Supplement

Critical Assessment of Massive Data Analysis (CAMDA)

One of the CAMDA 2018 Challenges was *The MetaSUB Forensics Challenge* where hundreds of novel city microbiome profiles from MetSUB have been provided. The challenge was to construct urban microbiome fingerprints and identify the geographical origin of mystery samples. Some of the mystery samples were from already known locations (classification problem) while other were from the new, unknown location (prediction problem). The prediction problem is still not solved and inside CAMDA and MetaSUB communities we are actively investigating different paths and the required data types and sizes.

In terms of classification problem results from CAMDA 2018 have shown, that already existing approaches can provide accuracy of 90% or more. In fact, all conceptual groups can classify that well. On next pages the descriptions of presented at CAMDA 2018 methods are provided.

Taxonomy/species centric: *Harris 2019, Walker 2019, and Ryan 2019*

Functional approach: *Casimiro-Soriguer 2019 and Zhu 2019*

K-mer based approach: *Kawulok 2019*

Massive metagenomic data analysis using abundance-based machine learning

Zachary N. Harris, Eliza Dhungel, Matthew Mosior & Tae-Hyuk Ahn

Biology Direct 2019 **14**:12

With over half of the world's population living in urban areas, mass-transit systems like subways and buses represent some of the most shared environments in the world. While effectively transporting millions of people around urban areas, these transit systems also serve as a home to a rich community of microbes that share DNA and RNA with their human passengers. These unseen genetic interactions play important roles in public health and disease outbreak, yet little is known about mass-transit biomes. Fortunately, metagenomics, defined as the direct analysis of genetic material found in an environment, offers an opportunity to better understand the microbial communities present in our mass-transit systems.

To analyse the genetic interactions and community compositions of mass-transit biomes, The MetaSUB International Consortium has collected whole-genome sequencing (WGS) data from multiple cities across the world. Further, the Conference on Critical Assessment of Massive Data Analysis (CAMDA) has released their Metagenomics Forensic Challenge - predict the geographic origin of a metagenomic sample when no reference samples from that location are known. For this challenge, CAMDA has provided access to the MetaSUB WGS data.

The WGS (also called as shotgun metagenomics) data can be analysed using several different approaches, but the methodological approaches can be divided into two categories: read-based and assembly-based. Read-based metagenomics analysis is useful for quantitative community profiling and identification of organisms especially if relevant references are available. MetaPhlAn2 identifies clade-specific marker genes for evidence of the associated clade presence. This allows for rapid assignment relative to a small database as compared to a full database including many whole genomes and fast mapping aligner, Bowtie2. Assembly-based workflows attempt to assemble the reads from one or more samples, group (bin) the contigs from these samples into genomes, then analyse the genes and contigs. Megahit, MetaSPAdes, and IDBA-UD are the most widely used k-mer based assemblers for high-throughput MPS metagenomic data. Most metagenomic classification tools match reads or assembled contigs against a database of microbial genomes to identify the taxon of each sequence. Several strain-level resolution taxonomic profilers were recently developed.

To distinguish the metagenomic profiling among different cities and also predict unknown samples precisely based on the profiling, two different approaches are proposed using machine learning techniques; one is a read-based taxonomy profiling of each sample and prediction method, and the other is a reduced representation assembly-based method. Among various machine learning techniques tested, the random forest technique showed promising results as a suitable classifier for both approaches. Random forest models developed from read-based taxonomic profiling could achieve an accuracy of 91% with 95% confidence interval between 80 and 93%. The assembly-based random forest model prediction also reached 90% accuracy. However, both models achieved roughly the same accuracy on the testing test, whereby they both failed to predict the most abundant label.

Our results suggest that both read-based and assembly-based approaches are powerful tools for the analysis of metagenomics data. Moreover, our results suggest that reduced representation assembly-based methods are able to simultaneous provide high-accuracy prediction on available data. Overall, we show that metagenomic samples can be traced back to their location with careful generation of features from the composition of microbes and utilising existing machine learning algorithms. Proposed approaches show high accuracy of

prediction but require careful inspection before making any decisions due to sample noise or complexity.

Recently CAMDA has provided access to the 2019 MetaSUB WGS data. Previous work has shown that the taxonomic information derived from this data can be used as features in machine learning models to predict an unlabeled sample's city of origin. Thus, we hypothesise that geographic locations in soil and transit biomes will be distinguishable given taxonomic composition and abundance profiles. Further, we assert that the classification of "unknown" samples with no reference samples is possible given additional geographic inputs. Based upon prior research, the random forest machine learning model was implemented for both continent and city sample classification. Best parameters for the model were determined before implementation using random and grid search. Accuracy scores for continent and city prediction were 98% and 95% respectively. The continent of Europe was by far the hardest to classify along with the various European cities being the hardest to properly classify as well. This is perhaps due to the large number of samples received from European cities along with other factors such as high prevalence and use of public transportation throughout Europe.

Identification of city specific important bacterial signature for the MetaSUB CAMDA challenge microbiome data

A. R. Walker and S. Datta

Biology Direct 2019 **14**:11

Whole genome sequencing of metagenomic samples collected from subways station in several countries has been used to train and predict sample provenance. As part of CAMDA, in our second year follow up of the "MetaSUB Forensic Challenge" we have implemented appropriate machine learning techniques and statistical inferences on this massive dataset in order to find distinguishing microbiome signatures associated with different city subway swab samples. As these analyses were done within a time constrained manner and also only part of the entire data was released for the analysis through the CAMDA challenge initiative, we might not have obtained the maximal information from the data. However, it was evident that lot more results are obtainable from the full data and more in-depth analysis. Room for improvements may involve better sequencing protocols, choosing better open reference OTU picking algorithms, and better training and predictions stages of the data. Up to this point this work has produced two publications and a third ongoing draft all with better results. Here we provide a brief summary of the results that we obtained so far, in our two previous publications.

In our first work[136] the most critical issues were the disparity in the quality of the sequencing data, DNA amplification protocols, and number of samples per city. This led to a highly unbalanced design and analysis, which required a bootstrap analysis in order to verify the how consistent the mean species diversity was across cities. This analysis showed opposing results when comparing both values of the "q" parameter in the species diversity equation. The classification error in this work ranged between 1-34% the city and taxonomic rank ("order", "family", "genus") and was highly influenced by the number of samples in the city and the quality of the DNA sequences. The network analysis showed interesting similarities in the nodes-connections for both east cities (Boston and New York) when compared with Sacramento (West coast) across all taxonomic ranks.

In our second publication[139] the unsupervised PCA analysis has shown interesting clustering of samples and a large proportion (>65%) of the variability in the data explained by the first three PC components. We have tried to predict the unknown cities on the basis of the classifier models built on the bacterial signatures from the known cities. Random Forest Support Vector machines were both decent in terms of the classification accuracies. The percentage accuracy, when predicting city provenance, ranged between 70%-90% depending on the dataset. Additionally, relative abundance analysis of the compositional taxonomic data revealed that some species are specific to some regions and played a very important role during the prediction of the unknown provenance city microbial samples. Upon improvement of the sequencing data in the 2018 challenge, the taxonomical classification of the data yielded a considerable larger number of OTUs reaching the taxonomic rank "species". This was definitely an improvement over the data from the 2017 challenge, which ultimately was reflected in overall better results. As stated in our 2019 paper, we firmly believe that in this work there is still much more room for improvement and as an example, implementing a variable selection step considering not only species as taxonomic ranks but also family and order could make a significant difference in the prediction results. Finally, as a closing remark, the results from our work showed an effective method to process, optimise, and classify the metagenomic samples by origin, but still there are scopes to improve upon the results by carefully tuning all possible sources of errors.

Application of machine learning techniques for creating urban microbial fingerprints

Feargal Joseph Ryan

Biology Direct 2019 **14**:13

In the CAMDA challenge I built a random forest classifier which was capable of identifying city of origin of a MetaSUB metagenomic sample with a high degree of accuracy[141]. The basic principles which led to the approach that was finally used were that it should be capable of assigning as many sequences reads as possible to a useable/informative count feature and it should be easily reproducible by other researchers. This was achieved through the use of the Kaiju metagenomic classifier which performs a 6-frame translation of each read and then classifies this against the nr database[202]. Such an approach is beneficial as sequence may be highly conserved at the amino acid level but divergent at the nucleotide level. The main limitation of this approach is the computational requirements, specifically the very high amount of RAM required for indexing the nr database for use with Kaiju. From this, the count of reads classified to each possible taxonomic rank for any domain of life were included as a count feature. Initially count features were filtered simply for presence in a certain percentage of samples in order to remove counts with majority zeroes. Next, t-Distributed Stochastic Neighbour Embedding (t-SNE) performed on the resulting count matrix showed a startling level of clustering by city of origin, similarly random forest was also capable of accurately identifying the city of origin in many cases. While this approach was largely successful it highlighted some key limitations in sample origin prediction with current resources. Most importantly, even using amino acid translation to the largest sequence database available, many samples were still primarily unclassified. Samples may contain organisms which perfectly predict a city of origin but that is of no use if researchers are unable to identify that organism. Recent work in the human microbiome field has highlighted the utility of in-silico approaches alone to mine for uncultured or unrepresented organisms that may be useful in constructing a database of representatives for sample origin prediction[180]. Furthermore incorporation of community diversity statistics in a database independent manner may aid prediction[203]. Finally, the impact of temperature, humidity and more generally season/climate on sample prediction has yet to be assessed and is potentially a large confounder.

Antibiotic resistance and metabolic profiles as functional biomarkers that accurately predict the geographic origin of city metagenomics samples

Carlos S. Casimiro-Soriguer, Carlos Loucera, Javier Perez Florido, Daniel López-López and Joaquin Dopazo

Biology Direct 2019 **14**:15

The availability of hundreds of city microbiome whole genome shotguns (WGS), is fostering the development of increasingly accurate predictors of the origin of a sample based on its microbiota composition. Currently, bacterial abundance profiles are used as features for the classification. However, bacterial genomes carry a wealth of information in their genomes regarding functionality, virulence, antibiotic resistance, etc. In fact, the value of existing functional profiling tools, which provide an extra perspective on biological interpretation, remains unexplored for classification purposes of environmental microbiota and, more specifically, urban metagenomes.

Here we study different transformations of the conventional bacterial strain or gene abundance profiles to functional profiles that account for bacterial metabolism and other interesting bacterial functionalities such as antibiotic resistance. These profiles are used as features for city classification in a machine learning algorithm that allows the extraction of the most relevant features for the classification.

From the sequencing reads we followed the MOCAT2 pipeline that maps the sequences to the corresponding genes of the different strains and carry out the functional annotation process using different databases (here KEGG -metabolism- and CARD -antibiotic resistance- modules). Thus, we generated KEGG and CARD functional profiles for each sample. The training set samples are labeled according to their city of origin. Then, to classify each sample into one of the known cities a classification pipeline was developed which mainly consists of: i) A base learner with decision trees, ii) An ensemble of base learners via Scalable Tree Boosting and, iii) A Bayesian optimisation framework for tuning the hyper parameters. Functional profiles can be used alone or combine them by means of a pipeline that can fuse different functional profiles by learning an approximation of the latent space by means of Canonical Correlation Analysis and then applying the machine learning pipeline already proposed.

The KEGG profiles classification accuracy is 0.73 and the CARD accuracy is 0.8, however, this accuracy increased to 0.9 by using the fusion pipeline, demonstrating that a more comprehensive functional description of the samples provides a better classification. The classification of new “problem samples” and “problem cities” was quite accurate as well.

Interestingly, this method provides not only an accurate classification but also provides full interpretability of the results in terms of interesting bacterial functionalities. The most relevant features were extracted from the classification pipeline from each run of the experiment, cross referencing the nested loop for the best set of hyperparameters and a final fit with all training data, by averaging the feature importance of each base learner of the ensemble. In the case of KEGG, a total of 44 features were found relevant for the classification. Just to mention an interesting interpretational result: three features related to antibiotic resistance classify Offa city in Nigeria, due to ABC-2 type transporter protein of *Staphylococcus aureus*, a pathogen of recognised higher incidence rates in sub Saharan Africa than those reported from developed countries. Moreover, two single-nucleotide polymorphisms prevalent in sub Saharan populations have recently demonstrated to be associated with susceptibility to *S. aureus* infection.

We demonstrate here that the use of functional profiles not only predict accurately the most likely origin of a sample but also to provide an interesting functional point of view of the biogeography of the microbiota. Interestingly, we show how cities can be classified based on the observed profile of metabolism or antibiotic resistances.

Fingerprinting cities: differentiating subway microbiome functionality

Chengsheng Zhu, Maximilian Miller, Nick Lusskin, Yannick Mahlich, Yanran Wang, Zishuo Zeng and Yana Bromberg

Biology Direct 2019 **14**:19

Accumulating evidence suggests that the human microbiome impacts individual and public health. City subway systems are human-dense environments, where passengers often exchange microbes. The MetaSUB project participants collected samples from subway surfaces in different cities and performed metagenomic sequencing. Previous studies focused on taxonomic composition of these microbiomes and no explicit functional analysis had been done till now. As a part of the 2018 CAMDA challenge, we functionally profiled the available ~400 subway metagenomes and built predictor for city origin. In cross-validation, our model reached 81% accuracy when only the top-ranked city assignment was considered and 95% accuracy if the second city was taken into account as well. Notably, this performance was only achievable if the similarity of distribution of cities in the training and testing sets was similar. To assure that our methods are applicable without such biased assumptions we balanced our training data to account for all represented cities equally well. After balancing, the performance of our method was slightly lower (76/94%, respectively, for one or two top ranked cities), but still consistently high. Here we attained an added benefit of independence of training set city representation. In testing, our unbalanced model thus reached (an over-estimated) performance of 90/97%, while our balanced model was at a more reliable 63/90% accuracy. While, by definition of our model, we were not able to predict the microbiome origins previously unseen, our balanced model correctly judged them to be NOT-from-training-cities over 80% of the time. Our function-based outlook on microbiomes also allowed us to note similarities between both regionally close (Ofa and Ilorin) and far-away (Boston and Porto, Lisbon and New York) cities. Curiously, we identified the depletion in mycobacterial functions as a signature of cities in New Zealand, while photosynthesis related functions fingerprinted New York, Porto and Tokyo.

Environmental metagenome classification for constructing a microbiome fingerprint

Jolanta Kawulok, Michal Kawulok, Sebastian Deorowicz

Biology Direct 2019 **14**:20

In our full paper, we focus on the MetaSUB Forensics Challenge (organized within the CAMDA 2018 conference) which consists in predicting the geographical origin of metagenomic samples (environmental classification). Most of the existing methods for environmental classification are based on taxonomic or functional classification. We demonstrate that it is not necessary to identify the organisms or their functions to perform effective environmental classification. Hence, we do not need large databases of annotated metagenomic reads (like the NCBI (nt) nucleotide database), which substantially decreases the amount of data we have to process. Furthermore, this makes it possible to exploit the organisms specific to each location, even if their genetic material is not included in the databases. In the paper, the microbiome fingerprint is defined as a set of DNA fragments (*k*-mers) derived from organisms living in a given place.

In our method, we exploit our CoMeta program and the Mash program which are applied to classify the extracted unknown metagenomes to a set of collections of known samples. We construct separate groups (G_i) of metagenomic reads for each city to compare the samples on the basis of their similarity, measured directly in the space of the metagenomic reads. Moreover, we use the CoMeta program to cluster the samples based on their mutual similarities, which allows us to identify several groups that have been derived from the same origin.

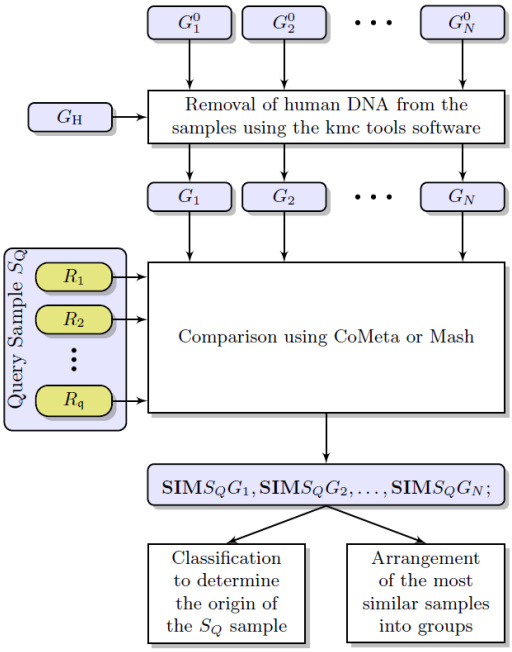

Figure 1 The processing pipeline for classifying metagenomic reads

A simplified diagram of our classification scheme is shown in Figure 1. At first, *N* groups (classes, samples for clustering, G_i), containing some reads (reference sequences for classification) are created and the reads from the query sample are compared with them. Before comparing, the human fragments (G_H) are removed from the groups using the KMC software. Each read R_i derived from a query sample (S_Q) is compared against each class using CoMeta or Mash. From the comparison, we obtain their mutual similarity values $\mathbf{SIM}S_QG_i$, which are used to determine the origin of an environmental sample or to arrange the most similar samples into groups.

The MetaSUB Challenge embraces three complementary independent test sets and a *primary dataset* (i.e., the reference set with all the metadata provided, including geographical origin of the data). To evaluate our method, we perform leave-one-out cross validation for the primary dataset. For the first test set, we classify the samples against the primary dataset to check whether they are classified correctly. The second and the third test sets come from cities that are not included in the primary dataset. Hence, we only present the similarities between them and the classes from the primary dataset. In addition, we show the mutual similarities between 16 samples in the third test set. These similarities are used to arrange the samples into groups.