Metagenomic Analysis of the Saliva Microbiome with Merlin

Pedro Barbosa[1], Oscar Dias[1], Joel P. Arrais[2], and Miguel Rocha[1]

[1] CEB/IBB, School of Engineering, University of Minho, Portugal
mrocha@di.uminho.pt
[2] Dep. Informatics Engineering / CISUC, University of Coimbra, Portugal
jpa@dei.uc.pt

Abstract. In recent years, metagenomics has demonstrated to play an essential role on the study of the microorganisms that live in microbial communities, particularly those who inhabit the human body. Several bioinformatics tools and pipelines have been developed for the analysis of these data, but they usually only address one topic: to identify the taxonomic composition or to address the metabolic functional profile. This work aimed to implement a computational framework able to answer the two questions simultaneously. Merlin, a previously released software aiming at the reconstruction of genome-scale metabolic models for single organisms, was extended to deal with metagenomics data. It has an user-friendly and intuitive interface, being suitable for those with limited bioinformatics skills. The performance of the tool was evaluated with samples from the Human Microbiome Project, particularly from saliva. Overall, the results show the same patterns reported before: while the pathways needed for microbial life remain relatively stable, the community composition varies extensively among individuals.

Keywords: Metagenomics, Annotation, Human microbiome.

1 Introduction

For most of the history of life, microorganisms were the only inhabitants on Earth, and they still keep dominating the planet in many aspects. Microbial life has also an important role in human health, agriculture and ecosystem functioning. For example, the human microbiome harbors over 100 times more genes than our genome [1] and has been linked to several diseases, such as obesity and inflammatory bowel disease [2]. Such discoveries were possible with the appearance of culture-independent methods, such as the 16S ribosomal rRNA or whole-metagenome shotgun (WMS) sequencing approaches. While the former primarily focuses on identifying the organisms that compose an environmental sample and their proportions, WMS extends the potential of metagenomics by allowing gene annotation and downstream metabolic analysis of microbial communities, either from assembled contigs or unassembled reads.

Along with whole-community screenings, bioinformatics challenges have arisen and several tools have been released to analyse WMS metagenomic samples at

J. Sáez-Rodríguez et al. (eds.), *8th International Conference on Practical Appl. of Comput. Biol. & Bioinform. (PACBB 2014)*, Advances in Intelligent Systems and Computing 294,
DOI: 10.1007/978-3-319-07581-5_23,

the taxonomic and functional levels. Community profiling is usually done by using extrinsic information from genome databases, but *unsupervised* approaches also exist, featuring the binning of the sequences based on intrisic features (e.g. GC composition, k-mer distribution or codon usage). Examples of such tools are LikelyBin [3] and CompostBin [4]. Homology-based classification relies on database searches, where the major strategy for taxon assignment is the selection of the best hits. However, this type of classification needs to be interpreted carefully, since the evolutionary distance between the DNA fragments and the hit is unknown. CARMA [5], MetaPhlAn [6] or MEGAN [7] are some of the similarity based tools, showing complementary features to improve the classification.

Regarding functional annotation and metabolic reconstruction, there are also plenty of choices. First, the user can choose to perform the analysis directly from the reads or based on the assembly. While the first can be more sensitive, as a greater number of sequences is classified, the second approach is suitable to overcome the bias of higher abundance of longer genes. In both cases, the strategy is to use a search engine (e.g BLAST [8], RAPsearch2 [9]) to scan the reads or genes predicted from the contigs against protein sequence databases, such as NCBI nr [10], SwissProt [11] or KEGG Orthology [12], or protein domain databases such as NCBI Conserved Domain Database (CDD) [13]. Pathway reconstruction relies in finding the most likely set of pathways in the metagenome, usually through a gene-pathway-centric view where the biochemical functions of the community members are treated as a whole. KEGG and SEED [14] are the common resources for analysing these broader functional units. Given the described methodologies, some standalone tools (again MEGAN, HUMAnN [15]) and web services (MG-RAST [16], CAMERA [17]) have been developed.

Indeed, many efforts have been done towards a proper analysis of environmental samples, but there is still a lack of choices to perform an integrative analysis of microbial communities at taxonomic and functional levels, simultaneously. Moreover, if the user is not interested in running a web service, using some of the available standalone programs can be a hurdle since they are usually command-line based and require libraries dependencies to be run.

In this context, the main goal of this project focused on developing an user-friendly tool capable of performing a taxonomy description, as well as a robust metabolic reconstruction of a microbial community. The work was done by adapting a previously developed software, originally designed to construct genome-scale metabolic models for single organisms, Merlin. For evaluation purposes, saliva samples from the Human Microbiome Project (HMP) were used.

2 Methods and Implementation

Merlin is an open-source application implemented in $Java^{TM}$ and was built on top of the AIBench (`http://www.aibench.org`) software development framework [18]. It utilizes a relational MySQL database to locally store the data and uses different Java libraries, such as NCBI Entrez Utilities Web Service Java Application Programming Interface (API) and KEGG Representational State

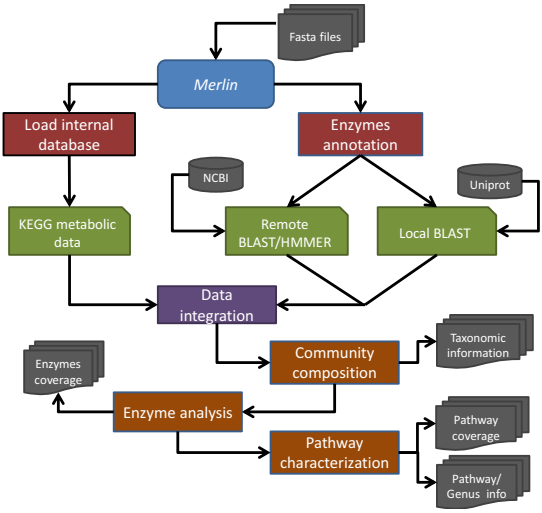


Fig. 1. Schematic representation of Merlin architecture for metagenomic analysis

Transfer (REST) API to access several web services. It requires a FASTA file of genes coding sequences as input in nucleotide or aminoacid format. The annotation of the samples is done via similarity searches using BLAST (either remote or local) and the metagenomics workflow is based on these results (Figure 1).

2.1 Taxonomic Analysis

The purpose of this operation is to assign a taxonomic label to each gene, as well as to describe the overall community composition. Thus, Merlin classifies each gene at the phylum and genus levels based on the list of homologues obtained from BLAST. Afterwards, given a classification for each gene, it calculates the proportions of each taxon in the whole set of genes.

The assignments are performed giving a weight to the number of times each phylum and genus are found within the homologues list. Merlin privileges the first five hits, since those are likely to be taxonomically more related to the target gene (Figure 2). In the end, a gene will be assigned with a taxonomic label only if it fulfills the following criteria:

- The number of homologues is higher than the minimum number required (default value is 5).
- The phylum score is higher than the defined threshold (default value is 0.5).
- The genus score is higher than the defined threshold (default value is 0.3).
- The phylum and genus are congruent.

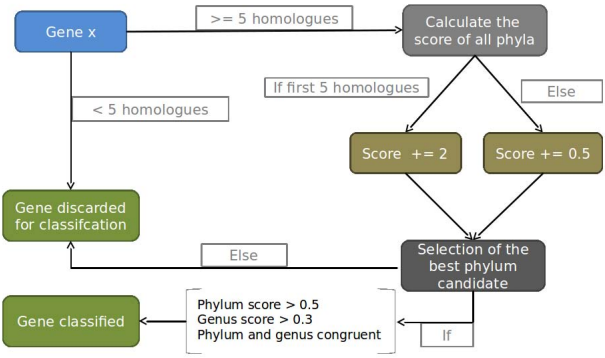


Fig. 2. Schematic representation of the taxonomic routine for gene classification employed in Merlin. The figure represents the schema for phylum classification but for genus the procedure is similar. Default values for the parameters can be changed.

2.2 Functional Analysis

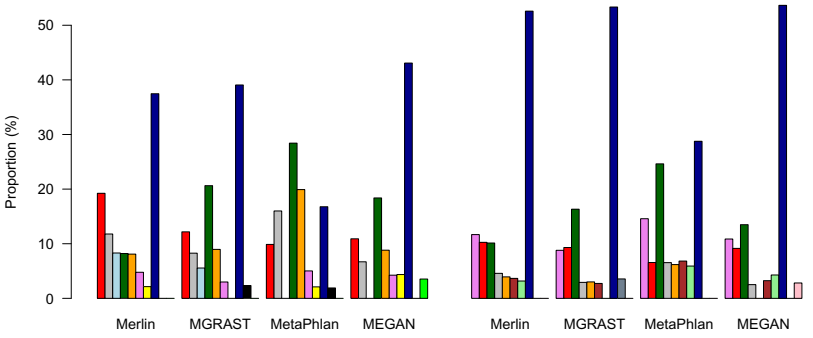
The goal of this module is to identify the metabolic potential of the microbial community. Merlin uses a routine to identify enzymes from the BLAST results giving a weight to the number of times an Enzyme Commission (EC) number is found within the list of homologues for each gene [19]. For metabolic pathway inference, Merlin uses KEGG as the resource for representation and performs hypergeometric tests to find statistical significance in pathway identification. This method is commonly used in pathway analysis studies [20] and identifies enriched pathways compared with a background distribution. Merlin uses the identified enzymes from the whole set of existing enzymes in KEGG for each sample as background distribution. Then, for each pathway, it calculates the probability that the number of enzymes observed in the enzymes list that compose a pathway occurred by chance. If not, a pathway is assumed to be present.

3 Case Study

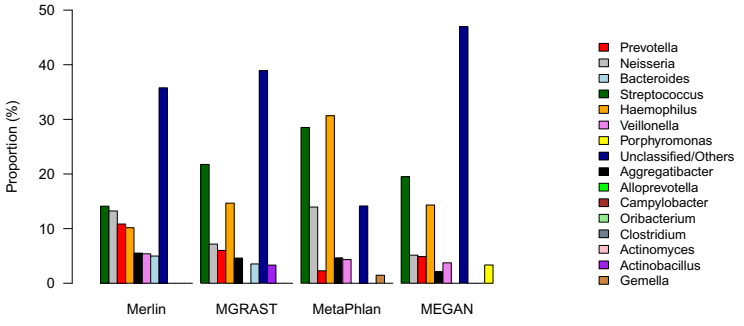
The assemblies from the saliva samples as part of the HMP were downloaded from `http://hmpdacc.org/HMASM/`. MetaGeneMark [21] was used to predict the putative genes for each sample. There were five samples from saliva available, but only three passed the quality control tests. For each sample, a remote BLAST against NCBI nr and a local one against SwissProt were performed. The results are displayed in Table 1. The differences between the two approaches are clear as the annotation against NCBI nr provides better results: a big fraction of genes present similarities. On the other hand, the local BLAST against SwissProt ran much faster, despite the large number of genes that remained unannotated ($\approx$ 60%), which was expected due to the database small size.

Table 1. Remote BLAST against NCBI nr vs Local BLAST against SwissProt for the HMP samples ran in Merlin. BLASTp set with e-value of 1^{-10}.

	NCBI nr		SwissProt	
Sample	Processed genes	With similarities	Processed genes	With similarities
SRS019210	49663	45023	49665	20231
SRS015055	46188	41073	46189	19176
SRS013942	41906	38508	41906	18677



(a) SRS019120 (b) SRS015055



(c) SRS013942

Fig. 3. Genus distribution of the most abundant taxa in each sample by different tools

3.1 Taxonomic Composition of Saliva Microbiome

Regarding the samples annotated using SwissProt, and given the low number of homologies found (Table 1), it becomes clear that this is not the best approach to taxonomically characterize metagenomes. Since the Merlin routine is highly dependent of the BLAST results, poor outputs on this step compromised the performance of the algorithm. Furthermore, using SwissProt as the reference database creates biased results because few organisms are well represented there, inducing the taxonomic assignments towards these organisms.

Using annotations against NCBI nr, Merlin was able to assign a taxonomic label in more than half of genes in each sample. The proportion could be increased if the default value for the minimum number of homologues required was changed, but it was decided to keep a conservative approach. Concerning the phylum analysis, three clearly stand out: *Bacteroidetes*, *Firmicutes* and *Proteobacteria*, despite none dominates the microbiome. The genus composition was also assessed for a comparison between samples and different tools (Figure 3). The high percentage of unclassified sequences in all cases is evident. Although the darkblue bars also represent organisms with residual abundance, results show the potential of metagenomics on unveiling new forms of life.

While the *Prevotella*, *Streptococcus*, *Veillonella*, *Neisseria* and *Haemophilus* are the overall most abundant genera in all samples, no consistency was found between the tools. The different proportions of each taxon on the different tools can be explained considering the way each method works (assembly/read based, BLAST all sequences/only marker genes used for classification). Thus, it is not possible to say which tool is the best. Furthermore, previous studies of the oral flora at the genus level reveal a diverse microbiome composition [22], which is in agreement with the pattern observed here. Overall, Merlin appears to be a good alternative for taxonomic studies of metagenomes.

3.2 Functional Capabilities

The enzymes encoded by each method were compared to those obtained in the IMG/M-HMP web server [23]. Table 2 shows a large discrepancy between assignments based on SwissProt and NCBI nr as the latter presents a smaller number of identified enzymes. Propagated errors on enzymes annotation in NCBI might be the main reason for these. The numbers regarding SwissProt annotations seem to agree in cardinality with those stored in IMG/M-HMP. Further tests confirmed that the majority of enzymes overlap between the two approaches, demonstrating the good results of Merlin.

Functional pathways were predicted in saliva samples using hypergeometric tests based on the number of enzymes encoded in each. The results obtained by HUMAnN were used to compare with those produced by Merlin. The number of metabolic pathways identifed ranged from 37 to 56 over the different samples and methods. As expected, the samples annotated against NCBI-nr harbored less pathways, since the number of encoded enzymes was smaller too (Table 2).

Table 2. Comparison of the complete EC numbers annotated by IMG/M and Merlin in each sample

	IMG/M-HMP		Merlin SwissProt		Merlin NCBI nr	
Sample	Encoded	Unique	Encoded	Unique	Encoded	Unique
SRS019210	12143 (24.34%)	957	10058 (20.25%)	977	4871 (9.81%)	605
SRS015055	11988 (25.81%)	997	9776 (21.17%)	977	2287 (4.95%)	506
SRS013942	10642 (25.46%)	954	8922 (21.29%)	957	2739 (6.54%)	507

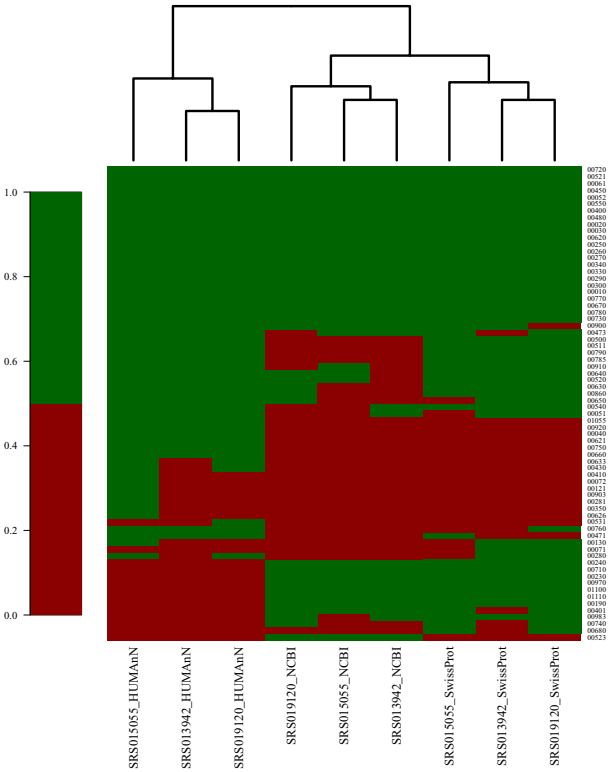


Fig. 4. Presence of metabolic pathways in the samples from saliva across different annotations. Columns represent the samples. Rows represent the binary value for pathway coverage. Green colors stands for present pathways whilst red cells account for the absent ones. Pathways with no enzymes in their constituents were filtered out from HUMAnN results for an easier comparison. Furthermore, as HUMAnN calculates coverage as a likelihood that ranges from 0 to 1, values higher than 0.5 were treated as present (binary value 1) and those with values lower than 0.5 were handled as absent(binary value 0).

To inspect if the inferred pathways were concordant across different methods, an heatmap was constructed (Figure 4).This data representation enabled to cluster the samples according to their similarity in terms of coverage. It is clear that the method used for pathways assignment influences more the clustering of the samples than the sample itself. Furthermore, only 70 different pathways out of 154 possible (with complete EC numbers) were found over all methods and samples, which also confirms the stable functional capability of the human microbiome described before [15] [24].

4 Conclusions and Future Work

An extension of Merlin, an user friendly tool for metabolic reconstruction, was presented. It enables the analysis of metagenomes based on an assembly-based approach. The performance of the software was evaluated with saliva samples from the HMP and the taxonomic profile predicted in Merlin was in agreement with other tools, despite some differences in the proportions. The functional characterization showed a conserved pool of pathways through different samples, although Merlin sometimes presented less pathways than expected because the routine is highly dependent on the enzymes annotation.

There are also some aspects that should be improved in the future. The most relevant one is to implement annotations against KEGG Orthology, or any other catalog of orthologs. This feature would increase the speed of the process maintaining high sensitivity for the taxonomic analysis.

Merlin is freely available from `http://www.merlin-sysbio.org` where a tutorial with more detailed information about the methods is also provided.

Acknowledgments. The work is partially funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within projects ref. COMPETE FCOMP-01-0124-FEDER-015079 and Strategic Project PEst-OE/EQB/LA0023/2013, and also by Project 23060, PEM - Technological Support Platform for Metabolic Engineering, co-funded by FEDER through Portuguese QREN under the scope of the Technological Research and Development Incentive system, North Operational.

References

1. Qin, J., Li, R., Raes, J., et al.: A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285), 59–65 (2010)
2. Greenblum, S., Turnbaugh, P.J., Borenstein, E.: Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proceedings of the National Academy of Sciences of the United States of America 109(2), 594–599 (2012)
3. Kislyuk, A., Bhatnagar, S., Dushoff, J., et al.: Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics 10(1), 316 (2009)
4. Chatterji, S., Yamazaki, I., Bai, Z., Eisen, J.A.: CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In: Vingron, M., Wong, L. (eds.) RECOMB 2008. LNCS (LNBI), vol. 4955, pp. 17–28. Springer, Heidelberg (2008)
5. Gerlach, W., Stoye, J.: Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Research 39(14), e91 (2011)
6. Segata, N., Waldron, L., Ballarini, A., et al.: Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods 9(8), 811–814 (2012)

7. Huson, D.H., Mitra, S., Ruscheweyh, H.J., et al.: Integrative analysis of environmental sequences using MEGAN4. Genome Research 21(9), 1552–1560 (2011)
8. Altschul, S., Gish, W., et al.: Basic Local Alignment Search Tool. J. Mol. Biol. 215(3), 403–410 (1990)
9. Zhao, Y., Tang, H., Ye, Y.: RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics 28(1), 125–126 (2012)
10. Pruitt, K.D., Tatusova, T., Brown, G.R., et al.: NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research 40(Database issue), D130–D135 (2012)
11. Consortium, U.: Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Research 41(Database issue), D43–D47 (2013)
12. Kanehisa, M., Goto, S., Furumichi, M., et al.: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research 38(Database issue), D355–D360 (2010)
13. Marchler-Bauer, A., Lu, S., Anderson, J.B., et al.: CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research 39(Database issue), D225–D229 (2011)
14. Overbeek, R., Begley, T., Butler, R.M., et al.: The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research 33(17), 5691–5702 (2005)
15. Abubucker, S., Segata, N., Goll, J., et al.: Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Computational Biology 8(6), e1002358 (2012)
16. Meyer, F., Paarmann, D., D'Souza, M., et al.: The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9(1), 386 (2008)
17. Sun, S., Chen, J., Li, W., et al.: Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Research 39(Database issue), D546–D551 (2011)
18. Glez-Peña, D., Reboiro-Jato, M., Maia, P., et al.: AIBench: a rapid application development framework for translational research in biomedicine. Computer Methods and Programs in Biomedicine 98(2), 191–203 (2010)
19. Dias, O., Rocha, M., Eugenio, F., et al.: Merlin: Metabolic Models Reconstruction using Genome-Scale Information. Computer Applications in Biotechnology 11(1), 120–125 (2010)
20. Evangelou, M., Rendon, A., Ouwehand, W.H., et al.: Comparison of methods for competitive tests of pathway analysis. PloS One 7(7), e41018 (2012)
21. Zhu, W., Lomsadze, A., Borodovsky, M.: Ab initio gene identification in metagenomic sequences. Nucleic Acids Research 38(12), e132 (2010)
22. Keijser, B., Zaura, E., Huse, S., et al.: Pyrosequencing analysis of the Oral Microflora of healthy adults. Journal of Dental Research 87(11), 1016–1020 (2008)
23. Markowitz, V.M., Chen, I.M.A., Chu, K., et al.: IMG / M-HMP: A Metagenome Comparative Analysis System for the Human Microbiome Project. PLoS One 7(7), 1–7 (2012)
24. The Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature 486(7402), 207–14 (June 2012)