

Abstrak

DNA (*Deoxyribose Nucleid Acid*) merupakan sebuah rangkaian-rangkaian protein asam nukleotida yang berada didalam tubuh makhluk hidup dimana DNA akan sangat identik dengan pewarisan sifat. DNA dalam tubuh makhluk hidup berbentuk pasang-pasangan, sehingga dibutuhkan analisis biologi untuk mencocokkan similaritas antar data DNA tersebut. Dikarenakan data DNA yang terlalu besar, maka *Big Data* dapat menjadi salah satu jawaban untuk melakukan komputasi kecocokan similaritas pada data tersebut. *Big Data* merupakan sebuah teknik yang dapat melakukan komputasi berskala besar dengan memiliki beberapa *framework* yang mendukung untuk mencari similaritas *sequence* biologi. Hadoop merupakan *framework* yang sangat tepat untuk menjalankan *Big Data* dengan memiliki beberapa macam tools yang dapat digunakan. Untuk pemrosesan data-data DNA tersebut, Bioinformatika hadir dengan beberapa teknik yang dapat digunakan. Salah satu teknik yang digunakan untuk memproses data DNA tersebut adalah MSA (*Multiple Sequence Alignment*) dimana salah satu algoritmanya memiliki nilai akurasi yang sangat tinggi yaitu algoritma T-COFFEE (*Tree Based Consistency Objective Function for Alignment Evaluation*). T-COFFEE merupakan sebuah algoritma *multiple sequence* yang sangat cocok untuk mencari similaritas pada data DNA dengan berfokus pada nilai akurasi yang sangat tinggi. Disamping memiliki nilai akurasi yang tinggi, T-COFFEE membutuhkan waktu yang sangat lama untuk pemrosesannya sehingga apabila diimplementasikan pada data DNA yang jumlahnya besar, maka akan memerlukan waktu yang lama. Pengimplementasian T-COFFEE pada paralelisasi hadoop dapat mereduksi waktu yang dibutuhkan T-COFFEE untuk melakukan komputasi dengan memiliki nilai speedup yang lebih baik dibandingkan dengan tanpa menggunakan paralelisasi hadoop.

Kata kunci : DNA, Big Data, MSA, T-COFFEE, paralelisasi hadoop

