Systems Metagenomics: applying systems biology thinking to human microbiome analysis

Golestan Sally Radwan[1] and Hugh Shanahan[2]

[1,2] Royal Holloway University of London, Egham TW20 0EX, UK
golestan.radwan.2016@live.rhul.ac.uk

Abstract. Metagenomics is the science of analysing the structure and function of DNA samples taken from the environment (eg soil, human gut) as opposed to a single organism. So far, researchers have used traditional genomics tools and pipelines applied to metagenomics analysis such as species identification, sequence alignment and assembly. In addition to being computationally expensive, these approaches lack an emphasis on the functional profile of the sample regardless of species diversity, and how it changes under different conditions. It also ignores unculturable species and genes undergoing horizontal transfer.

We propose a new pipeline based on taking a "systems" approach to metagenomics analysis, in this case to analyse human gut microbiome data. Instead of identifying existing species, we examine a sample as a self-contained, open system with a distinct functional profile. The pipeline was used to analyse data from an experiment performed on the gut microbiomes of lean, obese and overweight twins. Previous analysis of this data only focused on taxonomic binning. Using our systems metagenomics approach, our analysis found two very different functional profiles for lean and obese twins, with obese ones being distinctly more diverse. There are also interesting differences in metabolic pathways which could indicate specific driving forces for obesity.

Keywords: systems metagenomics, population, human microbiome, stress response, obesity, function, protein families.

1 Introduction

A key goal of metagenomics is to measure the diversity of a microbial sample and hence estimate the effects of certain stresses on the organisms present. This is traditionally achieved through techniques of taxonomic binning or phylogenetic classification and requires several steps of pre-processing and assembly. In addition to being computationally expensive, these approaches lack an emphasis on the functional profile of the sample regardless of species diversity, and how it changes under different conditions. It also ignores unculturable species and genes undergoing horizontal gene transfer [1].

We propose a new pipeline for analysing short-read data which focuses on the most dominant functions of the sample, treating it as a system of genes/proteins rather than a set of individual species. Specifically, we use a k-mer approach to identify overrepresented protein family motifs in the raw short-read data [2]. Via a series of filtering and statistical methods, we exclude weak hits and false positives, then use GO-term mapping to infer the functions of the remaining motifs. By comparing two similar datasets under changing conditions we can thereby hypothesise which protein families most influence the observed function of the microbial community under study. Our method dispenses with assembly and global multiple sequence alignment, instead only performing a local alignment on chosen sequences after multiple stages of filtering and analysis, thereby minimising computational cost. The entire pipeline can run on a standard laptop with 8 GB of memory in under 3 hours per dataset.

The pipeline was used to analyse data from an experiment performed on the gut microbiomes of lean, obese and overweight twins. Previous analysis of this data focused on taxonomic binning [3]. Our analysis found two very different functional profiles for lean and obese twins, with obese ones being distinctly more diverse. The obese patient data showed almost 3 times as many prominent functional groupings based on GO-term analysis of molecular function and biological process as lean or overweight ones.

This approach can also be used to identify overrepresented novel protein families in these samples which may play a role in the gut microbiome. We have identified around 185 novel candidates which warrant further experimental research.

2 Pipeline Overview

The new pipeline consists of the following steps:

Translation: After stripping metadata from the short read files the nucleic acid sequence is directly translated into a protein sequence. There is only 1/6 chance that the sequence will be in the correct read frame and hence much data will be lost but is compensated for in the model for the data.
Frequency Vectors: Frequency vectors are computed for each k-mer (each k-mer is a protein sequence fragment). A k-mer length of 6 was chosen following several experiments including looking up protein family motifs in databases such as PRINTS [4] as well as synthetic genomes. Henceforth, these 6-mers are referred to as ‘submotifs’. A rolling window technique was used to extract the submotifs one at a time. Any submotif with an unidentified base in it was discarded. Submotifs were then counted and written into frequency vectors along with their respective occurrences.
Null Model calculation: A statistical model was constructed to help quantify whether any given submotif would have occurred by chance with a specific frequency or whether its frequency might indicate actual over-representation of a particular submotif. This null model choice was built on the work outlined in [5].
Read extraction and MSA: All submotifs that fall below the significance threshold of 1.0 are eliminated. The remaining submotifs are sorted based on the value of the log-odds ratio (i.e most over-represented first). Then, a search is conducted on the amino acid reads to extract all short reads that contain this particular submotif. Each

set of reads pertaining to one submotif are then passed onto MUSCLE [6] to construct a multiple sequence alignment (MSA).

pHMM construction and analysis: The resulting MSA for each submotif is then used as a basis to construct a profile Hidden Markov Model (pHMM) using the hmm-build tool from the HMMER suite [7]. The resulting model is searched against a protein family database, in this case UniProt [8], to produce possible hits against known protein families.

Family identification and novel detection: The resulting hits are sorted based on E-values and a cutoff of E<=0.01 was used as a threshold for possible matches.

GO term: GO terms for each significant UniProt hit are identified and a frequency table tabulated. For the highest frequency terms a comparison is made between the two different environments examined in this paper.

Dataset: for this paper, the dataset from [2] was used. This was a study conducted on lean, obese and overweight twins with a sample size of 46. The raw short reads in FASTQ format were downloaded from the Sequence Reads Archive through the European Bioinformatics Institute's Metagenomics portal [9].

3 Results

Of the top 40 GO terms found in both datasets, we found 31 that were common to both sets and 9 that were unique to each set. Figure 1 shows those terms plotted as a function of the difference in their frequencies (i.e. Obese – Lean frequency) and are plotted as a strictly decreasing function.

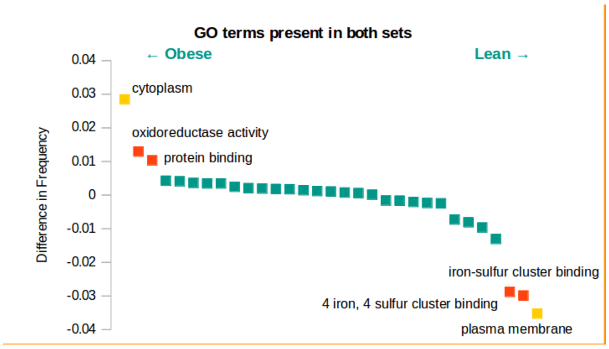

Figure 1 Relative frequency of GO terms found in both data sets. Yellow data-points indicate Cellular Components, orange ones indicate Molecular Functions, and green ones indicate GO terms occurring with similar frequency in both sets.

4 Discussion & future work

Our new approach obviates the need to perform assembly on a noisy data set and no clear set of reference genomes [10]. In this study we found a greater functional diversity in the obese rather than lean data sets. In addition, the obese data set does not exhibit any abundant functional groupings with a stress response, contrary to the findings in [11]. In the obese dataset, tRNA synthesis is playing an important role which is consistent with the findings of [12]. Other terms are related to electron transport and amino acid biosynthesis. In the lean data set we see terms related to biosynthesis.

These preliminary results are promising though much more work needs to be done. While the analysis of abundant motifs is computationally efficient the matching of those motifs to known sequence data is very intensive. The present search mechanism is based on building profile HMM's on the short reads and then looking for matches in UniProt. A corresponding search based on previously known profile HMM's and then querying those against the collated motifs has to be completed. Finally, in this present analysis we have focused on identifying known functional groupings; on the other hand the above approach could also be used to identify abundant motifs that have not been observed previously. Given the wide variety of metagenomic samples this is a potentially fruitful method for identifying new protein families.

References

1. Breitwieser, Florian P., Jennifer Lu, and Steven L. Salzberg. "A review of methods and databases for metagenomic classification and assembly." *Briefings in bioinformatics* (2017).
2. Turnbaugh, Peter J., et al. "A core gut microbiome in obese and lean twins." *nature* 457.7228 (2009): 480.
3. Simpson, Jared T. "Exploring genome characteristics and sequence quality without a reference." *Bioinformatics* 30.9 (2014): 1228-1235.
4. Attwood, Terri K. "The PRINTS database: a resource for identification of protein families.". Briefings in bioinformatics 3.3 (2002): 252-263.
5. Masatoshi Nei and Wen-Hsiung Li. "Mathematical model for studying genetic variation in terms of restriction endonucleases (molecular evolution/mitochondrial DNA/nucleotide diversity)". Genetics, 76(10):5269–5273, 1979.
6. Edgar, Robert C. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic acids research 32.5 (2004): 1792-1797.
7. Finn, Robert D., Jody Clements, and Sean R. Eddy. "HMMER web server: interactive sequence similarity searching." Nucleic acids research 39.suppl_2 (2011): W29-W37.
8. Boutet, Emmanuel, et al. "Uniprotkb/swiss-prot." Plant bioinformatics. Humana Press, 2007. 89-112.
9. https://www.ebi.ac.uk/metagenomics/
10. Simpson, Jared T. "Exploring genome characteristics and sequence quality without a reference." Bioinformatics 30.9 (2014): 1228-1235.
11. Sanz, Yolanda, Arlette Santacruz, and Paola Gauffin. "Gut microbiota in obesity and metabolic disorders." *Proceedings of the Nutrition Society* 69.3 (2010): 434-441.
12. Isokpehi, Raphael D., et al. "Genomic evidence for bacterial determinants influencing obesity development." International journal of environmental research and public health 14.4 (2017): 345.