Application of Markov Model in Simple DNA Sequence Determination

Anindya Prameswari Ekaputri / 13518034[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]13518034@std.stei.itb.ac.id

***Abstract*—DNA is an important part of an organism body—it holds the genetic informations about the organism, the instructions about how to maintain and develop the organism, and how some body organs should function. The DNA stores those informations on a code based on four chemical bases, and the order of those bases determines what information is stored on the DNA. There must be certain ways the DNA is coded so we can examine and get some useful informations of it. Because the DNA is very long and it would take very long if we read every chemical bases one by one, we need a model, a predicting algorithm to help us register those bases. The markov model can help us understand the main idea of how the DNA sequence is created.**

***Keywords*—bioinformatics, DNA sequencing, graph, markov model**

I. Introduction

Human live for some reasons. Probably it is the love of your life for you or my family for me, but we all have the same reason why we are still alive to this day: our body.

We have a body that works on its own without us asking them to function—do you ask your heart to beat today? No, but it is still beating nonetheless. And the question applies not only to our hearts, but also each of our organs: our stomach, liver, lungs, gastrointestines, kidneys, and many more. There are plenty of our body that function automatically, although we do not give the direct instructions, they are still working, moving, producing this enzyme, that protein, hormones, etc.

If it is not us that gives the instructions, then who does? *What* does? How does our body knows what to do, what to produce?

For protein production, the answer lies within our cells. Every cells in our body have chromosomes in their nucleus, and it is the chromosomes which gives the instruction what to do. It does not give the instructions verbally, of course, for cells do not have mouth, but it instructs through a code written in our DNA.

The DNA itself is formed by four chemical bases: adenine, guanine, cytosine, and thymine. We can simply say that those chemical bases are the language of our DNA code. Their sequence determines what instruction the DNA holds.

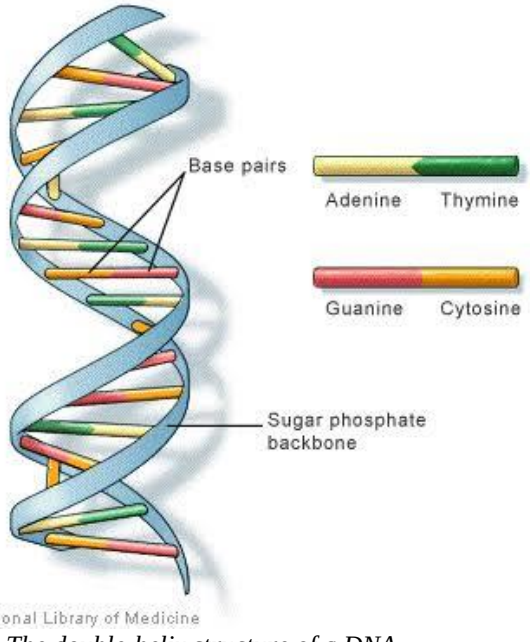
But how does the sequence is formed? Is it randomly picked, or is there some kind of a pattern behind the sequence? Is there a model to represent the making of a DNA sequence? How is the probability of this base being produced, is it more likely or less likely than the other base?

II. Deoxyribonucleic Acid

A. Definition and Structure

DNA is a shortened word for deoxynucleic acid, that is an agent in any living organism body which carries genetic instructions for the development, functioning, growth, and reproduction of the organisms. DNA are stored by every cell in the organism body in their nucleus, but few also have their DNA in mitochondria for energy converting.

The instruction is coded with four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The sequence of these bases determine what the instructions are, it is similar to how letters are used to form a word. These bases will also pair up with each other, A with T and G with C, and the pair will form a spiral ribbon-like structure of the DNA.



Picture 1: The double-helix structure of a DNA
Image source: https://gontornews.com/2019/03/02/mengenal-struktur-dna-2/

The spiral form is called a double-helix. This form will help during the replication process to create more DNA, copying genes, and producing amino acid to form proteins. The DNA will split up into two single strands and new chemical bases will

line up, creating a new DNA strand according to the sequence of its original DNA sequence. So, the strands act as a mold for new DNA strands. [6]

B. Function

As what is listed above, the DNA is responsible for carrying instruction to develop, grow, function, and reproduce the organism. In order to do that, the code formed by the chemical bases will be read and converted into a template to create proteins, which are the complex molecules that do the most work in our body.

The copying mechanism is also similar to how DNA copies itself. The DNA will split into two single strands, and some new chemical bases will line up to form a new strand. The difference is that this new strand does not act as a new DNA strand, but rather as a new instruction used in forming an amino acid. In amino acids, the chemical bases are no longer A, T, G, and C, but A, U, G, and C [7] .The sequence of these chemical bases will be read three-by-three by a translator, and it will 'call' an amino acid based on the three chemical bases.

Then, a new arrangement of amino acids will be formed. This arrangement of amino acids will then form a more complex structure, which is the protein itself. There are twenty types of amino acids that can be arranged into many different orders and create a wide range of proteins.

Table 1: Table of 20 amino acids created by codons

First letter \ Second letter	U	C	A	G	Third letter
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr **UAA Stop** **UAG Stop**	UGU Cys UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

table source:
https://www.khanacademy.org/science/biology/gene-expression-central-dogma/central-dogma-transcription/a/the-genetic-code-discovery-and-properties

This protein creating mechanism exists on every living organism and the same species will share the same code to create the same protein. This protein-making code in the DNA takes up 99% of our 'DNA dictionary' and it is the same for everyone, so what makes us different is actually just the 1% of our DNA.

C. DNA Sequencing

DNA sequencing is a process of determining the order of DNA chemical bases. The sequence is important to differentiate which parts of the DNA carry a genetic information and which parts carry some instructions. The accuracy of DNA sequencing is crucial because a slight change could lead into a disease.

As mentioned before, we have a 'DNA dictionary', its formal name is a *genome*. Genome is a complete set of genetic instructions made by the DNA, it contains all the informations needed to build and develop and organism. Human genome is made of 3.2 billion bases of DNA, which is very huge. In comparison, if we were to write our DNA dictionary, the book will be 61 metres high [8].

In short, it is impossible to write and determine every single letter of our DNA, which is why scientists break down DNA into smaller parts before examining them. However, for complete understanding of our genome, of course we need the full genome transcript. It was a pain to do DNA sequencing as it is very slow and complex, but due to our technology advancement, now we can list 3.2 billion DNA bases in a matter of hours.

DNA sequencing is not used directly to treat patients and cure diseases, but it helps physicians to quickly identify what type of cancer a patient has. DNA sequencing is also used to identify the genetic causes of rare disease, get the genetic details about 30 cancer types, and screen newborns for diseases.

III. GRAPHS

A. Definitions

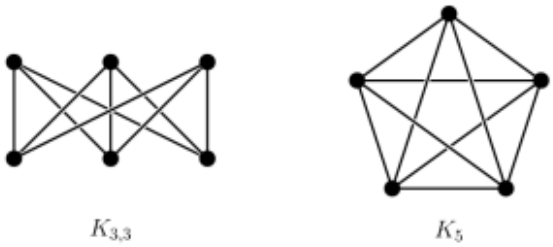
Graphs are mathematical models to represent relations and connections. It is usually used to help visualizing mathematical problems, circuit design, communication networks, ecology, engineering, operations research, counting, probability, set theory, information theory, sociology, and many more [1].

The basic components of a graph are vertex and edge. Vertices (or nodes) are the points in graph, they are usually representing a state, an area, an object, something that is described with the graph. The vertices are usually kept in a set that cannot be empty, so a graph should contain at least one vertex.

The second property of a graph is the edges. Edges are lines (so they are also called lines) that connects the vertex, it indicates this vertex and that vertex are related. Edges are also defined in a set, but the edges set can be empty, which means there exist a graph that contains only unrelated vertices.

Take two vertices, for example *A* and *B*. If *A* and *B* is connected by an edge *c*, they are said to be *adjacent.* Therefore, *A* is a neighbor of *B* and *B* is a neigbor of *A*. The edge *c* is also said to be *incident with* the vertex *A* and *B*, which means the *A* and *B* are the endpoints of *c*.

Vertex has a property named *degree.* A degree describes how many edges are incident with the vertex. For example, if vertex *A* have three edges connecting *A* with other vertices, then the degree of *A* is three.

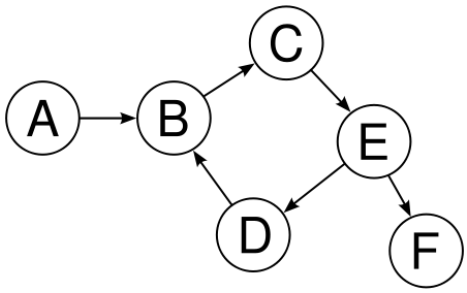


Picture 2: Example of graphs
source: https://grafen.wordpress.com/2011/02/03/kazimierz-kuratowski/

The components listed above are the basic ones, the standard ones found on a basic graph. There are a lot of variations to the components, for example a *loop* is an edge that starts and ends within the same vertex. There is also an *isolated point*, a vertex that has no edge. For the sake of simplicity, this paper will only explain the variations that is used in a markov model.

B. Directed Graph

Directed graph has an arrow at the end of their edges. In undirected graphs, which vertex is the start and which one is the end is not important, but directed graph counts this matter into consideration. If there are two edges, *x* and *y*, where the edge *x* starts at vertex *A* and ends at vertex *B* but the edge *y* is the other way around (starts at vertex *B* and ends at vertex *A*), then *x* and *y* are two different edges. It means to go from *A* to *B* we must go by the edge *x*, but to go from *B* to *A* we have to go by the edge *y*; while on undirected graph, which edge we take would not matter.



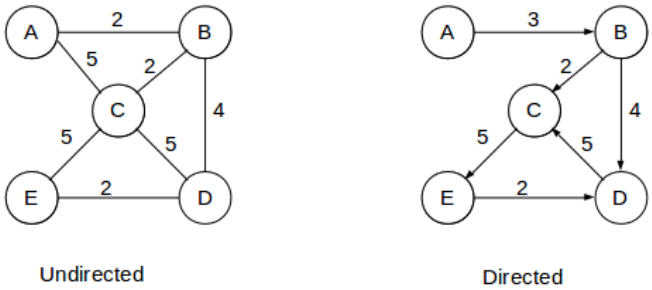
Picture 3: A directed graph
source:https://computersciencewiki.org/index.php/The_web_as_a_directed_graph

With this variation, the edge (which now can be called an *arc*) has two additional properties: head and tail. The head is where the edge ends, meanwhile the tail is where the edge starts.

C. Weighted Graph & Labeled Graph

Weighted graph has a value on its edges. The value could be travelling time, distance, production cost, anything. Upon weighted graphs, we can count how long or how much a path is.

Weighted graphs are also commonly called labeled graph, but the two actually refers to different graphs. While weighted graphs have values on their edges, labeled graph could also have labels on their vertices. The labels not only in non-negative numbers, it could also be names; for example a city name, a state, etc.

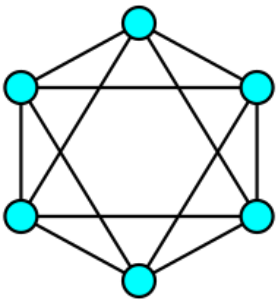


Picture 4: Labeled undirected and directed graph
source: https://study.com/academy/lesson/weighted-graphs-implementation-dijkstra-algorithm.html

D. Connected Graph

Two vertices said to be connected when there is a path, a way to go from vertex *A* to vertex *B*. They do not have to be adjacent, the path can go through the vertex *C* and *A & B* are still said to beconnected.

When a graph is connected, it means for each vertex in the graph can get accessed from whichever vertices in the graph.

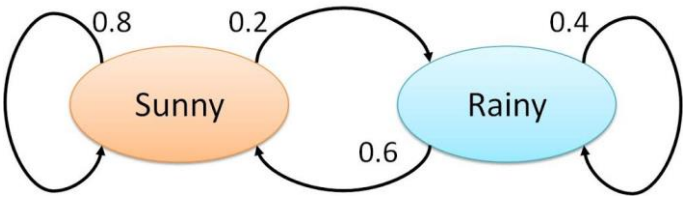


Picture 5: A connected graph
source: https://en.wikipedia.org/wiki/K-vertex-connected_graph

IV. MARKOV MODEL

Markov model, also known as markov chain, are mathematical system that is created by Andrey Markov. It is drawn as a directed, labeled, and connected graph. The vertices describe the states or situations of an object, while the edges means we can 'hop' from a situation to another. The weight of the edges describes how likely it is to jump to the other state. If the weight of the edges that are out from one vertex are summed, they should add up to one [2].

Here is a simple example for better understanding. We will predict some day's weather based on a markov model.



Picture 6: An example of a markov model
source: https://sgmustadio.wordpress.com/2011/02/17/primer-

on-hidden-markov-models-part-1/

Let us assume that the weather on some days depends on how is the weather the day before. So, if it rains yesterday, it is more likely to rain again today, even though there are still a chance for today to be sunny; so does the other way around: if it's sunny today, it is more likely to be sunny tomorrow, but it does not mean tomorrow will not rain.

In the graph above, there are two states: a day is sunny and a day is rainy. The chance of staying in the sunny state is more likely, but it does not cover the chance of jumping into the rainy state. Here is one transcript that can be produced with the markov model above:

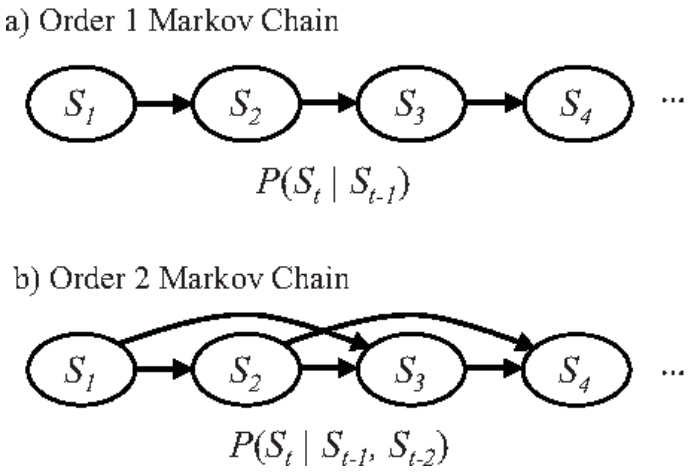
R S S S S S S R S S R S S S S S S S R S S S S
S S S S S S S R S R S S S S R R R S R S S S S

Compare it with a transcript produced by a markov model with 50:50 chance.

R S S S R R S R R R S S S R S S S R R S S R S
R R S S R R S R S S R R S S R S S R S S R R R S

We can see that the one generated with 50:50 chance model is more random and does not hold the 'depending on the state before' property.

Markov model also has an order property. The models can be zero ordered, first ordered, second ordered, and so on [5]. The order describes how many states are considered in making the decision which state to go next. Zero order markov model means the model has no dependency on the previous state, so all states have the same possibility. First order model means it will consider only one previous state, and the result model will look like the weather sequencing model that has just explained above.



Picture 7: Examples of some ordered markov chain
source: https://www.semanticscholar.org/paper/Generating-Maps-Using-Markov-Chains-Snodgrass-Onta

V. Generating DNA Sequence with A Markov Model

Now, we will be generating a DNA sequence with first order markov model.

Although this model is capable in visualizing the production of DNA sequence, it is not the best model which portrays the process. However, it could gives us the main idea about how the DNA in our body are coded.

The markov model used in this simulation is the zero and first order markov model. Because we are just going through the basic, we are going to use simple models. For more advanced use, it is recommended to use a higher order ones, for example use a third order markov model to read the protein generating instructions which are read in three-tuple.

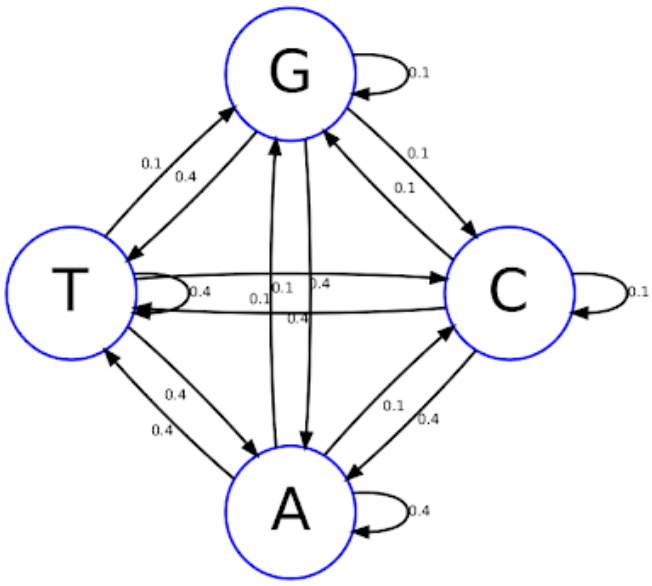
The probability of jumping into another state is also randomly decided on this one. Describing the probability of the next chemical base of a DNA is a complex topic and it needs a more advanced model than the markov model. But as we are just going through the basic to get the main idea, we will use a 'mock up' probability numbers and just see how the process went.

In the zero order model, we have our probabilites like so:

Table 2: Zero order markov model transition matrix

next state > / previous	A	T	G	C
A	0.4	0.4	0.1	0.1
T	0.4	0.4	0.1	0.1
G	0.1	0.1	0.4	0.4
C	0.1	0.1	0.4	0.4

When drawn, the table above would result in the markov model below:



Picture 8: Zero order markov model in DNA sequencing
source: https://www.r-bloggers.com/introduction-to-markov-chains-and-modeling-dna-sequences-in-r/

This model is stating that A and T are more likely to be found in DNA strands more than G and C. It does not matter what is

the previous state, the probability of getting an A-base or a T-base is always 40%, bigger than the probability of getting a C-base or G-base which is just 10%.

The model above would produce a strand like this:

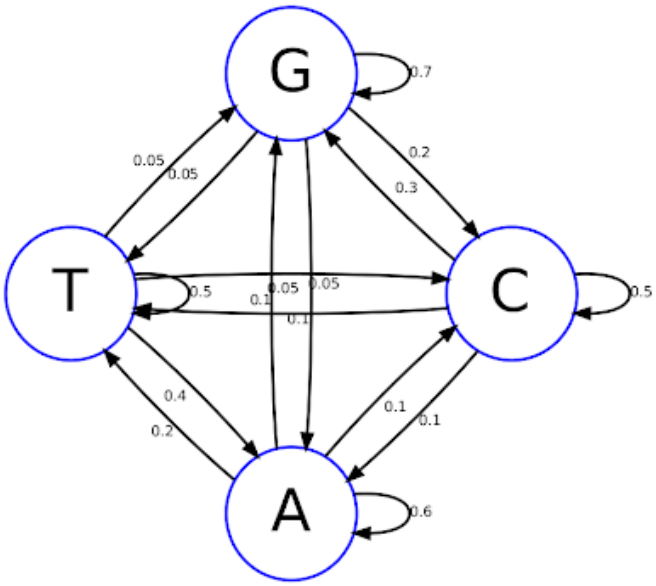
```
G G T T A A A T T A A A A G G C A A T C C G G
G T T T A T A T A A A A T T A T T C T A C C G
```

Let us see how changing the order changes the result. The table for a first order markov model is:

Table 2: First order markov model transition matrix

next state > / previous	A	T	G	C
A	0.6	0.2	0.1	0.1
T	0.4	0.5	0.05	0.05
G	0.05	0.05	0.7	0.2
C	0.1	0.1	0.3	0.5

The first order markov model would look like this:



Picture 9: First order markov model in DNA sequencing
source: https://www.r-bloggers.com/introduction-to-markov-chains-and-modeling-dna-sequences-in-r/

With the model above, we could produce a DNA strand like this:

```
G G G G G G G T T T G G G T G T T G T T T G
G T A A A A A C A A T G G G T G G G T G T T T
```

Let us compare our transcripts with a DNA strand containing a start codon, codons, and a donor site:

```
C G C C A T G C C C T T C T C C A A C A G G T
G A G T G A G C C T C C C A G C C C T G C C C
```

In a glance, we can see that there are mostly C, then G on the real DNA transcript, with occasional T and A. The transcript produced by our first order model contains mostly G and T, which means it does not represent the real DNA, but it get the idea that a DNA strand will more likely to produce some certain bases more than some others.

On the other side, the transcript produced by our zero order markov model is a bit more on the random side, similar to the 50:50 markov model we have demonstrated with the weather model. It contains mostly A and T, which is the opposite of our real DNA sample where A and T are the least ones produced.

Those mistakes are understandable because the numbers of probability used in these models are generated at random, not their real appearing probability, but we can see that this markov model gets the general idea of DNA sequencing.

What chemical bases are being produced now depends on the previous base produced. Those bases also does not get produced at the same rate, some bases are more likely to be produced more than the others.

So, DNA sequence can be determined by using a markov model. However, this model is still very simple and not fully representing a real organism's DNA. Using a more advanced model and algorithm would give more accurate result, for example the hidden markov model.

The hidden markov model is similar to the markov model, but as the name states, it has some hidden states. Using the same weather example above, we do not decide the model states by the weather itself, but with another indicator instead.

Say that we want to know what is the weather outside, but we are in a windowless room, which means we can not see how it is outside. Instead of looking at the sky directly, we will be able to predict the weather outside with the room temperature.

If it is hot, then it is more likely to be sunny rather than raining, and the other way around when it is cold. If the temperature is rather warm, it could be a 50:50 chance for it to be sunny or rainy.

Applied to the DNA sequence modelling, as mentioned above, this hidden markov model allows us to take more facors into account, for example, whether the DNA is the coding one or the non-coding one. We would be able to further examine whether this part of a DNA takes part in the protein making process or determining the genetic properties of an organism [10].

VI. Conclusion

DNA sequence can be determined by using a markov model. However, this model is still very simple and not fully representing a real organism's DNA. Using a more advanced model and algorithm would give more accurate result. Few recommended ones are the hidden markov model, which can take more factors into account, and have several more fitting algorithms for DNA sequencing, for example the forward-backward algorithm and viterbi algorithm.

VII. Acknowledgment

First of all, it is thanks to God that by His guidance and encouragement I can finish this paper on time. I would also like to thank my teacher, Rinaldi Munir, an inspiring person who taught me so many things that I know would be important for my future studies. This paper is specially dedicated to my

mother who inspires me to learn the bioinformatics field (*this paper is me testing the waters*), and to my soon-to-be pharmacists friends (Aristo, Mangantjo, and Fikri, hello!) who had an antimicrobial resistance awareness campaign right before I was about to write this paper. I got this topic idea from you. Thank you!

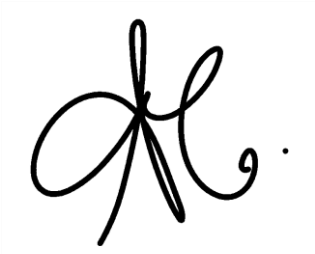
REFERENCES

[1] Kenneth H. Rosen, John G. Michaels, Jonathan L. Gross, et.al., *Handbook of Discrete and Cominatorial Mathematics*, Florida, CRC Press, 1999, ch. 8.
[2] Victor Powell, "Markov Chains Explained Visually", www.setosa.io/ev/markov-chains, accessed 4 December 2019.
[3] Avril Coghlan, " Hidden Markov Models: A little more about R", www.a-little-book-of-r-for-bioinformatics.readthedocs.io/en/latest/src/chapter10.html, accessed 4 December 2019.
[4] R. Munir, *Matematika Diskrit*, Bandung, Departemen Teknik Informatika Institut Teknologi Bandung, 2003.
[5] Benjamin Tovar, "Introduction to Markov Chains and Modeling DNA Sequences in R", www.r-bloggers.com/introduction-to-markov-chains-and-modelling-dna-sequences-in-r, accessed 5 December 2019.
[6] School of Life Sciences, "DNA Structure", www.askabiologist.asu.edu/dna-shape-and-structure, accessed 5 December 2019.
[7] Khan Academy, "The Genetic Code & Codon Table", www.khanacademy.org/science/biology/gene-expression-central-dogma/central-dogma-transcription/a/the-genetic-code-discovery-and-properties, accessed 6 December 2019.
[8] Your Genome, "What is a genome?", www.yourgenome.org/facts/what-is-a-genome, accessed 5 December 2019.
[9] Bioinformatics Organization, "Hidden Markov Model", www.bioinformatics.org/wiki/Hidden_Markov_Model, accessed 6 December 2019.
[10] Mathisca de Gunst, "Hidden Markov Model", www.ibi.vu.nl/teaching/a4g/materials/lect11.pdf, accessed 4 December 2019.
[11] www.markov.yoriz.co.uk, accessed 6 December 2019.

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, 4 Desember 2019



Anindya Prameswari Ekaputri
135 18 034