CROP: Clustering 16S rRNA for OTU Prediction

Xiaolin Hao[1], Rui Jiang[2], Ting Chen[1]*

[1] Molecular and Computational Biology, University of Southern California, University Park, Los Angeles, CA 90089

[2] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua Uni-versity, China, 100089

Email: tingchen@usc.edu

Abstract:

In metagenomics studies, hierarchical clustering is commonly used for clustering 16S rRNA to determine both the number of operational taxonomic units (OTUs) and their relative abundance in a sample at different similarity levels. Such results have been frequently used to profile the diversity of organisms in an environmental sample. However, this approach is highly sensitive to PCR/sequencing errors and can result in significant overestimation of organism richness.

To address the challenge, we propose an unsupervised Bayesian clustering method termed Clustering 16S rRNA for OTU Prediction (CROP). CROP uses Gaussian Mixture Model which considers the reads in data to be independently drawn from a mixture of Gaussian distributions, and further adopts a Markov Chain Monte Carlo approach to find the optimal clustering result.

By restricting the variance of Gaussian distributions, we can find clusters based on the natural organization of data at various dissimilarity levels without setting hard cut-off thresholds as required by hierarchical clustering methods. After applying CROP to several datasets, we demonstrate that CROP is robust against sequencing errors and produces more accurate and biologically meaningful results than conventional hierarchical clustering methods.

My research focuses on developing mathematical models and corresponding software to help biologists distinguishing species (i.e.: bacteria) and their relative abundance in given environmental samples using extracted DNA sequence data.

My software can accurately measure the complex organism community structure under various environmental conditions, which provides insights into a wide variety of important topics. For instance, bacteria distribution in the ocean is related to climate change and pollution level, while when inhabiting human skin, it reflects risk of certain diseases.

- My undergraduate work in network security area was elected as “Excellent Student Thesis” in Tsinghua University and was registered as a Chinese patent.
- After starting my research in bioinformatics area for only two years, I have published a journal paper which gained wide academic attention.

In 2010, I interned at Lili English as a summer camp coordinator. Our team designed and organized the company’s first U.S. summer camp program. This avant-garde program was proved successful and made a net profit of $200,000. We had to double the program’s capacity to meet its increasing demands afterwards.

During my one-month internship in the Laboratory of Pattern Recognition, Chinese Academy of Sciences as a software developer, I was responsible for designing a user interface to integrate research results of four different groups.

Due to lack of communication, those groups have their own coding styles which impede efficient integration. Therefore, I have to let them standardize their codes. However, as a new member and at a much younger age, I don’t have the required authority.

In order to fulfill my job on time, I carefully read their papers to get familiar with their work and chit chatting with them to gain trust and welcome.

After one week, I built up a friendly relationship with them and successfully held a large group discussion meeting, during which I demonstrated the importance and advantages of standardizing codes and got their agreement. Meanwhile, my deep understanding about programming and their research gained their respect.

Later, through lively and frequent discussions, we worked as a team to revise codes and finished the user interface on time. Since it is highly compatible and extendable due to its standardized coding manner, this work is still being used by the laboratory to exhibit their newest research results.

McKinsey version:

During my one-month internship in the Laboratory of Pattern Recognition, Chinese Academy of Sciences as a software developer, I was responsible for designing a user interface to integrate research results of four different groups.

Due to lack of communication, those groups have their own coding styles which impede efficient integration. Therefore, I have to let them standardize their codes. However, as a new member and at a much younger age, I don’t have the required authority.

In order to fulfill my job on time, I carefully read their papers to get familiar with their work

and chit chatting with them to gain trust and welcome. After one week, I built up a friendly relationship with them and successfully held a large group discussion meeting, during which I demonstrated the importance and advantages of standardizing codes and got their agreement. Meanwhile, my deep understanding about programming and their research gained their respect.

Later, through lively and frequent discussions, we worked as a team to revise codes and finished the user interface on time. Since it is highly compatible and extendable due to its standardized coding manner, this work is still being used by the laboratory to exhibit their newest research results.