Proposition de sujet de Master 2

1 Intitulé du stage

Caractérisation fine de la biodiversité via les paramètres d'un modèle statistique : exemple sur la diversité moléculaire des arbres de Guyane.

2 Contexte scientifique

La biodiversité, comprise comme fruit de l'histoire évolutive, est au coeur de la biologie, comme l'indique le sous-titre de l'ouvrage d'Ernst Mayr[1]. Elle est une notion intuitive mais difficile à définir précisément. Plusieurs types de diversités sont reconnus : diversité génétique, spécifique, fonctionnelle. De même, on distingue la diversité α, en un lieu comme la diversité d'une communauté, et diversité β, entre lieux différents. Il existe de nombreux indices qui permettent chacun de quantifier tel ou tel aspect de la diversité, de prendre en compte telle ou telle information. La quantification de la diversité fonctionnelle par exemple repose sur les dissimilarités entre valeurs des traits de vie.
L'un des indices les plus utilisés en écologie des communautés est l'indice de Shannon, qui relie la diversité avec l'entropie de l'ensemble des comptages par espèce (ou fréquence) des individus d'une communauté, élargie par Hill à une famille d'indices reliés à l'entropie de Rényi. Plus récemment, ces indices ont été enrichis en tenant compte d'informations complémentaires, comme la distance entre les catégories (ici, les espèces), comme l'entropie quadratique de Rao.
Tous ces indices sont *scalaires*, c.a.d. unidimensionnels. Cela présente un avantage opérationnel : il est alors simple de comparer la diversité de deux communautés, ou d'une communauté avant et après un acte de gestion et d'en étudier l'impact. Cependant, résumer la diversité d'une communauté par un nombre est réducteur. La notion même de diversité pousse à la caractériser selon plusieurs axes, ou plusieurs directions, non nécessairement parallèles. De telles approches multidimensionnelle ont été récemment proposées[2].
Dans ce stage, nous nous concentrons sur la diversité spécifique et proposons d'étudier une approche différente pour une caractérisation multidimensionnelle de la structure de cette biodiversité. Elle repose sur une hypothèse construite en trois étapes :

- la structure de la biodiversité est contenue dans la structure du tableau de distances entre individus,
- ce tableau de distances peut être modélisé comme réalisation d'un modèle statistique paramétrique des distances entre des groupes auxquels les individus appartiennent.

[1] Mayr, E., *The Growth of Biological Thought; Diversity, Evolution, Inheritance*, Harvard University Press, 1985.

[2] Villégier, S. Mason, N. W. H. & Mouillot, D. - 2008 - New Multidimensional functional diversity indices for a multifaceted framework in functional ecology. *Ecology*, **89(8):**2290-2301 ; Naeem, S. *et al.* - 2016 - Biodiversity as a multidimensional construct: a review, framework and case study of herbivory's impact on plant biodiversity. *Proc. R. Soc. B.*, **283:**20153002, http://dx.doi.org/10.1098/rspb.2015.3005.

- l'ensemble des paramètres d'un tel modèle est alors un bon résumé de la structure de la diversité spécifique présente.

L'idée sous-jacente est que, de même que la moyenne et l'écart type sont un bon résumé d'une distribution gaussienne (elles permettent même de la reconstruire exactement), les paramètres du modèle statistique produisant le tableau de distances en sont un bon résumé.
On associe donc à une communauté non plus une seule valeur, comme pour les indices scalaires, ou un ensemble de valeurs estimées séparément, comme dans les indices multidimensionnels, mais un jeu de paramètres estimés ensemble qui forme un tout insécable. Notre hypothèse est que cette description va être plus riche et donc plus discriminante.
Le modèle retenu est le modèle SBM[3] (Stochastic Block Model en anglais), qui est un modèle très flexible pour la recherche de structure dans une matrice de distances. Vis à vis des indices multidimensionnels, cette approche présente deux avantages :

- le résumé est construit selon un algorithme, et n'est pas un choix "au dire d'expert"
- comme le SBM modélise l'ensemble du tableau de distances, les paramètres estimés du modèle couvrent l'ensemble de la diversité.

3 Question de recherche

Etudier la possibilité d'une caractérisation fine de la diversité spécifique d'une communauté par les paramètres d'un modèle SBM de la structure des distances moléculaires entre les individus.

4 Méthode de travail

Le premier attendu du stage est une synthèse bibliographique qui, en s'appuyant sur les indices scalaires maintenant bien établis, portera sur les développements récents d'indices multidimensionnels. Ce travail sera utile tant pour la communauté que pour la suite du stage.

Il s'agira ensuite de comparer certains de ces indices et les modèles SBM dans leur capacité à discriminer des patterns de biodiversité différents. Pour cela, nous disposons d'un jeu de données construit à partir d'une collection de 2000 arbres situés dans la parcelle expérimentale 'Piste de Saint Elie', en Guyane[4] et pour lequel les distances entre séquences ont été calculées pour chaque paire d'individus. Ce jeu de données sera utilisé pour créer plusieurs sous-jeux de données correspondant à des patterns de diversité différents.

La caractérisation de la biodiversité par modèle SBM se fera à partir des valeurs des paramètres du modèle. La contre-partie est qu'il n'y a plus de manière simple pour ordonner des communautés par diversité croissante à partir d'un vecteur de valeurs. Il faudra donc étudier les options possibles pour 'ordonner' / comparer deux communautés décrites par deux modèles SBM.

5 Informations pratiques

- Nom et contact des encadrants :

[3] K. Nowicki & T. Snijders. Estimation and prediction for stochastic block-structures, *J. Am. Stat. Assoc.*, **96:**1077-1087, 2001

[4] H. Caron, J.-F. Molino, D. Sabatier, P. Chaumeile, C. Scotti-Saintagne, J.6M. Frigério, I. Scotti, A. Franc and R. J. Petit, Chloroplast DNA variation in a hyperdiverse tropical tree community. *Ecology and Evolution* ,**3909**(8):4897–4905, 2019

- Alain Franc (unité BioGeCo, INRAE Bordeaux, alain.franc@inrae.fr)
- Nathalie Peyrard (unité MIAT, INRAE Toulouse, nathalie.peyrard@inrae.fr)

- Laboratoire d'accueil : unité BioGeCo, INRAE Bordeaux ou unité MIAT, INRAE Toulouse (avec des visites régulières au laboratoire qui ne sera pas le laboratoire d'accueil)
- Lieu du stage : centre INRAE de Pierroton ou centre INRAE de Toulouse
- Montant de la gratification : environ 3500 € sur l'ensemble du stage (6 mois, de janvier à juin)
- Exigeances particulières : formation principale en écologie, avec des notions de modélisation et une expérience du logiciel R.
- Date limite de candidature : mi-octobre 2020
- Date de retour sur les candidatures : mi-novembre 2020