INTISARI

ANALISIS PERBANDINGAN METODE HIERARCHICAL CLUSTERING UNTUK MULTIPLE SEQUENCE ALIGNMENT

Oleh

Aryasakti Wirasena

18/424181/PA/18286

Multiple sequence alignment atau MSA adalah sebuah proses penting dalam bidang bioinformatika yang melakukan pengolahan data sekuens protein asam amino atau basa DNA. Salah satu implementasi MSA adalah Clustal, yang membagi proses MSA menjadi tiga tahap. Salah satu tahap Clustal adalah pembuatan *guide tree* menggunakan algoritma *hierarchical clustering*, yang menentukan urutan pemrosesan sekuens dalam tahap *progressive alignment*. Karena terdapat beberapa metode *hierarchical clustering*, diperlukan perbandingan performa beberapa metode *hierarchical alignment* dalam membentuk *guide tree* untuk proses MSA.

Penelitian ini membandingkan metode *single linkage*, *complete linkage*, dan *average linkage* dalam membuat *guide tree* untuk melakukan MSA pada tiga buah data sekuens dari BaliBASE 4. Guide tree yang dihasilkan dari masing-masing metode dijadikan input untuk program ClustalX untuk menghasilkan alignment. Metrik penilaian column scoring digunakan untuk menilai kualitas sebuah alignment.

Hasil penelitian menunjukkan bahwa metode *single linkage* cenderung menghasilkan *alignment* dengan *quality score* yang lebih tinggi dari kedua metode lainnya. Akan tetapi pada sebagian besar kasus keunggulan *single linkage* tidak begitu signifikan dan dapat dikalahkan oleh metode lain pada beberapa kasus. Metode single linkage dinilai paling efektif jika data yang digunakan mengandung sejumlah outlier.

Kata kunci: Multiple sequence alignment, Clustal, Hierarchical clustering

ABSTRACT

ANALYSIS AND COMPARISON OF HIERARCHICAL CLUSTERING METHODS FOR MULTIPLE SEQUENCE ALIGNMENT

By

Aryasakti Wirasena

18/424181/PA/18286

Multiple sequence alignment or MSA is an important process in the field of bioinformatics that performs processing on a sequence of amino acid protein or sequence of DNA bases. One implementation of MSA is the Clustal program, that divides the MSA process into three stages. One of those stages is guide tree generation using a hierarchical clustering algorithm which determines in what order the sequences will be processed in the progressive alignment stage. Since there are many hierarchical clustering methods, there is a need to compare the performace of hierarchical clustering methods in generating a guide tree for MSA.

This study compares the single linkage, complete linkage, and average linkage methods in generating a guide tree for MSA on three sequence data taken from BaliBASE 4. The guide trees generated by each method is used as input to the ClustalX program to aid in creating an alignment. The built-in column scoring metric is used to determine the quality of an alignment.

Results show that the single linkage method tends to generate an alignment with higher quality score in comparison to the other two methods. However, on most cases the difference of score between single linkage and the other methods are insignificant and sometimes outdone by the other methods on other cases. Single linkage may be more effective if the data used contains a number of outliers.

Keywords: Multiple sequence alignment, Clustal, Hierarchical clustering