 30% de femmes en vigueur dans les comités de concours de la fonction

³ Programme-Cadre de Recherche et de Développement.

⁴ Institut National des Sciences Mathématiques et de leurs Interactions.

⁵ Conseil National des Universités.

publique ? Mettre plus de femmes dans les comités n'augmente pas forcément le recrutement des femmes mais met plus à l'aise les candidates.

Alors, faut-il instaurer des quotas ? Il y a par exemple une volonté du ministère pour que le pourcentage de femmes dans la composition du CNU reflète celui de la section concernée. En revanche, avoir des quotas pour les promus compliquerait les discussions du CNU. Cette question reste donc délicate. D'abord, les femmes ne se reconnaissent pas dans les quotas, elles les trouvent dévalorisants. Ensuite, avec seulement 36 femmes PR 25, si on impose des quotas, les mêmes femmes seront systématiquement sollicitées partout. Un autre argument contre les quotas : dans les pays (Suède, Suisse, Canada...), où sont imposés des quotas de femmes dans les comités, la situation n'est pas meilleure que celle de la France.

Voici deux suggestions : commencer systématiquement par choisir les femmes à mettre dans ces comités plutôt que de choisir en dernière minute une femme qui devra satisfaire trop de contraintes incompatibles. Et puis, prendre le temps de réfléchir à des femmes adaptées à accomplir ces tâches – mine de rien, il y en a plein... et bizarrement on demande toujours aux mêmes. Une autre recommandation est de mettre les femmes à l'aise dans ces comités : « faire la fille » est rarement agréable, les femmes de rang B des comités d'évaluation ont souvent l'impression que leur avis est moins pris en compte que les autres.

D'autres moments importants qui sont clairement un frein à la carrière des femmes, avant même de postuler, sont les congés maternité. Ils doivent être mieux pris en compte. Quelques propositions concrètes seraient faciles à mettre en œuvre, comme une décharge systématique d'un demi-service par congé, la neutralisation d'un an sur le CV, l'indication systématique (demandée dans les formulaires) des congés lors des demandes de délégation, CRCT, primes... Les collègues et/ou DU doivent être vigilants pour aider les jeunes mamans à reprendre leurs activités de recherche dans ces périodes cruciales. Peut-être pourrait on mettre en place un « référent parité » dans chaque laboratoire ?

Évidemment, les maternités ralentissent les carrières des femmes et elles sont donc souvent plus âgées lorsqu'elles postulent PR. Les comités doivent donc accepter qu'il y ait de la diversité dans les déroulements de carrière. La tendance est encourageante puisque l'âge moyen de recrutement PR augmente ces dernières années : en 2011, en section 25, l'âge moyen est de 41,6 pour les femmes et 38,4 pour les hommes ; en section 26, l'âge moyen est de 38 pour les femmes et 39 pour les hommes.

Continuons par une revendication statistique : les chiffres fournis par L. Broze sont le fruit d'un travail laborieux. Il serait bon d'avoir régulièrement des données officielles détaillées et d'avoir systématiquement des données sexuées en matière de recrutement, d'évaluation, etc.

Enfin, nous proposons de créer un indice de parité académique, sur le modèle de l'indice de mobilité académique, pour chaque laboratoire. Le but n'est bien sûr pas de montrer du doigt les laboratoires peu féminisés (statistiquement, la féminisation est meilleure à Paris qu'en province. Est-ce à cause du plus grand bassin d'emploi : 25% des chercheurs sont en poste à Paris ?) mais de faire prendre conscience à tous du faible nombre de femmes dans le milieu. Ces chiffres pourraient également apparaître dans les dossiers d'évaluation par l'AERES⁶.

⁶ Agence d'Évaluation de la Recherche et de l'Enseignement Supérieur.

Conclusion

Ce qui nous semble le plus inquiétant dans ce bilan de la situation des femmes en mathématiques est la stagnation des pourcentages depuis 20 ans. En particulier, les femmes PR 25 sont en voie de disparition. Faire attention ne suffira pas dans les années à venir, il faut désormais des actions très volontaristes à tous les niveaux... Et ce de la part de toute la communauté, hommes et femmes. De ce point de vue, le nombre d'hommes présents à cette journée était très encourageant.

Nous remercions Benoîte de Saporta et Rozenn Texier-Picard de leurs notes sur cette journée qui nous ont permis de rédiger ce compte-rendu.

Table ronde « Égalité des chances : mythe ou réalité ? »

Une table ronde sur le thème « Égalité des chances : mythe ou réalité ? » a été organisée et animée par Christian Mauduit (université de la Méditerranée) à l'occasion des Journées Annuelles de la SMF au CIRM en juin 2011. Les intervenants en étaient Martin Andler (président de Animath, université de Versailles Saint-Quentin), Éric Barbazo (président de l'Association des Professeurs de Mathématiques de l'Enseignement Public), Antoine Bodin (directeur du CE-DEC, Écoles de la Deuxième Chance, Marseille), Jean-Pierre Demailly (membre de l'académie des Sciences, UJF, Grenoble I) et Véronique Lizan (présidente de *femmes et mathématiques*, Université de Toulouse II Le Mirail).

Christian Mauduit présente le thème général et propose au public de réagir au fil des interventions. L'égalité des chances, au-delà du mythe ou de la réalité, est aussi un discours politique qui nous vient de l'Europe. Antoine Bodin commente de nombreuses données chiffrées de l'OCDE concernant les inégalités liées aux origines sociales et géographiques, au genre, etc., en situant la France par rapport aux autres pays du monde. Il explique ce que peuvent nous apprendre les études PISA. Véronique Lizan parle des différents aspects des inégalités liées au genre. Au delà du constat numérique, il faut prendre en compte les raisons des résistances au changement, les changements de dénomination (mixité, égalité, etc.) correspondant aussi à des changements idéologiques. Éric Barbazo présente le point de vue des enseignants, notamment à partir des résultats de l'enquête faite par l'AP-MEP dans 230 lycées lors de la mise en place de la réforme du lycée cette année, où horaires globalisés, accompagnement personnalisé, etc., posent pleinement la question de l'égalité des chances. Jean-Pierre Demailly explique le rôle important de l'enseignement primaire, les lacunes y apparaissant ne pouvant être rattrapées plus tard. Il décrit le projet d'expérimentation « Savoir Lire, Écrire, Compter, Calculer ». Martin Andler explique comment les activités périscolaires peuvent contribuer à l'égalité des chances, comment les actions de diffusion de la culture scientifique peuvent encourager les jeunes qui ne se lanceraient pas spontanément dans les études supérieures à le faire. Il présente le projet Cap'Math. Antoine Bodin conclut

la discussion avec le travail des Écoles de la Deuxième Chance destinées à des jeunes ayant quitté le système scolaire sans aucun diplôme.

Le compte-rendu qui suit, rédigé par Valérie Girardin à partir des notes de Jean-Marie Barbaroux et Yann Lefeuvre, a été soumis aux différents intervenants.

Les chiffres sur l'égalité des chances

Antoine Bodin commence par dire que tous les chiffres, quelle que soit la façon de les regarder et quoi que l'on puisse penser des études chiffrées, montrent que l'égalité des chances est un mythe, aussi bien socialement ou géographiquement qu'en terme de genre. Par exemple, en 1995, les enfants d'ouvriers (ou d'inactifs) représentaient 38% des élèves de sixième; 7 ans plus tard, ils ne représentaient plus que 9% des inscrits en CPGE (classes préparatoires aux grandes écoles). Aux mêmes dates, les enfants de cadres ou professions libérales représentaient 16% des élèves de sixième, mais 55% des CPGE (Source DEPP-MEN). Le différentiel de bacheliers d'une région à l'autre atteint 20%. Seule une fille sur quatre titulaire d'un baccalauréat avec mention demande à entrer en CPGE contre, dans les mêmes conditions, un garçon sur deux. En 2005, 2,25% des jeunes français de 20 à 39 ans étaient titulaires d'un diplôme de master de sciences et techniques, mais parmi ceux-ci seulement 28% étaient des femmes, ce qui, en ce qui concerne le genre, classe la France à la place 21 sur 27 en Europe (source Eurostat).

En mathématiques comme dans les autres domaines, les données fournies par PISA sont de moins en moins favorables à la France. Alors que 16,6% des élèves français étaient classés en difficulté en 2003 dans la majeure mathématiques de l'étude, ils étaient 22,5% en 2009. On peut remarquer que l'Allemagne ayant entre-temps adapté son enseignement, le même indicateur est passé pour elle de 21,6% à 18,6%. En France, le pourcentage d'élèves classés performant (niveaux PISA 5 et 6) a légèrement diminué dans la même période. La différence des résultats des élèves des déciles supérieurs et inférieurs de l'indice socio-économique et culturel utilisé par PISA (premier et neuvième décile de l'indice ESCS) est l'un des plus importants des pays de l'OCDE.

Ces résultats peuvent pour une partie être expliqués par la structure même du système scolaire : les études PISA privilégient la traduction en mathématiques de problèmes de la vie quotidienne plutôt que les exercices de mathématiques au sens classique. De plus, elles concernent les élèves de 15 à 16 ans. En France, en 2003, seulement la moitié des élèves interrogés était en seconde générale dont le cursus correspond mieux aux demandes des études PISA que ceux des filières technologiques ou professionnelles. Les résultats des élèves n'ayant jamais redoublé y rejoignent ceux des élèves des pays les plus performants. Néanmoins, pour l'ensemble, la France était très médiocrement placée en 2003 et son niveau a nettement baissé entre 2003 et 2009. En termes d'égalité des chances il faut souligner qu'en France la variance des résultats est maximale entre les établissements, alors que, dans la plupart des pays qui réussissent mieux dans les enquêtes PISA, cette variance est maximale à l'intérieur des établissements.

L'égalité des chances et le genre

Véronique Lizan retrace une histoire de l'égalité des chances et du genre à l'école dont deux points essentiels sont la loi Haby de 1975 sur l'égalité des chances et la loi d'orientation de 2005 qui vise à favoriser mixité et égalité entre hommes et femmes, notamment en matière d'orientation. Le MEN publie depuis cinq ans des données précises sur « filles et garçons sur le chemin de l'égalité à l'école ». Une « Convention pour l'égalité entre les filles et les garçons, les femmes et les hommes, dans le système éducatif » existe depuis 2006, l'objectif est affiché, même si, dans les faits, il n'est pas prioritaire. Dans le primaire, les filles redoublent moins souvent que les garçons. En France comme en Europe, les filles sont généralement meilleures en français, de niveau égal aux garçons en sciences (et particulièrement en mathématiques dans le primaire) et sont plus diplômées à la fin de leur cursus scolaire. Au lycée, comme les filles choisissent plus souvent les options générales et les garçons les options technologiques, la mixité est rarement réalisée avec par exemple 92% de filles en Sciences médico-sociales contre 6% en Informatique et système de production. Si la série S compte 45% de filles, le choix des options y dépend aussi du genre. Les filles se dirigent plus vers les séries ES et L, et les garçons vers les séries S et STI. Lorsqu'ils se jugent très bons en mathématiques (respectivement en français), 8 garçons sur 10 vont en S (respectivement 1 en L), alors que seulement 6 filles sur 10 (respectivement 3 en L) font le même choix. De nombreuses interrogations émanent de la salle sur le moyen de tester et la signification du fait de se juger bon en mathématiques.

Les études cherchant les explications de ces différences sont nombreuses. On peut dégager le poids de l'histoire (les femmes et les sciences y sont peu liées), la moindre confiance en soi à niveau égal au moment des orientations vers 14-15 ans, le poids des stéréotypes (sur les professions à responsabilité, la conciliation de la vie privée et de la vie professionnelle) et du curriculum caché (la finalité des études apparaît ne pas être la même pour les filles et pour les garçons).

Pour amener les filles vers les sciences, il faut lutter contre l'ensemble de ces facteurs, ce que ne font pas les médias à l'intention des adolescents qui, au contraire, continuent à véhiculer les idées reçues les plus absurdes. Les amener vers les mathématiques dès l'enseignement primaire est primordial, la vision des mathématiques étant un facteur déterminant dans le choix de poursuivre des études scientifiques. Différents moyens sont utilisés pour cela, de l'information aux parents et élèves à la sensibilisation et formation des professionnels de l'éducation. La valeur de l'exemple, du « modèle », est mise en avant.

Les réactions de la salle sont vives, il est difficile de comprendre ce que les mathématiques auraient de spécifique qui ferait fuir les filles, c'est sans doute la perception qu'elles en ont, les représentations qu'on leur en donne ou le contexte dans lequel elles leur sont présentées qui sont en cause.

Le rôle des réformes de l'enseignement en France dans l'égalité des chances

Eric Barbazo rappelle que l'introduction des mathématiques modernes avait aussi pour but d'éliminer les inégalités sociales. Les études PISA montrent clairement que le système actuel les augmente. Depuis 2007, tous les niveaux de

l'enseignement ont été touchés par des réformes. Dans le primaire, les nouveaux programmes sont très critiqués. Au collège, la mise en place de l'enseignement par compétences, et de son évaluation pour arriver à un « socle commun de compétences » qui devra être validé par tous les élèves, constitue une transformation en profondeur. Au lycée, le cadrage national diminue au profit d'une liberté locale, par exemple à travers les décisions de dédoublements de classes pour travaux pratiques, ou le contenu des deux heures par semaine consacrées à « l'accompagnement personnalisé » (heures d'enseignement en dehors des programmes officiels, en général en classes entières).

L'enquête menée par l'APMEP sur la mise en place de cette réforme en seconde cette année montre une disparité importante, par exemple entre les lycées comportant des classes préparatoires et les autres. L'APMEP en conclut que les dérèglements liés à cette nouvelle réforme vont accentuer les inégalités par la disparité des horaires (plus ou moins deux heures de mathématiques par semaine sur les trois années du lycée, soit l'équivalent d'une année). Le recours aux cours particuliers très usité en France accentue ces inégalités sociales.

Aucune formation n'est proposée aux enseignants sur le thème de l'égalité des chances. Ils vont pourtant devoir s'adapter aux nouvelles conditions, entre socle de compétences et accompagnement personnalisé. Un recadrage national des heures consacrées à chaque discipline paraît indispensable pour une équité géographique et sociale entre établissements.

L'égalité des chances dans l'enseignement primaire

Jean-Pierre Demailly rappelle que l'expérience SLECC est née du constat fait autour de 2000 que l'école était loin de fonctionner de manière optimale : des méthodes d'enseignement peu efficaces sont souvent proposées – voire imposées – aux enseignants par l'institution, ce qui entraîne une spirale descendante depuis les années 1970. Le projet consiste à assurer aux élèves le plus tôt possible, à savoir dès l'école maternelle, un enseignement très structuré d'une année sur l'autre, dans une progression pas à pas visant à éviter les lacunes pédagogiques. Il nécessite une forte synergie des différents apprentissages ; ainsi, par exemple, celui de l'écriture est coordonné avec celui de la lecture – et le précède même, et les quatre opérations sont abordées simultanément dès le CP. L'approche des nombres est faite de façon très progressive, avec beaucoup de manipulations et de petits problèmes à résoudre, en introduisant d'abord les nombres de 1 à 20, puis petit à petit de 1 à 100 au cours du CP. L'interdisciplinarité se manifeste dès la maternelle, le dessin est par exemple lié à l'écriture, les activités corporelles (gymnastique, danse) pouvant quant à elles contribuer à l'acquisition de la latéralisation, conduire à des procédures de comptage, de partage, etc. Des manuels ont été écrits dans cette optique pour tous les niveaux depuis la maternelle ; en particulier des manuels de calcul sont disponibles jusqu'au CE2 (les niveaux CM1-CM2 étant en préparation). Les classes où l'expérience est menée depuis 2005 sont réparties aléatoirement en fonction de la présence d'enseignants volontaires, et incluent des zones ouvrières ou rurales en bon nombre. Une insistance particulière est portée sur le travail individuel écrit. Lors de l'évaluation de CE1 de 2009, la moyenne nationale faisait apparaître plus de la moitié des élèves ne maîtrisant pas la « barre médiane » de 40 items sur 60, ceci aussi bien en mathématiques qu'en français. Dans les quatre classes SLECC

de CE1 testées en 2009, tous les élèves sans exception sont parvenus à maîtriser plus de 40 items : même si l'échantillon était réduit, ce résultat est très significatif.

Une démarche analogue a été testée en collège et en lycée, pour tenter de redonner des bases à des élèves en difficulté, avec un enseignement là encore centré sur les connaissances fondamentales. Les essais se sont malheureusement avérés peu concluants, sans doute en partie à cause d'un refus psychologique très présent à l'adolescence d'accepter de se replonger dans des savoirs perçus comme élémentaires (même si par ailleurs ils ne sont pas maîtrisés!). La difficulté d'une remédiation collective est également apparue très crûment. Une approche spécifique et individualisée en dehors du temps de classe serait nécessaire pour remédier à ces problèmes une fois qu'ils sont installés; ceci nécessiterait à l'évidence des moyens adaptés.

Le rôle des activités périscolaires dans l'égalité des chances

Martin Andler rappelle que 3,3% des élèves se classent au niveau supérieur de l'étude PISA et que la spécialité mathématiques du bac S réunit 3,6% d'une classe d'âge, la répartition géographique en étant très inégale. Par ailleurs, si les filles représentent 57,4% des reçus au bac général, elles ne représentent que 47% des reçus au bac S (40% en spé math), 27% des élèves en CPGE, et 14% des reçus à l'École Polytechnique.

Les activités périscolaires incluent les activités et conférences culturelles (comme les Promenades mathématiques), celles des clubs, ateliers, les projets scientifiques (Maths en jeans, Hippocampe, etc.), les concours en temps limité (Olympiades, Kangourou, rallyes, etc.), les stages pendant les vacances (comme MathC2+). Les publics visés sont particulièrement les jeunes de milieu défavorisé et les filles. Le but de ces différentes activités est de limiter la fuite vers les établissements les plus cotés, de gérer l'hétérogénéité, de mettre en confiance les publics fragiles, de remédier à l'isolement des élèves motivés (les « bons en maths » sont parfois stigmatisés) en les rassemblant de temps en temps, et de renouveler les méthodes d'enseignement. Plus généralement, elles visent à donner aux jeunes une maîtrise de ce qu'ils font, à les encourager à travailler, à développer leur autonomie, à diversifier les critères de réussite.

L'ensemble des projets et activités existants a été réuni pour répondre à travers le projet Cap'Maths à un appel d'offre du grand emprunt sur l'égalité des chances.

Le rôle des formations en dehors du système scolaire

Antoine Bodin rappelle que ce sont 17% des jeunes d'une classe d'âge qui quittent le système scolaire sans aucun diplôme. Il faut leur apporter une restructuration de leurs connaissances, ce que fait par exemple le réseau des Écoles de la Deuxième Chance.

La LOF qui influe sur les crédits alloués aux établissements en fonction de leur taux de réussite ne participe pas à ce mouvement. La notion de socle commun qui s'impose partout revient à ajuster les résultats à des taux de réussite pré-établis, sans souci d'une évaluation juste des connaissances. Il est à craindre que les Écoles de la Deuxième Chance entre autres voient bientôt arriver des jeunes ayant validé ce socle commun de compétences sans pour autant maîtriser les compétences correspondantes. Ce constat alarmiste sert de conclusion à la table ronde.



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Professor of Mathematics

The Department of Mathematics at ETH Zurich (www.math.ethz.ch) invites applications for a professor position in Mathematics. We are seeking for candidates with an outstanding research record and a proven ability to direct research of high quality. Willingness to teach at all university levels and to participate in collaborative work both within and outside the school is expected.

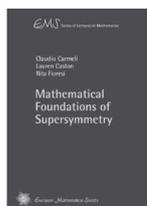
The new professor will be responsible, together with other members of the Department, for teaching undergraduate (German or English) and graduate courses (English) for students of mathematics, natural sciences and engineering.

Please apply online at www.facultyaffairs.ethz.ch. Your application should include your curriculum vitae and a list of publications. The letter of application should be addressed to the **President of ETH Zurich, Prof. Dr. Ralph Eichler. The closing date for applications is 30 November 2011.** With a view towards increasing the number of women in leading academic positions, ETH Zurich specifically encourages women to apply.



New books published by the
European Mathematical Society

www.ems-ph.org



Claudio Carmeli (University of Genova, Italy), Lauren Caston (RAND Corporation, Santa Monica, CA, USA) and Rita Fiorese (University of Bologna, Italy)

Mathematical Foundations of Supersymmetry (EMS Series of Lectures in Mathematics)

ISBN 978-3-03719-097-5. 2011. 301 pages. Softcover. 17 x 24 cm. 42.00 Euro

Supersymmetry is a highly active area of considerable interest among physicists and mathematicians. It is not only fascinating in its own right, but there is also indication that it plays a fundamental role in the physics of elementary particles and gravitation.

The purpose of the book is to lay down the foundations of the subject, providing the reader with a comprehensive introduction to the language and techniques, with a special attention to giving detailed proofs and many clarifying examples. It is aimed ideally at a second year graduate student. After the first three introductory chapters, the text divides into two parts: the theory of smooth supermanifolds and Lie supergroups, including the Frobenius theorem, and the theory of algebraic superschemes and supergroups. There are three appendices, the first introducing Lie superalgebras and representations of classical Lie superalgebras, the second collecting some relevant facts on categories, sheafification of functors and commutative algebra, and the third explaining the notion of Fréchet space in the super context.



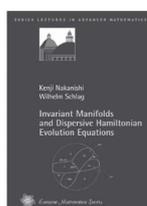
Ronald Brown (Bangor University, UK), Philip J. Higgins (Durham University, UK) and Rafael Sivera (Universitat de València, Spain)

Nonabelian Algebraic Topology. Filtered spaces, crossed complexes, cubical homotopy groupoids (EMS Tracts in Mathematics, Vol. 15)

ISBN 978-3-03719-083-8. 2011. 703 pages. Hardcover. 17 x 24 cm. 98.00 Euro

The main theme of this book is that the use of filtered spaces rather than just topological spaces allows the development of basic algebraic topology in terms of higher homotopy groupoids; these algebraic structures better reflect the geometry of subdivision and composition than those commonly in use.

The structure of the book is intended to make it useful to a wide class of students and researchers for learning and evaluating these methods, primarily in algebraic topology but also in higher category theory and its applications in analogous areas of mathematics, physics and computer science. Part I explains the intuitions and theory in dimensions 1 and 2, with many figures and diagrams, and a detailed account of the theory of crossed modules. Part II develops the applications of crossed complexes. The engine driving these applications is the work of Part III on cubical ω -groupoids, their relations to crossed complexes, and their homotopically defined examples for filtered spaces. Part III also includes a chapter suggesting further directions and problems, and three appendices give accounts of some relevant aspects of category theory. Endnotes for each chapter give further history and references.



Kenji Nakanishi (Kyoto University, Japan) and Wilhelm Schlag (University of Chicago, USA)

Invariant Manifolds and Dispersive Hamiltonian Evolution Equations (Zürich Lectures in Advanced Mathematics)

ISBN 978-3-03719-095-1. 2011. 259 pages. Hardcover. 17 x 24 cm. 38.00 Euro

This monograph is based on recent research by the authors and the proofs rely on an interplay between the variational structure of the ground states on the one hand, and the nonlinear hyperbolic dynamics near these states on the other hand. A key element in the proof is a virial-type argument excluding almost homoclinic orbits originating near the ground states, and returning to them, possibly after a long excursion.

These lectures are suitable for graduate students and researchers in partial differential equations and mathematical physics. For the cubic Klein–Gordon equation in three dimensions all details are provided,

including the derivation of Strichartz estimates for the free equation and the concentration-compactness argument leading to scattering due to Kenig and Merle.



Jean-Yves Girard (CNRS, Marseille, France)

The Blind Spot. Lectures on Logic

ISBN 978-3-03719-088-3. 2011. 552 pages. Hardcover. 16.5 x 23.5 cm. 68.00 Euro

These lectures on logic, more specifically proof theory, are basically intended for postgraduate students and researchers in logic. The question at stake is the nature of mathematical knowledge and the difference between a question and an answer, i.e., the implicit and the explicit. The problem is delicate mathematically and philosophically as well: the relation between a question and its answer is a sort of equality where one side is “more equal than the other”, and one thus discovers essentialist blind spots. Starting with Gödel’s paradox (1931) – so to speak, the incompleteness of answers with respect to questions – the book proceeds with paradigms inherited from Gentzen’s cut-elimination (1935). Various settings are studied: sequent calculus, natural deduction, lambda calculi, category-theoretic composition, up to geometry of interaction (GoI), all devoted to explicitation, which eventually amounts to inverting an operator in a von Neumann algebra.

European Mathematical Society Publishing House
orders@ems-ph.org

Seminar for Applied Mathematics, ETH-Zentrum
FLI C4, CH-8092 Zürich, Switzerland

LIVRES

Opera de Cribro

H. IWANIEC ET J. FRIEDLANDER

American Mathematical Society, Colloquium Publications, Volume 57, 2010. 527 p.

ISBN : 978-0-8218-4970-5. \$103

Comme son titre latin l'annonce, cet ouvrage, écrit en anglais, traite de méthodes de crible. Il était temps de faire le point sur ces puissantes techniques d'arithmétique après les bouleversements qu'elles ont subis depuis la parution en 1974 de l'incontournable livre d'Halberstam-Richert *Sieve Methods* dans lequel j'ai personnellement appris les fondements de ces méthodes. Bien sûr étaient parus les remarquables ouvrages de Bombieri, de Cojocaru-Murty, de Diamond-Halberstam, de Greaves, de Harman, de Hooley, de Motohashi... et, en 1991, le *Collected Papers* de Selberg et son chapitre fondateur *Lectures on sieves*. À cette liste, il convient d'ajouter certains chapitres de récents livres de théorie analytique des nombres. Les ouvrages cités précédemment ne pouvaient que servir d'introduction, ou ne traitaient qu'un aspect du crible, là où l'auteur avait travaillé avec succès pour mener à des résultats subtils mais fort ardu.

On ne pouvait rêver d'auteurs plus qualifiés que John Friedlander et Henryk Iwaniec pour présenter une vision globale, précise et pertinente sur l'évolution du crible, sur ses splendeurs et ses misères. Rappelons qu'une des nombreuses fiertés de ces deux auteurs est d'avoir prouvé, en 1998, l'existence d'une infinité de nombres premiers p de la forme $p = a^2 + b^4$ (a et b entiers), avec, de plus, le décompte asymptotique de tels $p \leq x$ lorsque x tend vers l'infini. Ce résultat historique (120 pages à *Annals of Mathematics*) est, à mon avis, l'ultime étape avant la preuve qu'une infinité de premiers apparaissent comme conducteurs de courbes elliptiques sur \mathbb{Q} , et, à plus longue échéance, qu'une infinité de premiers p est de la forme $n^2 + 1$.

Qu'est-ce qu'un problème de crible ? C'est une question de passage du local au global, qui, pour sa description requiert les quelques notations suivantes : on part de $A \geq 2$, d'une suite finie d'entiers \mathcal{A} appartenant à $[1, A]$, de cardinal $|\mathcal{A}|$. Pour tout entier $d \geq 1$, \mathcal{A}_d est la sous-suite de \mathcal{A} constituée des éléments divisibles par d . On introduit un ensemble \mathcal{P} de nombres premiers p (on réserve la lettre p , avec ou sans indice, aux nombres premiers), z est un paramètre ≥ 2 , enfin $P(z)$ est le produit des éléments de \mathcal{P} inférieurs à z .

Il s'agit d'obtenir des renseignements sur la *sifting function* $S(\mathcal{A}; \mathcal{P}, z)$ qui dénombre les éléments de \mathcal{A} , dont aucun diviseur premier $< z$ n'appartient à \mathcal{P} (autrement dit les éléments premiers à $P(z)$) à partir de la donnée de divisibilité (renseignement local)

$$(1) \quad |\mathcal{A}_d| = \frac{\omega(d)}{d} |\mathcal{A}| + r_d,$$

où d est un entier ≥ 1 et $d \mapsto \omega(d)$ est une fonction *multiplicative*, autrement dit vérifiant la condition $\omega(dd') = \omega(d)\omega(d')$ pour tous d et d' premiers entre eux. De plus r_d est appelé *terme reste*, c'est-à-dire que, en un sens à préciser, r_d est *petit* par rapport au terme principal $\frac{\omega(d)}{d}|\mathcal{A}|$ lorsque $d \leq D$, pour un certain D (appelons ceci le *premier axiome*).

À partir des données de \mathcal{A} , de $|\mathcal{A}|$, de D et de la fonction ω , peut-on construire une formule générale renseignant sur $S(\mathcal{A}; \mathcal{P}, z)$? Si on ne se trouve pas dans une situation aberrante, on pressent quelque chose comme

$$(2) \quad S(\mathcal{A}; \mathcal{P}, z) \asymp \prod_{p < z, p \in \mathcal{P}} \left(1 - \frac{\omega(p)}{p}\right) |\mathcal{A}|.$$

Toute la difficulté consiste à correctement définir le symbole \asymp , en fonction des paramètres $s := (\log D)/(\log z)$ et κ la valeur moyenne de la fonction ω , en supposant qu'elle existe et qu'elle soit finie. On appelle alors κ la *dimension* du crible.

Pour approcher $S(\mathcal{A}; \mathcal{P}, z)$, le réflexe immédiat est d'itérer le principe d'inclusion-exclusion pour parvenir à la formule dite du crible d'Ératosthène-Legendre (μ est la fonction de Möbius)

$$(3) \quad S(\mathcal{A}; \mathcal{P}, z) = \sum_{d|P(z)} \mu(d) |\mathcal{A}_d|,$$

puis d'injecter la formule (1). Mais on se doute que la situation n'est pas si simple, sinon il n'aurait pas fallu attendre 1896 pour connaître le comportement asymptotique de la fonction $\pi(x) := |\{p; p \leq x\}|$, à savoir $\pi(x) \sim x/\log x$, ($x \rightarrow \infty$) et ce, par des méthodes d'analyse complexe : c'est le théorème des nombres premiers de Hadamard et La Vallée Poussin. Pour étudier $\pi(x)$ par le crible, on est tenté d'appliquer (3) avec $\mathcal{A} := \{a; 1 \leq a \leq x\}$. On prend aussi pour \mathcal{P} l'ensemble de tous les premiers, on pose $z = \sqrt{x}$ et on constate que, à un terme négligeable près, $S(\mathcal{A}; \mathcal{P}, z)$ vaut $\pi(x)$. La formule (1) est écrite avec le choix $\omega(d) = 1$ et le terme reste vérifie $|r_d| \leq 1$. Par contre, l'égalité (3), par la prolifération des termes qu'elle contient, est incapable de prouver la majoration $\pi(x) = O(x/\log x)$ que Chebychev avait obtenue en 1852 et, pire encore, ne permet pas de prouver qu'il y a une infinité de premiers.

C'est à Viggo Brun que l'on doit, autour de 1920 d'avoir fait vraiment démarrer les méthodes de crible (*The Big Bang* comme l'écrivent Friedlander et Iwaniec). V. Brun a trouvé comment raccourcir l'égalité (3) en un nombre étonnamment restreint de termes, le prix à payer étant que (3), suivant les cas, se métamorphose en une majoration ou une minoration, mais toutes deux pertinentes ! La technique de Brun, bien que toujours incapable d'exhiber une infinité de nombres premiers, redonne la majoration de Chebychev $\pi(x) = O(x/\log x)$, et surtout, établit la suprématie du crible, ou si on préfère de la combinatoire, sur les méthodes analytiques dans certaines circonstances : majoration asymptotique du cardinal des nombres premiers jumeaux (c'est-à-dire des p tels que $p+2$ soit aussi premier) ou encore des $p \leq x$, $p \equiv a \pmod q$ lorsque q est grand par rapport à x (théorème de Brun-Titchmarsh).

De façon évidemment schématique, on peut alors sérier les progrès réalisés dans l'étude générale $S(\mathcal{A}, \mathcal{P}, z)$ (majoration ou minoration) suivant trois axes.

– Poursuivre l'idée de Brun en construisant des ensembles de plus en plus sophistiqués \mathcal{D}^+ et \mathcal{D}^- pour raccourcir (3) en vue de l'encadrement

$$(4) \quad \sum_{\substack{d|P(z) \\ d \in \mathcal{D}^-}} \mu(d)|\mathcal{A}_d| \leq S(\mathcal{A}; \mathcal{P}, z) \leq \sum_{\substack{d|P(z) \\ d \in \mathcal{D}^+}} \mu(d)|\mathcal{A}_d|.$$

C'est le crible de Rosser (ou β -crible), dont la délicate construction dépend de la valeur de κ et dont l'outil de base est la positivité du cardinal de tout ensemble fini !

– Exploiter la puissante idée de Atle Selberg (1944), qui se démarque totalement de celle de Brun, en constatant que pour toute famille de réels $(\lambda_d)_{d \leq \sqrt{D}}$ avec $\lambda_1 = 1$, on a l'inégalité

$$S(\mathcal{A}; \mathcal{P}, z) \leq \sum_{a \in \mathcal{A}} \left(\sum_{\substack{d|P(z) \\ d|a}} \lambda_d \right)^2,$$

puisque'un carré de réel est toujours positif ! Il reste à injecter (1), puis à minimiser une certaine forme quadratique, les λ_d s'exprimant alors en fonction des $\omega(d)$.

– Utiliser l'égalité, dite de Buchstab,

$$S(\mathcal{A}; \mathcal{P}, z) = S(\mathcal{A}; \mathcal{P}, z_1) - \sum_{\substack{z_1 \leq p < z \\ p \in \mathcal{P}}} S(\mathcal{A}_p; \mathcal{P}, p) \quad (z_1 < z),$$

qui permet de basculer d'une majoration à une minoration et vice versa. Encore le principe d'inclusion-exclusion !

Friedlander et Iwaniec nous invitent à un parcours raisonné et pertinent parmi ces trois méthodes : comparaison des puissances des deux premières, convergence d'une itération infinie du processus de Buchstab, combinaison de ces diverses méthodes. Leur compte-rendu est parfaitement limpide. On voit que la situation est stable dans les cas les plus intéressants pour les applications, à savoir celui des cribles linéaires et semi-linéaires (correspondant aux dimensions 1 et 1/2) et que certaines formules de majoration et de minoration de $S(\mathcal{A}, \mathcal{P}, z)$ sont optimales. En particulier, réapparaît l'impossibilité de détecter des nombres premiers par le crible comme décrit ci-dessus. Une version simplifiée de l'exemple de Selberg consiste à choisir pour \mathcal{A} la suite $\tilde{\mathcal{A}}$ des entiers $a \leq A$ ayant un nombre pair de facteurs premiers et $z = A^{\frac{1}{2}}$. Des méthodes analytiques standard sur les fonctions L de Dirichlet prouvent qu'on a $\kappa = 1$ et $D = A^{1-\varepsilon}$ ($\varepsilon > 0$ quelconque). Évidemment, la partie gauche de (2) vaut 0 alors que le second membre est non nul ! Voici une des misères du crible : alors qu'il a été créé pour cela, le crible, comme il est présenté, est incapable de détecter des nombres premiers !

Par contre, rien n'est simple dans les cas restants. Il faudrait discuter suivant la valeur de s et de κ . Les auteurs ont l'intelligence de nous épargner les détails de tous les raffinements et de renvoyer en fin de volume le bagage nécessaire sur les équations différentielles avec retard.

C'est une excellente initiative que d'avoir réservé un des tout premiers chapitres au crible de Bombieri : le lecteur est alors convaincu de l'abîme de difficulté qui sépare, même dans les conditions optimales d'application du crible linéaire, le problème qui consiste à détecter dans \mathcal{A} des nombres premiers, de celui des

nombres entiers ayant au plus deux facteurs premiers. Une autre façon de voir ce gouffre est de constater que, alors que le théorème de Chen (1973) affirme que tout entier pair N assez grand s'écrit comme $N = p_1 + p_2$ ou $N = p_1 + p_2 p_3$, on est incapable de proposer une hypothèse raisonnable de théorie analytique des nombres, qui entraînerait la véracité de la conjecture de Goldbach (tout entier pair N assez grand s'écrit comme $N = p_1 + p_2$).

Les bouleversements que j'ai évoqués au début de ce rapport sont évidemment présentés, examinés toujours de façon plus agréable que lors de leur première apparition dans la littérature. Le premier bouleversement concerne l'exigence que le terme reste r_d soit *petit*. Dans les cribles jusqu'à la fin des années 1970, on demandait que la somme $\sum_{d \leq D} |r_d|$ soit négligeable par rapport à $|\mathcal{A}|(\log z)^{-\kappa}$. On remarque que les r_d apparaissent en module, et on doit à Iwaniec, en 1980, d'avoir prouvé qu'on peut tenir compte des oscillations du signe de r_d , en exploitant une subtilité de la construction des ensembles \mathcal{D}^\pm dans la formule (4). Cette propriété s'est avérée cruciale dans la preuve de l'existence d'une infinité de n tels que $n^2 + 1$ ait au plus deux facteurs premiers (Iwaniec 1978), et dans la preuve de l'existence de p dans tout intervalle de la forme $[x, x + x^\theta]$ pour x suffisamment grand et pour certains θ inférieurs à ceux directement déduits des théorèmes de densité des zéros de la fonction ζ de Riemann.

Une autre révolution est la méthode de Goldston-Pintz-Yildirim qui démontre l'existence d'intervalles minuscules entre nombres premiers consécutifs, bien en deçà de la conséquence directe du théorème des nombres premiers qui affirme que $\log p$ est l'écart moyen entre le premier p et le suivant. Grâce aux travaux de GPY on est *presque* à la conjecture des nombres premiers jumeaux, mais ce *presque* cache un abîme... Là aussi, Friedlander et Iwaniec donnent de tout ceci une savoureuse présentation, se plaçant résolument dans le paysage du crible de Selberg.

Enfin, on sait que le crible peut maintenant produire des nombres premiers, à condition d'ajouter une autre hypothèse sur la suite \mathcal{A} : elle doit se comporter de façon harmonieuse par rapport à la turbulente fonction de Möbius. De façon plus précise, on doit observer des compensations sur une certaine forme bilinéaire contenant en son cœur une somme, pour $a \in \mathcal{A}$, de la fonction $\mu(a)$, ou d'une fonction proche structurellement. Appelons cette condition le *deuxième axiome*. La suite $\tilde{\mathcal{A}}$ introduite ci-dessus, ne satisfait pas à cet axiome, bien évidemment ! Reste à exhiber des situations où le deuxième axiome est vérifié et les auteurs ainsi nous dévoilent quatre splendeurs du crible :

- l'équirépartition, modulo 1, de l'ensemble des fractions $\frac{\nu}{p}$ pour $p \equiv 1 \pmod{4}$ et ν racine de la congruence $\nu^2 + 1 \equiv 0 \pmod{p}$ (Duke-Friedlander-Iwaniec),
- le décompte asymptotique des p de la forme $p = a^2 + p_1^2$ (Fouvry-Iwaniec),
- le décompte asymptotique des p de la forme $p = a^2 + b^4$ (évoqué plus haut),
- le décompte asymptotique des p de la forme $p = a^3 + 2b^3$ (Heath-Brown).

Les deuxième et troisième exemples ont une interprétation directe dans le langage des irréductibles π de l'anneau $\mathbb{Z}[i]$: il y a donc une infinité, et ce avec le bon ordre de grandeur asymptotique, de tels π , non associés à des $p \equiv 3 \pmod{4}$, dont la partie imaginaire est un premier (de \mathbb{Z}) ou un carré. Friedlander et Iwaniec ne peuvent donner qu'une preuve complète du deuxième de ces exemples, car les trois autres nécessitent une très longue préparation, que ce soit dans la théorie des formes modulaires, dans la théorie des caractères de l'anneau $\mathbb{Z}[i]$ ou dans l'arithmétique de

l'anneau $\mathbb{Z}[\sqrt[3]{2}]$. De tels outils, cruciaux dans les preuves correspondantes eussent été hors sujet dans le présent ouvrage. Je ne puis que recommander sur ce thème des formes récentes du crible, les articles de Deshouillers, de Michel et de Kowalski au Séminaire Bourbaki.

Le crible pondéré – celui qui permet de détecter dans \mathcal{A} des entiers ayant au plus r facteurs premiers, où $r \geq 2$ est un entier fixé – donne lieu à un chapitre remarquablement limpide : on n'a pas les meilleurs résultats... mais presque. Tout aussi lumineux est le court chapitre consacré au grand crible. Ce diamant de la théorie analytique des nombres, trouvé en 1941 par Linnik, est un outil très puissant de la théorie analytique des nombres, qu'on ne cesse d'étudier pour lui-même (voir le livre de Ramaré) ou d'exploiter dans des domaines inattendus (voir le livre de Kowalski, pour des applications algébriques ou géométriques). L'abord de Friedlander et Iwaniec en est classique mais terriblement efficace, menant entre autres au théorème de Bombieri-Vinogradov sur la répartition en moyenne des nombres premiers dans les progressions arithmétiques, qui, lors des applications, remplace efficacement l'hypothèse de Riemann généralisée aux fonctions L de Dirichlet.

Passons au difficile théorème de Linnik qui, rappelons-le, affirme l'existence d'une constante absolue L , telle que pour tout couple (a, q) d'entiers premiers entre eux, il existe un p vérifiant $p \equiv a \pmod{q}$ et $p \leq q^L$. Autrement dit, ce théorème annonce quand apparaît le premier p dans la suite des nombres premiers congrus à a modulo q , suite que l'on sait être infinie depuis Dirichlet. Les auteurs suivent une démarche à mon avis inédite, où le crible intervient de façon cruciale au détriment de l'étude subtile de la localisation des zéros des fonctions L de Dirichlet. Ainsi, s'affranchit-on du phénomène de répulsion : si une fonction L a un zéro extrêmement proche du point 1, alors les autres zéros sont repoussés vers la droite critique.

Malheureusement, les auteurs, pour des raisons évidentes de longueur, ne peuvent qu'effleurer les trésors d'ingéniosité nécessaires pour majorer, selon divers types de moyenne, le terme reste r_d . Il leur eut fallu parcourir la théorie des sommes d'exponentielles qu'elles viennent de la géométrie algébrique ou de la théorie des formes modulaires... Qu'il me soit permis de signaler que, dans certaines situations, on a un meilleur contrôle de r_d que celui directement impliqué par l'hypothèse de Riemann généralisée aux fonctions L de Dirichlet.

J'ai beaucoup apprécié le choix des applications. Au lieu de se confiner aux classiques nombres premiers jumeaux, nombres presque premiers représentés par des polynômes, conjecture de Goldbach, nombres premiers dans les progressions arithmétiques... questions historiquement à l'origine du crible, les auteurs ont recensé des applications moins connues, où on sait exploiter à l'extrême une propriété de divisibilité. Citons entre autres

- la majoration très explicite du nombre de solutions à l'équation $N = aX^2 + bY^2 + cZ^2$ avec a, b et c entiers donnés,
- l'écriture de tout entier assez grand dans une certaine classe de congruence, comme somme de deux normes d'idéaux de K , K étant une extension cubique cyclique fixée de \mathbb{Q} ,
- l'étude des propriétés multiplicatives du nombre de points de la réduction modulo p de la courbe elliptique $y^2 = x^3 - x$,
- le dénombrement de points rationnels sur une certaine surface de Châtelet.

Tout ceci donne un éclairage très dynamique du crible. Malheureusement, les méthodes présentées ci-dessus semblent, pour l'instant, inopérantes pour construire, une infinité de p supersinguliers, pour une courbe elliptique donnée sur \mathbb{Q} , non CM. On ne dispose toujours que de la preuve de Elkies, preuve reposant sur une construction à la Euclide.

Mais l'histoire du crible n'est pas close : au cours de la rédaction de ce livre, il est apparu sous une autre forme, celle du *crible en expansion* (voir un autre exposé de Kowalski à Bourbaki), il a contribué à la preuve de la conjecture d'équirépartition de masse pour la surface $SL(2, \mathbb{Z}) \backslash \mathbb{H}$ (Holowinsky et Soundararajan) et il se manifeste maintenant en théorie des groupes (Lubotzky et Meiri).

Vous l'aurez compris, j'ai une profonde admiration pour cet ouvrage qui, sur un sujet que je pense bien connaître, m'a dévoilé des raccourcis, élégants et insoupçonnés. La présentation en est méticuleuse, avec un soin scrupuleux sur les uniformités dans les formules. Les commentaires, tant historiques que méthodologiques sont riches et pertinents. Ce livre, qui écrase tout ce qui a été écrit précédemment sur le sujet, devient immédiatement une référence et le restera pour longtemps. On en suivra les notations, on s'y reportera pour des énoncés clairs et on le recommandera aux débutants. Les auteurs ont une telle virtuosité dans leur sujet qu'ils s'autorisent de délectables facéties (*Kvadrater er Positive*). Mais, par dessus tout, j'ai été très sensible au ton général de ce chef d'œuvre : les auteurs font de l'arithmétique avec plaisir et fierté.

Et ils ont raison.

Etienne Fouvry
Université Paris-Sud