

RAPPORT DE DOCTORAT

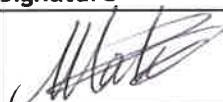

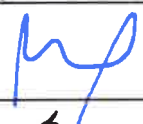

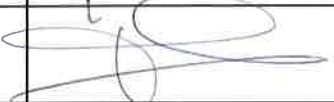



à déposer si possible trois jours ouvrés après la soutenance et au plus tard un mois après la soutenance
au service de scolarité de l'établissement de préparation de la thèse, daté et signé

Madame Morgane PIERRE-JEAN

ECOLE DOCTORALE : Structure et Dynamique des Systèmes Vivants

ETABLISSEMENT DE PREPARATION DU DOCTORAT: université d'Evry-Val-d'Essonne

La candidate a présenté de façon claire et didactique une sélection judicieuse de ses travaux portant sur le développement de méthodes statistiques pour l'analyse du nombre de copies d'ADN en cancérologie. Au cours de son travail, elle a su s'appuyer sur un éventail de techniques très différentes. Sa thèse comporte à la fois des contributions méthodologiques en statistique, des développements algorithmiques et logiciels et des applications en biologie. Elle a été impliquée dans plusieurs collaborations inter-disciplinaires. Ses réponses aux questions ont montré à la fois une grande maturité et une intégrité scientifique. Le jury est convaincu que la candidate a toutes les capacités pour être un très bon chercheur ou enseignant-chercheur. Le jury lui décerne donc sans hésiter le grade de docteur en sciences de la vie et de la santé de l'Université Paris Saclay.

Civilité, Nom et Prénom	Titre	Fonction dans le Jury	Signature
Mme Catherine MATIAS	Directeur de recherche	Directeur de thèse	
M. Jean-Phillipe VERT	Directeur de recherche	Rapporteur <input type="checkbox"/> Président	
M. Pierre NEUVIAL	Chargé de recherche	CoDirecteur de thèse	
Mme Guillemette MAROT	Maître de conférences	Examineur <input type="checkbox"/> Président	
M. Cyril DALMASSO	Maître de conférences	Examineur <input type="checkbox"/> Président	
M. François RADVANYI	Directeur de recherche	Examineur <input checked="" type="checkbox"/> Président	
Mme Anne-Laure BOULESTEIX	Professeur des universités	Rapporteur <input type="checkbox"/> Président	
Mme Nancy ZHANG	Professeur associé	Rapporteur <input type="checkbox"/> Président	



CENTRE DE BIO-INFORMATIQUE - CENTRE FOR COMPUTATIONAL BIOLOGY

Rapport sur le manuscrit de thèse

« Development of statistical methods for DNA copy number analysis in cancerology »

présenté par Morgane Pierre-Jean

Les cellules cancéreuses accumulent des anomalies génomiques lors de la tumorigenèse. L'étude de ces anomalies peut permettre de caractériser finement chaque tumeur au niveau moléculaire pour, par exemple, envisager un traitement personnalisé à chaque patient. Le développement rapide depuis une quinzaine d'années de technologies dites « à haut débit » en génomique, notamment les puces à ADN et le séquençage à haut débit, permet de mesurer finement et quantitativement certaines anomalies, comme les changements dans le nombre de copies de l'ADN. Les données brutes produites par ces technologies fournissent cependant une mesure indirecte et bruitée de l'information biologique sous-jacente, et nécessitent un traitement informatique préalable à toute analyse biologique. Les travaux présentés par Morgane Pierre-Jean visent à développer de tels traitements informatiques, en s'appuyant sur une modélisation statistique des données, pour deux applications principales : la détection de ruptures dans les profiles de nombres de copies d'ADN tumoraux, et la déconvolution de ces profils en sous-populations clonales.

Après une rapide introduction générale, le manuscrit se compose de deux parties principales, correspondant aux deux applications principales, suivies d'une troisième partie exposant les méthodes bio-informatiques utilisées pour réaliser les analyses de données réelles et de quelques annexes.

La première partie présente donc des travaux concernant la segmentation de profils génomiques mesurant les nombres de copies d'ADN. Un premier chapitre présente un état de l'art de méthodes existantes ; il est clair, complet, et détaille à la fois les modèles mathématiques et les stratégies algorithmiques mises en œuvre dans chaque méthode. Le deuxième chapitre présente une contribution originale, publiée dans le journal *Briefings in Bioinformatics*, visant à comparer la performance de différents modèles de segmentation. Pour cela un outil de simulation réaliste de données est proposé, ainsi qu'un protocole pour tester la performance de méthodes de segmentation sur ces données. Ces travaux sont implémentés dans des logiciels sous forme de package R, et facilement utilisables par la communauté scientifique. L'outil de simulation proposé s'appuie sur des données réelles tout en offrant la capacité de faire varier le rapport signal/bruit des données, ce que n'offrent pas les outils existants. Par construction, les données simulées ne semblent pas avoir de structure spatiale le long du profil (en dehors des points de changements), et il serait intéressant de vérifier si les données réelles n'ont également pas de structure spatiale ou si cela constitue une limitation de la méthode. Un troisième chapitre, enfin, présente un algorithme efficace pour implémenter une méthode de segmentation non paramétrique proposée par Harchaoui et Cappé (2007), en améliorant la méthode de programmation dynamique utilisée et en approximant un noyau défini positif par un produit scalaire en faible dimension. Ce travail original est intéressant car il permet d'appliquer cette méthode de segmentation sur des données de grandes tailles, étudiées dans le cadre de cette thèse, et d'obtenir des

résultats expérimentaux prometteurs. Il pourrait également ouvrir la voie d'autres approximations en dimension finie pour la segmentation non-paramétrique, comme l'utilisation de *random features* pour approximer des noyaux par exemple.

La deuxième partie présente des travaux concernant l'analyse de l'hétérogénéité des tumeurs, notamment la détection et la quantification de populations sous-clonales. Après un chapitre introductif présentant de façon claire et synthétique des méthodes existantes pour résoudre le problème considéré, un nouveau modèle est présenté. Ce modèle généralise des modèles existants en ajoutant des contraintes spécifiques et en permettant d'utiliser l'ensemble des informations capturées par les puces à ADN, notamment la fréquence de l'allèle minoritaire (mais pas les mutations somatiques capturées par le séquençage). Des expériences sur données simulées montrent des résultats très prometteurs pour la nouvelle méthode, qui est ensuite appliquée sur deux jeux de données réelles : des données publiques de cancer de l'ovaire, et des données non publiques sur les cancers du sein triple-négatifs dans le cadre d'une collaboration avec l'Institut Curie. Les résultats sur données réelles n'ont pas (encore) donné lieu à des validations expérimentales, et il est donc assez difficile d'évaluer la véracité des prédictions présentées ; ces applications sur données réelles démontrent tout de même que la méthode proposée peut effectivement fonctionner dans ce cas, et générer des hypothèses biologiques intéressantes.

La troisième partie présente des méthodes bio-informatiques utilisées pour analyser les données réelles. Elle commence par un travail sur l'estimation du degré d'hétérozygotie en l'absence d'échantillon sain, et se poursuit par un protocole pour estimer les données de nombre de copie d'ADN à partir de données de séquençage. Ces travaux sont intéressants et illustrent une partie des difficultés pratiques que Morgane Pierre-Jean a dû rencontrer pour pouvoir appliquer ses méthodes sur des données réelles, notamment pour choisir et développer des outils adaptés à des données particulières.

L'ensemble du travail est rigoureux et conséquent. Il montre la capacité de la candidate à s'attaquer à des problématiques réels et importantes, en s'appuyant sur un bagage informatique et mathématique solide, et à collaborer avec des experts d'autres domaines. J'ai apprécié l'analyse critique et honnête de l'état de l'art et des méthodes développées, ainsi que le souci de rendre reproductibles les travaux effectués et de publier les algorithmes sous forme de logiciels documentés ; cela me semble attester de l'éthique scientifique de la candidate, et plus généralement de la qualité de son approche scientifique. Pour ces raisons, je suis favorable à la soutenance de thèse de Morgane Pierre-Jean en vue d'obtenir le grade de Docteur en sciences de la vie et de la santé de l'Université Paris Saclay.

Jean-Philippe Vert
Directeur de Recherche, MINES ParisTech





LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

MEDIZINISCHE FAKULTÄT
INSTITUT FÜR MEDIZINISCHE INFORMATIONSVERARBEITUNG,
BIOMETRIE UND EPIDEMIOLOGIE - IBE



Prof. Dr. Anne-Laure Boulesteix
IBE – Universität München
Marchioninistraße 15
81377 München
Allemagne

Tél: 089/4400-77598
Fax: 089/4400-77491

boulesteix@ibe.med.uni-muenchen.de
WWW: <http://ibe.web.med.uni-muenchen.de/>

RAPPORT SUR LA THÈSE SOUMISE PAR Madame Morgane PIERRE-JEAN

“Development of statistical methods for DNA copy number analysis in cancerology”

Madame Morgane Pierre-Jean soumet ici un travail en vue de l'obtention du diplôme de doctorat de l'Université Paris-Saclay, spécialité Sciences de la Vie et de la Santé. Elle propose et évalue dans sa thèse des nouvelles méthodes statistiques et bioinformatiques pour l'analyse du nombre de copies d'ADN, en particulier pour la segmentation et le *dictionary learning*. La thèse traite à la fois des aspects méthodologiques de ces procédures d'analyse tenant compte de la forte structure des données le long du génome, de leur implémentation logicielle et de leur évaluation systématique dans le cadre d'études de simulation de grande ampleur ainsi que d'applications à des données réelles dans le domaine de la cancérologie.

Résumé chapitre par chapitre

La thèse se compose de neuf chapitres structurés en trois grandes parties, suivis d'une conclusion générale. Le *Chapitre 1* est une rapide introduction sur le contexte biomédical des méthodes étudiées, traitant en particulier des thèmes suivants : génétique, cancer, nombre de copies d'ADN, méthodes de mesure et défis pour le statisticien/bioinformaticien. Le *Chapitre 2* donne une vue d'ensemble des méthodes

existantes de segmentation et présente une méthode bidimensionnelle permettant de réaliser conjointement la segmentation du nombre total de copies et de la fraction d'allèle B. Le *Chapitre 3* présente une étude comparative de grande ampleur des méthodes de segmentation du nombre de copies d'ADN et accorde une attention particulière au problème de la génération de données simulées réalistes et aux critères d'évaluation à utiliser. Le *Chapitre 4*, qui clôt la première partie dédiée aux méthodes de segmentation, propose une méthode de segmentation non-paramétrique par noyau dont l'idée est de détecter le long du génome des différences dans la distribution (et pas seulement dans la moyenne).

La deuxième partie, consacrée à la clonalité des tumeurs, s'ouvre sur le *Chapitre 5* qui traite de méthodes existantes permettant la détection de l'hétérogénéité des cancers tandis que le *Chapitre 6* présente une nouvelle méthode statistique pour l'étude de l'hétérogénéité des cancers sur la base des altérations du nombre de copies d'ADN. Cette méthode est appliquée dans le *Chapitre 7* à deux jeux de données réels, l'un public et l'autre mis à disposition par des partenaires de coopération cliniques.

Enfin, la troisième partie présente des pipelines bioinformatiques : l'une pour l'estimation de la diminution du taux d'hétérozygotie en l'absence de référence normale traitée dans le *Chapitre 8* et l'autre pour l'estimation du nombre de copies d'ADN à partir de données de séquençage traitée dans le *Chapitre 9*.

Evaluation

Forme

Ce mémoire volumineux est bien structuré et rédigé de manière claire. Mme Pierre-Jean écrit en anglais dans un langage correct et précis. La qualité des formules mathématiques, de la bibliographie et des figures satisfait aux exigences des meilleures revues des domaines scientifiques concernés.

Fond

Méthodes

Le manuscrit traite de nombreuses méthodes statistiques complexes que Mme Pierre-Jean décrit et analyse avec précision. Elle fait preuve d'une bonne maîtrise des outils statistiques et bioinformatiques impliqués dans les méthodes qu'elle considère ainsi que des problématiques biomédicales motivant son travail. Les nouvelles méthodes répondent à des réels besoins rencontrés en cancérologie tout en étant assez générales pour pouvoir, éventuellement sous une forme adaptée, ouvrir la porte à d'autres applications en dehors du cadre de la génomique/cancérologie.

Implémentation

Mme Pierre-Jean a systématiquement implémenté les méthodes considérées sous forme de packages R publiquement disponibles. Ces packages constituent un output important de la thèse. Par ailleurs elle accorde tout au long de son travail une grande importance aux aspects liés à l'implémentation et l'efficacité calculatoire des algorithmes, proposant en particulier des astuces pour limiter les ressources nécessaires.

Points forts à remarquer

Mme Pierre-Jean a dédié un soin tout particulier aux comparaisons entre méthodes (nouvelles et existantes). En particulier, elle accorde une attention importante au design d'études de simulation et au choix de critères de comparaison adaptés à la question scientifique considérée, évitant ainsi l'écueil classique consistant à ne se focaliser que sur le développement de nouvelles méthodes en les évaluant de manière superficielle et biaisée. D'autre part, les approches proposées pour la comparaison de méthodes pourront être réutilisées dans des études ultérieures (en particulier grâce au package publié) et présentent ainsi un intérêt propre au-delà de leurs résultats spécifiques dans le contexte étudié.

Un autre point très positif à noter est le large spectre de connaissances nécessaires à ce travail, à l'interface entre statistiques, mathématiques, bioinformatique, biologie et médecine. Mme Pierre-Jean soumet ici un travail prenant en compte les exigences et spécificités de chacun de ces domaines scientifiques et fait preuve d'une impressionnante capacité à travailler dans un contexte interdisciplinaire.

Publications / participation à des conférences

Mme Pierre-Jean a publié un article en tant que premier auteur dans la revue renommée *Briefings in Bioinformatics* (Impact Factor : 8.399). Elle est également premier auteur d'un article soumis à publication dans une revue de statistique et a présenté ses travaux lors de plusieurs conférences nationales et internationales de haut niveau. Mme Pierre-Jean fait preuve d'une bonne activité de publication. Le mémoire comprend en outre plusieurs chapitres dont le contenu n'a pas encore été publié mais pourra sans aucun doute faire l'objet de publications à l'avenir.

Conclusion

En conclusion, il s'agit d'un très bon travail de thèse alliant rigueur méthodique et innovation scientifique sur des thématiques pertinentes du point de vue médical. Les développements méthodologiques proposés répondent directement à des besoins

concrets des biologistes/médecins et constituent une contribution significative à l'état de l'art.

Aussi, j'émet un avis favorable pour la soutenance du travail en vue de l'obtention du grade de docteur en Sciences de la Vie et de la Santé de l'Université Paris-Saclay.

A Munich, le 28 octobre 2016

A handwritten signature in black ink, appearing to read 'A. Boulesteix', with a long horizontal stroke underneath.

*Prof. Dr. Anne-Laure Boulesteix
Université de Munich (Allemagne)*



Department of Statistics
The Wharton School
University of Pennsylvania
400 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104.6340
215.898.8222 phone
215.898.1280 fax

November 18, 2016

Review of Doctoral Thesis by Morgane Pierre-Jean

I had the pleasure to read the Doctoral Thesis by Morgane Pierre-Jean on "Development of statistical methods for DNA copy number analysis in cancerology." The thesis contains a detailed and comprehensive review of this field, and two original methods (described in Chapters 4 and 6). The methods are illustrated and tested on both simulated and real data sets. It is clear that Morgane has gained a good comprehension of published literature, and his new methods represent an exciting advance on existing state of the art. I find this to be an impressive collection of research work for a doctoral thesis, and ready for defense.

Morgane's thesis bridges the subjects of DNA copy number segmentation and tumor heterogeneity. This is a growing field that receives a lot of attention. Copy number aberrations are one of the fundamental genomic features in cancer, and play important roles in tumor progression and disease treatment. Segmentation of the genome by copy number is a difficult problem that, despite years of effort, could still benefit from statistical method innovation. The first original contribution of this thesis is a kernel based method for segmentation of DNA copy number based on the bivariate sequence total copy number ratio and B-allele frequency. Most existing methods segment DNA copy number only using by total copy number, since the B-allele frequency at each variant position depends on the genotype, and thus do not exhibit a simple shift in mean at breakpoints. Morgane's new method bypasses this difficulty by a nonparametric framework that detects change in distribution.

The second original contribution of Morgane's thesis is a comprehensive method for assessment of inter-tumor and intra-tumor heterogeneity using copy number profiles. Not all patients who have the same clinical diagnosis have the same tumor – inter-tumor heterogeneity refers to the difference in genetic profiles of tumors from patients with the same disease. Similarly, not all cells within a tumor have the same genetic profile – intra-tumor heterogeneity refers to the difference in genetic makeup between cells within a tumor. Such heterogeneity, at multiple levels, contribute to varied patient response to treatment and failures of targeted therapies. Morgane proposed a new framework for assessment of heterogeneity at multiple levels, and for finding shared copy number aberrations across tumor sub-clones and across samples. I think this is a very useful and innovative advance over existing methods.

More details and questions on each chapter are below.

Chapters 1-3 gave a comprehensive of existing segmentation algorithms, as applied to DNA copy number segmentation, and their strengths and weaknesses. Chapter 3 in particular describes an informative benchmark study of existing one-dimensional segmentation algorithms. One question I had

here was, when applying the two dimensional Gaussian model to $(\log R, \text{DoH})$, how do you deal with the fact that DoH is bounded between 0 and 1? Or, does it not matter too much?

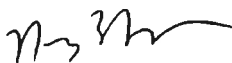
Chapter 4 describes an original research result: A non-parametric segmentation method using kernels that looks for change-points in the distribution of the signal. One nice feature of kernels is that it is straightforward to extend to multidimensional data by taking a weighted sum of the kernel functions, which is still a kernel. One issue that wasn't clear to me was how α , the weighting parameter, was actually chosen in the simulation and real studies. Maybe I missed something in my fast reading. I also find it a little difficult to understand why the kernel method does not result in more false positives, as Figure 4.5 indeed shows that it obtains good results. Intuitively, the BAFs change in a very specific way when there is a copy number gain or loss: The heterozygote BAFs split into a mixture of two means that are symmetric around 0.5. This is a strong restriction on how the distribution should change in the case of a true signal, and modeling this knowledge may reduce false positives? Anyway, this is a minor point.

Chapter 5 focuses on copy number segmentation and deconvolution in samples that are genetic mixtures, that is, samples containing cells that have different copy number profiles. One thing here that may need more explanation: In equation 5.4, why should the w_i 's be sparse? Chapter 6 builds on the background review of Chapter 5 to propose a new method of cancer subclone identification and integer copy number estimation. I think the proposed method is very interesting and can potentially be quite useful. My only question here is regarding the assumption that the latent profiles are shared across samples. When is this assumption reasonable? What if the samples come from different patients? Since the tumors in different patients evolve independently and may have different mutation histories, it seems that this assumption may be too restrictive. Or, by "profile" what we really mean are individual copy number aberrant loci and not whole genome profiles?

Chapter 7 contains some analysis of real data, which are quite interesting. The first example, on the ovarian data set, seems to illustrate the promise of the method, but it is treating each patient separately. Following along my question from Chapter 6, what if we combine all patients together? It would be interesting to see an inter-tumor heterogeneity analysis on this data. I find the analysis of the breast cancer samples also quite interesting, and partially helps me understand how the inter-tumor heterogeneity is done. It would be interesting to cluster these patients and see if there is any correlation with clinical outcome. Also, how do the results compare to a simple, straightforward clustering on the total copy number estimates?

In summary, I think this thesis is ready for defense, and the results, if polished, should be worthy of publication in very good journals. I would be happy to discuss further if more details are desired.

Sincerely,



Nancy R. Zhang
Associate Professor
Department of Statistics
The Wharton School
University of Pennsylvania