INTISARI

COMPUTATIONAL TOOL UNTUK EVALUASI HASIL *METAGENOMIC ASSEMBLY*

Oleh

Fitri Hasanah Amhar
12/334729/PA/14961

Perkembangan Bioinformatika memungkinkan para ahli melakukan identifikasi *genome-genome* pada suatu lingkungan yang disebut identifikasi *metagenome*. Data *metagenome* yang diperoleh dari suatu lingkungan akan diperoleh potongan DNA-nya melalui alat *sequencing*. Selanjutnya potongan DNA ini digabungkan oleh *assembler* agar dapat diketahui spesies asalnya. Terdapat berbagai *assembler* dengan metode penggabungan yang berbeda satu sama lain. Perkembangan komputer dengan kecepatan berpikir dan kemampuan otomatisasi pemecahan masalah sangat dibutuhkan untuk menyelesaikan persoalan biologi, termasuk dalam hal mengevaluasi hasil *metagenomic assembly* untuk menemukan hasil penggabungan terbaik.

Praktisi Biologi membutuhkan data hasil *metagenomic assembly* yang paling akurat dan mendekati *reference genome*. Karenanya dibutuhkan *computational tool* untuk melakukan evaluasi hasil *metagenomic assembly*. Pada penelitian ini, digunakan parameter N50, N-len(x), dan Nm50 untuk melakukan evaluasi hasil *metagenomic assembly* serta penghitungan *cover rate* dan *chimeric rate* yang menunjukkan keakuratan hasil *assembly*. Evaluasi data *metagenome* dengan dua simpul *genome* yang mengacu pada *reference genome* berbeda, atau disebut sebagai bagian *chimeric*, dilakukan dengan menghilangkan bagian tersebut terlebih dahulu sebelum menerapkan penghitungan parameter evaluasi terhadap datanya.

Parameter evaluasi ini diujikan pada *simulated dataset* dari hasil *assembler* Xgenovo, MetaVelvet-SL, dan IDBA-UD serta *real dataset* dari hasil *assembler* RayMeta, dan SOAPdenovo2. Hasil penelitian menunjukkan bahwa *computational tool* untuk evaluasi hasil *metagenomic assembly* telah berhasil dibangun. Parameter N50 dapat digunakan untuk mengevaluasi *simulated dataset* yang tidak mengandung bagian *chimeric*. Parameter Nm50 digunakan untuk mengevaluasi *simulated dataset* yang mengandung bagian *chimeric*. Parameter N-len(x) digunakan untuk mengevaluasi *real dataset* yang tidak memiliki *reference genome* dan ukurannya sangat besar.

Kata kunci : evaluasi hasil *metagenomic assembly*, N50, N-len(x), Nm50, *chimeric*

ABSTRACT

A COMPUTATIONAL TOOL FOR EVALUATING METAGENOMIC ASSEMBLY RESULT

by

Fitri Hasanah Amhar
12/334729/PA/14961

The development of bioinformatics allows the expert to identify genomes in any environment, it's called metagenome identification. Metagenome data which is obtained from an environment will pass the sequencing tool to gain the pieces of its DNA. The pieces of DNA are assembled by the assembler in order to know its origin species. There are various assemblers with its own method of merging that differ each other. The development of computer with the speed of thinking and automation problem solving ability is needed to resolve biological problem, including evaluation of metagenomic assembly result to find the best assembly merging result.

Biologists need the most accurate metagenomic assembly result data and the closest to the reference genome. Therefore, it takes computational tool for evaluating metagenomic assembly result. In this study, N50, N-len(x), and Nm50 are used as parameter to evaluate the metagenomic assembly result. Cover rate and chimeric rate calculation are used to show the accuracy of the assembly. Evaluation of metagenome data with two genome nodes that refer to different reference genome, called chimeric node, is done by eliminating its chimeric part first before applying the parameters calculation to evaluate the data.

The evaluation parameters were tested in simulated dataset from Xgenovo, MetaVelvet-SL, and IDBA-UD assemblers and also real dataset from the RayMeta and SOAPdenovo2 assemblers. The results showed that the computational tool for evaluating metagenomic assembly result has been successfully developed. N50 parameter can be used to evaluate simulated dataset that does not contain any chimeric node. Nm50 parameter is used to evaluate simulated dataset containing the chimeric node. N-len(x) parameter is used to evaluate the real dataset which do not have any reference genome and has very large size.

Keyword : evaluating metagenomic assembly result, N50, N-len(x), Nm50, chimeric