BAB I
PENDAHULUAN

1.1 Latar Belakang

DNA barcoding merupakan penggunaan *barcode Deoxyribonucleic acid* (DNA) atau bagian spesifik dari DNA (Weitschek et al., 2014). Idealnya, satu gen akan efektif dalam semua pengelompokan organisme atau takson yang berbeda, namun, bagian-bagian tertentu dari DNA ditemukan lebih efektif dalam takson yang berbeda (Purty & Chatterjee, 2016). Untuk hewan, *barcode* yang paling efektif adalah fragmen sekitar ~650 *base pairs* (bp) dekat terminus 5′- dari gen *mitochondrial cytochrome c oxidase I* (COI) (Saddhe & Kumar, 2017). Dalam jamur, *barcode* yang lebih tepat adalah urutan ribosom *internal transcribed spacer* (ITS) (Schoch et al., 2012). Dalam spesies tumbuhan, ada beberapa kesulitan dengan pencacahan *barcode*, salah satunya adalah tingkat substitusi nukleotida COI yang rendah (Fazekas et al., 2008). *Consortium for the Barcode of Life* (CBOL) telah merekomendasikan bahwa gen *chloroplast ribulose-1,5-bisphosphate carboxylase large subunit* (*rbcL*) (Gielly & Taberlet, 1994) dan *Maturase K* (matK) (Hilu & Liang, 1997)digunakan sebagai *barcode* tumbuhan (China Plant BOL Group et al., 2011). Tantangan lainnya dalam *DNA barcoding* bagi tumbuhan adalah tingkat keberhasilan identifikasi yang lebih tinggi pada hewan dibandingkan dengan tumbuhan (Fazekas et al., 2009).

Metode untuk mendapatkan *DNA barcode* memiliki banyak variasi tergantung pada tingkat klasifikasinya dalam bidang biologi (takson) (Wong et al., 2014). Secara umum, proses ini melibatkan pengumpulan sampel, isolasi DNA, pencocokan primer tertentu, *polymerase chain reaction* (PCR), analisis kromatogram, memenuhi standar DNA barcode, pengajuan ke *The Barcode of Life Data System* (BOLD), analisis data dan validasi, publikasi dan *hosting* data, dan akhirnya pengguna akhir (Purty & Chatterjee, 2016).

DNA barcoding dapat sangat membantu dalam kehidupan nyata (Fišer Pečnikar & Buzan, 2014). Pendekatan ini dapat digunakan untuk identifikasi hama untuk tujuan *biosecurity* guna melindungi dari potensi spesies invasif (Madden et al., 2019). Pemerintah dapat menggunakan *barcode* DNA untuk memantau perdagangan hewan ilegal pada spesies yang dilindungi (Gonçalves et al., 2015). *Barcode* DNA telah dijelaskan sebagai penambahan yang kuat untuk identifikasi kayu meskipun kualitas DNA yang didapat berkualitas menengah hingga rendah (Jiao et al., 2020). Selain untuk tujuan identifikasi, *barcode* DNA dapat digunakan untuk mengelompokkan spesimen ketika ada keraguan dalam morfologi, seperti karena kurangnya deskripsi fitur morfologi (Tänzler et al., 2012). Ini juga dapat digunakan sebagai alat untuk menentukan apakah spesies yang tidak dikenal harus dikelompokkan dengan suatu kelompok spesies yang diketahui atau sebagai spesies baru berdasarkan *barcode* DNA (Rossini et al., 2016). Ini juga dapat digunakan sebagai pelengkap dataset taksonomik lainnya dalam proses pembatasan batas spesies (Lukhtanov et al., 2016).

Terdapat beberapa kategori pendekatan komputasional untuk menganalisis *barcode* DNA: metode *tree based*, *similarity based*, dan berbasis karakter (Sandionigi et al., 2012). Pendekatan lainnya meliputi kombinasi dan *alignment-free* (Little & Stevenson, 2007; Weitschek et al., 2014; Yang et al., 2022). Setiap pendekatan ini memiliki kelebihan dan kekurangannya masing-masing. Misalnya, metode berbasis kesamaan dan *tree* sangat bergantung pada *sequence alignment*. Metode diagnostik atau berbasis karakter memiliki keberhasilan lebih banyak dibandingkan pendekatan *similarity based* dan *tree*, namun akurasinya masih kurang dibandingkan dengan pendekatan berbasis *supervised machine learning* (Weitschek et al., 2014). Salah satu pendekatan semacam itu memetakan urutan barcode ke dalam vektor berdasarkan frekuensi *k-mer* dan menggunakan pengklasifikasi *random forest* untuk mengidentifikasi sekuens (Meher et al., 2016). Beberapa pendekatan komputasional kontemporer yang digunakan dalam *DNA barcoding* berbentuk *machine learning* (Soueidan & Nikolski, 2015). Hal ini disebabkan oleh kompleksitas dan variabilitas studi yang terlibat dengan genomik.

Dinilai sebagai revolusi untuk penemuan taksonomik, DNA barcoding baru dua dekade lalu diformalkan sebagai alat *history tool* yang lebih luas (DeSalle & Goldstein, 2019). Klasifikasi formal organisme dalam sains Barat berasal dari sekitar tahun 1753 dengan karya Carl Linnaeus (Stevens, 2003). Namun, klasifikasi berbagai organisme sendiri selalu muncul dalam berbagai budaya manusia sepanjang sejarah. Klasifikasi atau lebih tepatnya taksonomi yang diusulkan oleh Linnaeus mengklasifikasikan organisme ke dalam berbagai peringkat dengan masing-masing peringkat menjadi lebih spesifik. Sejak pertama kali dirancang, desain ini telah mengalami banyak revisi dan penyesuaian (Avise & Liu, 2011). Beberapa sumber yang digunakan oleh komunitas ilmiah mendefinisikan hirarki dalam peringkat berikut dalam urutan yang paling hingga paling tidak homogen: *realm*, *subrealm*, kingdom, *subkingdom*, *phylum*, *subphylum*, *class*, *subclass*, *order*, *suborder*, *family*, *subfamily*, *genus*, dan *subgenus* (Greuter et al., 2000; Parker et al., 2019; Schoch et al., 2020; Walker et al., 2019). Namun, karena inkonsistensi sistem tersebut (Avise & Liu, 2011) dan faktor lainnya (Sigwart et al., 2018), ketidaksesuaian dan perselisihan dalam taksonomi muncul (Das, 2012; Docker, 2015; Ji et al., 2020; Ristaino, 2020) seperti dalam kasus *Leguminosae* (Mondal & Mondal, 2011; Patel & Panchal, 2014).

Leguminosae adalah kelompok yang besar dari tanaman berbunga yang penting secara pertanian. Kelompok ini terdiri dari berbagai spesies termasuk tanaman herbal, semak-semak, dan pohon (Patel & Panchal, 2014). Manusia menggunakan kacang polong dalam berbagai cara, termasuk sebagai sumber makanan pokok, pakan hewan, dan pupuk. Selain itu, kacang polong juga digunakan untuk mensintesis banyak produk termasuk perasa, obat, racun, dan pewarna. Kelompok *Plantae* ini juga bermanfaat bagi tanaman lain dengan mengubah nitrogen atmosfer menjadi senyawa nitrogen yang berguna dalam proses biokimia. *Leguminosae* adalah kelompok terbesar ketiga dalam tanaman berbunga setelah *Orchidaceae* (Doyle & Luckow, 2003) dan mencakup 650 genera dengan 18.000 spesies (Polhill & Raven, 1981). Dhakad (2018) menggambarkan kelompok ini sebagai memegang peran penting dalam keanekaragaman hayati dalam ekosistem dan mendominasi sebagian besar jenis vegetasi di dunia. Selain itu,

Leguminosae juga memegang peran penting dalam komposisi hutan dan manajemen tujuan berkelanjutan.

Klasifikasi *Leguminosae* sebagai satu keluarga telah menjadi pengelompokan taksonomi yang diperselisihkan dengan para ahli mengambil beberapa sikap berbeda terhadap masalah ini. Kelompok pertama ahli setuju bahwa *Leguminosae* harus diklasifikasikan sebagai ordo yang berbeda dan diklasifikasikan menjadi tiga famili yang berbeda yaitu *Fabaceae* (*Papilionoid*), *Caesalpiniaceae*, dan *Mimosaceae* (Cronquist, 1981; Hou et al., 1996; I. C. Nielsen, 1992). Kelompok kedua ahli memiliki pandangan bahwa *Leguminosae* adalah suatu famili dengan tiga subfamili yaitu *Mimosoideae*, *Caesalpinioideae*, dan *Papilionoideae* (Hsuan, 1983; Lewis, 2005; Takhtajan, 1980). Perubahan penamaan yang diusulkan oleh beberapa ahli juga dicatat dalam *International Code of Botanical Nomenclature*, salah satunya adalah perubahan *Fabales* menjadi *Fabaceae* (Mondal & Mondal, 2011). Beberapa penelitian terkini yang membahas perselisihan ini oleh Mondal & Mondal (2011) dan Patel & Panchal (2014) setuju bahwa ketiga kelompok tersebut berbeda. Namun, Patel & Panchal (2014) menekankan bahwa perbedaan tersebut dibuat sebagai subfamili yang berbeda dari keluarga yang sama.

Disisi lain Borges dkk (Borges et al., 2013) menjelaskan bahwa klasifikasi tradisional dari famili Leguminosae ke dalam tiga subfamili, yaitu Caesalpinioideae, Mimosoideae, dan Papilionoideae, telah dipertanyakan oleh penelitian terbaru. Peneliti berpendapat bahwa sistem klasifikasi ini sudah usang dan menyesatkan dari perspektif evolusioner, karena tidak mewakili dengan tepat hubungan filogenetik dalam famili tersebut. Mereka juga mengusulkan adanya rekonstruksi ulang terhadap famili tersebut.

Studi ini bertujuan untuk membantu mengklarifikasi perselisihan tentang taksonomi *Leguminosae* dengan memanfaatkan *machine learning* dalam bentuk *hierarchical clustering* dan *barcode* DNA. *Hierarchical clustering* adalah teknik yang *unsupervised* untuk melakukan analisis eksplorasi data. Tujuan utama teknik ini adalah untuk membangun pohon gabungan biner (F. Nielsen, 2016). Teknik ini adalah jawaban pertama atas batasan metode *similarity based* (Sandionigi et al., 2012). Sebuah *dendrogram* merupakan gambar visual dari pengelompokan hirarkis,

memberikan informasi yang kaya untuk evaluasi kualitatif maupun kuantitatif (F. Nielsen, 2016). Oleh karena itu, visualisasi yang dibuat dari *hierarchical clustering* dapat digunakan untuk menilai studi ini. Banyak penelitian dengan *DNA barcoding* terus menggunakan teknik *hierarchical clustering* karena keberadaannya di mana-mana dan relatif sederhana (An et al., 2022; Lucas et al., 2012; Nikitina et al., 2021, 2022; Papa et al., 2021; Xu et al., 2021; Zhao et al., 2018). Dalam penelitian ini, *hierarchical clustering* dengan *barcode* DNA akan digunakan untuk mencapai dua tujuan. Pertama, kegunaan *hierarchical clustering* diuji dengan *distance method* yang berbeda untuk masalah ini. Kedua, metode yang telah divalidasi digunakan untuk menentukan pengelompokan dalam taksonomi *Leguminosae*. Ini dilakukan untuk menentukan pandangan mana tentang taksonomi *Leguminosae* yang didukung oleh hasil *hierarchical clustering*. Singkatnya, studi ini bertujuan untuk menjelaskan apakah *Leguminosae* harus diklasifikasikan sebagai tiga keluarga yang berbeda, subfamili, atau permutasi lainnya.

Machine learning adalah bidang interdisipliner. Bidang ini mengambil wawasan dari berbagai disiplin ilmu, termasuk kecerdasan buatan, probabilitas dan statistik, teori kompleksitas komputasional, teori kontrol, teori informasi, filsafat, psikologi, dan bahkan neurobiologi. Dalam berbagai domain, algoritma *machine learning* terbukti sangat berguna, misalnya, dalam domain pengenalan suara berbasis algoritma pembelajaran mesin mengungguli pendekatan lain yang telah dicoba (Mitchell, 1997). Pendekatan *machine learning* memiliki kelebihan belajar berdasarkan hasil *training* dan tidak perlu secara manual memperhitungkan banyak variasi yang ditemukan dalam data genetik. Kategori pendekatan dalam literatur tidak memiliki skema penamaan yang konsisten. Beberapa artikel merujuk *machine learning* sebagai kategori terpisah dari *tree based*, *distance based*, dan berbasis karakter (He et al., 2019; Meher et al., 2016; Yang et al., 2022), meskipun beberapa pendekatan dalam kategori lain juga merupakan *machine learning*, meskipun sebagian besar *unsupervised* seperti *hierarchical clustering* atau dengan kata lain pendekatan *tree based* (Sandionigi et al., 2012; Soueidan & Nikolski, 2015).

1.2 Rumusan Masalah

Rumusan masalah dari latar belakang masalah yang telah dipaparkan pada sub bab sebelumnya adalah sebagai berikut:

1. Bagaimana model komputasi proses *clustering* berdasarkan DNA Barcode menggunakan *hierarchical clustering*?
2. Bagaimana perbandingan performa dari beberapa *distance method* yang dipilih untuk *DNA barcoding* pada *hierarchical clustering*?
3. Bagaimana performa *hierarchical clustering* dalam melakukan *clustering* dalam kasus *DNA barcoding*?
4. Bagaimana klasifikasi kelompok *Leguminosae* berdasarkan hasil *hierarchical clustering* yang dilakukan?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang dijelaskan sebelumnya, makan tujuan dari penelitian ini adalah sebagai berikut

1. Merancang dan mengimplementasikan *hierarchical clustering* untuk melakukan *clustering* kelompok tumbuhan.
2. Membandingkan beberapa *distance method* yang dipilih untuk menentukan *distance method* yang terbaik dalam melakukan *clustering* berdasarkan data kelompok yang tidak bermasalah.
3. Melakukan eksperimen, analisis, dan evaluasi performa *hierarchical clustering* pada *DNA barcoding*.
4. Melakukan klarifikasi dari kasus sengketa pada kelompok *Leguminosae.*

1.4 Manfaat Penelitian

Penelitian ini dilakukan dengan harapan dapat memberikan manfaat sebagai berikut:

1. Menginvestigasi penggunaan *hierarchical clustering* pada studi kasus menentukan taksonomi tanaman yang bermasalah.

2. Memberikan alternatif metode dengan menggunakan *hierarchical clustering* dalam melakukan klarifikasi saat menentukan taksonomi tanaman.

3. Membuat model komputasi yang dapat dikembangkan pada penelitian selanjutnya.

1.5 Batasan Masalah

Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Sekuens DNA yang dikenali adalah nukleo basa (adenine(A), guanine(G), cytosine(C), thymine(T)).

2. Sekuens DNA yang digunakan sebagai barcode adalah *internal transcribed spacer* (ITS).

3. Data utama yang digunakan diambil dari *National Center for Biotechnology Information* (NCBI) menggunakan library *rentrez*.

4. Data yang digunakan menggunakan masukan dengan format FASTA.

5. Spesies yang digunakan dalam studi kasus merupakan kelompok dari famili *Leguminosae* atau famili *Fabaceae* (*Papilionoid*), *Caesalpiniaceae*, and *Mimosaceae*.