IMPLEMENTASI METODE HIERARCHICAL CLUSTERING PADA DATA GENETIK MIKROARRAY

Humasak T.A. Simanjuntak[1)]

[1)] *Sistem Informasi, Institut Teknologi Del*
Jl Sisingamangaraja, Sitoluama, Laguboti, Toba Samosir, 22381
Email : humasak@gmail.com[1]

Abstrak

Teknologi Mikroarray menghasilkan data genetik mikroarray, yang jumlahnya akan semakin banyak dari hari ke hari. Jika dikelola dengan baik, data genetik mikroarray yang dihasilkan akan sangat bermanfaat dalam bidang kesehatan. Namun, jumlah data yang terus bertambah akan mempersulit dalam pengambilan informasi yang berguna.

Pada kajian ini, data genetik mikroarray dikelola dengan menggunakan metoda Agglomerative Hierarchical Clustering. Dengan metoda ini, data genetik mikroarray dikelompokkan ke dalam clustering-clustering. Gen-gen dengan ekspresi gen yang mirip akan dikelompokkan ke dalam satu clustering. Hasil metoda Agglomerative Hierarchical Clustering adalah dendogram yang menggambarkan clustering-clustering dan hubungan antar clustering tersebut.

Algoritma yang digunakan pada Hierarchical Clustering adalah Hierarchical Clustering Algorithm. Pada algoritma ini dibutuhkan Proximity Matrix yang menyajikan jarak antar gen. Jarak antar gen dapat dihitung dengan menggunakan tiga fungsi yaitu Euclidean Distance, Manhattan Distance dan Pearson Correlation. Gen–gen dengan jarak terdekat akan digabungkan dalam satu clustering. Untuk menghitung jarak antar cluster yang baru terbentuk dengan gen-gen lain dapat digunakan tiga algoritma yaitu Single Linkage, Complete Linkage dan Average Linkage. Clustering yang dihasilkan dengan menggunakan algoritma penghitungan jarak antar clustering yang berbeda, akan mempengaruhi kepadatan hasil clustering dan kesamaan hasil clustering.

Kata kunci: *Clustering, Agglomerative Hierarchical Clustering, Proximity Matrix, Mikroarray.*

1. Pendahuluan

Teknologi microarray menghasilkan data genetik mikroarray. Mikroarray adalah array dari molekul DNA yang memungkinkan eksperimen pencangkokan dilakukan secara paralel. Dengan menggunakan mikroarray dapat diketahui tingkat ekspresi gen secara bersamaan. Pertumbuhan jumlah data genetik mikroarray yang dihasilkan sangat cepat dari hari ke hari. Untuk mendapatkan informasi yang berguna dari data dalam jumlah besar ini, maka perlu dilakukan pengolahan data. Informasi yang diperoleh dari data genetik mikroarray dapat dimanfaatkan dalam bidang kesehatan, misalnya untuk menemukan gen tertentu yang dapat menyebabkan suatu penyakit [1].

Salah satu cara yang dapat digunakan untuk memperoleh informasi yang berguna dari data genetik mikroarray adalah dengan melakukan *clustering*. *Clustering* adalah proses pengelompokan kumpulan dari objek-objek yang anggotanya memiliki kesamaan menjadi *class-class* dari objek yang sama. Tujuan *clustering* adalah mengelompokkan sekumpulan data atau objek yang memiliki kesamaan dalam satu *cluster* dan berbeda dengan objek pada *cluster* lain [2].

Salah satu metoda pada *clustering* adalah *Hierarchical Clustering Method*. Pada *Hierarchical Clustering Method*, *clustering* dilakukan dengan mengelompokkan objek-objek ke dalam *tree of clustering*. Dengan menggunakan metoda ini, penyajian data genetik mikroarray dalam jumlah yang besar dapat lebih mudah dipahami. Oleh karena itu, tujuan utama kajian ini adalah melakukan implementasi *cluster*ing pada data genetik mikroarray dengan metoda Hierarchical *Clustering* untuk memperoleh pengelompokan gen-gen yang memiliki kesamaan dan membandingkan setiap *clustering* yang dihasilkan.

2. Landasan Teori

2.1 *Clustering*

Clustering merupakan pengelompokan data objek ke dalam *class* yang memiliki kesamaan. Kesamaan (*similarity*) diantara objek–objek ditaksir menurut *Distance Measure*. Hasil dari *clustering* adalah *cluster*, yang memiliki sekumpulan objek dengan kesamaan tertentu dalam satu *clustering* dan berbeda dengan objek pada *cluster* lain. Aktivitas untuk melakukan *clustering* disebut *Clustering analysis* [2].

Clustering termasuk *unsupervised classification* (tidak mengetahui label dari class dan jumlah class yang akan dibentuk). Salah satu penggunaan *clustering* adalah pada data genetik mikroarray [2].

Metoda *clustering* dapat diklasifikasikan atas 5 kategori, antara lain, *Partitioning Method, Hierarchical Method, Density-based Method, Grid-based Method, Model-based Method.* Yang menjadi fokus pada kajian ini adalah *Hierarchical Method.*

2.2 *Hierarchical Clustering Method*

Hierarchical Clustering Method (HCE) merupakan salah satu metoda yang digunakan dalam melakukan *clustering*. *Clustering-clustering* yang dibentuk dari HCE berupa sebuah hirarki (tree). Pada hierarchial *clustering*, data tidak dibagi menjadi sebuah *clustering* dalam satu langkah tetapi dalam serangkaian *clustering*. Pembagian dapat dilakukan dari sebuah *clustering* yang berisi semua objek ke beberapa *clustering* dalam jumlah n dimana masing-masing mengandung sebuah objek [2].

Kualitas dari *hierarchial clustering method* dipengaruhi oleh kemampuan metode ini untuk membuat keputusan yang tepat dan sesuai ketika menggabung atau memisah *clustering*. Ketika keputusan untuk menggabung atau memisah telah dilakukan, maka keputusan tersebut tidak dapat dirubah lagi.

Metoda *Hierarchial clustering* diklasifikasikan dalam dua jenis yaitu *Agglomerative* dan *Divisible Hierarchial Clustering* tergantung pada komposisi hirarkis dilakukan dari bentuk bottom-up atau top-up. Metode yang digunakan pada kajian ini adalah Metoda agglomerative Hierarchical *Clustering*. Metode ini dilakukan dengan serangkaian pengelompokan objek (banyak objek menjadi satu bagian). Dimulai dengan menempatkan setiap objek pada *clustering* dan kemudian menggabungkan struktur *clustering* tersebut menjadi *clustering* yang lebih besar sampai semua objek berada pada satu *clustering* yang sama atau sampai memenuhi kondisi yang telah ditentukan. Algoritma ini dilakukan dengan pendekatan bottom-up [2].

Algoritma yang digunakan pada Hierarchial *Clustering* Method adalah *Hierarchical Clustering Algorithm*. Pada algoritma tersebut, jika diberikan sejumlah N objek yang akan di-*clustering* dengan proximity matrix N*N, maka [4]:

a. Diawali dengan membuat setiap objek menjadi satu *clustering*. Jika tersedia sejumlah N objek yang akan di-*clustering*, maka akan terbentuk *clustering* sebanyak N yang berisi hanya satu objek.
b. Cari pasangan *clustering* yang terdekat dan digabungkan menjadi satu *clustering*. *Clustering clustering* dapat digabungkan menjadi satu *clustering* jika memiliki objek dengan kesamaan terdekat terhadap objek pada *clustering* lain.
c. Hitung *distance* atau kesamaan antara *clustering* yang baru dibentuk dan *clustering-clustering* lain yang sudah ada sebelumnya.
d. Ulangi langkah b dan c sampai semua objek yang di-*clustering* berada pada satu *clustering*.

2.2.1 Penghitungan Jarak Antar Gen

Pada algoritma *Hierarchical Cluctering Method* di atas, yang menjadi masukkan adalah proximity matrix. Proximity Matrix diperoleh dengan melakukan penghitungan jarak antar gen. Penghitungan jarak antar gen dapat dilakukan dengan tiga cara yaitu *Euclidean Distance, Manhattan Distance* dan *Pearson Correlation.*

a. Euclidean Distance

Euclidean Distance mengukur jarak ketidaksamaan antar dua profil. Semakin besar angka yang dihasilkan, maka semakin kecil kesamaan antar profil tersebut dan sebaliknya. Fungsi Euclidean Distance mengukur jarak lurus (*the 'as-the-crow-flies' distance*). Rumus untuk jarak antara sebuah titik X (X1, X2,…) dan sebuah titik Y (Y1,Y2,...)[5] adalah :

$$dE(Xi,Yi) = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2} \quad \text{..... (1)}$$

Dengan n adalah jumlah variabel, dan Xi, Yi adalah nilai dari variabel ke-i pada titik X dan Y secara berturut-turut. Memperoleh Euclidean Distance antara dua titik meliputi penghitungan akar pangkat dua dari jumlah pengkuadratan selisih antara nilai-nilai yang sesuai [5].

b. Manhattan Distance

Fungsi Manhattan Distance menghitung jarak yang ditempuh untuk sampai dari satu titik ke titik lain jika sebuah *grid-like path* diikuti. Manhattan Distance antara dua titik adalah jumlah dari selisih komponen-komponen yang sesuai. Rumus untuk jarak antara sebuah titik X (X1, X2,…) dan sebuah titik Y (Y1,Y2,...) [5] adalah :

$$dM(Xi,Yi) = \sum_{i=1}^{n} |Xi - Yi| \quad \text{..... (2)}$$

Dengan n adalah jumlah variabel, dan Xi, Yi adalah nilai dari variabel ke –i pada titik X dan Y secara berturut-turut [5].

c. Pearson Correlation

Pearson correlation menghitung kesamaan (*similarity*) antara dua profil dengan melihat kesamaan naik turunnya ekpresi dua profil. Dua profil dikatakan mirip, jika level ekspresi profil meningkat dan menurun pada waktu yang sama [6]. Rumus untuk jarak ketidaksamaan antara sebuah titik X (X1, X2,…) dan sebuah titik Y (Y1,Y2,...) [7] adalah :

$$dPC(Xi, Yi) = 1 - \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum (Xi - \bar{X})^2 \sum (Yi - \bar{Y})^2}} \text{.....(3)}$$

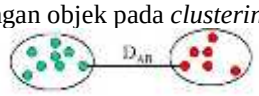
Dengan n adalah jumlah variabel, dan Xi, Yi adalah nilai dari variabel ke –i pada titik X dan Y secara berturut-turut adalah rata-rata variabel pada titik X dan adalah rata-rata variabel pada titik Y [7].

Sesuai dengan rumus penghitungan jarak (3), hasil dari penghitungan dengan Pearson Correlation menghasilkan ketidaksamaan (*dissimilarity*) karena pada rumus tersebut telah dilakukan 1- koefisien correlation.

2.2.2. Penghitungan Jarak Antar *Clustering*

Proximity matrix yang diperoleh sebagai masukkan untuk algoritma Hierarchical *Clustering* Method, tepatnya pada langkah ketiga dapat dilakukan dengan tiga cara berbeda, yaitu [1]:

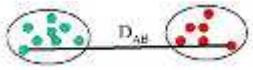
a. *Single-Linkage*, distance antar *clustering* ditentukan melalui jarak minimum antara objek pada satu *clustering* dengan objek pada *clustering* lain.



D_{AB} = min (d (A,B))

Gambar 1 Menghitung distance antar *clustering* dengan *Single-Linkage*

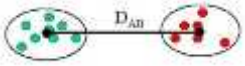
b. *Complete-Linkage*, distance antar *clustering* ditentukan melalui jarak maksimal antara objek pada satu *clustering* dengan objek pada *clustering* lain.



D_{AB} = max (d (A,B))

Gambar 2 Menghitung distance antar *clustering* dengan Complete-Linkage

c. *Average-Linkage*, *distance* antar *clustering* ditentukan melalui rata-rata jarak antar semua objek yang ada pada satu *clustering* dengan semua objek pada *clustering* lain.



$$D_{AB} = \frac{dA + dB}{2}$$

Gambar 3 Menghitung distance antar *clustering* dengan *Average-Linkage*

Proximity Matrix N*N adalah D = [d(i,j)]. Setiap melakukan *clustering* diberikan *sequence number* yang dimulai dengan 0, 1,...,(n-1), dan L(k) yang merupakan level dari *clustering* k yang dibentuk. Sebuah *clustering* dengan *sequence number* disimbolkan dengan m dan kedekatan antara r dan s adalah d[(r),(s)].

2.3 Data Genetik Mikroarray

Microarray adalah array dari molekul DNA yang memungkinkan eksperimen pencangkokan dilakukan secara paralel. Dengan menggunakan microarray dapat mengetahui tingkat ekspresi dari ribuan gen secara bersamaan. Keluaran dari teknologi mikroarray adalah matriks dengan baris yang berasosiasi dengan gen. Setiap baris merepresentasikan pola ekspresi gen. Setiap kolom merepresentasikan eksperimen. Masukan pada baris data matriks adalah nilai rasio, absolut atau nilai distribusi [1].

Gen-gen yang akan di-*clustering* adalah ekpresi gen ragi Saccharomyces cerevisiae yang diambil pada waktu yang berbeda. Data tersebut diambil pada waktu tertentu pada siklus pembelahan sel setelah sinkronisasi oleh alpha factor (18 titik waktu) dan dengan *temperature-sensitive* cdc 15 mutant (CDC15, 27 titik waktu), *centrifugal elutriation* (14 titik waktu), CLN3 *induction* (3 titik waktu), sporulasi (13 titik waktu), dan *diauxic shift* (7 titik waktu).

Pada Tabel 1, diberikan 6 record eksperesi gen pada 6 ekperimen, yang merupakan data genetik mikroarray dari Eisen Lab sebagai data ekspresi yang akan digunakan dalam kajian ini.

Tabel 1 Data Genetik Mikroarray - Eisen Lab

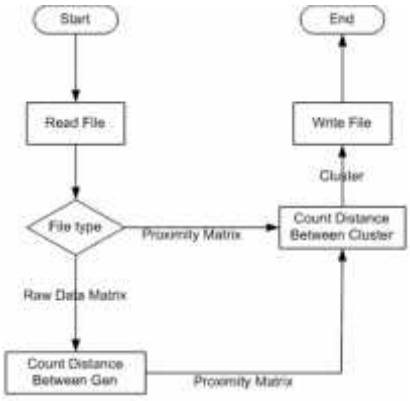
YORF	NAME	GWEIGHT	Cell-cycle Alpha-Factor 1	Cell-cycle Alpha-Factor 2	Cell-cycle Alpha-Factor 3	Cell-cycle Alpha-Factor 4	Cell-cycle Alpha-Factor 5	Cell-cycle Alpha-Factor 6
EWEIGHT			1	1	1	1	1	1
YHR051W	YHR051W COX6 oxidative phosphorylation cytochrome-c oxidase subunit VI S0001093	1	0.03	0.3	0.37	0.38	-0.14	-0.12
YKL181W	YKL181W PRS1 purine, pyrimidine, tryptophanphosphoribosylpyrophosphate synthetase S0001664	1	0.33	-0.2	-0.12	-0.01	0.07	-0.07
YHR124W	YHR124W NDT80 meiosis transcription factor S0001166	1	0.36	0.08	0.06	-0.3	0	-0.23
YHL020C	YHL020C OPI1 phospholipid metabolism negative regulator of phospholipid biosynthesS0001012	1	-0.01	-0.03	0.21	-0.1	0.06	0.25
YGR072W	YGR072W UPF3 mRNA decay, nonsense-mediated unknown S0003304	1	0.2	-0.43	-0.22	-0.36	-0.15	-0.42
YGR145W	YGR145W unknown unknown; similar to MESA gene of Plasmodium fS0003377	1	0.11	-1.15	-1.03	-0.76	0.03	-0.3

Pada Tabel 1, baris pada tabel merepresentasikan gen-gen dan kolom pada tabel merepresentasikan pengamatan. Setiap baris (gen) memiliki sebuah pengenal (identifier) yang selalu berada di kolom pertama.

3. Analisis Pembangunan Aplikasi Hierarchical *Clustering*

Sebelum dilakukan implementasi Hierarchical *Clustering*ing Method pada data genetik mikroarray, maka terlebih dahulu dilakukan analisis terhadap aplikasi yang dibangun. Implementasi dilakukan dengan

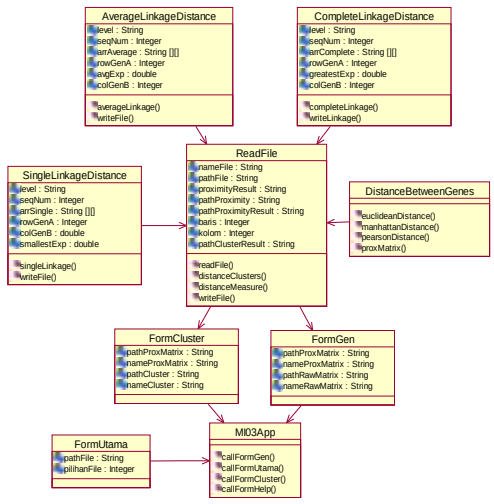
menggunakan bahasa pemrograman Java dan NetBeans sebagai *development tool.* Alur proses pada aplikasi Hierarchical *Clustering* MI03, adalah:



Gambar 4 Flowchart Hierarchical *Clustering* MI03

File yang menjadi masukan untuk proses *clustering* adalah file dengan format .txt dengan menggunakan tab-delimited. File yang menjadi masukan dapat berisi data genetik mikroarray (raw data matrix) atau proximity matrix. Aplikasi membaca file masukan. Jika file yang dimasukkan adalah raw data matrix, maka akan dilakukan penghitungan jarak antar gen untuk menghasilkan proximity matrix. Proximity Matrix digunakan sebagai masukan untuk proses *clustering*. Jarak antar *clustering* dihitung dengan menggunakan tiga pilihan algoritma yaitu Single-Linkage, Complete-Linkage atau Average-Linkage. Jika file masukan adalah Proximity Matrix, maka dilakukan penghitungan jarak antar *clustering*. Kemudian, hasil *clustering* ditulis (disimpan) ke dalam sebuah file bertipe text. File yang dihasilkan menampikan *clustering* yang terbentuk dengan menggunakan *new line* sebagai pemisah antar *clustering*.

Rancangan kelas yang digunakan dalam implementasi Hierarchical *Clustering* MI03 ditunjukkan pada Gambar 5. Dalam implementasi ini, dibutuhkan sebanyak sembilan kelas.



Gambar 5 *Class Diagram Hierarchical Clustering* MI03

4. Pembahasan

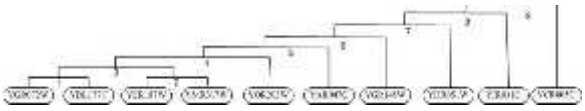
Hasil dari proses *clustering* dengan menggunakan aplikasi Hierarchical *Clustering* MI03 dibahas dari segi kepadatan *clustering* yang terbentuk, kecenderungan terbentuknya *clustering* yang sama, dan pengaruh jumlah data yang di-*clustering* terhadap hasil *clustering*.

Hasil *clustering* diperoleh dengan menggunakan semua kombinasi rumus penghitungan jarak antar gen dan penghitungan jarak antar *clustering*.

4.1 Kepadatan *Clustering*

Berdasarkan gambar *tree of clustering* yang dihasilkan, maka hasil *clustering* yang diperoleh berbeda-beda. Tetapi, dalam hal kepadatan *clustering* yang dihasilkan, yang sangat mempengaruhi adalah pemilihan algoritma penghitungan jarak antar *clustering* yaitu Single Linkage, Complete Linkage atau Average Linkage.

Pengukuran jarak antar *clustering* dengan menggunakan metoda Single Linkage menghasilkan *clustering* yang lebih jarang. Artinya, kemungkinan pembentukan *clustering* yang saling lepas pada setiap *sequence number* semakin kecil. Hasil *clustering* dengan menggunakan kombinasi Euclidean Distance–Single Linkage pada data 10 gen dapat dilihat pada Gambar 6.

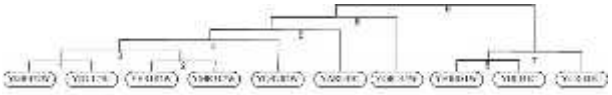


Gambar 6 Hasil *clustering*ing Euclidean Distance – Single Linkage pada 10 gen

Pada Gambar 6 dapat dilihat bahwa pembentukan *clustering* yang lepas pada satu sama lain terbentuk di dua *sequence number*, yaitu YGR072W/YDL177C dan YER187W/YMR317W. Tetapi pada proses *clustering* yang berikutnya, kedua gen yang saling lepas tersebut digabungkan menjadi satu *clustering*. Kemudian *clustering* yang baru terbentuk ini, digabungkan kembali dengan *clustering* yang terdiri dari satu gen saja.

Hal tersebut juga terjadi pada kombinasi Single Linkage yang dikombinasikan dengan Manhattan Distance dan Pearson Correlation. Hal ini terjadi karena Single Linkage adalah metoda yang melihat jarak terkecil antara *clustering* yang baru terbentuk dengan *clustering* lainnya. Sehingga akan semakin besar kemungkinan untuk menggabungkan *clustering* yang baru terbentuk dengan *clustering* lainnya, karena penggabungan *clustering* selalu melihat jarak terkecil antar *clustering* (*clustering* dengan kemiripan yang tinggi).

Tetapi dengan menggunakan metoda Complete Linkage menghasilkan *clustering* yang lebih padat. Artinya, pembentukan *clustering* yang saling lepas satu sama lain akan semakin besar. Pada setiap tahap proses *clustering*, sering terjadi proses *clustering* pada *clustering* yang berbeda-beda.



Gambar 7 Hasil *clustering* Euclidean Distance – Complete Linkage pada 10 gen

Pada Gambar 7 dapat dilihat bahwa pembentukan *clustering* yang saling lepas satu sama lain terbentuk di tiga *sequence number* yaitu YGR072W/YDL177C, YER187W/YMR317W, dan YHR051W/YIR031C. Kemudian, *clustering* ini digabungkan kembali dengan gen-gen lainnya.

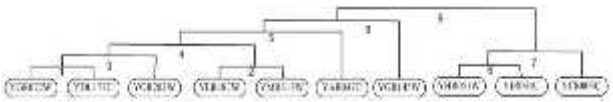
Demikian pula hasil dengan menggunakan Complete Linkage yang dikombinasikan dengan Manhattan Distance dan Pearson Correlation. Hal ini terjadi karena Complete Linkage adalah metoda yang melihat jarak terjauh antara *clustering* yang baru terbentuk dengan *clustering* lainnya. Sehingga semakin kecil kemungkinan untuk menggabungkan *clustering* yang baru terbentuk dengan *clustering* lainnya. Jarak antara *clustering* yang baru terbentuk dengan *clustering* lain semakin besar. Padahal, penggabungan *clustering* dilakukan dengan melihat jarak terkecil.

Metoda penghitungan Average Linkage cenderung sama dengan Complete Linkage yaitu menghasilkan *clustering* yang padat. Dari hasil *clustering* yang telah ditampilkan, dapat disimpulkan bahwa yang mempengaruhi kepadatan hasil *clustering* yang terbentuk adalah metoda pengukuran jarak antar *clustering* dan bukan metoda pengukuran jarak antar gen.

Kesamaan pola hasil *clustering* yang jarang juga terjadi pada hasil *clustering* pada data 20 gen dengan kombinasi Euclidean–Single Linkage dan Euclidean Distance-Complete Linkage

4.2 Kecenderungan Hasil *Clustering* yang sama

Apabila menghitung jarak antar gen dengan menggunakan rumus Euclidean Distance, maka hasil *clustering* cenderung lebih mirip dengan Manhattan Distance. Sedangkan jika menggunakan Pearson Correlation, hasil *clustering* cenderung berbeda dengan Euclidean Distance dan Manhattan Distance. Salah satu contoh hasil *clustering* akhir dengan menggunakan metode Complete Linkage untuk percobaan terhadap data 10 gen ditunjukkan pada Gambar 8.



Gambar 8 Hasil *clustering* Manhattan Distance-Complete Linkage pada 10 gen

Hasil *clustering* antara Manhattan Distance dengan Euclidean Distance cenderung lebih mirip karena kedua rumus tersebut sama-sama menghitung jarak ketidaksamaan antar gen. Sedangkan hasil *clustering* dengan menggunakan Pearson Correlation berbeda dengan kedua rumus lainnya.

Proximity Matrix yang dihasilkan dengan penghitungan Euclidean Distance dan Manhattan Distance merepresentasikan jarak ketidaksamaan antar gen. Semakin kecil angka yang merepresentasikan jarak antar gen, semakin besar kesamaan kedua gen tersebut, dan sebaliknya. Sedangkan proximity matrix yang dihasilkan dengan penghitungan Pearson Correlation merepresentasikan jarak kesamaan antar gen. Semakin besar angka yang merepresentasikan jarak antar gen, semakin besar kesamaan kedua gen tersebut.

Hasil *clustering* dengan menggunakan metode penghitungan jarak antar *clustering* cenderung berbeda untuk setiap metode yaitu Single Linkage, Complete Linkage dan Average Linkage. Apabila data yang digunakan semakin banyak, maka semakin kecil kemungkinan kesamaan hasil *clustering* diantara ketiga metode tersebut. Hal ini disebabkan karena apabila data semakin banyak, maka variasi jarak antar gen semakin tinggi.

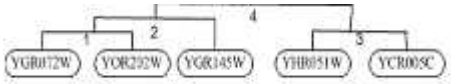
Ketika menggunakan Single Linkage, Complete Linkage, dan Average Linkage hasil *clustering* cenderung sama hanya pada *sequence number* pada awal proses. Pada hasil selanjutnya, masing-masing metode menghasilkan *clustering* yang berbeda.

4.3 Jumlah Data

Jumlah data gen yang di-*clustering* dapat mempengaruhi kesamaan hasil akhir yang diperoleh dengan menggunakan Single Linkage, Complete Linkage dan Average Linkage. Semakin banyak data yang digunakan, maka semakin kecil kemungkinan *clustering* yang dihasilkan sama pada hasil *clustering* antara Single Linkage, Complete Linkage dan Average Linkage. Sebaliknya, apabila data gen yang digunakan sedikit, maka semakin besar kemungkinan *clustering* yang dihasilkan akan sama pada *clustering* antara Single Linkage, Complete Linkage dan Average Linkage.

Pada kajian ini, hasil akhir *clustering* terhadap 5 gen dapat diperoleh 7 *clustering* yang sama, sedangkan pada 10 gen dapat diperoleh 4 *clustering* yang sama.

Salah satu contoh hasil akhir *clustering* yang sama antara Single Linkage, Complete Linkage dan Average Linkage pada percobaan 5 gen dapat dilihat pada Gambar 9.

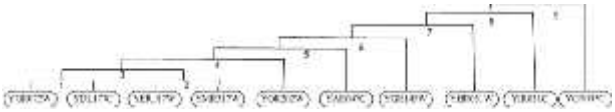


Gambar 9 Euclidean Distance – Average Linkage, Manhattan Distance - Complete Linkage dan Average Linkage

Pada hasil *clustering* 5 gen dengan Single Linkage, Complete Linkage dan Average Linkage, maka diperoleh 7 *clustering* yang sama, dimana hasil akhir *clustering* Euclidean Distance–Average Linkage sama dengan hasil akhir *clustering* pada Manhattan Distance–Complete Linkage dan Manhattan Distance–Average Linkage. Hasil akhir *clustering* Euclidean Distance–

Single Linkage sama dengan hasil akhir *clustering* pada Manhattan Distance–Single Linkage dan hasil akhir *clustering* Pearson Correlation–Complete Linkage sama dengan hasil akhir *clustering* pada Pearson Correlation–Average Linkage

Salah satu contoh hasil *clustering* yang sama antara Single Linkage, Complete Linkage dan Average Linkage pada percobaan 10 gen ditunjukkan pada Gambar 10.



Gambar 10 Euclidean Distance–Single Linkage dan Complete Linkage

Pada hasil *clustering* 10 gen dengan Single Linkage, Complete Linkage dan Average Linkage, maka diperoleh 4 *clustering* yang sama dimana hasil *clustering* Euclidean–Single Linkage sama dengan hasil *clustering* Euclidean Distance-Complete Linkage, dan hasil akhir *clustering* Euclidean Distance-Average Linkage sama dengan hasil *clustering* Manhattan Distance-Average Linkage.

Hasil tersebut menunjukkan apabila data semakin banyak, maka variasi jarak antar gen akan semakin tinggi sehingga penghitungan jarak antar *clustering* akan berbeda khususnya pada Single Linkage dan Complete Linkage.

4.4 Metode Linkage

Pada data genetik mikroarray Eisen Lab yang digunakan pada kajian ini, ekspresi gen pada setiap eksperimen cenderung berdekatan (fluktuasi ekspresi gen teratur). Hal ini mengakibatkan jarak antar gen tidak terlalu besar atau dapat dikatakan kemiripan antar gen tinggi. Apabila ingin mengetahui gen-gen yang mirip yaitu gen yang memiliki jarak berdekatan dengan gen yang lain, maka metode linkage (penghitungan jarak antar *clustering*) yang lebih baik digunakan adalah Single Linkage.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan implementasi yang telah dilakukan, maka dapat disimpulkan bahwa:

1. Clustering yang dilakukan oleh Aplikasi Hierarchical *Clustering* MI03 dengan menggunakan Single Linkage pada data genetik mikroarray Saccharomyces cerevisiae memberikan hasil yang lebih baik.
2. Hasil akhir *clustering* dengan kombinasi algoritma yang berbeda dapat sama tetapi urutan gen yang dijadikan satu *clustering* untuk setiap *sequence number* pada proses *clustering* berbeda.
3. Pemilihan metode penghitungan jarak antar *clustering* mempengaruhi kepadatan hasil *clustering*. *Clustering* yang dihasilkan oleh Complete Linkage lebih padat jika dibandingkan dengan Single Linkage. *Clustering* yang dihasilkan Average Linkage cenderung sama dengan Complete Linkage.

5.2 Saran

Berdasarkan implementasi yang telah dilakukan, saran yang dapat disampaikan untuk pengembangan selanjutnya adalah agar aplikasi yang dihasilkan pada kajian ini dapat digunakan untuk melakukan *clustering* data mikroarray dengan format ekspresi yang berbeda dengan data mikroarray yang dihasilkan oleh Eisen Lab.

Daftar Pustaka

[1] http://compbio1.utmem.edu/MSCI814/Module10.htm, Yan Cui: "*Module 10: Microarray Data Analysis I*", 27 Februari 2009
[2] J. Han and K Micheline, *Data Mining Concepts and Technique*, Morgan Kaufmann, 2001.
[3] L, Rosni, *Studi dan Implementasi Teknik Clusteringing untuk Data Genetik Microarray*, Institut Teknologi Bandung, 2007.
[4] http://home.dei.polimi.it/matteucc/*Clustering*/,S.C.Johnson: "*Hierarchical Clustering Algorithm*", 23 Maret 2009.
[5] http://www.improvedoutcomes.com/index.html/: "Euclidean", 6 Mei 2009.
[6] Korol Abraham: "*Microarray Cluster Analysis and Application*", 2009.
[7] http://www.improvedoutcomes.com/index.html/ : "Pearson Correlation and Pearson Squared", 6 Mei 2009.

Biodata Penulis

Humasak T.A. Simanjuntak, lahir di Pematang Siantar, 26 April 1983. Pada tahun 2007, menyelesaikan studi S1 Teknik Informatika pada Institut Teknologi Bandung dan menyelesaikan program S2 (*Master of Information System Development*) di HAN University of Applied Sciences, The Netherland pada tahun 2010. Mulai tahun 2007 sampai sekarang menjadi tenaga pengajar di Institut Teknologi Del. Beberapa mata kuliah yang diajarkan adalah *Database Administration,Database System, Data Processing, E-Business*, Framework Component Based Software Development.