EasyChair Preprint

№ 1016

Discovering the Inter-species Interaction among Microorganisms Based on Iterative Random Forest Algorithm

Xinzhe Pang and Xingpeng Jiang

May 26, 2019

Discovering the Inter-species Interaction among Microorganisms Based on iRF Algorithm

Xinzhe Pang, M.S.[1], and Xingpeng Jiang, Ph.D.[2]
[1] Central China Normal University, China, pangxinzhe@mails.ccnu.edu.cn
[2] Central China Normal University, China, xpjiang@mails.ccnu.edu.cn

I. Introduction

Microorganisms are widely distributed in various parts of the human body[1]. They interact and interweave to form a complex and diverse microbial community ecosystem[2]. Changes in the microbial community lead to abnormalities in the body's local state and even cause disease. Therefore, it is important to study the microbial composition and interaction of different populations for disease treatment and health monitoring. In this paper, we proposed a microbial inter-species interaction discovery strategy using an iterative random forest (iRF) algorithm, and fusion phylogenetic distance. Finally, we applied this algorithm to the microbial dataset of human intestinal cirrhosis. The experimental results show that this method can be used to quickly identify the interaction between microbial species. In addition, we verified these interactions, which indicated that the inter-species interactions extracted by the iRF algorithm are effective.

II. Relevant Theories of Learning

Our work is mainly developed from three theories, which constructed the theoretical basis of our algorithm from two aspects of computational theory and biological significance. The specific theory is described as follows, in which the first two theories describe the source of our algorithm, and the third theory introduces the fusion phylogenetic information theory.

A. Random Forest Theory

The Random Forest (RF)[3] algorithm is a non-parametric integrated learning tool based on tree structure. It combines the idea of adaptive nearest neighbors with Bagging[4] ideas to make it highly adaptive. The "grouping characteristics"[5] of the tree structure enable RF to better process and interpret the correlation and interaction between features. In addition, RF can also select and order variables through variable importance metrics. Due to this nature of random forests, it is well suited for processing biological information data[6,7].

B. Random Intersection Tree

The Random Intersection Tree (RIT)[8] algorithm is a new method for finding high-order relationships among variables in high-dimensional and large-scale data sets. It starts with a maximum interaction that contains all the variables, determines whether a variable appears in a randomly selected observation of a target category, and gradually deletes the variable. In general, the stronger the interaction between variables, the greater the probability of being retained.

C. Fusion Phylogenetic Distance

Inferring the interactions between microbial species requires an understanding of where each species comes from. The Phylogenetic tree is a directed graph of a tree-like structure that summarizes the kinship between different species, that is the evolutionary distance. A common phenomenon is closely related species in phylogeny, and their functions are usually similar[9]. The non-independence of this function or distribution of traits is defined as phylogenetic autocorrelation, commonly referred to as phylogenetic signals[10]. Valuable information such as phylogenetic-related patterns and variability in life evolution processes can be obtained by studying phylogenetic signals. In microbial community research, phylogenetic information can be used to discover the distribution characteristics of microbial species and some associations between species.

III. Enabling Technological Advances

The main research work of this paper is the discovery and prediction of the interaction between microbial species in human intestinal microbial ecosystem[11].

A. Iterative Random Forest Algorithm

Iterative Random Forest (iRF)[12] is an interaction discovery algorithm that relies on increasing feature weighted RF, and then one by one to perform soft-dimension dimension of feature space and extract stable decision path simultaneously. It incorporates the Random Intersection Tree Algorithm (RIT) to resolve the fitted random forest (RF) and in the process identifies feature combinations that are universal in the RF decision path.

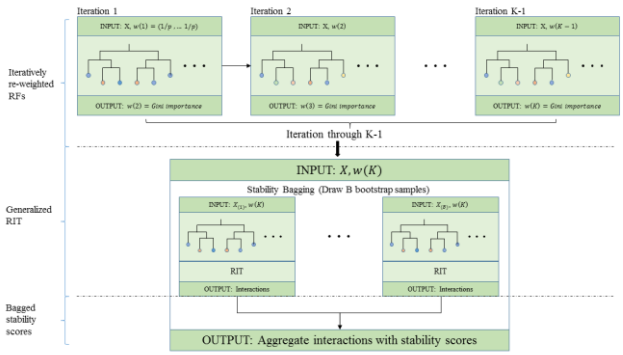


Figure 1: Diagram of the iRF Algorithm

B. Calculating Phylogenetic Distance

The integration of phylogenetic information in the discovery of interactions between microbial species makes the discovered interactions more biologically significant. Patristic Distance[13] is the sum of the lengths of the branches between two nodes in the phylogenetic tree. As shown in Figure 2, assuming that the phylogenetic tree is a rooted tree, and taking node A and node D as examples, the Patristic distances of v_A and v_D are defined as:

$$d_{AD}^{Tree} = c_{AA} + c_{DD} - 2c_{AD} \quad (1)$$

Where c_{AA} represents the path length of node A to ROOT, c_{DD} represents the path length of node D to ROOT, and c_{AD} represents the path length of node A and the nearest common ancestor (MRCA) of node D to ROOT.

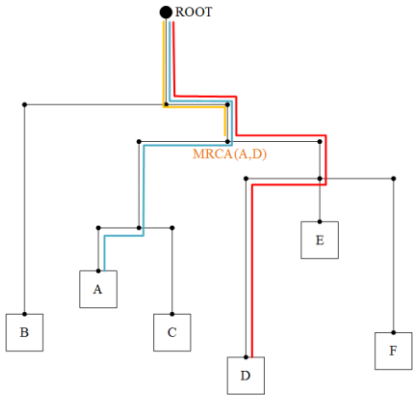


Figure 2: Diagram of Patristic Distance in phylogenetic tree

IV. Real World Applications

A. Microorganisms and diseases

The microbial interaction prediction method based on the iterative random forest (iRF) algorithm integrates the microbial phylogenetic distance information, making the results more biologically significant. This method puts the microbial interaction extraction into a supervised learning framework, which ensures the reliability of the relationship extraction process and the stability of the results. Integrating microbial phylogenetic distance data ensures that the discovered interactions have real biological significance. Finally, the microbial metabolic network model was used to verify the interaction relationship from the metabolic point of view, which provided a candidate set for the biological experiment design of microbial co-culture.

Understand the interactions between microbial species and explore specific types of relationships, so as to deepen the understanding of microbial community ecosystems, improve the ability to utilize and transform microbial communities, and provide new inspiration for ecological restoration, health monitoring and disease treatment and drug development.

B. Other complicated issues

Identifying complex interactions can be applied not only to the study of disease-related problems, but also to other complex issues such as teaching effectiveness analysis, resource allocation, and relationship coordination. The analysis process used in this paper has two potential assumptions: First, all microorganisms coexist in a resource-poor environment, and they need to make full use of various means to survive. Second, the types of microorganisms are different. They have different niches in complex ecosystems, so their own life activities are also different. Therefore, our method also has important reference significance in the extraction and research of other complex relationship problems.

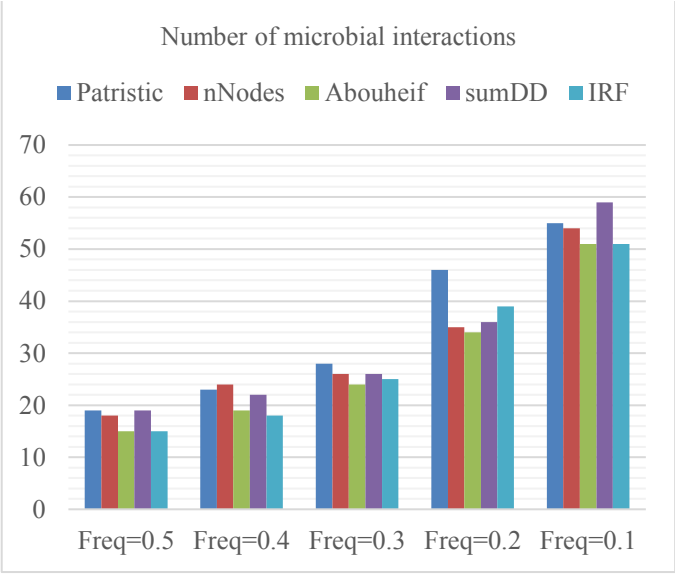
V. Evidence of Potential Impacts

The data used herein was derived from the human cirrhosis intestinal microbiological dataset[11], which included 118 cirrhosis patient sampling data and 114 healthy person sampling data, and 167 microbial species. Then we build a phylogenetic distance matrix based on the phylogenetic tree and fuse it with the abundance data matrix.

The iRF algorithm inherits the relevant parameters from its two basic algorithms (RF and RIT). We know that the predictive performance of RF is highly resistant to the choice of parameters. Therefore, we use the default parameters of the randomForest package in the R language. For the RIT algorithm, we create $M = 500$ cross-trees with depths of $D = 5$ and $n_{child} = 2$. Experiments have shown that the prediction accuracy of the iRF and the identified interactions are fairly robust to these parameters.

A. Analysis of results

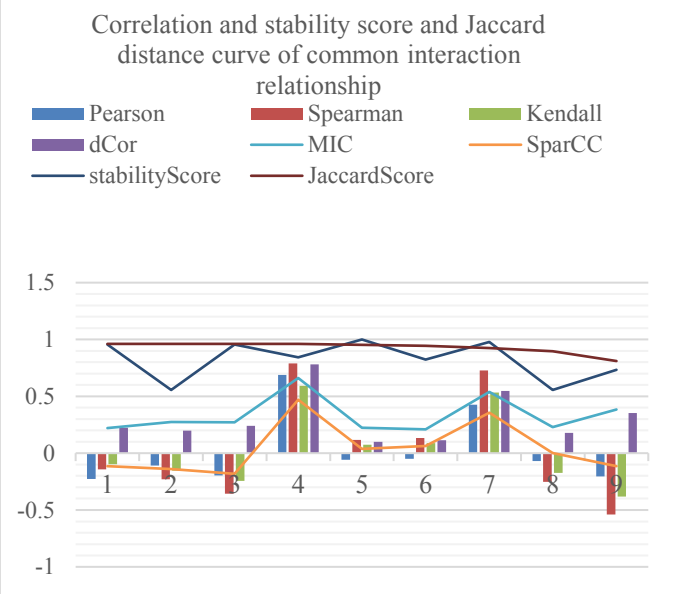
We analyzed the difference in the interaction between the iterative random forest algorithm that integrated the four phylogenetic distances and the original iterative random forest algorithm



Experiments show that the Phy-iRF algorithm can find more inter-species interactions than the iRF algorithm. That is, the Phy-iRF algorithm has found 19 interactions, and the iRF algorithm has found 15 interactions, 9 of which are the same relationship. In addition, we also reviewed the relevant data according to the PubMed literature resource database, which proves that these relationships do exist in the scientific literature.

B. *Statistical and biological analysis*

In order to verify the statistical significance of the discovered inter-species interactions, we compared these interactions with the symbiosis patterns of the species. We calculated the Jaccard distance, Pearson correlation coefficient, Spearman correlation coefficient, Kendall correlation coefficient, dCor correlation coefficient, MIC correlation coefficient, SparCC correlation coefficient between these species relationships, and compared these correlation coefficients with the stability scores of the interaction relationship.



In addition, we measured the correlation between the co-occurrence score of the inter-specific relationship and the metabolic interaction score based on the microbial metabolic data obtained by the KEGG database[14]. Here, we analyzed the metabolic competition[15] and metabolic cooperation[16] relationship of the microbial paired model, and the correlation between them metabolic interactions and the co-occurrence of microbial species.

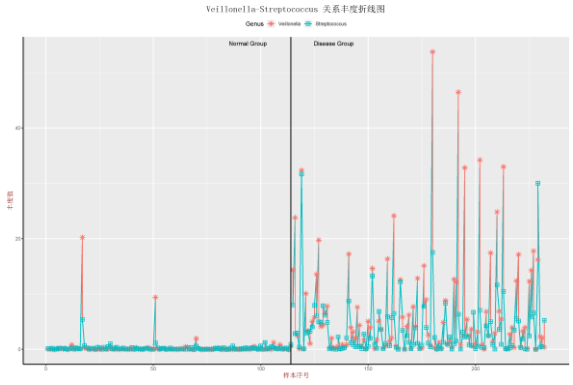


Figure 3: A line diagram of the abundance relationship between Veillonella and Streptococcus in the sample (Spearman = 0.79)

VI. SUMMARY

In this paper, the iterative random forest algorithm with integrated phylogenetic distance and the original random forest algorithm are used to mine high-order interactions on cirrhotic disease microbial dataset. Experiments show that the integrated phylogenetic distance can indeed find more microbial interactions, and the correlation calculations of these interactions show that both algorithms can identify a combination of relationships with obvious correlation. Based on the perspective of species metabolism, the metabolic scores between microbial species were calculated, and the causes and significance of microbial species relationship formation were analyzed from two aspects: resource constraints and ecological roles.

For other combinations of relationships with resource constraints and role differences, we can also use this idea to improve our understanding and analysis of such complex problems, helping us to clarify complex interactions and ultimately solve problems.

REFERENCES

[1] Thomas S, Izard J, Walsh E, et al. The Host Microbiome Regulates and Maintains Human Health: A Primer and Perspective for Non-Microbiologists.[J]. Cancer Research, 2017, 77(8): 1783-1812.

[2] Klitgord N, Segre D. Environments that Induce Synthetic Microbial Ecosystems[J]. PLOS Computational Biology, 2010, 6(11).

[3] Breiman L: Random Forests. Mach Learn 2001, 45(1):5-32.

[4] Breiman L: Bagging predictors. Mach Learn 1996, 24(2):123-140.

[5] Gilhodes J, Zemmour C, Ajana S, Martinez A, Delord JP, Leconte E, Boher JM, Filleron T: Comparison of variable selection methods for high-dimensional survival data with competing events. Comput Biol Med 2017, 91:159-167.

[6] Chen C, Schwender H, Keith J M, et al. Methods for Identifying SNP Interactions: A Review on Variations of Logic Regression, Random Forest and Bayesian Logistic Regression[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(6): 1580-1591.

[7] Chen X, Ishwaran H. Random forests for genomic data analysis.[J]. Genomics, 2012, 99(6): 323-329.

[8] Shah RD, Meinshausen N: Random Intersection Trees. J Mach Learn Res 2014, 15:629-654.

[9] Keck F, Rimet F, Bouchez A, Franc A: phylosignal: an R package to measure, test, and explore the phylogenetic signal. Ecol Evol 2016, 6(9):2774-2780.

[10] Penny D: The comparative method in evolutionary biology. Journal of Classification 1992, 9(1):169-172.

[11] Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L et al: Alterations of the human gut microbiome in liver cirrhosis. Nature 2014, 513(7516):59-64.

[12] Basu S, Kumbier K, Brown JB, Yu B: Iterative random forests to discover predictive and stable high-order interactions. Proc Natl Acad Sci U S A 2018, 115(8):1943-1948.

[13] Patristic Distance. In: Encyclopedia of Genetics, Genomics, Proteomics and Informatics. Springer, Dordrecht:Springer Netherlands;2008:1454-1454.

[14] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012, 40(Database issue):D109-114.

[15] Kreimer A, Doronfaigenboim A, Borenstein E, et al. NetCmpt: a network-based tool for calculating the metabolic competition between bacterial species[J]. Bioinformatics, 2012, 28(16): 2195-2197.

[16] Levy R, Carr R, Kreimer A, et al. NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation[J]. BMC Bioinformatics, 2015, 16(1): 164-164.