Qingpeng Zhang
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
+1 (678) 901 9911
qingpeng@msu.edu

Department of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093

February 02, 2015

Dear Dr. Knight,

I am excited to apply for the postdoc position in your group. I am strongly interested in developing and applying computational methods to guide large scale efforts of using sequencing technologies as a tool to answer biological questions. I have more than ten years of experience in bioinformatics and have been working with metagenomic data as part of my PhD. I specialize in developing scalable tools for processing and analyzing complex shotgun metagenomes. Currently, I am a graduate student in Dr. C. Titus Brown's group at Michigan State University and expect to receive my doctorate this May.

In my years of PhD research working on metagenomic data, I have come across many great papers from your group, from the earlier one about the clustering metric UniFrac to the recent review on personal human microbiome. The tools developed in your group and your insights to the future of microbiome research benefit my work significantly. For example, QIIME is integrated into the pipeline of my recent work on efficient diversity analysis, using biom format to store the IGS tables. The package scikit-bio is also heavily used to facilitate the analysis. I even happened to come across a problem and submitted a bug ticket to the developers.

Based on my previous work on an efficient k-mer counting approach and digital normalization, recently I am working on a novel method to enable efficient and scalable microbial diversity analysis without the requirement of assembly and reference sequences. A novel concept - IGS (informative genomic segment) is proposed to represent the unique information in a metagenomics data set. The abundance of IGSs in different samples can be retrieved by mapping the reads to the de Bruijn graph database built from separate samples. After we get the samples-by-IGS matrix, existing package like QIIME can be used to do different kinds of diversity analysis. I have evaluated this method on multiple metagenomes from a variety of environments (e.g., human microbiome, soil in collaboration with James Tiedje, ballast water viromes in collaboration with Joan Rose). I believe that the IGSs can be used as a cornerstone for diversity analysis of whole shotgun metagenomics data sets just like OTU for 16S rRNA data sets. The new assembly-free, reference-free framework to do diversity analysis based on IGS will be a beneficial complement to the QIIME package, which is mainly focused on analysis based on OTU. Also, there is more work to do to tackle the computational challenge to accommodate the much larger and more complicated samples-by-IGS matrix, including, but not limited to, adding new feature to QIIME, biom and/or sickit-bio. Going forward, I plan to integrate those methods I have been working on to solve more problems in metagenomics , like binning approaches, functional annotation, and phylogenetic analysis, with the adoption of more machine learning methods, data structures or algorithms while necessary.

About the future of microbiome science I come up with many ideas previously. But I read the news about 2016 budget submitted by the president today and it can not be better to cite several numbers from it to emphasize my points. Firstly $200 million will be allotted to the precision medicine initiative. According to a director in the White House, "precision medicine" is a term for " tailoring treatments to an individual's genetic makeup, microbiome, and other factors" So obviously microbiome science research will play an important role in facilitating precision medicine or personalized medicine. Also $130 million will be allotted to an effort to create a research cohort consisting 1-million-volunteers. This reminds me of Human Microbiome Project, Human Gut project and the biobank effort in San Diego. Another interesting number is

that the biggest bump within DOE goes to their Advanced Scientific Computing Research program, which would see an increase of 14.8%, to $621 million. This may not be directly related to microbiome science, but it emphasizes the crucial role computing will play to advance scientific research. To analyze and interpret the big biological data like that from the 1-million-volunteers cohort and make sense of the variations and patterns inside, powerful, efficient, scalable computational methods will be highly demanded.

As discussed above, this is a great time to work on microbiome science with the help of efficient and powerful computing. I am greatly enthusiastic about the possibility of working in your group. With your supervision and the collaboration with members in your group , I believe I can contributing my skills and knowledge to facilitate the microbiome science research, and make a difference in a broader field like precision medicine subsequently. I have attached my CV and research statement for your review. I would greatly appreciate the opportunity to talk to you more about the position. Please do not hesitate to contact me with any questions.

Sincerely,

Qingpeng Zhang

Attached: curriculum vit, research statement