Early Detection of Down syndrome Through the use of String Matching Algorithms

Tirta Wening Rachman - 13512004[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]tirtawening@gmail.com

***Abstract*— Down syndrome is a genetic condition that causes delays in physical and intellectual development. It is the most frequently occurring chromosomal disorder, and it occurs in one in every 691 live births[1]. Living with Down syndrome is extremely hard to experience as one's physical and intellectual development is underdeveloped. Since Down syndrome is a condition that is caused by a genetic mutation, it is entirely possible to detect whether or not a fetus has Down syndrome or not. This can be done through DNA testing. Furthermore, because of the DNA code can be represented by the letters A, G, T, and C, it can be digitalized and the detection of Down syndrome can be performed through the use of string matching algorithms.**

***Index Terms*— Down syndrome, String Matching Algorithms, Bioinformatics, DNA**

I. INTRODUCTION

Down syndrome is a genetic condition that causes delays in physical and intellectual development. It is the most frequently occurring chromosomal disorder, and it occurs in one in every 691 live births[1]. Down syndrome is not related to race, nationality, religion or socioeconomic status.

Fig 1. The face of a baby with Down syndrome
Source: U.S. National Center on Birth Defects and Developmental Disabilities

Living with Down syndrome is extremely hard to experience as one's physical and intellectual development is underdeveloped. It can be especially stressful for parents to raise a child with Down syndrome, especially if it is detected later in the child's life. Early treatments since birth can be extremely beneficial to the child's development. Thus, it is extremely crucial for parents to detect if their child developed Down syndrome or not since before the child's birth.

Also, if Down syndrome is detected early in the pregnancy, some parents may be able to opt out early for abortion since the fetus is not yet developed. This however, is subject to a large debate whether or not the law should mandates that the child lives or may be aborted.

Since Down syndrome is a condition that is caused by a genetic mutation, scientists now are able to detect whether or not a fetus has Down syndrome or not. This can be done through DNA testing. DNA are the blueprint of our body, and they are coded with only four chemical bases. These four bases each are given their own names, and the first letter of each base determines the "code" of the DNA. Therefore, it is fully possible to digitalize our DNA.

Remembering the fact that the detection of Down syndrome can be done by finding certain pattern of chemical bases in out DNA and the fact that our DNA can be digitalized, it is entirely possible to detect Down syndrome pre-birth through the use of string matching algorithms.

In the near future, bioinformatics may as well be a lifesaver for some, if not many individuals.

II. THEORIES

A. DNA and Chromosome

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria. Nuclear DNA are the types that are usually observed in DNA testing.

The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people[5]. The order, or sequence, of these bases determines the information available for building

and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences.

DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder.

An important property of DNA is that it can replicate, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell. However, sometimes a genetic mutation occurs that results in an abnormal replication of the DNA. Most of the times these genetic mutations are harmless. Unfortunately, some cases of genetic mutation results in complications like Down syndrome or the sickle cell disease.

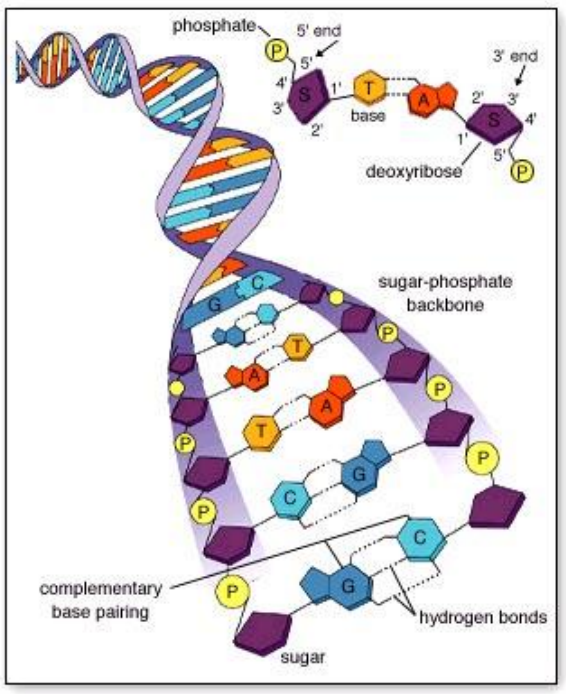


Fig 2. Illustration of the DNA chain
Source: Justin Taylor

Strands of nucleic DNA are would up tightly into a unit that is called a chromosome. Within each chromosome, DNA are tightly coiled around a lump of protein called histones that serves as an axis. A healthy human being have 23 pairs of chromosomes and each one are unique. They contain our genetic blueprints determine our physical qualities.

B. Down syndrome

Down syndrome is a genetic disorder caused by the presence of a third copy of the 21st chromosome. Victims of Down syndrome experience delayed physical growth and intellectual abilities. The average IQ of an adult experiencing down syndrome is 50. It is the intellectual equivalence of an 8 year old child[1]. Other than their impaired intellectual abilities, people with Down syndrome may have some or all of the following physical characteristics: a small chin, slanted eyes, poor muscle tone, a flat and wide face, a short neck, and a protruding tongue due to a small mouth and large tongue.

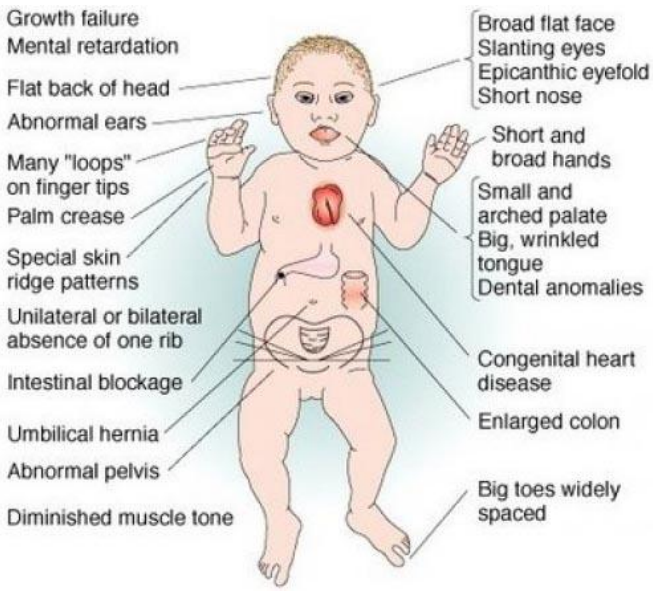


Fig 3. Symptoms of Down syndrome
Source: The Hong Kong Down syndrome Association

Other than their distinct physical appearances, people experiencing Down syndrome are also experiencing hearing and vision disorders. Cataracts and chronic ear infections are extremely common with people experiencing Down syndrome.

Although the overall risk of cancer is not changed, there is an increased risk of leukemia and testicular cancer and a reduced risk of solid cancers. Solid cancers are believed to be less common due to increased expression of tumor suppressor genes present on chromosome 21.

Males with Down syndrome usually are not able to father children, while females have lower rates of fertility relative those who are unaffected. Fertility is estimated to be present in only 40% of women experiencing Down syndrome[2]. Furthermore, menopause typically occurs at an earlier age. The poor fertility in men is thought to be due to problems with sperm development. However, it may also be related to not being sexually active. As of 2006 there have been three recorded instances of males with Down syndrome fathering children and 26 cases of women having children in North America and Europe. Without assisted reproductive technologies, approximately half of the pregnancies of someone with Down syndrome will also have the syndrome.

C. Pregnancy and Early Detection of Genetic Disorders of the Fetus

Genetic disorders, are often diagnosed through physiological and psychological tests. However, when the person being observed is not yet born, it is impossible to perform a physiological and/or psychological test. The only way to do so is through DNA testing.

Obtaining the child's DNA is extremely tricky. Almost all procedures are invasive and some procedures even includes high risks of miscarriages. The two most common methods are through amniocentesis and CVS.

Amniocentesis (also referred to as amniotic fluid test or AFT) is a medical procedure in which a small amount of amniotic fluid, which contains fetal tissues, is sampled from the amnion or amniotic sac surrounding a developing fetus, and the fetal DNA is examined for genetic abnormalities. Amniocentesis is usually done when a woman is between 16 and 22 weeks pregnant. This test is developed by a person by the name of Richard Dedrick.

CVS (Chorionic Villus Sampling) is a diagnostic procedure which involves removing some chorionic villi cells from the placenta at the point where it attaches to the uterine wall. There are two ways that samples are collected. Trans-cervical, where an ultrasound guides a thin catheter through the cervix to one's placenta and the chorionic villi cells are gently suctioned into the catheter; and trans-abdominal, where an ultrasound guides a long thin needle through the abdomen to one's placenta and the needle draws a sample of tissue and then is removed. CVS is usually done when a woman is between 10 and 13 weeks pregnant.

D. Brute Force Algorithm

The Brute Force Algorithm is the most general and simple way of approaching a problem. The application of this algorithm is no particular to string matching. It can be used to solve numerous of real world problems given enough computing power.

The idea is, to generate all of the possible scenarios then see if any of them constitutes as the solution.

In string matching, the Brute Force algorithm simply checks all of the characters within the pattern and match it to the text, if one of the doesn't match, the search moves one character to the right to the point where it reaches the end of the string, or the complete match is found.

The following is the implementation of the Brute Force algorithm in Java:

```
public static int brute(String text,String
pattern) {
    int n = text.length(); // n is length of
text
    int m = pattern.length(); // m is length
of pattern
    int j;
    for(int i=0; i <= (n-m); i++) {
        j = 0;
        while ((j < m) && (text.charAt(i+j)
               == pattern.charAt(j))){
            j++;
        }
        if (j == m)
            return i; // match at i
    }
        return -1; // no match
} // end of brute()
```

[3]

The Brute Force algorithm achieves a worst-case complexity of *O(mn)*, a best-case complexity of *O(n)*, and an average-case complexity of *O(m+n)*.

E. Knuth-Morris-Pratt Algorithm

The Knuth-Morris-Pratt algorithm was conceived by Donald Knuth and Vaughan Pratt and independently by James H.Morris in 1977. This algorithm was the first linear-time string matching algorithm at that time. Rather than checking each and every character, it minimalizes the complexity by utilizing the border function.

The border function preprocesses the pattern to find matches of prefixes of the pattern with the pattern itself. The border function b(k) is defined as the size of the largest prefix of P[1..k] that is also a suffix of P[1..k]. It is also known as the failure function or fail function.

The following is the implementation of the Knuth-Morris-Pratt algorithm in Java:

```
private static int[]computeFail(String pattern)
{
    int fail[] = new int[pattern.length()];
    fail[0] = 0;

    int m = pattern.length();
    int j = 0;
    int i = 1;

    while (i < m) {
        if (pattern.charAt(j) ==
            pattern.charAt(i)) {
            fail[i] = j + 1;
            i++;
            j++;
        } else if (j > 0) {
            j = fail[j - 1];
        } else {
            fail[i] = 0;
            i++;
        }
    }
    return fail;
}

private static int kmpMatch(String text,
String pattern) {
    int n = text.length();
    int m = pattern.length();
    int fail[] = computeFail(pattern);

    int i = 0;
    int j = 0;
    text = text.toLowerCase();
    pattern = pattern.toLowerCase();
    while(i<n){
        if(pattern.charAt(j) == text.charAt(i)){
            if(j == m-1)
                return i-m+1;
            i++;
            j++;
        }else if(j>0)
            j = fail[j-1];
        else
            i++;
```

```
    }
    return -1;
}
```

[3]

The Knuth-Morris-Pratt algorithm is able to achieve a linear complexity of O(m+n), extremely efficient when compared to the Brute Force algorithm. This algorithm is good for processing very large text files. However, Knuth-Morris-Pratt Algorithm gets increasingly inefficient as the size of the alphabet increases, as the value of the border function gets lower and lower.

F. Boyer-Moore Algorithm

The Boyer-Moore algorithm was developed by Robert S. Boyer and J Strother Moore in 1977. It is currently the most used algorithm for practical search of english literature, and it also serves as a benchmark for other algorithms. The algorithm preprocesses the string being searched for (P), but not the string being searched in (T). It employs the looking-glass technique and the character-jump technique.

In the Boyer-Moore algorithm, situations are divided into three cases:

Case 1:

If P contains x somewhere, then try to shift Pright to align the last occurrence of x in P with T[i].

Case 2:

If P contains x somewhere, but a shift right to the last occurrence is not possible, then shift P right by 1 character to T[i+1].

Case 3:

If cases 1 and 2 do not apply, then shift P to align P[1] with T[i+1].

The following is the implementation of the Boyer-Moore algorithm in Java:

```
private static int[] buildLast(String pattern){
    int last[] = new int[256];

    for(int i=0;i<256;i++)
        last[i] = -1;

    for(int i=0; i<pattern.length(); i++)
        last[pattern.charAt(i)] = i;

    return last;
}

private static int bmMatch(String text, String
pattern){
   int last[] = buildLast(pattern);
   int n = text.length();
   int m = pattern.length();
   int i = m-1;

   if(i > n-1)
       return -1;
   int j = m-1;
   do{
       if(pattern.charAt(j) == text.charAt(i))
           if(j == 0)
               return i;
           else{
               i--;
               j--;
           }
       else{
           int lo = last[text.charAt(i)];
           i = i + m - Math.min(j, 1+lo);
           j = m-1;
       }
   }while(i <= n-1);
   return -1;
}
```

[3]

The Boyer-Moore algorithm achieves a worst-case complexity of *O(nm+A)*. However, if the size of the alphabet (A) is large, the Boyer-Moore algorithm gets increasingly fast. Therefore, the Boyer-Moore algorithm is extremely efficient in matching patterns of the Latin text.

G. DNA Code Compression

One strand of DNA is about 1 meter long, and all of your DNA in your body uncoiled would be the distance from the earth to the sun multiplied by one thousand. That makes the code written in our DNA extremely long (3 billion sequences) [5]. Therefore, in the case of applying pattern matching in scanning our DNA for genetic disorders, it is exceedingly important to compress the data as much as possible, to reduce the time needed to calculate the result.

DNA are coded with four letters: A, G, T, and C. They are then grouped by the size of three where each of the group corresponds to one particular type of protein. Since DNA are grouped by three, it would be possible to compress the size of the amino acid chain by the factor of three.

This compression can be done by utilizing two of the special characteristics of the DNA itself. First, the fact that DNA is coded using four distinct letters. We can replace these letters with an integer. Second, the fact that DNA is coded in groups of three. These three integers can be “merged” into one by using a hash function. The steps of compression are the following:

1. Replace each code of the DNA with the corresponding letter. A with 0, C with 1, T with 2, and G with 3.
2. Divide the DNA pattern into groups of three, consecutively.
3. For each group, multiply the leftmost integer by a factor of 42, multiply the integer in the middle by a factor of 4, multiply the leftmost integer by a factor of 1, and then add all of them together.
4. This algorithm will results in the pattern being one third the size of what it was before. Also it will be in integer form and it will be easier to process.

[4]

III. Implementation and Analysis

A. DNA Testing Through the use of String Matching Algorithms

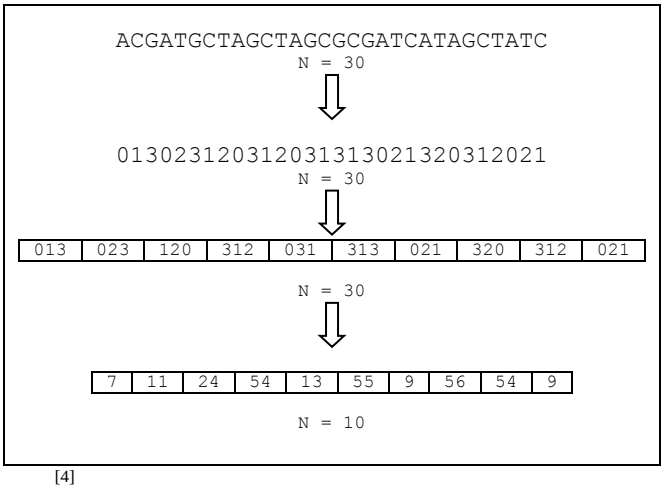
First, it is determined that our blueprints are coded with four bases, and they each corresponds to a letter. To utilize string matching algorithms, the biological code have to be first digitally encoded to make them ready for processing.

This digitalization of our DNA, is extremely complicated. First, one's cheek cells are extracted from one's mouth. Then, they are submerged in a solution to break down their cell membrane and nuclear wall. This is done so that the DNA is exposed and ready for extraction. After the DNA is fully exposed, alcohol is poured on to the solution. The alcohol will float on top because of its low density. The strands of DNA will also rise to the top mixing with the alcohol. Finally, the strands of DNA can be extracted for further testing.

After the DNA is isolated, only then can the DNA sequence can be read. The Human Genome Project have complied a few methods on extracting the DNA code from the DNA. In the current time, it still take a significant amount of time and effort to extract these codes. However, in the near and foreseeable future, there should be a significant increase in the speed and significant decrease in the effort of which DNA pattern can be read.

Following the digitalization of the DNA sequence, the reading of the DNA can be executed. To decrease the size of the pattern, a hash function is used so that the code is compressed (as mentioned above).

As an example, the compression of a short strand of DNA is described below:



[4]

Finally, the DNA sequence can be read and matched to see if the gene contains certain characteristics or not.

B. Detection of Down syndrome Before Birth through DNA Testing

The DNA of a person experiencing Down syndrome is different than those who are normal. Down syndrome is caused by the extraneous existence of the 21st chromosome, also known as trisomy 21.

The most common way of detecting Down syndrome is trough physiological and psychological tests. However, performing physiological and psychological test on an unborn fetus is impossible.

If a mother is pregnant with a fetus experiencing Down syndrome, she is only able to detect it through DNA testing. The DNA of the fetus is first extracted trough either the CVS or amniocentesis procedure. Through either of these procedures, then the DNA of the unborn fetus can be examined. The procedure of the examination is the same as a regular DNA examination once the extraction of the fetus's cell is performed.

As mentioned above, the strands of DNA are first extracted, then observed. After the DNA sequence have been digitalized, it is now possible to apply pattern matching algorithms to determine whether or not the fetus have Down syndrome.

A fetus with Down syndrome have a third copy of chromosome 21. Therefore, it would be extremely easy to apply computation to see if the third copy of chromosome 21 exists.

All of the algorithms mentioned in this paper is applicable to solve for the detection of Down syndrome. For this case, the pattern (P) that is being matched is the whole chromosome 21 itself. The text (T) is the whole DNA. The goal is not only to obtain one match, but to find three matches. If three matches are found, then the fetus tested positive for Down syndrome. If two matches are found the fetus tested negative for Down syndrome.

While each algorithms mentioned in this paper is applicable, they all performed in a different way. The Brute Force algorithm for example, examines all possibilities and take a lot of computational power to perform. Some algorithms are more efficient than others.

C. Determining the Most Effective Algorithm

The Brute Force algorithm achieves a worst-case complexity of $O(mn)$, a best-case complexity of $O(n)$, and an average-case complexity of $O(m+n)$. The Brute Force algorithm makes all of the comparison necessary to perform the task. Some comparison however, are extraneous and not needed. Therefore, the Brute Force algorithm is not the most effective algorithm to tackle the problem of pattern matcing in our DNA.

The Boyer-Moore algorithm achieves a worst-case complexity of $O(nm+A)$. However, if the size of the alphabet (A) is large, the Boyer-Moore algorithm gets increasingly fast. The Boyer-Moore algorithm is a major improvement when compared to the Brute Force algorithm. Unfortunately, the Boyer-Moore algorithm is only effective when the alphabet is large and the chance of a mismatch is extremely high. Our DNA's alphabet consists of only four letters: A, G, T, and C. Since the alphabet is very small, it is not optimal to apply the Boyer-Moore algorithm to solve this problem.

The Knuth-Morris-Pratt algorithm is able to achieve a linear complexity of $O(m+n)$, extremely efficient when compared to the Brute Force or the Boyer-Moore

algorithm. This algorithm is also good for processing very large text files. The only downside is that the Knuth-Morris-Pratt Algorithm gets increasingly inefficient as the size of the alphabet increases, as the value of the border function gets lower and lower.

Due to the fact that the size of the alphabet of our DNA is very small (only four), and the fact that the text in question is exceedingly large, the most effective algorithm to use for early detection of Down syndrome is the Knuth-Morris-Pratt algorithm.

IV. CONCLUSION

In conclusion, the urgency of detecting pre-birth whether or not a fetus has Down syndrome is exceedingly important since Down syndrome occurs in one in every 691 live births[1]. The occurrence of Down syndrome can be detected trough either the CVS (Chorionic Villus Sampling) or Amniocentesis procedure. After the cells are extracted, the DNA is then taken out of the cells and the DNA sequence is digitalized and compressed by utilizing a hash function. Once digitalized and compressed, pattern matching algorithms can be applied to it and a result can be obtained. It is also concluded that the most effective algorithm to be applied in the DNA testing is the Knuth-Morris-Pratt Algorithm.

REFERENCES

[1] http://www.ndss.org/Down-Syndrome/Down-Syndrome-Facts/ Accesed at May 17th 2014.

[2] Nelson, Maureen R. (2011). Pediatrics . New York: Demos Medical. p. 88. ISBN 978-1-61705-004-6H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[3] Munir, Rinaldi "Diktat Kuliah IF3051 Strategi Algoritma," Teknik Informatika, Institut Teknologi Bandung, Bandung, 2014

[4] Jamasoka, Septu "Modifikasi String dan Pattern untuk Mempercepat Pencocokan Rantai Asam Amino pada Rantai DNA" Teknik Informatika, Institut Teknologi Bandung, 2011

[5] Matthews, Harry R. DNA Structure Prerequisite Information. 1997.

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, 17 Mei 2014

Tirta Wening Rachman
13512004