INTISARI

METAGENOMIC ASSEMBLY MENGGUNAKAN RANDOM FOREST

Oleh

Jason Kurniawan
13/347220/PA/15176

Kemajuan teknologi telah banyak mempengaruhi perkembangan ilmu biologi molekular modern dan genetika. Dengan teknologi *Next Generation Sequencing*(NGS), *Whole Genome Sequencing* (WGS) atau pengurutan keseluruhan genom dari suatu organisme dapat dilakukan secara masif dan paralel dengan waktu yang lebih cepat dan murah. Hal ini juga menyebabkan pertumbuhan data genom yang besar sehingga membutuhkan metode komputasi untuk mengolahnya. Data yang berasal dari teknologi NGS merupakan potongan-potongan genom yang saling tumpang tindih (*overlap*) atau biasa disebut dengan *reads*. Potongan-potongan tersebut perlu disusun kembali menjadi *sequence* yang lebih panjang, proses penyusunan ini disebut dengan *assembly*. Pada data genom yang diambil dari suatu lingkungan atau komunitas, informasi genom tidak hanya berasal dari satu spesies. Dibutuhkan metode tertentu untuk melakukan *assembly* pada data *metagenome* yang biasa disebut dengan *metagenomic assembly*.

Pada penelitian ini, dilakukan pengembangan kembali *tool metagenomic assembly* MetaVelvet-SL di mana *Random Forest* digunakan sebagai algoritma *machine learning* yang digunakan pada modul *Supervised Learning*. Pada penelitian ini dilakukan pembandingan nilai akurasi dan *F-measure* pada klasifikasi kandidat *chimeric node* dengan penelitian sebelumnya yang menggunakan *Support Vector Machines*. Pada penelitian ini juga dilakukan pembandingan hasil *assembly* dengan menghitung skor N-50 dan Nm50. Dari hasil eksperimen yang dilakukan, *Random Forest* terbukti mengungguli performa *Support Vector Machines* pada kasus klasifikasi *chimeric node* pada 4 level taksonomi, yakni Ordo, Family, Genus, dan Species. Pada level taksonomi Family dan Species skor Nm50 *Random Forest* mampu menggungguli *Support Vector Machines*.

Kata Kunci: *Metagenomic Assembly*, *Supervised Learning*, *Random Forest*, klasifikasi

ABSTRACT

METAGENOMIC ASSEMBLY USING RANDOM FOREST

By

Jason Kurniawan
13/347220/PA/15176

Advances in technology have influenced the development of modern molecular biology and genetics. With technology Next Generation Sequencing (NGS), Whole Genome Sequencing (WGS) or sequencing the entire genome of an organism can be done in parallel with the massive and faster time and cost. It also led to the growth of large genomic data that require computational methods to process it. The data coming from NGS technologies are pieces of the overlapping genome or commonly called the reads. The pieces need to be reassembled into longer sequence, the preparation process is called assembly. At the genomic data extracted from a environtment or community, genome information not only reserved from one species. It takes a certain method to do assembly in the metagenome data which is called the metagenomic assembly.

In this study, we conducted redevelopment tool metagenomic assembly MetaVelvet-SL where Random Forest is the machine learning algorithm that used on the Supervised Learning module . In this research, benchmarking classification accuracy and F-measure values in the candidate node chimeric previous research which uses Support Vector Machines. In this study, we also conducted benchmarking results of the assembly by calculating a score of N-50 and Nm50. From the results of the conducted experiments, Random Forest proved to surpass the performance of Support Vector Machines in case the classification node chimeric at 4 level taxonomy: Order, Family, Genus and Species. The accuracy value is not give a significant impact on the resulted N-50 score. On the taxonomy level of the Family and Species, the Nm50 scores of Random Forest are able to outperform Support Vector Machines.

Keywords: Metagenomic Assembly, Supervised Learning, Random Forest, classification