BAB 1

PENDAHULUAN

1.1. Latar Belakang Masalah

Untuk jangka waktu yang lama semenjak awal mula kemunculan ilmu mikrobiologi, wawasan dan pengetahuan yang mengisi ranah ilmu ini berasal dari hasil pengkulturan (pembudidayaan) murni mikroba yang diketahui dapat dikembangbiakkan di dalam laboratorium melalui prosedur *sequencing* materi genetik dan perakitannya. Peluang adanya varian-varian mikroba lain di luar apa yang dapat dibudidayakan layaknya demikian tidak dipertimbangkan, karena pada masa itu, segala teori terkait muncul atas dasar keyakinan bahwa tidak ada satupun mikroorganisme yang dapat diklasifikasikan tanpa terlebih dahulu berhasil melalui pengkulturan, namun dasar itu mulai luntur pada tahun 1980-an (Handelsman, 2004, p. 669). Gagasan-gagasan baru mulai timbul bahwa ruang lingkup dunia mikroba non-kultur ini jauh lebih luas dari dugaan dan bahwa dunia yang tidak kasat mata ini dapat dipelajari (Pace dkk., 1985, 1986, 1995). Seiring tahun demi tahun berlalu, berbagai metode mulai dikembangkan untuk memperoleh perspektif baru terhadap seluk-beluk mikroorganisme yang tidak dapat dikulturkan (non-kultur) ini.

Dalam upaya menganalisis materi genetik mikroorganisme non-kultur, pengambilan sampel dilakukan secara langsung dari lingkungan. Sampel yang diambil tersebut dapat mengandung pecahan-pecahan (fragmen) materi genetik (genom) dari beragam spesies yang berbeda. Ketika prosedur *sequencing* dan

perakitan dilakukan terhadap campuran fragmen ini secara serentak, ketidakcocokan antara genom satu spesies dengan yang lainnya akan menghasilkan *chimeric contigs* yang berujung pada fenomena *interspecies chimerae*, sehingga ragam spesies dari sampel tersebut tidak dapat diketahui (Overbeek, Kusuma dan Buono, 2013; Simangunsong, 2015). Istilah *contig* ini sendiri diambil dari kosakata Bahasa Inggris '*contiguous*' ('berdekatan' dalam Bahasa Indonesia) dan didefinisikan sebagai untaian fragmen-fragmen genom (DNA) dari suatu spesies yang saling berdekatan dan secara bersama-sama mewakili satuan bagian dari DNA (Gregory, 2005). *Chimeric contigs* kemudian diartikan sebagai satu untaian *contig* yang tersusun atas fragmen genom dari dua atau lebih spesies yang berbeda (Scholz dkk, 2020). Untuk meminimalisir peluang terjadinya *chimeric contigs* ini, perlu diterapkan teknik *binning* agar tiap fragmen dari gabungan tersebut dapat dipisahkan sebaik mungkin dari satu sama lain.

Teknik *binning* ini sendiri memiliki dua pendekatan, yakni *binning* berdasarkan homologi dan *binning* berdasarkan komposisi. Pendekatan homologi dilakukan dengan cara menjajarkan sekuensi fragmen metagenom sampel terhadap data sekuensi dari NCBI dan menyimpulkan hasilnya pada tingkat taksonomi. Sementara itu, pendekatan komposisi menggunakan hasil ekstraksi ciri berupa pasangan basa (*base pair*) sebagai masukan dalam pembelajaran model (Simangunsong, 2015).

Ada dua cara pembelajaran untuk model *machine learning*, yakni pembelajaran dengan contoh (*supervised learning*) dan pembelajaran dengan observasi (*unsupervised learning*). Dalam konteks klasifikasi, *supervised learning*

telah memiliki informasi pengkategorian dalam basis pembelajarannya, sedangkan *unsupervised learning* hanya memiliki data latih sebagai basis pembelajaran. Metode yang digunakan dalam penelitian ini termasuk dalam ranah *unsupervised learning*.

Dalam penelitian ini, akan dilakukan pengelompokan fragmen metagenom terhadap data dari sumber NCBI menggunakan metode *k-mer* untuk ekstraksi fitur, metode *linear discriminant analysis* (LDA) untuk reduksi dimensi data, dan algoritma *agglomerative (bottom-up) hierarchical clustering* untuk pengelompokannya. Metode *k-mer* dipilih karena metode ini bekerja dengan cara menghitung jumlah frekuensi kemunculan untaian-untaian pendek (*substring*; polimer) sepanjang *k* huruf dalam satu untaian genom, yang nantinya akan menonjolkan perbedaan karakteristik yang bersifat tidak ambigu berdasarkan perbedaan pada jumlah frekuensi tiap polimer antar sampel individu (Rosen *et al.*, 2008). Metode LDA dipilih untuk digunakan dalam penelitian ini karena metode ini bertujuan untuk mencoba menjelaskan suatu variabel terikat/dependen sebagai keluaran (tingkat taksonomi genus dari data yang diteliti) berdasarkan nilai variabel-variabel bebas/independen sebagai masukan (untaian genom milik data sampel). Sementara itu, metode *agglomerative hierarchical clustering* dipilih atas dasar alur kerjanya yang bersifat *bottom-up* karena analisis fragmen metagenom dimulai dari satuan pasangan basa yang kemudian membentuk beragam untaian panjang dari setiap fragmen berdasarkan frekuensi kemunculan kombinasi pasangan basa tertentu.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang sudah dijabarkan, maka perumusan dari masalah yang terdapat dalam penelitian ini berupa:

1. Bagaimanakah metode *agglomerative hierarchical clustering* diterapkan untuk mengelompokkan fragmen metagenom ke dalam pohon filogenetik tingkat genus?
2. Seberapa besarkah akurasi dari hasil pengelompokan ini?

1.3. Batasan Masalah

Karena kehidupan mikrobial di alam semesta ini beserta pengelompokannya sungguh luas, maka diterapkan beberapa batasan masalah untuk penelitian ini sebagai berikut.

1. Fragmen metagenom yang diteliti berasal hanya dari mikroorganisme jenis bakteri.
2. Mikroorganisme pada penelitian ini dikelompokkan pada tingkat taksonomi genus.
3. Data sampel yang digunakan bersumber dari *National Centre for Biotechnology Information* (NCBI) pada tautan https://www.ncbi.nlm.nih.gov/genome/browse/#!/prokaryotes/.

1.4. Tujuan Penelitian

Tujuan yang hendak dicapai oleh penelitian ini sebagai berikut.

1. Merancang sebuah model *agglomerative hierarchical clustering* dengan metode reduksi dimensi *linear discriminant analysis* untuk mengelompokkan fragmen-fragmen metagenom yang diteliti.
2. Mengukur tingkat akurasi yang diperoleh dan membandingkannya dengan penelitian lain yang terkait.

1.5. Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah agar model ini dapat digunakan sebagai salah satu metode untuk mengelompokkan mikroorganisme jenis bakteri yang tidak dikenali sebelumnya ke dalam kelompok-kelompok yang ada pada tingkat taksonomi genus. Selain itu, keluaran hasil dari model ini juga diharapkan dapat menjadi acuan untuk dicocokkan dengan hasil dari model-model lain yang menggunakan algoritma yang berbeda.