Statistique et évolution génomique des populations

Olivier François, Professeur Ensimag, Grenoble INP. Laboratoire TIMC-IMAG, Université Joseph Fourier Grenoble, CNRS UMR 5525, Equipe Biologie Computationnelle et Mathématique.
Propos recueillis par Emilie Neveu.

Quels sont les processus étudiés ? Quelles sont les applications environnementales - hors médecine, astronomie et géophysique interne ?

Mes recherches portent sur la statistique appliquée en génétique des populations et sur l'évolution génomique. Je développe des méthodes statistiques et numériques pour l'inférence de la structure génétique des populations, la démographie et l'adaptation locale.

L'échelle moléculaire est importante. Nous étudions, par exemple, les déplacements d'espèces de plantes alpines face aux changements climatiques à partir des données moléculaires, et de génotypes. Pour cela, nous devons identifier les signatures de l'adaptation dans les gènes. Avec pour but de répondre à la question : Quelles traces la sélection peut-elle laisser sur les génomes des organismes ?

On retrouve les mêmes problèmes en médecine avec, par exemple, l'étude du séquençage génétique des agents infectieux et leurs réponses adaptatives à l'environnement. Dans ce cas, l'environnement est contrôlé par le clinicien via le traitement administré au patient. De nombreuses méthodes ont d'abord été inventées pour la génétique humaine et médicale puis importée en écologie. Exclure du projet les études en médecine et à l'échelle moléculaire n'est pas pertinent, puisque les écologues étudient justement les liens entre gènes et environnement, utilisant les mêmes méthodes que la génétique humaine.

Mon groupe de recherche distribue également des logiciels qui permettent le calcul de coefficients d'ascendance génétique et la détection de l'adaptation locale en utilisant des marqueurs génétiques et des variables géographiques et environnementales.

Quels sont les modèles mathématiques ?

Mes approches sont principalement basées sur les méthodes de Monte Carlo par chaînes de Markov, le « machine learning » et les approches bayésiennes, et utilisent des données géographiques et environnementales. Le but est de fournir des procédures d'estimation statistique en génétique des populations.

J'étudie aussi des applications à la théorie de coalescence, les propriétés mathématiques des généalogies et la forme des arbres phylogénétiques.

Les méthodes doivent être adaptées à chaque cas.

En général, on ne peut pas se contenter d'analyser les données à l'aide de simples calculs de corrélation, et de nombreuses corrections doivent être faites. Par exemple, lorsqu'on cherche à associer des gènes à une maladie, il faut prendre en compte l'analyse d'ascendance génétique. La réduction de dimension est elle aussi utilisée pour expliquer la variabilité des données.

Des modèles de démographie, de médecine, d'écologie sont nécessaires pour aider à l'interprétation des données, par exemple pour les études de relations inter-espèces, de prédiction de mouvements d'espèces selon les changements climatiques. La modélisation est la composante principale dans les études d'association gène-environnement, pour étudier quels sont les gènes d'adaptation au climat, et les liens entre le climat, l'agriculture, et les maladies apparaissant. Il est prouvé qu'il existe des gènes, des mutations spécifiques pour des maladies ayant une composante génétique comme l'alcoolisme ou l'intolérance au gluten (la maladie cœliaque) liées aux changements environnementaux.

Il est à noter que le modèle d'analyse de données adapté sera différent selon que l'on cherche à prédire ou à expliquer. Pour expliquer, nous allons chercher le lien le plus fort entre les sorties et les entrées sans forcément prendre en compte des sorties qui ont un impact prédictif plus important.

Travailler en méthodologie demande aussi beaucoup de développement logiciel. Il est important de pouvoir diffuser ces méthodes pour qu'elles soient utilisées par des biologistes ou des écologues.

De ce point de vue-là, avez-vous des besoins en HPC ?

En bio-informatique, nous devons faire face à de gros calculs et gérer des centaines de Téra de données. Mais, il s'agit aussi d'accélérer les méthodes. Voilà pourquoi de nombreux algorithmes en parallèle se développent. Le machine learning permet de faire beaucoup de choses là où les méthodes de Monte Carlo sont trop lentes.

Je m'inspire aussi beaucoup des méthodes d'apprentissage. Il faut bien sûr adapter la méthode, et analyser son fonctionnement numérique sur les exemples spécifiques de biologie et d'écologie.

Vous nous parlez des méthodes d'apprentissage. Y a t'il un lien avec le Big Data ?

« Big Data » est un « buzz word » qui a en fait été inventé par les génomistes quand le décryptage de l'ADN a eu lieu. Il est repris maintenant dans le contexte des données sociales.

Donc, oui il y a un lien avec le Big Data.

Vous travaillez en statistiques dans un laboratoire composé de moitié de médecins. Votre équipe est un peu à part dans le paysage français, non ? Pourquoi ?

Oui, en France, les département de mathématiques sont parfois isolés des applications. Les UFR et les CNU, auxquels sont attachés les laboratoires peuvent être des freins institutionnels à l'interdisciplinarité. Il nous est par exemple difficile de recruter des postes de mathématiciens dans l'équipe, alors que cela serait très utile. Le CNRS permet une plus grande flexibilité.

Les mathématiciens purs ou appliqués ont des préoccupations et des objectifs parfois trop éloignées des biologistes. Il y a une difficulté à trouver des intérêts communs et une difficulté à communiquer, à cause d'un vocabulaire différent qui peut agir comme une barrière. Pour les mêmes raisons, l'évaluation des projets interdisciplinaires est difficile.

Pour créer une réelle interaction, des mathématiciens devraient être placés dans les laboratoires de sciences appliquées. Cela demande un réel effort des mathématiciens car les biologistes font les maths qui leur sont utiles de toute façon. Il y a un grand besoin de double affiliation.

Pourquoi ce besoin de double affiliation ?

La même personne doit faire des mathématiques et de la biologie. Sinon la barrière disciplinaire va perdurer.

Pour travailler sur des projets réellement appliqués, les mathématiciens doivent se rapprocher des biologistes, et ne plus les considérer comme de simples utilisateurs. De nombreux biologistes et écologues font l'effort de faire des mathématiques, ils développent même des logiciels. Celui de Wilfried Thuiller, par exemple, est déjà enseigné dans les masters. Alors qu'il est a priori plus facile pour quelqu'un formé en mathématiques d'apprendre certains aspects de la biologie, que pour un biologiste d'apprendre les mathématiques.

Cependant, il ne faut pas tomber dans l'arrogance, et croire qu'en tant que mathématicien, je vais pouvoir apporter des solutions toutes faites. Cela demande du temps et de l'investissement personnel pour comprendre les réels besoins d'un biologiste.

Dans le même temps, il faut maintenir un niveau en mathématiques élevé pour être reconnu par la communauté. De fait aux Etats-Unis, les chercheurs à l'interface entre deux disciplines publient dans des revues de biologistes mais aussi de mathématiciens.

Aux Etats-Unis c'est différent ?

Oui, à l'étranger en général, il existe des petits groupes mixtes, des laboratoires ou des départements de mathématiques dans les facultés de médecine. Par exemple, Noah Rosenberg, de l'Université de Stanford, est un mathématicien dans un département de biologie. Tout le système est légèrement différent. Notamment un élève brillant en mathématiques sera incité à faire une thèse à l'interface entre deux disciplines, par exemple la biologie.

En France, nous sommes très en retard de ce point de vue et c'est en partie dû au système très rigide des UFR et au problème de recrutement des postes interdisciplinaires.

Cependant, il existe des facilités avec le CNRS et les intersections STII, ou les nouvelles équipes hébergées par le Collège de France, notamment SMILE (Stochastic Models for the Inference of Life Evolution) menée par Amaury Lambert, UPMC, au Center for Interdisciplinary Research Biology (CIRB). Ce sont ces structures qui gagneraient à être développées en France. Et ce n'est pas anodin, nombre de mes étudiants partent à l'étranger après leur thèse et ne reviennent pas.

Le CIRB ou Center for Interdisciplinary Research in Biology (CIRB) est une nouvelle structure de recherche Collège de France / CNRS / INSERM hébergée par le Collège de France dans le centre de Paris. Neuf équipe de différents horizons ont récemment créé cette structure dans l'esprit de promouvoir de nouvelles collaborations en biologie et à travers les disciplines qui la composent. Au long terme, les neuf équipes spécialisées dans les maladies infectieuses, les neurosciences et la recherche cardio-vasculaire, vont être rejoints par un nombre similaire d'équipes, essentiellement de nouvelles équipes, incluant des chimistes, des physiciens et des mathématiciens ayant un profond intérêt pour la biologie. Ce nouveau Centre va bénéficier de la proximité de nombreux autres laboratoires et d'un milieu intellectuel extraordinairement riche, proposant des conférences qui vont couvrir tous les domaines de la connaissance. Le CIRB a développé de fortes interactions avec les institutions à l'extérieur du Collège de France, en particulier l'ENS et l'Institut Curie.