Chloé-Agathe Azencott

Dossier de qualification aux fonctions de maître de conférence: Pièces complémentaires

Table des matières

Projet d'enseignement

Si j'ai jusqu'à présent relativement peu enseigné, j'ai eu maintes fois en temps que chercheur l'occasion de confirmer l'importance à mes yeux des activités de transmission (écriture d'articles, présentations orales) et d'encadrement (notamment des élèves de master dont j'encadre les travaux de recherche, mais aussi des doctorants avec lesquels j'ai collaboré). Je prends en outre beaucoup de plaisir à expliquer aussi bien mes travaux que les bases théoriques et considérations pratiques sur lesquels ils s'appuient. C'est pourquoi je souhaite faire de l'enseignement une part conséquente de mon activité professionnelle.

Ayant reçu de multiples bourses de recherche pendant ma thèse à UC Irvine, je n'ai pu exercer en tant que moniteur (*teaching assistant*) que pendant deux trimestres, selon le minimum requis. J'ai complété cette expérience en suivant les cours de pédagogie offerts aux doctorants. Ils m'ont notamment permis d'apprendre à définir les objectifs d'un cours et à mettre au point un plan de cours, des exercices et des questions d'examen qui en découlent. Actuellement en post-doctorat au Max-Planck, institut de recherche sans activité d'enseignement, je participe autant que possible aux activités d'enseignement du directeur de mon laboratoire à l'Université de Tübingen. Ces expériences m'ont permis de me confronter à l'enseignement, depuis l'élaboration d'un cours jusqu'à l'évaluation des étudiants, et je crois avoir beaucoup appris d'elles. En particulier, j'ai appris à impliquer les étudiants, par exemple lors du module « Probabilités et statistiques en informatique », où j'ai su adapter les exercices et le déroulement de mes travaux dirigés pour évoluer depuis une première séance durant laquelle les étudiants ne parlaient quasiment pas vers des scéances réellement interactives.

Un projet d'enseignement ne peut s'élaborer qu'en fonction de l'établissement dans lequel il est destiné à se développer. Néanmoins j'envisage les activités suivantes :

- A court terme, je suis disposée à enseigner non seulement des modules correspondant à mon domaine d'expertise (apprentissage statistique, théorie des graphes, bio- et chémo-informatique), mais aussi les cours d'introduction à l'informatique. Mon expérience de communication avec des biologistes, chimistes, médecins et pharmaciens dans le cadre de ma recherche me prépare à la prise en charge de cours pour non-spécialistes.
- A long terme, je serais intéressée par le développement d'un cursus d'informatique niveau Master centré sur la recherche thérapeutique, qui fournisse aux étudiants les bases de sciences de la vie nécessaire à la compréhension des grands problèmes de bio- et chémo-informatique, et leur permette d'approfondir les outils informatiques nécessaires (apprentissage statistique, mais aussi par exemple bases de données, masses de données, et visualisation).

J'estime aussi qu'il est important d'apprendre le plus tôt possible à lire des publications scientifiques et envisage pour cela la mise en place d'un séminaire similaire à celui pour lequel j'ai enseigné à Tübingen, dans lequel les étudiants de L3 et master seraient invités à lire et analyser des publications récentes.

Enfin, étant bilingue, je souhaiterais prendre en charge un module d'informatique en anglais, qu'il s'agisse d'un cours particulièrement conçu pour familiariser les étudiants avec la langue anglaise scientifique, d'un cours offert en anglais parallèlement au même cours en français, ou d'un cours en anglais dans le cadre d'un master international.

Projet de recherche

Je souhaite continuer à développer de techniques d'apprentissage automatique permettant de faciliter la recherche thérapeutique. En particulier, je prévois d'organiser mes travaux autour du thème de la compréhension de l'effet des mutations du génome, notamment dans le cadre de maladies humaines. Encore une fois, un projet de recherche dépend de l'établissement dans lequel il s'inscrit, mais je peux ici esquisser quelques pistes que j'ai l'intention de poursuivre.

1. Effets de sous-ensembles structurés de loci génétiques

Dans l'immédiat, la poursuite de mes travaux de recherche est centrée sur la découverte de sous-ensembles structurés de mutations génétiques associées à un phénotype. En particulier, il s'agira d'extensions de la méthode SConES (présentée dans la section 3.3 du document principal) permettant de détecter des réseaux de loci génétiques conjointement associés à un phénotype.
Ces extensions prennent la forme d'un développement théorique, permettant notamment :

- La généralisation de SConES à n'importe quelle contrainte sous-modulaire, ce qui permet dans le cas pratique qui nous intéresse de définir des structures plus générales que celles déterminées par un réseau (par exemple, la combinaison de sous-réseau d'un réseau biologique définis par des interactions, et de groupes définis par des voies métaboliques) ;
- La généralisation de SConES à divers types de tests d'associations, y compris des tests de corrélation type HSIC [1] qui permettent une plus grande flexibilité. L'extension à des tests permettant de prendre en compte des effets multiplicatifs reste un défi, car la formulation actuelle, ne bénéficiant plus de la séparabilité que permet la linéarité, devient trop complexe et en temps et en espace pour être applicable ;
- Une étude statistique de SConES. Il serait en effet intéressant de dériver du test d'association utilisé une valeur p associée aux réseaux prédits par SConES. Pour que pouvoir interpréter ces valeurs p, il faudra alors comprendre comment prendre en compte les comparaisons multiples effectuées lors du déroulement de l'algorithme.

J'envisage aussi des développements plus applicatifs, en particulier :

- l'étude de la définition de réseaux biologiques entre SNPs. Les réseaux que SConES utilise actuellement reposent sur la notion de proximité d'un SNP avec un gène ; deux SNPs sont connectés s'ils sont proche du même gène, ou s'ils sont proche de deux gènes eux-même connectés dans un réseau biologique. Ce type de construction est assez limitatif, en particulier parce qu'il ne permet pas de prendre des effets de régulation en compte. Ce projet demandera l'aide de biologistes, soit parmi mes collaborateurs actuels, soit dans le cadre de la mise en place de nouvelles collaborations ;
- l'application de SConES à divers cas d'études, en particulier la broncho-pneumopathie chronique obstructive, dans le cadre de ma participation à l'étude COPDGene [2], qui collecte et analyse des données génomiques pour des dizaines de milliers de patients souffrant de cette affliction. L'extension de la méthode à des phénotypes corrélés multiples permettrait aussi de combiner cette étude avec une étude sur l'asthme.

Ces travaux seront dans un premier temps axés sur l'étude de mutations ponctuelles (SNPs), mais une extension à d'autres types de mutations, tels que duplications (en particulier, variabilité du nombre de copies d'un gène, ou CNVs, *Copy Number Variations*), délétions ou insertions, est envisageable.

1. A. Gretton et al., *Measuring statistical dependence with Hilbert-Schmidt norms*, Algorithmic Learning Theory, pp. 63–77, 2005
2. `http://www.copdgene.org/`

A plus long terme, la question de l'*apprentissage* du réseau de SNP sous-jacent me paraît intéressante et pertinente. Ce problème peut être reformulé comme trouver l'arête à ajouter au graphe pour acroître maximalement la valeur de l'objectif de SConES.

2. Causes et effets de la modification de fonctions protéiques

Je prévois à moyen terme de développer un axe de recherche complémentaire à ce premier, en étudiant plus précisément les causes de la modification de fonctions protéiques.
De tels changements peuvent être la conséquence d'au moins deux types d'effets génétiques :

- *effets génétiques « directs »* : la mutation d'un gène entraîne un changement structurel de la protéine qu'il encode. Les effets de ces mutations dites faux sens sont les mieux étudiés ;
- *effets génétiques « indirects »* : une mutation en dehors d'un gène entraîne un changement pour la protéine qu'il encode, par exemple un changement en abondance via des effets de régulation. Ces effets sont bien moins connus, du moins à large échelle, mais les larges jeux de données GWA et d'études eQTL (visant à déterminer les loci génétiques qui régulent l'expression de protéines ou d'ARNs) devraient permettre de les étudier de manière plus systématique et c'est sur ce point que je me propose de me concentrer plus particulièrement. Remarquons aussi que les réseaux de régulations plus précis et complets qui seraient ainsi mis au point pourraient être utilisés dans le cadre de l'algorithme SConES présenté ci-dessus.

Cet axe rejoindra le premier décrit ci-dessus lorsque l'on intégrera des annotations développées suite à l'étude de ces différents cas à des analyses multi-locus de data GWA. C'est une extension naturelle du travail mené actuellement par Fabian Aicheler dans le cadre de son Master Recherche sur l'enrichissement de la représentation de SNP par des annotations liées à leur potentiel délétère.

Enfin, à plus long terme, la question des effets de ces modifications est aussi ouverte. Il s'agit non plus de prédire si la fonction d'une protéine change, mais comment ce changement affecte la santé ou les signes cliniques du patient. Si l'on peut essayer d'effectuer de telles prédictions à partir de données concernant seulement la protéine et ses modifications, on peut raisonnablement attendre de meilleurs résultats de méthodes qui prennent en compte l'environnement de la protéine modifiée. Une approche systémique pourrait s'appuyer la propagation (type marche aléatoire, cheminement, ou noyaux pour graphes) d'informations concernant ces modifications sur un réseau d'interactions entre protéines.

3. Medecine personnalisée

Au fur et à mesure que les jeux de données GWA grandissent, il devient de plus en plus raisonnable d'envisager de les utiliser dans le cadre du développement de diagnostiques, pronostiques et traitements personnalisés (couramment regroupés sous l'appellation de « médecine personnalisée »). En particulier, certains consortiums et groupes de recherche avec lesquels je collabore actuellement, tels que COPDGene [3] aux États-Unis ou le groupe de recherche de Marcus Ising au Max Planck de Psychiatrie à Munich [4], commencent à recueillir, outre les profils génétiques et d'états cliniques, des données sur la réponse des patients à certains traitements.

Dans ce contexte, je projette de tirer parti de mon expérience combinée en chémoinformatique et en génomique pour développer de nouvelles techniques permettant d'intégrer le point de vue moléculaire aux analyses génomiques proposées précédemment. Il faudra pour cela étudier les spécificités de

3. `http://www.copdgene.org/`
4. `http://www.mpipsykl.mpg.de/en/research/groups/ising/index.html`

l'utilisation de phénotypes complexes tels que la réponse à un traitement, en sachant prendre en compte les différents facteurs environnementaux influençant ces réponses (en s'appuyant sur les travaux visant à corriger l'impact de tels facteurs même quand ils sont cachés). Un autre obstacle sera le fait que ces données sont nécessairement incomplètes au sens où il est rare, voire impossible, que plusieurs thérapies soient testées successivement et dans des conditions comparables sur plusieurs patients. L'utilisation de la description chimique de différents traitements pour quantifier leurs similarités pourrait alors pallier ces difficultés.

Tableau des acronymes

UCI	University of California, Irvine
MLCB	Machine Learning and Computational Biology (groupe de recherche)
ICML	International Conference on Machine Learning (conférence)
MLCB	Machine Learning and Computational Biology (colloque)
MLSB	Machine Learning in Systems Biology (colloque)
NETTAB	Network Tools and Applications in Biology (colloque)
COPD	Chronic Obstructive Pulmonary Disease
CPU	Computing Processing Unit
CROC	Concentrated Receiver-Operator Characteristic
eQTL	Expression Quantitative Trait Loci
IRV	Influence-Relevance Voter
GLIDE	GPU-based LInear regression for the Detection of Epistasis
GPU	Graphics Processing Unit
GWA(S)	Genome-Wide Association (Study)
HSIC	Hilbert-Schmidt Independence Criterion
ROC	Receiver-Operator Characteristic
SConES	Selection of Connected Explanatory SNPs
SNP	Single Nucleotide Polymorphism

Liste des pièces jointes

Documents administratifs

Sont joints à ce dossier :
- Une copie de la déclaration de candidature ;
- Un justificatif des activités d'enseignement à University of California Irvine ;
- Un justificatif des activités d'enseignement à Universität Tübingen ;
- Une copie de la lettre attestant de la réception de la bourse de doctorat IBM en 2009 ;
- Une copie de la lettre d'offre de bourse de la fondation Alexander von Humboldt en 2010.

Lettres d'appréciation

Les trois membres de mon comité de thèse soutiennent ma candidature et feront parvenir leurs lettres de recommandation aux rapporteurs par courrier électronique.
- Pierre Baldi (directeur de thèse et président du jury) ;
- Padhraic Smyth ;
- Sheryl Tsai.

Est jointe à ce dossier une lettre de recommandation de Karsten Borgwardt, mon superviseur actuel.