# 1. PENDAHULUAN

## 1.1 Latar Belakang

Bioinformatika merupakan perpaduan disiplin ilmu Biologi dan Informatika ntuk pencarian informasi *(retrevial),* penyimpanan data *(storage),* manipulasi dan distribusi informasi molekul biologi seperti DNA, RNA dan protein[1]. Menemukan kemiripan pada perbandingan *sequence*, termasuk identitas, kecocokan dan homologi memberikan petunjuk tentang fungsi genetik *sequence* yang baru[2]yang banyak digunakan oleh ahli biologi. Penyejajaran sekuen (*sequence alignment*) adalah pensejajaran dua atau lebih *sequence* sehingga memperoleh tingkat kesamaan *(sequence similarity)* untuk menemukan kesamaan paling maksimum antar residu yang membantu mendapatkan informasi kemiripan fungsi, struktur, keturunan dan analisis filogenetik.

Berdasarkan jumlah *sequence* yang dibandingkan, algoritma *alignment* dibedakan menjadi *pairwise sequence alignment* dan *multiple sequence alignment*, sedangkan berdasarkan cara alignment yang dilakukan dibedakan menjadi *global alignment* dan *local alignment.* Pada *pairwise sequence alignment* perbandingan *sequence* kueri dilakukan secara berpasangan dengan semua *sequence* dalam basis data. Pencarian pada basis data harus memenuhi kriteria antara lain *sensitivity*, *selectivity* dan *speed*[1].

*Dynamic programming* adalah metode kuantitatif yang banyak digunakan untuk *sequence alignment* dengan hasil akurat dan reliabel namun ketika menangani *sequence* yang memiliki panjang dan jumlah yang banyak, *Dynamic Programming* terlalu lambat saat *resource* komputasi terbatas[3]. Diperlukan metode pencarian khusus untuk mempercepat proses komputasi perbandingan *sequence.* Metode heuristik merupakan salah satu solusi untuk mempercepat pencarian dengan memeriksa sebagian kecil dari kemungkinan *alignment*[1][4].

*Basic Local Alignment Search Tool* (BLAST) adalah versi heuristik dari *pairwise local alignment* algoritma Smith-Waterman[5]. Algoritma BLAST menggunakan teknik *seed-and-extended*[6] dalam strategi optimasi metode *word* dimana setiap *word* biasanya berisi tiga residu *sequence* protein dan 11 residu untuk *sequence* DNA sehingga sulit jika mengelola data kueri yang sangat banyak.
Keterbatasan skalabilitas membuat proses kurang efektif karena membatasi sensitivitas *alignment.* Munculnya teknologi *Big Data* seperti Hadoop dapat memberikan efisiensi komputasi sehingga kumpulan data *sequence* berukuran besar dapat diproses lebih cepat. Hadoop melakukan eksekusi secara paralel ke beberapa *worker* dengan tugas distribusi data pada HDFS dan tugas komputasi oleh MapReduce. Dengan begitu, implementasi komputasi dengan memanfaatkan *framework* hadoop berpotensi mempercepat waktu komputasi BLAST karena pada prinsipnya operasi perbandingan berpasangan adalah saling independen sehingga bisa diparalelkan.

## 1.2 Topik dan Batasannya

Penelitian ini menerapkan BLAST yang melakukan *sequence alignment* didalam framework Hadoop mapreduce. Dalam program HadoopBlast menerapkan *distributed cache,* untuk penyimpanan sementara program dan basis data BLAST yang dapat mengurangi jumlah pembacaan dari lokasi HDFS. Program BLAST dilakukan pada proses "mapper" dan hanya menggunakan mapper hadoop, juga dilakukan analisis performa komputasi BLAST Hadoop.

## 1.3 Tujuan

Tujuan dari penelitian ini adalah memodelkan aplikasi bioinformatika BLAST yang dapat melakukan *sequence alignment* dan digunakan pada framework Hadoop juga efisiensi antara *stand-alone* BLAST dan dengan menggunakan Hadoop.

## 1.4 Organisasi Tulisan

Jurnal ini terdiri dari lima bab utama, bab pertama merupakan Pendahuluan yaitu latar belakang, topik dan batasannya serta tujuan penelitian. Bab kedua yaitu Studi Terkait, berisi teori-teori penunjang penelitian yang dilakukan, meliputi: *sequence alignment*, metode-metode pemrosesan *sequence* biologi, BLAST dan penjelasan mengenai *framework* Hadoop (HDFS dan Mapreduce). Bab ketiga berisi Sistem yang Dibangun, menjelaskan mengenai rancangan sistem yang dibuat. Bab keempat yaitu Evaluasi, terdiri dari hasil rancangan sistem yang telah dibangun sesuai rancangan pada bab tiga. Pada bagian terakhir, bab kelima berisi kesimpulan dari hasil pengujian sistem yang dibangun serta saran untuk penelitian selanjutnya.