# Graph Application in Genome Circular Ideogram

Nada Afra Sabrina 13517118
*Program Studi Teknik Informatika*
*Sekolah Teknik Elektro dan Informatika*
*Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia*
*nadaafrs@students.itb.ac.id*

***Abstract*—Genomes are whole of genetic information in a cell. Genomes sequencing are developed to use in many sectors, especially in medical sector. Circular ideogram is one of the ways to represent genome sequences. Circular ideogram have same concept with graph in represent data. With genome circular ideogram, we can compare every sequence of genome easily.**

***Keywords*—Graph, Circular ideogram, DNA sequencing, Genome.**

## I. Introduction

Living organism built from cells. A cell is smallest unit of life. Besides perform the living process of organisms, cells also store the genetic information that inherited the characteristic of organism. All genetic information in a cell is called genome. The genetic information stored in nitrogenous bases that arranged to build *deoxyribonucleic acid* or DNA. DNA is a part of nucleic acid.

Nowadays technology makes it possible to perform DNA sequencing. DNA sequencing is process to get the arrangement of nitrogenous bases. The result of DNA sequencing can be classified to make genome sequencing. Genome sequencing can be served in genome circular ideogram.

Circular ideogram is a representation of indirect graph. Graph is a set of vertices that connected by edges. In genome circular ideogram, the outer line contains vertices that representing genetic information. Genetic information that show same characteristic is connected by edges inside the circle. This paper will discuss about concept of graph in genome circular ideogram and the usability of the genome circular ideogram.

## II. Basic Theory

### A. Graph

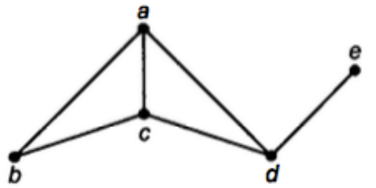
A graph can be defined by G = (V, E), V is a set of vertices and E is a set of edges. In graph, set of vertices (V) can't be empty, in other hand, E can be empty. Two vertices that connected by an edge, then the two vertices neighbor each other. Graph can be presented in many ways:

1. Adjacency Matrix
   If vertices ( i, j) are neighbor to each other, then matrix element in column i and row j, also matrix element in column j and row i, annotated by number 1. If the vertices are not connected, then annotated by 0.

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Picture 2.1 matrix representation graph
(from Munir, 2010, p.382)

2. Enumeration of Vertices and Edges
   Vertex symbolize by number or alphabet. Then edge symbolize by mentioning 2 vertices that connected by the edge. For example V= {A, A1, B1, C2, 5}, and E = {(A, A1),(A, C1)}.
3. Visualization with nodes and lines
   Vertex placed in a node, then the edge connected the neighbor vertices.



Picture 2.2 graph representation
(from Munir, 2010, p.387)

### B. Undirected Graph

Edges of undirected graph do not have direction orientation. Therefore, if we have an edge (i, j), it will has same meaning with edge (j, i). An example of relationship that can be presented by undirected graph is graph of close friends in a class.

### C. Circular ideogram

A circular ideogram is an graphical method to visualize data in circular layout. Circular ideogram for visualize data is popularized by Martin Krzywinski and his team in 2007. They created a software named *circos* to facilitate circular ideogram making for genomes comparison. *Circos* is capable to presenting data as scatter, line and histogram plots, heat maps, tiles, text and connectors.

Previous methods that used to visualize genomic date are linearly arranged ideograms that connected by lines to represent rearrangement. The weakness of linearly arranged ideograms is the challenge in displaying large volumes of genomic rearrangement data.

Originally, circular ideogram used to presented genome of bacteria that are circular in reality. Martin Krzywinski develop this method for other genomes that not circular in reality until he makes a software to facilitate his research, *circos*. *Circos*' initial application was to visualize fingerprint profiling (Krzywinski et al. 2007) of cancer genomes.

## *D. DNA Sequencing*

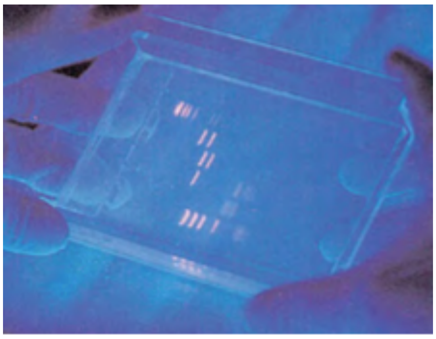
Genetic information arranged from nitrogenous bases. There are 4 nitrogenous bases that exist in DNA, Thymine (T), Adenine (A), Guanine (G), and Cytosine (C). Every nitrogenous base has different chemical structure. Because of chemical structure, nitrogenous bases exist in fixed pair, A with T, and G with C.

Arrangement of many nitrogenous bases built a pattern. The pattern will carry the genetic information of organism. DNA's patterns determine the product of protein synthesis. Protein synthesis is a process to produce protein that will used by cell itself or distributed to another cells. Protein synthesis is the most important process of living organism besides cell respiration.

A little mutation that changes the DNA arrangement can conduct a big problem for the organism. The example is a cancer cell. A cancer cell is a normal cell that mutated, the mutation that occ urs in very small fragment of DNA can changes the whole characteristic of cell.

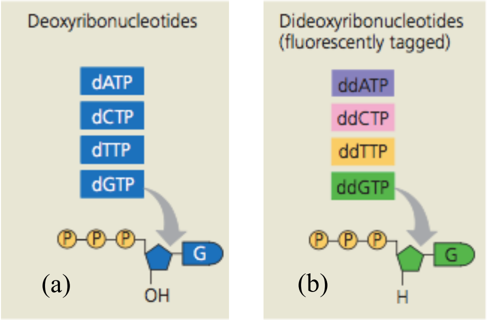
Scientist tried to know the arrangement of nitrogenous bases in DNA. To know the arrangement of DNA, scientists develop DNA sequencing method. There are many DNA sequencing methods, such as gel electrophoresis, dideoxy chain termination method, and southern blotting method.

The principal of gel electrophoresis are separating nucleic acids or proteins by the difference of size, electrical charge, or other physical properties. Gel that used in this technique made if a polymer, a long strand of monomers arrangement. The gels act as separator of nucleic acids and proteins. With electric charge, nucleic acid that carries negative charges will move to the positive. When the mixture undergoes electrophoresis it forms a band pattern characteristic of the starting molecule to restriction enzyme that used in this technique. The pattern of restriction enzyme can be used to identify small DNA molecules.



Picture 2.3 Gel Electrophoresis
(from Reece, 2011, p.405)

The principal of dideoxy chain termination method is using fluorescently tagged dideoxyribonucleotides. The color of fluorescent tag is different based on type of nitrogenous bases. First step is denatured the DNA strand into single strand. This single strand act as a template for many labeled strands that arranged from deoxyribonucleitides and terminated by dideoxyribonucleitides. Labeled strands separated by its length using polyacrylamide gel. After that, with laser and detector, the nitrogenous bases can be identified by the color of fluorescent tag.



Picture 2.4 (a) Deoxyribonucleitides that will be paired with DNA strand, (b) Dideoxyribonucleitides with fluorescent tag as the terminal of the labeled strand.
(from Reece, 2011, p.408)

## *E. Genome*

Genomes are whole of genetic information in a cell. Genomes have various size, number of genes, and gene density for every species. The size of genomes is represent by number of base pairs. The unit of genome size is Mb (Million base pairs). Size of genome does not affect the number of genes. Is possible, genome with longer size has fewer genes inside. Gene density is the division between number of genes and length of DNA.

There is other structure that connected with genome, chromosome. In genome sequence, genetic information separated by chromosome types. Chromosomes only exist when the cell perform a cell division. There are two types of cell division, mitosis and meiosis. A chromosome is condensed of DNA strands. The number and the shape of chromosomes in every organism are different. For example human has 23 pairs, snail has 24 pair and potato has 48 pairs of chromosomes.

Picture 2.3 Human Chromosomes Set
(from Reece, 2011, p.250)

In picture 2.3, we can see that human genome has 22 pairs chromosomes that symbolized by number, a chromosome X, and a chromosome Y. Both of chromosomes X and Y are sex chromosome. The other chromosomes types called autosomes. Sex chromosomes determine the gender of organism, for human, XX for females and XY for males.

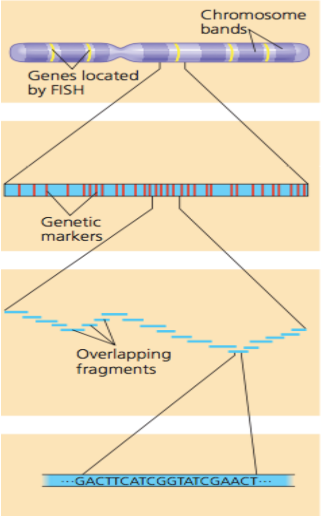
## III. Graph Implementation in Genome Circular Ideogram

### A. Genome Sequencing as Data Source

Genome sequencing is a process to identify the genetic information arrangement in a whole cell. If DNA sequencing produce the genetic information arrangement of a piece DNA, therefore genome sequencing is way to combine all the pieces of genetic information into a whole cell genetic information.

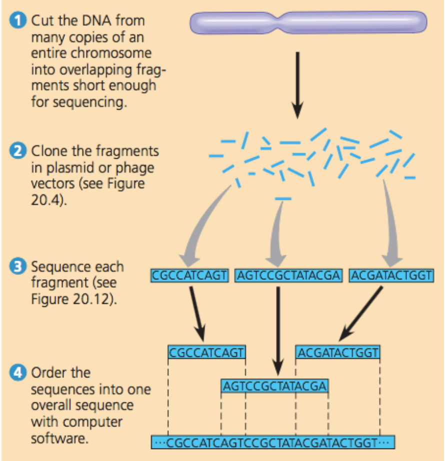
The position of DNA fragments in a chromosome is the most important thing in entire genome sequencing. There are some stages to approach the genome sequencing:

1. Annotate the chromosome banding pattern and location of specific genes using fluorescence *in situ* hybridization (FISH).
2. Perform the linkage mapping with ordering genetic markers. Genetic markers are some pattern of gene that easily identified by enzyme. Such as Restriction Fragment Length Polymorphism (RFLP) and Single Nitrogenous Polymorphism (SNP).
3. Perform the physical mapping with ordering large overlapping fragments and ordering smaller fragments. This step is necessary because DNA sequencing technology only support for small fragments, mostly less than 1000 base pairs.
4. Perform DNA sequencing for each small fragments and assemble the partial sequences into the complete genome sequence.



Picture 3.1 The stage of genome sequencing. The stage represented from top of the picture.
(from Reece, 2011, p.427)

Today's most popular method to perform a genome sequencing are *whole-genome shotgun approach* method, invented by molecular biologist J. Craig Venter. In this method, we don't have to perform linkage mapping nor physical mapping. This method only perform DNA sequencing and use a powerful computer program to reassemble many overlapping DNA sequences into a single sequence.


1 Cut the DNA from many copies of an entire chromosome into overlapping fragments short enough for sequencing.
2 Clone the fragments in plasmid or phage vectors (see Figure 20.4).
3 Sequence each fragment (see Figure 20.12).
CGCCATCAGT
AGTCCGCTATACGA
ACGATACTGGT
4 Order the sequences into one overall sequence with computer software.
CGCCATCAGT
ACGATACTGGT
AGTCCGCTATACGA
···CGCCATCAGTCCGCTATACGATACTGGT···

Picture 3.2 Whole genome shotgun approach method
(from Reece, 2011, p.428)

Perform a genome sequencing is not easy. There are many sophisticated tools and chemical substances needed such as centrifuge, enzyme, fluorescence molecules, etc. But now, everyone can get the genome sequence information from internet. There are many web or application that provide genome sequence data such as *Integrated Genome Viewer* (IGV) and *BEDTools*. The applications not only provide human genome sequence but also some animals and plants genome sequences.

### B. Genome Data Visualization in Circular ideogram

Genome sequences information are contain long strand of nitrogenous base pairs. Without further data processing, is hard to make the genome sequences usable, especially to comparing a genome sequence with the other one. Circular ideogram is an alternative to presenting the genome sequencing data in simple way.

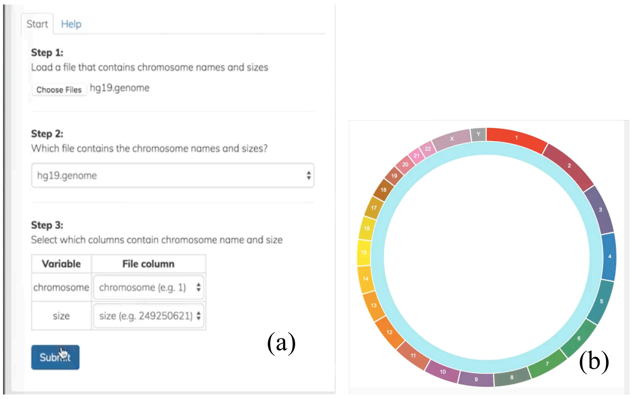
Genome circular ideogram is consist of vertices that represent data from genes that classified by the chromosomes. Data that represented by circular ideogram can vary, depend on the purpose of making. Edges in a genome circular ideogram connecting vertices that have similar genetic patterns. This makes it possible to comparing genes in a genome or comparing more than one different genomes, therefore genome sequence information can be used easily.

Everyone can make they own genome circular ideogram. Genome sequence information can be downloaded easily. To make a genome circular ideogram in circular ideogram application, the most important thing are genome sequences information in format .genome, .bed, .bedpe, or other supported files. The files can be downloaded via *Integrated Genome Viewer* (IGV), *BEDTools*, or similar application. This is how to make a genome circular ideogram:

1. Once the application opened, it will display Start tab. Click the 'choose files' button, after that choose a file

that contains chromosome names and sizes.

2. Step 2 is mentioning which file contains the chromosome name and size, this step will automatically fill in if there is only one file in step one. Otherwise, you can choose which file that contains chromosome name and size.
3. Step 3 will automatically fill in if step 2 has been filled.
4. Click ‘submit’ button.



Step 1:
Load a file that contains chromosome names and sizes
hg19.genome

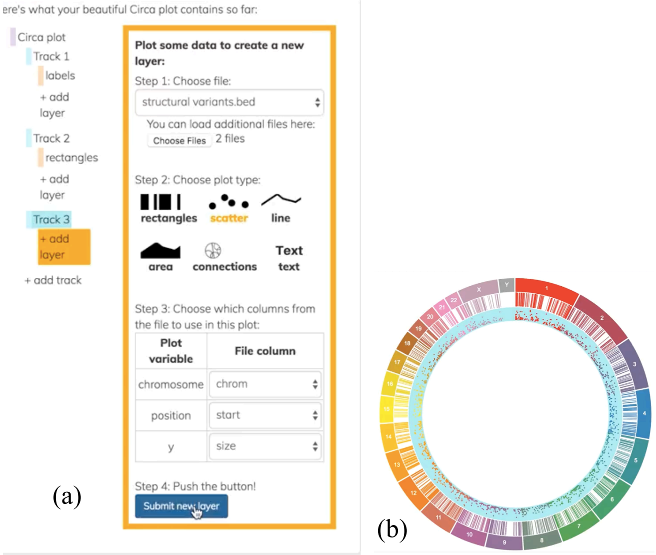
Step 2:
Which file contains the chromosome names and sizes?
hg19.genome

Step 3:
Select which columns contain chromosome name and size

| Variable | File column |
|---|---|
| chromosome | chromosome (e.g. 1) |
| size | size (e.g. 249250621) |

Picture 3.3 (a) Start tab in application (b) result (from https://www.youtube.com/watch?v=cKlJDhucul8 accessed 8 December 2018)

5. In Build tab, we can add some additional information with add new track and layer.



ere's what your beautiful Circa plot contains so far:

Circa plot
Track 1
labels
+ add layer
Track 2
rectangles
+ add layer
Track 3
+ add layer
+ add track

Plot some data to create a new layer:
Step 1: Choose file:
structural variants.bed
You can load additional files here:
2 files

Step 2: Choose plot type:

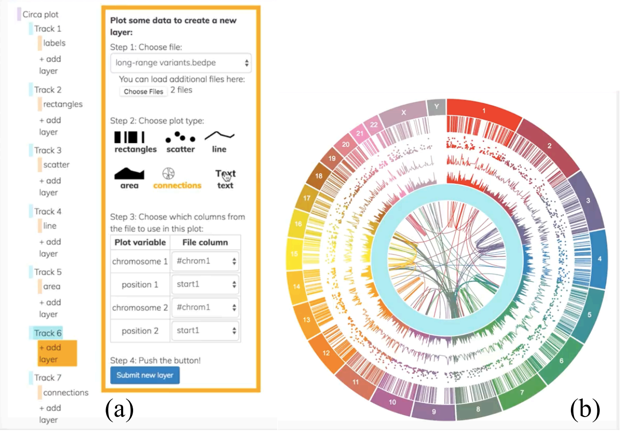
Step 3: Choose which columns from the file to use in this plot:

| Plot variable | File column |
|---|---|
| chromosome | chrom |
| position | start |
| y | size |

Step 4: Push the button!

Picture 3.4 (a) Built tab in application (b) result (from https://www.youtube.com/watch?v=cKlJDhucul8 accessed 8 December 2018)

6. The information can be the new layer of vertices. The file usually has ‘.bed’ extension. This vertices information can be presented in form rectangle, scatter, line, or area plot type. After choose the plot type, the next step is input information in plot variable.



Circa plot
Track 1
labels
+ add layer
Track 2
rectangles
+ add layer
Track 3
scatter
+ add layer
Track 4
line
+ add layer
Track 5
area
+ add layer
Track 6
+ add layer
Track 7
connections
+ add layer

Plot some data to create a new layer:
Step 1: Choose file:
long-range variants.bedpe
You can load additional files here:
2 files

Step 2: Choose plot type:

Step 3: Choose which columns from the file to use in this plot:

| Plot variable | File column |
|---|---|
| chromosome 1 | #chrom1 |
| position 1 | start1 |
| chromosome 2 | #chrom1 |
| position 2 | start1 |

Step 4: Push the button!

Picture 3.5 (a) Start tab in application (b) result (from https://www.youtube.com/watch?v=cKlJDhucul8 accessed 8 December 2018)

7. The information also can be the edges that connected the vertices. The file usually has ‘.bedpe’ extension. The edges can only presented by connection plot. After choose the connection plot, the next step is input information in plot variable. The plot variables for edge consist of chromosome 1, position 1, chromosome 2, and position 2. The information is the index of column, because the file ‘.bedpe’ presented the connection of the genome in table.

### *C. Concept of Graph in Genome Circular ideogram*

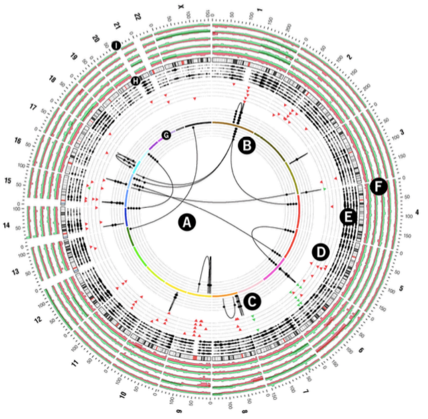
Circular ideogram have same concept with graph. There are vertices and edges in both graph and circular ideogram. We can make genome circular ideogram without edges because edges is an additional information that optional for user. But, we can’t make a circular ideogram without vertex, because the information of vertices that consist chromosome name and size used as start information to build a genome circular ideogram. This concept is same as graph that the set of vertices can’t be empty and in other hand the set of edges can be empty.

## IV. Use of Genome Circular Ideogram in Various Sectors

### *A. Detect Genomic Rearrangement in Cancer Genome*

Normal cells have a cell cycle. This cycle control when the cell division occur and send the signal to destroy the cell if the cell is no longer useful. Cancer cell lost this cell cycle control, therefore, the division of cancel cells will be uncontrolled and invade the other tissue. Cancer cells are “immortal”, because they cell will never send a signal to destroy themselves.

Most of the cancer cells have mutation in their genome. A small mutation in genome can conduct fatal consequence. Scientist tried to find a way to repair the mutation genome. The way to repair the mutated genomes are find the genome sequence of both healthy and cancer cell, then make a comparison to find location of the mutation. Genome circular ideogram can help to comparing the genome sequences of cancer and healthy cell. Genome circular ideogram can only identify the large-scale rearrangement in tumor cell because the data that used in genome circular ideogram are simplified data.



Picture 3.5 A whole-genome view of structural changes in five follicular lymphoma tumor samples
(from Krzywinski, 2009, Picture 1)

Picture above is a genome circular ideogram that comparing 5 mutated genomes. Every genome stacked up in every circular layer. The edges inside the circular ideogram showed the comparison of genomes. Mutation can be identified by an edge that does not connect vertices in same location for all genomes. The different genome might be a mutated genome.

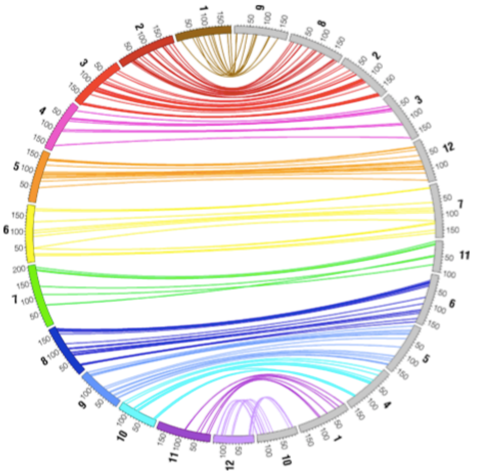
### B. Identification of Quality Plant Species

The growth rate of the organism are controlled by the cells of the organism. Cell produce protein, hormone, and other substances that influenced the growth of the plant. Scientist find out the genetic pattern from DNA sequences that affect the growth of plants by identifying hormones of enzymes that affected the plant growth rate. After that, scientists have to find which genetic patterns that regulate the hormone production.

The known genetic pattern makes it possible to predict the growth rate of the plant even from seed phase. The existence of the genetic pattern indicate higher growth rate. However, there are other factors that affect the growth rate, the example is existence of inhibitors that detain the production of hormones and enzymes, therefore predictions is not always accurate.

### C. Identify the Connection Between Different Species

Genome circular ideogram also can compare genome of different species to find how close the connection between them. Closer the relationship between two species, the more similarities in their genome sequence. This relationship will be useful in classification of living organism.



Picture 3.6 Genome circular ideogram that comparing spruce and loblolly pine genome. The spruce indicated by color vertices.
(from Pavy, 2012, p.6)

From the picture above, we can see there are many homologue chromosomes from different species. Besides proof the relationship between two species, from comparison above we can see that every organism perform similar process in its body that regulated by similar genes, despite the difference of the arrangement in genome sequence.

## V. Conclusion

Graphs are widely used in many sectors. Circular ideogram is one of the graph concept implementation that brought a big change especially in genome sequencing technology. By using the circular ideogram, now it possible to comparing genome from different cell. This comparison is useful in many sectors, for example is comparison of cancer cell and normal cell in medical sector and comparison of genomes from different species in taxonomy. However, the genome circular ideogram still has several disadvantages, especially for presenting detailed information from the sequence.

## VII. Acknowledgment

My deep gratitude to Allah SWT that makes me able to finished this paper. I would like to say thank you for my parents that always give me support, to Dr. Ir. Rinaldi Munir, M.T. for his dedication as the lecturer of IF2120 Discrete Mathematics in K01, and for all that give me support.

## References

[1] M. I. Krzywinski, J. E. Schein, I. Birol, et al, "Circos: An information aesthetic for comparative genomics, " New York: Cold Spring Harbor Laboratory Press, 2009.

[2] J. B. Reece, L. A. Urry, M. L. Cain, S. A. Wasserman, R. B. Jackson, et al, "Campbell BIOLOGY," San Francisco: Pearson Education, Inc., 2011.

[3] N. Pavy, B. Pelgas, J. Laroche, P. Rigault, N. Isabel, et al, "A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers," London: BioMed Central, 2012.

[4] R. Munir, Matematika Diskrit, Bandung: Departemen Teknik Informatika Institut Teknologi Bandung, 2003.

[5] W. Li, H. Shang, Q. Ge, C. Zhou, J. Cai, et al, “Genome-wide identification, phylogeny, and expression analysis of pectin methylesterases reveal their major role in cotton fiber development,” London: BioMed Central, 2016.

[6] OMGenomics, (2017). “Circa Tutorial: Different ways to plot your data,” from https://www.youtube.com/watch?v=cKlJDhucul8, retrieved at 8 December 2018.

## PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 9 Desember 2017

Nada Afra Sabrina 13517118