Révéler les signaux mineurs dans l'analyse du microbiote intestinal : une nouvelle transformation des données pour une interprétation améliorée.

David Martin*[1,2], Frédéric Derbré[2], and Valerie Monbet[1]

[1]Institut de Recherche Mathématique de Rennes – Université de Rennes, Institut National des Sciences Appliquées - Rennes, École normale supérieure - Rennes, Université de Rennes 2, Centre National de la Recherche Scientifique, Institut Agro Rennes ANgers – France

[2]Laboratoire Mouvement Sport Santé – Université de Rennes, École normale supérieure - Rennes, Université de Rennes 2, Structure Fédérative de Recherche en Biologie et Santé de Rennes – France

Résumé

L'exploration des données décrivant l'écosystème bactérien intestinal est complexe et nécessite l'application de méthodes appropriées. Les méthodes que nous déployons dans cette étude sont des méthodes d'analyses non-supervisées (i.e. clustering), ainsi que des méthodes d'analyses supervisées (i.e. classification).
Nous montrons que (1) les méthodes d'analyses précédemment évoquées sont principalement influencées par un ou plusieurs groupes d'espèces avec une forte abondance ou variance (nous les appelons signaux dominants et sont très proches du concept d'entérotype). Dans ce contexte, nous montrons aussi (2) que les signaux dominants masquent la détection d'espèces moins abondantes (i.e. les signaux mineurs) quelques soit les methodes.
Afin d'explorer l'impact des espèces moins exprimées dans les analyses, nous proposons une transformation des données préservant les signaux mineurs et minimisant l'influence des signaux dominants sur les analyses. Nos expériences numériques démontrent que cette transformation (1) génère un clustering étroitement associés à la santé de l'hôte et qu'elle (2) améliore la performance des algorithmes d'apprentissage automatique dans des contextes où le fléau de la dimensionnalité est forte ($n << p$).
Ces résultats sont associés avec une influence plus grande des signaux mineurs dans les analyses. Nous appliquons ensuite, cette transformation sur les données issues de plusieurs cohortes précédemment publiées. Nous démontrons que les signaux dominants masquent des informations liées à la santé de l'hôte et que les signaux mineurs permettent d'améliorer la performance prédictive de cette même variable. Alors que les signaux dominants (ici, l'expression abondante des espèces Prevotella) sont liés à cette même capacité d'exercice, la transformation des données prouve que les signaux dominants agissent, dans cette cohorte, comme des facteurs confondants.
Cela suggère que, malgré leur association significative avec la santé de l'hôte, les signaux dominants peuvent biaiser les interprétations. Nous proposons que d'autres recherches portant sur l'élaboration de modèles prédictifs, les testent avec et sans transformation, afin de confirmer si les signaux dominants sont réellement " des arbres qui cachent la forêt ".

*Intervenant