

EXPOSÉ

1 ACTIVITÉS D'ENSEIGNEMENT

Les paragraphes de cette section visent à présenter le contenu des enseignements dispensés au cours mon contrat doctoral avec monitorat.

1.1 UNIVERSITÉ D'EVRY

Au cours de mon monitorat, j'ai eu l'occasion de dispenser des enseignements à des publics très variés. En effet, lors de ma première année de monitorat, j'ai commencé par dispenser au niveau Licence des TD de probabilité et statistique.

J'ai ensuite eu l'occasion de dispenser des TD pour des Master 1 en biologie. L'objectif de ces cours était de familiariser les étudiants avec les bases et outils statistiques utilisés par biologistes. Les TD de ces enseignements étaient largement inspirés d'expériences concrètes en biologie et leur permettaient d'appliquer les notions théoriques vues en cours à travers le logiciel R.

Pendant mon année en tant qu'ingénieur de recherche, j'ai également dispensé une formation doctorale pour les biologistes. Cette formation visait à présenter les bases statistiques appliquées à des problèmes biologiques. Le but étant d'introduire ou de rappeler des notions de bases mais également de leur présenter le logiciel R.

2013-2014	PROBABILITÉ ET STATISTIQUE	
	Niveau	Licence 2 mention Informatique
	Charge	38h
	Contenu	Rappels de probabilité, variables aléatoires réelles, lois continues
2013-2015	DÉMARCHE STATISTIQUE 1	
	Niveau	Master 1 mention Biologie
	Charge	32h
	Contenu	Statistiques descriptives, Tests, initiation au langage R
2013-2015	DÉMARCHE STATISTIQUE 2	
	Niveau	Master 1 mention Biologie
	Charge	24h
	Contenu	Suite des Tests, Modèles linéaires classiques
Janvier 2013	BASES STATISTIQUES POUR LA BIOLOGIE	
	Niveau	Formation doctorale
	Charge	16h
	Contenu	Tests, initiation au langage R

1.2 ECOLE D'INGÉNIEUR (ENSIIE)

J'ai également été chargée de TD en modélisation statistique en 1ère année à l'Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (admission à partir de la banque de concours Centrale-SupElec). Ce cours visait à introduire les notions de modélisation statistique notamment en abordant les notions de maximum de vraisemblance. J'ai eu l'occasion d'élaborer des sujets de projets pour ce cours. Ce projet était l'occasion pour les étudiants de réaliser une petite étude descriptive et une modélisation d'un jeu de données.

2013-2014 MODÉLISATION STATISTIQUE 1

Niveau

1ère année école d'ingénieur

Charge

64h

Contenu

Statistiques descriptives, Maximum de vraisemblance, Estimation, Intervalles de confiance, Tests, initiation au langage R

2 ACTIVITÉS DE RECHERCHE

2.1 CONTEXTE

Mes travaux de recherche sont axés sur le développement de méthodes statistiques pour analyser le nombre de copies d'ADN en cancérologie. Dans une cellule normale, le nombre de copies d'ADN est égal à 2 (une copie provenant de chacun des parents), ce n'est pas le cas pour une cellule tumorale. En effet, on peut observer le long du génome des gains et des pertes de certaines parties de chromosomes. Ces altérations au niveau du nombre de copie d'ADN peuvent avoir un lien avec la résistance aux traitements. De plus, l'étude des altérations génétiques en cancérologie peuvent permettre d'adapter certains traitements. Il est possible de mesurer le nombre de copies d'ADN avec les expériences de puces à ADN ou de séquençage. Les données génomiques issues de ces expériences ont deux caractéristiques principales : leur grande dimension (le nombre de marqueurs dépassant de plusieurs ordres de grandeurs le nombre d'observations), et leur forte structuration (notamment via les dépendances entre marqueurs). La prise en compte de cette structuration est un enjeu clé pour le développement de méthodes performantes en grande dimension. Les signaux provenant d'expériences de puces SNP mesure la quantité d'allèle à un grand nombre de position le long du génome.

Formellement, pour chaque $j = 1, \dots, J$, notons θ_{Aj} et θ_{Bj} les intensités de la position j , respectivement pour les allèles A et B . θ_{Aj} et θ_{Bj} sont proportionnels aux quantités d'allèles. On définit la première dimension du signal par le nombre de copies total qui est proportionnelle à $\theta_j^t = \theta_{Aj}^t + \theta_{Bj}^t$ (la somme des quantités d'allèles A et B dans la tumeur notée t). Si un échantillon de référence est disponible, on peut mesurer le nombre total de copies dans la tumeur par

$$c_j = 2 \times \frac{\theta_j^t}{\theta_{Aj}^r + \theta_{Bj}^r} \quad (1)$$

où θ_{Aj}^r et θ_{Bj}^r sont les intensités dans l'échantillon de référence noté r .

La seconde dimension du signal issu des puces SNP est la fraction d'allèle B . La fraction d'allèle B est définie à la position j par :

$$b_j^t = \frac{\theta_{Bj}^t}{\theta_{Aj}^t + \theta_{Bj}^t} \quad (2)$$

et est comprise entre 0 et 1.

La Figure 1 représente un exemple d'un signal issu d'une cellule tumorale. La Figure 1a représente le nombre total de copies et la Figure 1b, la fraction d'allèle B .

2.2 CONTRIBUTIONS

2.2.1 MÉTHODES DE SEGMENTATION

Durant ma thèse nous avons développé plusieurs méthodes pour retrouver les ruptures dans les signaux de nombre copies d'ADN issus d'expériences de puces à ADN (lignes verticales rouge sur la Figure 1). Le modèle classique du nombre de copies d'ADN qui est usuellement utilisé dans la littérature par plusieurs méthodes suppose que le nombre de copies d'ADN est constant par morceaux dans la moyenne. Le modèle statistique pour S points de ruptures aux positions $m_S = (t_1, \dots, t_S)$ peut s'écrire de la façon

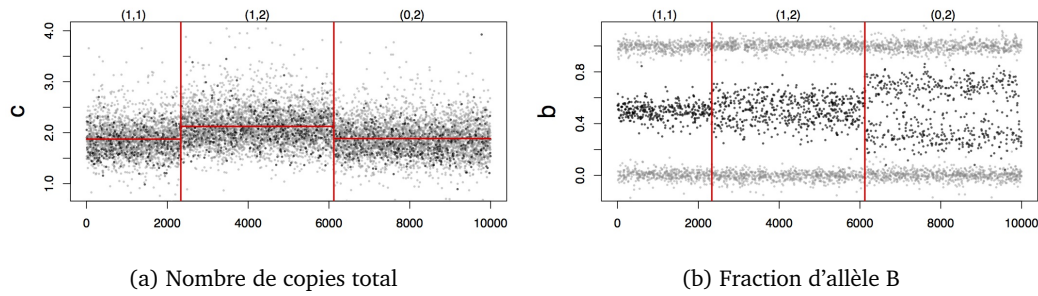


FIGURE 1 – Signaux issus d'une cellule tumorale

suivante :

$$\forall j = 1, \dots, J \quad c_j = \gamma_j + \epsilon_j \quad (3)$$

où $\forall s \in \{1, \dots, S+1\}, \forall j \in [t_{s-1}, t_s[\quad \gamma_j = \Gamma_s$ avec la convention suivante $t_0 = 1$ et $t_{S+1} = J+1$. Mes contributions portent sur la mise en place d'une méthode permettant de retrouver les ruptures du signal simultanément dans le nombre de copies d'ADN et la fraction d'allèle B contrairement aux méthodes classiques basées sur la segmentation du nombre de copies d'ADN uniquement. Deux méthodes ont été développées, la première est l'extension de la méthode de segmentation binaire récursive à deux dimensions, la seconde est basée sur les méthodes à noyau. Après avoir développé un schéma de simulation basée sur données des données réelles, nous avons démontré que l'intégration de l'information fournie par la fraction d'allèle B améliorerait la détection des ruptures dans les signaux. Ce travail a fait l'objet d'une publication dans Briefings in Bioinformatics.

L'autre méthode de segmentation basée sur les méthodes à noyau a fait l'objet de plusieurs présentations et le travail a été soumis dans Computational Statistics and Data Analysis.

2.2.2 HÉTÉROGÉNÉITÉ

Dans la deuxième partie de ma thèse, je me suis consacrée à l'étude de l'hétérogénéité tumorale à partir des données de nombre de copies d'ADN. En effet, il y a deux types d'hétérogénéité tumorale et son étude pourrait permettre de mettre en place des traitements personnalisés. Le premier type d'hétérogénéité est l'hétérogénéité inter tumorale, ceci signifie que pour un même cancer on peut observer différents types. Par exemple, il existe plusieurs types de cancer du sein. Le deuxième type est l'hétérogénéité intra-tumorale, ceci signifie que pour une même tumeur on a différents types de cellules qui présentent des altérations différentes. Ces différents types de cellules sont appelées sous-clones. Le modèle d'hétérogénéité vise à retrouver les altérations dans les sous-clones mais aussi la composition des échantillons tumoraux à partir de plusieurs échantillons de nombre de copies d'ADN. En supposant que les sous-clones sont partagés entre les échantillons observés et que les échantillons sont une combinaison linéaire de poids et de sous-clones, le modèle d'hétérogénéité mis en place a trois caractéristiques. La première est le fait d'utiliser une méthode de segmentation conjointe afin de réduire la dimension des signaux observés. La deuxième est le fait d'intégrer l'information de la fraction d'allèle B dans le modèle ce que les modèles existants ne faisaient pas. Enfin, la troisième est le fait de mettre des contraintes qui ont un sens biologique. La modélisation de l'hétérogénéité implique de résoudre un problème de factorisation matricielle où la première matrice représente les poids (la composition des différents échantillons observés) et la seconde matrice représente les différents sous-clones (les altérations pour chacun des sous-clones). Ce modèle a été appliqué à plusieurs jeux de données publics ou issus de collaborations détaillées dans la section suivante.

2.2.3 COLLABORATIONS

Dans le cadre de l'étude de l'hétérogénéité tumorale, j'ai eu l'occasion de collaborer avec plusieurs laboratoires. La première collaboration a été réalisée avec l'équipe RT² de Fabien Rey à l'Institut Curie.

Nous avons analysé des données issues de patientes atteintes de cancer du sein triple-négatif (étude de l'hétérogénéité inter-tumorale). Le but étant de retrouver des altérations communes entre les patientes. La seconde collaboration a été réalisée avec Joe Costello et Henrik Bengtsson de l'Université de San Francisco. Le modèle d'hétérogénéité a été appliqué à des données issues de patients atteints de glioblastome (tumeur du système nerveux). Le but étant de découvrir les altérations liées à la résistance aux traitements en analysant plusieurs échantillons issus d'un même patient (étude de l'hétérogénéité intra-tumorale).

2.2.4 IMPLÉMENTATIONS

Pour chacune des méthodes développées nous nous sommes attachés à réaliser des packages R avec une implémentation efficace et disponible librement. A ce jour deux packages concernant le schéma de simulation et les méthodes de segmentation sont disponibles sur github (<https://github.com/mpierrejean?tab=repositories>).

3 PUBLICATIONS ET COMMUNICATIONS

3.1 PUBLICATIONS

- [1] Morgane Pierre-Jean, Julien Chiquet, and Pierre Neuvial. Cancer clonality using DNA copy number data, In preparation for Biostatistics.
- [2] Morgane Pierre-Jean, Guillemette Marot, Rigai Guillem, and Alain Celisse. Non parametric DNA copy number segmentation using kernel methods, Submitted to Computational Statistics and Data Analysis (CSDA).
- [3] Morgane Pierre-Jean, Guillem Rigai, and Pierre Neuvial. Performance evaluation of DNA copy number segmentation methods. *Briefings in bioinformatics*, 16(4) :600–615, 2015.

3.2 CONFÉRENCES INTERNATIONALES

Juin. 2013 2èmes Rencontres R à Lyon
Mai. 2013 45èmes journées de la SFDS à Toulouse
Nov. 2012 Journée Annuelle du groupe Biopharmacie et Santé de la SFDS à Paris

3.3 CONFÉRENCES NATIONALE

Jan. 2013 SMPGD (Statistical Methods for Post-Genomic Data) à Amsterdam

3.4 SÉMINAIRES

Avril 2015 Statistics and Genomics Seminar UC Berkeley (USA) invitée par Sandrine Dudoit
Avril 2013 Statistics for Systems Biology à Evry

3.5 POSTERS

Jan 2017 SMPGD (Statistical Methods for Post-Genomic Data) à Londres
Nov 2016 Statistical Analysis of Massive Genomic Data à Evry
Jan. 2015 SMPGD (Statistical Methods for Post-Genomic Data) à Munich