proceedings

MDPI

Proceedings

Identification of Prevotella, Anaerotruncus and Eubacterium Genera by Machine Learning Analysis of Metagenomic Profiles for Stratification of Patients Affected by Type I Diabetes †

Diego Fernández-Edreira [1], Jose Liñares-Blanco [1,2,*] and Carlos Fernandez-Lozano [1,2]

[1] Department of Computer Science and Information Technologies, Faculty of Computer Science, Universidade da Coruña, Campus Elviña s/n, 15071 A Coruña, Spain; diego.fedreira@udc.es (D.F.-E.); carlos.fernandez@udc.es (C.F.-L.)

[2] CITIC-Research Center of Information and Communication Technologies, Universidade da Coruña, 15071 A Coruña, Spain

* Correspondence: j.linares@udc.es; Tel.: +34-881-01-1302

† Presented at the 3rd XoveTIC Conference, A Coruña, Spain, 8–9 October 2020.

Published: 27 August 2020

check for updates

Abstract: Previous works have reported different bacterial strains and genera as the cause of different clinical pathological conditions. In our approach, using the fecal metagenomic profiles of newborns, a machine learning-based model was generated capable of discerning between patients affected by type I diabetes and controls. Furthermore, a random forest algorithm achieved a 0.915 in AUROC. The automation of processes and support to clinical decision making under metagenomic variables of interest may result in lower experimental costs in the diagnosis of complex diseases of high prevalence worldwide.

Keywords: diabetes; machine learning; microbiome; metagenomics; data science

1. Introduction

It is known that diabetes type I (DTI) is a disease that is closely linked to changes in the microbiota [1]. Typically, works that study the metagenomic profile of a microbe in DTI uses only conventional statistical approaches [2]. Therefore, in this work a novel methodology to analyze DTI status using machine learning (ML) is proposed. In addition, new metagenomics genera are been identified with potential in the development of this disease.

2. Materials and Methods

OTUs genera faecal samples from 124 newborns were downloaded from Diabinmune project [2]. The experimental design starts removing near zero features and scaling the data; Random Forest (RF) [3] and glmnet [4] algorithms were used following a nested cross validation (CV) approach for training the models. A holdout was used for hyperparameter tuning (2/3 for training and 1/3 for testing) followed by a 10-fold CV for model validation (repeated 5 times).

3. Results

We have obtained 45 genera suitable for carrying out the study. Figure 1a showed the experimental results carried out. We found a statistical difference between the models and the best results were achieved with RF. Feature importance is shown in Figure 1b. *Prevotella* is the bacteria with the higher accumulated importance along with *Anaerotruncus*, *Scherichia*, *Eubacterium*, *Odoribacter* and *Collinsella*.

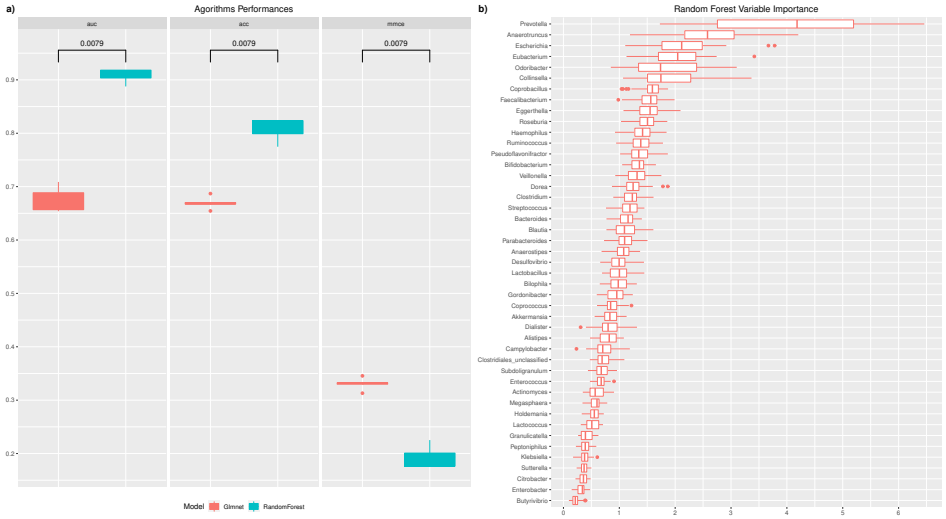

Figure 1. (**a**) Comparison of the 5 times 10-fold CV using a Wilcoxon test and (**b**) RF variable importance.

4. Discussion

We found in the literature that *Prevotella* and *Eubacterium* are strongly linked to DTI and *Anaerotruncus* with gestational diabetes. All of them are also correlated with instestinal dysbiosis processes [5,6]. In summary, we demonstrated the feasibility of a ML analysis of metagenomic profiles.

Author Contributions: Conceptualization, J.L.-B. and C.F.-L.; methodology, C.F.-L.; software, D.F.-E., J.L.-B. and C.F.-L.; formal analysis, D.F.-E.; Writing—Original Draft preparation, D.F.-E.; Writing—Review and Editing, D.F.-E., J.L.-B. and C.F.-L.; supervision, J.L.-B. and C.F.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the "Collaborative Project in Genomic Data Integration (CICLOGEN)" PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER)—"A way to build Europe." and the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431G/01, ED431D 2017/16), the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23) and Competitive Reference Groups (Ref. ED431C 2018/49). The funding body did not have a role in the experimental design; data collection, analysis and interpretation; and writing of this manuscript.

References

1. Tai, N.; Wong, F.S.; Wen, L. The role of gut microbiota in the development of type 1, type 2 diabetes mellitus and obesity. *Rev. Endocr. Metab. Disord.* **2015**, *16*, 55–65.
2. Kostic, A.D.; Gevers, D.; Siljander, H.; Vatanen, T.; Hyötyläinen, T.; Hämäläinen, A.M.; Peet, A.; Tillmann, V.; Pöhö, P.; Mattila, I.; et al. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. *Cell Host Microbe* **2016**, *20*, 121.
3. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
4. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
5. Siljander, H.; Honkanen, J.; Knip, M. Microbiome and type 1 diabetes. *EBioMedicine* **2019**, *46*, 512–521.
6. Hasan, S.; Aho, V.; Pereira, P.; Paulin, L.; Koivusalo, S.B.; Auvinen, P.; Eriksson, J.G. Gut microbiome in gestational diabetes: a cross-sectional study of mothers and offspring 5 years postpartum. *Acta Obstet. Gynecol. Scand.* **2018**, *97*, 38–46.