Can the Genomics of the Gut Microbiota in Stool Samples be Analyzed by MERLIN?

Mudyawati Kamaruddin*, Hardianti, Fatma

Postgraduate Program of Clinical Laboratory Science, Universitas Muhammadiyah Semarang, Jl. Kedungmundu Raya No. 18, Semarang 50273, Indonesia
*Corresponding author: mudyawati@unimus.ac.id

Ahmad Ilham

Department of Informatics, Faculty of Engineering, Universitas Muhammadiyah Semarang, Jl. Kedungmundu Raya No. 18, Semarang 50273, Indonesia

Ardi Pranata

Medical Laboratory Technology, Faculty of Nursing and Health Science Universitas Muhammadiyah Semarang, Jl. Kedungmundu Raya No. 18, Semarang 50273, Indonesia

Article histroy :
Received: 09 JAN 2023
Accepted: 23 JAN 2023
Avalaible online: 23 FEB 2023

Research article

Abstract: Metagenomics is important for studying microorganisms that live in microbial communities, particularly those inhabiting the human body. For instance, the gut microbiota is a community of microorganisms that reside in the human gut and interact with humans through secondary metabolites. These metabolites produced by the gut microbiota are extremely important and serve as precursors for various human needs, such as short chain fatty acids (SCFA). While there have been reports of functional secondary metabolites produced by different gut microbiota, none have been utilized on the Merlin platform. In this article, we will examine how the Merlin platform can analyze the gut microbiota community. Metabolic Models Reconstruction using Genome-Scale Information (MERLIN) is a bioinformatics tool that can analyze the functional microbial community as well as the taxonomy of these bacteria.

Keywords: GUT MICROBIOTA; METAGENOMIC; MERLIN; SECONDARY METABOLITE

1. Introduction

Metabolic Models Reconstruction using Genome-Scale Information (MERLIN) is a bioinformatics tool that can analyze the functional microbial community as well as the taxonomy of the bacteria. The MERLIN program consists of four steps to carry out this function:

- Loading the Internal Database;
- Enzyme Annotation;
- Transporter Annotation; and
- Compartment Prediction.

Metagenomics plays an essential role in studying microorganisms that live in microbial communities, particularly those that inhabit the human body (Kamaruddin et al., 2014, 2020). For instance, the gut microbiota is a community of microorganisms that live in the human gut and interact with humans through secondary metabolites. These metabolites are crucial and serve as precursors for various human needs, such as short-chain fatty acids (SCFA) (Minarti et al., 2020).

There have been reports of functional secondary metabolites produced by various gut microbiota, but none have yet been used on the MERLIN platform. This article will explore how the MERLIN platform can analyze the gut microbiota community (Fig. 1). We will explain the four stages of the MERLIN process in carrying out the analysis, as referred to in Fig. 2, to gain an understanding of the interrelationship of the MERLIN platform in analyzing the functional metabolites of the gut microbiota community (Holland et al., 2008).

2. Functional Genomic

Although functional genomics is currently a distinct field dedicated to determining gene functions, assigning functions to all genes in a sequenced genome is a challenging task that involves several steps, as shown in Fig. 1 (Diaz, OML., 2013).

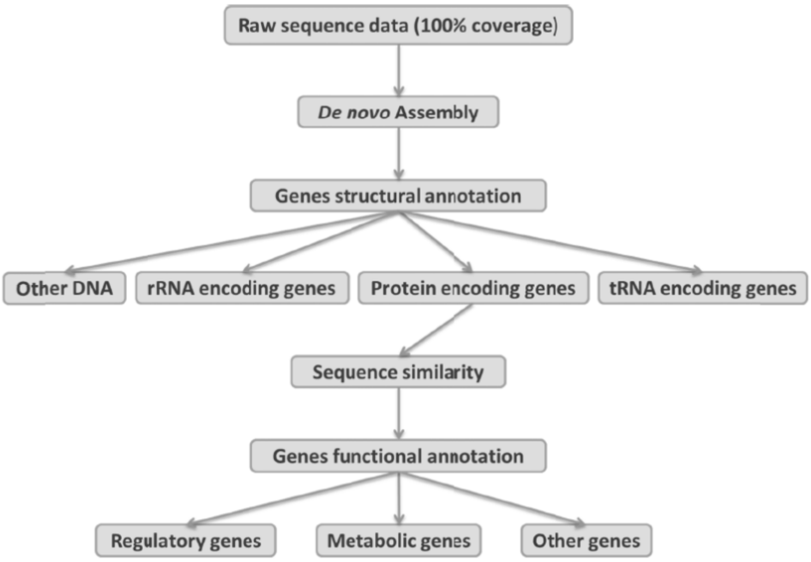

Fig 1. Fungsional genomic (Diaz, OML., 2013).

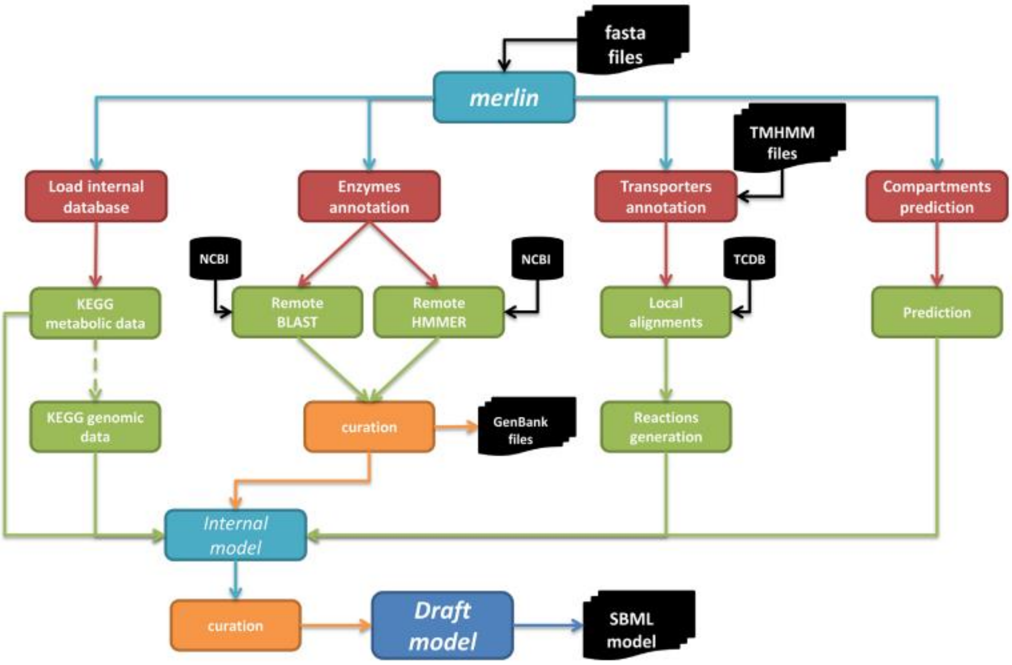

Fig 2. The architecture of merlin 2.0 is depicted schematically (Barbosa, 2013).

The raw genome contains almost all of the information regarding an organism's potential phenotype. Decoding this information, however, is not yet complete. High-throughput sequencing technologies produce a set of short sequence reads that need to be assembled, a process known as de novo genome assembly. The latest sequencing technologies enable the low-cost decoding of microbial genomes through a small number of experiments. However, the drawback to these improvements is a significantly shorter read length, making the sequence reassembly process more challenging. For every base added to the read, the sequence complexity increases four-fold, and the chance of detecting redundancy in a sequence pool decreases dramatically (Tamaki et al., 2007).

As a result, those working on de novo genome assembly welcome new sequencing approaches that can generate longer read lengths and improve data quality. Nevertheless, the cost of sequencing genomes has been decreasing year after year, leading to hundreds of gigabytes of data per genome. Consequently, new, more efficient algorithms are required to analyze and assemble a genome from scratch.

The next step in genome annotation is to identify all genes within a given genomic sequence. This stage is known as genome structural annotation, and it involves identifying all protein-encoding genes, different types of RNA, and other DNA within the genome. As experimental verification can be costly and time-consuming, this process is usually carried out using bioinformatics resources. Identifying the boundaries of protein-encoding open reading frames in prokaryotes and some minor eukaryotes (ORFs) is relatively easy. Identifying long ORFs within these genomes is mainly a matter of running a tool that identifies ORFs longer than a given threshold within all six frames (Barbosa, 2013).

In the first stage of Genome Annotation, data is retrieved from various databases such as NCBI or KEGG. Enzyme Commission (EC) and Transporter Classification (TC) numbers, associated genes and names, and gene products are all collected at this stage. Although genome annotation information for a wide range of organisms can be found in public databases, it should be noted that many annotations take a long time to complete and that the information gathered during this process does not always meet GSMM requirements. As a result, re-annotation is often necessary as the first step in the reconstruction process.

3. Load Interna Database

The "load internal database" step is used to construct the metabolic data backbone of the Merlin internal model. This step retrieves KEGG data, including compounds, glycans, drugs, reactions, modules, pathways, and enzymes. Merlin saves this information and creates a local database. When combined with information from other steps, this step can optionally retrieve genomic information for organisms annotated in KEGG genes. This data can be used to construct the draft GSMM on its own or in conjunction with information from other steps.

4. Enzim Annotation

The Basic Local Alignment Search Tool (BLAST), profile Hidden Markov Models (HMMER), or both are utilized to assign enzymatic functions to proteins encoded in the genome. NCBI provides all the data that Merlin 2.0 retrieves for each homologous gene identified in either the BLAST or the HMMER similarity searches, including the species' name and full lineage. This database also contains the locus tag gene identifiers for genomes downloaded from NCBI's FTP website. Each gene is processed individually, and the homology data retrieved for each homolog identified by the similarity alignment (regardless of which program is used to perform the searches) includes the locus identifier, expected value, score, and organism.

Subsequently, Merlin 2.0 remotely retrieves and collects data for each of the homolog genes. For this purpose, it retrieves information from the Entrez Protein database, including taxonomy, organelle (if available), chromosome (if available), locus tag, product (protein name), EC number (if available), and molecular weight. Finally, the downloaded data is saved in the MySQL relational local database of Merlin 2.0.

Merlin 2.0 assigns EC numbers and product names to each gene g using a routine. This routine assigns weights to each EC number ec (or product name pn) based on their frequency in the homolog gene records (frequency) and the taxonomy of the organisms to which such records belong (taxonomy). This procedure is described in Eq. (1). The weights for frequency (Score) and taxonomy (Score) are determined in Eq. (1).

$$\mathrm{Score}_{\mathrm{g}}^{\mathrm{ec}} = \alpha \times \mathrm{Score}_f + (1-\alpha) \times \mathrm{Score}_{\mathrm{t}} \tag{1}$$

The frequency score, in turn, calculates the number of occurrences of an EC number (or product name) across all homologs of a gene. This score is calculated by counting the number of homologous genes that encode an EC number (or product name) and dividing it by the total number of homologous genes (*n*), as shown in Eq. (2).

$$\mathrm{Score}_f = \frac{\sum_{i=1}^{n} (v_i)}{n} \tag{2}$$

where

$$v_i = \begin{cases} 1, & \text{if ec number exists in record } i \\ 0, & \text{otherwise} \end{cases}$$

The Taxonomy score is utilized to prioritize homologies that are closely related to the organism being studied. To calculate this score, we first determine the Taxonomy frequency, which is the sum of the number of common taxa between the organism being studied and the first n homology records. This value is then multiplied by a penalty factor, as shown in Eq. (3), to adjust for possible errors in annotation or incorrect assignments. The penalty factor reduces the score for EC numbers (or product names) that are assigned by only a few genes.

To calculate the denominator, we multiply the maximum taxonomy value ($\mathrm{Max}_{\mathrm{Taxonomy}}$), which is the number of taxa of the organism being studied, by the smallest number of genes encoding EC number ec (or product name pn) and the user-defined minimal number of homologies ($\mathrm{n}_{\mathrm{homologies}}$).

This classification allows you to determine whether the first n homology records of a given EC number ec (or product name pn) are taxonomically related to the case study. The taxonomic score is calculated as follows Eq. (3).

$$Score_{t(ec(pn))} = \frac{\sum_{i=1}^{n}\left(t_i \times v_{ec(pn)_i}\right) \times \left(1 - p_{ec(pn)} \times \beta\right)}{Max_{Taxonomy} \times min\left(\sum_{i=1}^{n} v_{ec(pn)_i}, n_{homologies}\right)} \tag{3}$$

where ti is the common taxa count for the hit I and β record and is a penalty parameter that is initially set to 0.15

In Eq. (4), the $\mathrm{p}_{\mathrm{ec(pen)}}$ is calculated by subtracting the frequency of genes encoding EC number ec (or product name pn) from the nhomologies. If the result is positive, the $\mathrm{p}_{\mathrm{ec(pen)}}$ penalty is multiplied by and subtracted from 1. Otherwise, the $\mathrm{p}_{\mathrm{ec(pen)}}$ penalty is zero.

$$p_{(ec(pn))} = \begin{cases} 0, & \sum_{i=1}^{n} v_{ec(pn)_i} \geq n_{homologies} \\ n_{homologies} - \sum_{i=1}^{n} v_{ec(pn)_i}, & \text{otherwise} \end{cases} \tag{4}$$

In merlin 2.0's 'Homology Data Viewer,' can directly configure the, α, β, and nhomologies parameters. The confidence score, which has a numeric value between 0 and 1, makes it simple to curate the EC numbers (or product names) assigned to a given gene. The user can also specify a minimum threshold score value for automatic annotation acceptance. However, all annotations can be curated and the automatic assignments can be changed. The annotated metabolic genome produced by this tool can be integrated into the internal model of merlin 2.0 or exported to files in the GenBank (*.gbk) or Excel formats(Barbosa, 2013).

5. Transporters Annotation

Models of transport reactions are frequently only included if there is evidence supported by experimental data or literature. However, this method generates a very small number of transporters and does not allow for Gene-Protein-Reaction (GPR) associations because the associated gene is frequently unknown.

As a result, we proposed a new methodology for identifying and annotating transportation systems. This methodology assigns TC family numbers to carriers and generates transport reactions for all metabolites transported by these carriers. It is based on the identification and classification of genes that encode transmembrane proteins, as transport proteins are thought to be found in membranes.

Because merlin 2.0 cannot access this process remotely, the user must first submit the genome amino acid fasta files to the TransMembrane Prediction using Hidden Markov Models (TMHMM)31 web server. The TMHMM tool is used to find protein-encoding genes that contain transmembrane domains. Then, merlin 2.0 compares protein sequences with at least n transmembrane helices (where n is a user-defined parameter with a default value of 1) to all protein sequences currently available in TCDB. The SW algorithm is used to find similar regions in two sequences. This algorithm compares segments of various lengths and optimizes the similarity measure when performing local sequence alignments. Because transmembrane domains are small sequences of about 20 amino acids in length found within protein sequences, this algorithm was preferred over global alignment algorithms (such as BLAST or Needleman-Wunsch) (Holland et al., 2008).

The SW similarity search results are saved in a relational database. This database establishes links between the genome of the organism under study and the TCDB records. These records frequently provide direct access to specific information, such as the UniProt Accession Number, organism, Protein Name, Length, and so on. However, the substrates and direction of transport are not directly provided to date, so these characteristics must be inferred from the information provided for each record.

Merlin 2.0 comes with a growing database that includes thousands (over 4200) of TCDB records that have already been annotated with metabolites and directions. To assign identifiers to the metabolites transported by each carrier annotated in merlin 2.0, several databases were used, including TCDB, KEGG, ChEBI, and the semanticSBML tool. Although our database does not contain all TCDB records, if similarities to unannotated TCDB records are discovered, such records can be annotated by the user and uploaded to merlin 2.0 via a specific operation (Diaz, OML., 2013).

Finally, the metabolites transported by each carrier identified in the genome uploaded to merlin 2.0 are deduced from the annotations of TCDB records that are similar to such carriers. To classify the assignment of metabolites and TC family numbers, merlin 2.0 employs an internal scorer similar to the one described above for EC numbers (and product names).

The methodology for predicting the subcellular localization of proteins and metabolites, based on WoLF PSORT for Eukaryotes and PSORTb v3.0 for Prokaryotes, was also described in the same article. The data provided by these tools is stored in a relational database. In WoLF PSORT, the protein localization in eukaryotic organisms is determined using a simple remote Java API provided by Paul (Diaz, OML., 2013).

6. Compartments Prediction

The genes are automatically assigned to the main compartment predicted by these programs. Furthermore, if secondary compartments have scores that differ from the main compartment by less than a user-defined percentage (the default value is 10%), the gene will be assigned to those compartments as well.

To annotate transportation systems, three criteria must be met. The first two are that the gene sequences have transmembrane domains and have similarities to TCDB records. The third is having a prediction of localization within a membrane. However, the WoLF PSORT and PSORTb 3 predictions combine intracellular membranes with intracellular organelle predictions, allowing assignment only to the cytoplasmic membrane or outer membrane for prokaryotes and the plasma membrane for eukaryotes. As a result, if a sequence met the first two requirements and WoLF PSORT predicted that it would be assigned to an intracellular organelle, it is assumed that the sequence encoded an intracellular transport system.

If, on the other hand, a regular enzyme is predicted to be assigned to a membrane, that enzyme is assigned to both sides of the membrane.

These modules enable the generation of compartment-specific transport reactions, as well as the establishment of associations between genes and reactions, allowing the construction of more robust and reliable models.

Because all TCDB records contain cross-references to UniProt, this identifier is also used as an unambiguous identifier for such records in merlin 2.0. Furthermore, taxonomic information for TCDB records is retrieved for metabolite classification and TC family numbers (Barbosa, 2013; Tamaki et al., 2007).

7. Conclussion

Merlin is a user-friendly Java application that reconstructs genome-scale metabolic models for each organism whose genome has been sequenced. Merlin includes tools for identifying and annotating genes that encode transport proteins, as well as creating transport reactions for those carriers. Also developed and integrated into merlin are tools for compartmentalization models that predict the localization of encoded proteins in the genome, and thus the localization of metabolites involved in the reactions induced by those proteins. Merlin includes a number of tools for curating genome annotation and model drafting, which greatly aids model validation and allows for the creation of an accurate model.

Metagenomics can be applied to gut microbiota using Merlin's work procedure.

Acknowladgements

We would like to thank to the Clinical Laboratory Science, Universitas Muhammadiyah Semarang for encouraging and supporting.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Barbosa, P. S. (2013). Taxonomic and functional analysis of metagenomes (Doctoral dissertation, Universidade do Minho (Portugal)).

Dias, O. M. L. (2013). Reconstruction of the Genome-scale Metabolic Network of Kluyveromyces lactis (Doctoral dissertation, Universidade do Minho (Portugal)).

Holland, R. C. G., Down, T. A., Pocock, M., Prlic, A., Huen, D., James, K., Foisy, S., Drager, A., Yates, A., Heuer, M., & Schreiber, M. J. (2008). BioJava: An open-source framework for bioinformatics. Bioinformatics, 24(18), 2096–2097. https://doi.org/10.1093/bioinformatics/btn397

Kamaruddin, M., Tokoro, M., Rahman, Md. M., Arayama, S., Hidayati, A. P. N., Syafruddin, D., Asih, P. B. S., Yoshikawa, H., & Kawahara, E. (2014). Molecular Characterization of Various Trichomonad Species Isolated from Humans and Related Mammals in Indonesia. The Korean Journal of Parasitology, 52(5), 471–478. https://doi.org/10.3347/kjp.2014.52.5.471

Kamaruddin, M., Triananinsi, N., Sampara, N., Sumarni, S., Minarti, M., & Ra, A. M. (2020). Karakterisasi DNA Mikrobiota Usus Bayi pada Persalinan Normal yang diberi ASI dan Susu Formula. Media Kesehatan Masyarakat Indonesia, 16(1), 116. https://doi.org/10.30597/mkmi.v16i1.9050

Minarti, Triananinsi, N., Nurqalbi, Sumarni, & Kamaruddin, M. (2020). Metagenomic Diversity of Gut Microbiota of Gestational Diabetes Mellitus of Pregnant Women. 13(01), 1–8. https://doi.org/10.31001/biomedika.v13i1.747

Tamaki, S., Arakawa, K., Kono, N., & Tomita, M. (2007). Restauro-G: A Rapid Genome Re-Annotation System for Comparative Genomics. Genomics, Proteomics & Bioinformatics, 5(1), 53–58. https://doi.org/10.1016/S1672-0229(07)60014-X