

Primer Set Selection Satisfying Resolution Conditions in PCR Amplification based on Greedy Algorithm

Pipin Kurniawati - 13513089
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
pipinkurniawati@students.itb.ac.id

Abstract—In molecular biology, one of the most important techniques used to amplify small segments of DNA, which is known to the scientific community as Polymerase Chain Reaction (PCR), is selecting the suitable set of primers. PCR cleverly exploits the DNA replication machinery in a cyclic reaction that creates an exponential number of copies of specific DNA fragments. Because significant amounts of a sample of DNA are necessary for molecular and genetic analyses, studies of isolated pieces of DNA are nearly impossible without PCR amplification. Among many applications of the PCR experiment such as DNA fingerprints, genome typing, etc., require PCR amplifications of many different target objects, and producing these amplifications in a series of experiments by grouping them can save experimental costs greatly. This saving can be achieved by designing a good collection of primers for PCR amplification process. In this paper I will outline the design of a Greedy approach for selecting set of primers, given all conditions needed in PCR amplification.

Index Terms—amplification, DNA fragment, PCR experiment, primer

I. INTRODUCTION

We are truly living in the age of information technology. Since the mid-90s, the world has seen a rapid advancement in IT. Civilization has finally found a way to take advantages of this situation by creating interdisciplinary field that could develop methods and software tools for understanding biological data. This tremendous field is often called bioinformatics.

In this modern era, the use of bioinformatics has significantly expanded. It is one of strongly related working fields in molecular biology. Major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by researchers. This deluge of molecular biology information has led to an absolute requirement for computerized databases to store, organize, and index the data, also for specialized tools to view and analyze the data.

The increasing reliability of bioinformatics, along with the discovery of structural model of DNA by the American James Watson and the Englishman Francis Crick, has led Kjell Kleppe, a researcher in Khorana's Lab, envisioned a replication process of a specific segment of DNA which latter were known to the scientific community as Polymerase Chain Reaction (PCR) amplification.

PCR is an efficient and rapid *in vitro* method for enzymatic amplification of specific DNA or RNA sequences from nucleic acids of various sources.^[1] PCR amplifies DNA sequences efficiently and this technique has been used for various purposes. A simple PCR reaction consists of a set of synthetic oligonucleotide primers that flank the target DNA sequence, target DNA, a thermostable DNA polymerase and deoxynucleotides (dNTPs). A repetitive series of cycles involving template denaturation primer annealing, followed by extension of the annealed primers, yields tremendous amounts of DNA. Because the strands synthesized in one cycle serve as a template in the next, a million-fold increase in the DNA amount is achieved in just 20 cycles.

Among many applications of the PCR experiment, DNA fingerprints, genome typing, etc., require PCR amplifications of many different target objects, and producing these amplifications in a series of experiments by grouping them can save experimental costs greatly. This saving can be achieved by designing a good collection of primers for PCR experiments.

Designing such a good primer set is a challenging problem, and this paper proposes greedy algorithm for finding a small collection of primers satisfying resolution conditions in PCR experiments, given some biological constraint.

Some promising results of preliminary computational experiments are given. In the field of genome informatics, this primer selection problem might be rather new, and more explanations are needed to state the results in more detail. Hence, from here in this introduction, I will try to explain the backgrounds and requirements of this problem, in the following subsections, and then state the results.

II. FUNDAMENTAL THEORIES

2.1 DNA Sequence and Primer

Deoxyribonucleic acid (DNA) is a molecule that encodes the genetic instructions used in the development and functioning of all known living organisms and many viruses. DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a "polynucleotide." Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group. There are four different types of nucleotides found in DNA, differing only in the nitrogenous base. The four nucleotides are given one letter abbreviations as shorthand for the four bases.

- A is for adenine
- G is for guanine
- C is for cytosine
- T is for thymine

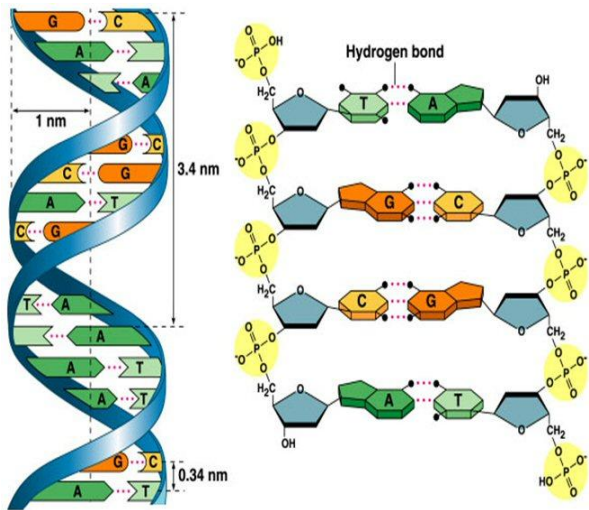


Figure 2.1 DNA Structure of Watson-Crick Model

Source : <http://kvhs.nbed.nh.ca/gallant/biology/biology.html>

The nitrogenous bases of the double helix are paired in specific combinations: adenine (A) with thymine (T), and guanine (G) with cytosine (C). Adenine and guanine are purines, nitrogenous bases with two organic rings, while cytosine and thymine are nitrogenous bases called pyrimidines, which have a single ring. Thus, purines (A and G) are about twice as wide as pyrimidines (C and T). A purine-purine pair is too wide and a pyrimidine-pyrimidine pair too narrow to account for the 2-nm diameter of the double helix. Always pairing a purine with a pyrimidine, however, results in a uniform diameter

Each base has chemical side groups that can form hydrogen bonds with its appropriate partner: Adenine can form two hydrogen bonds with thymine and only thymine; guanine forms three hydrogen bonds with cytosine and only cytosine. In shorthand, A pairs with T, and G pairs with C

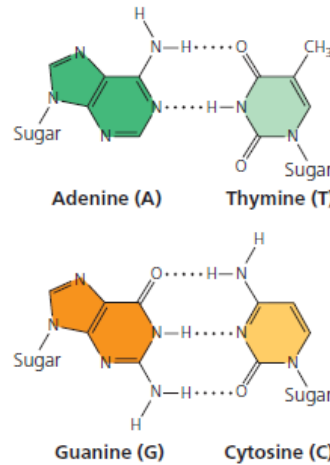


Figure 2.2 Base Pairing in DNA
Source : Campbell Biology 9th edition

Later Chargaff's Law says that in the DNA of any organism, the amount of adenine equals the amount of thymine, and the amount of guanine equals the amount of cytosine. Although the base pairing rules dictate the combinations of nitrogenous bases that form the "rungs" of the double helix, they do not restrict the sequence of nucleotides *along* each DNA strand. The linear sequence of the four bases can be varied in countless ways, and each gene has a unique order, or base sequence.

Another term that I would like to explain is primer. A primer is a strand of short nucleic acid sequences (generally about 10 base pairs) that serves as a starting point for DNA synthesis. It is required for DNA replication because the enzymes that catalyze this process, DNA polymerases, can only add new nucleotides to an existing strand of DNA. The polymerase starts replication at the 3'-end of the primer, and copies the opposite strand.

The two complementary strands of double-stranded DNA (dsDNA) are usually differentiated as the "sense" strand and the "antisense" strand. The DNA sense strand looks like the messenger RNA (mRNA) and can be used to read the expected protein code; for example, ATG in the sense DNA may correspond to an AUG codon in the mRNA, encoding the amino acid methionine. However, the DNA sense strand itself is not used to make protein by the cell. It is the DNA antisense strand which serves as the source for the protein code, because, with bases complementary to the DNA sense strand, it is used as a template for the mRNA.

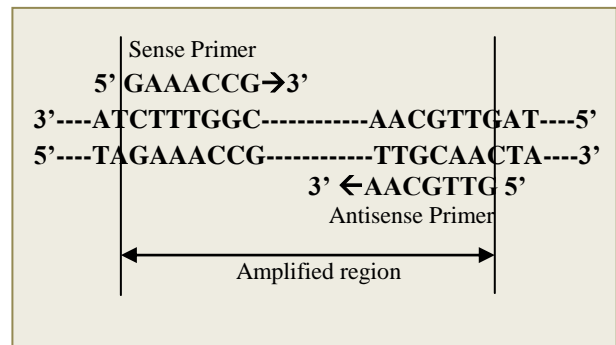


Figure 2.3 DNA Sequences and Primers

2.2 Polymerase Chain Reaction (PCR)

Polymerase chain reaction (PCR) is an *in vitro* DNA amplification protocol. It selectively amplifies a specific DNA sequence from any source (i.e. virus, bacteria, plant, human) hundreds of millions of times in a matter of hours. PCR is a technique necessary for several other genetic applications such as sequencing, RFLP analysis and microsatellite analysis.

PCR is very precise and can be used to amplify, or copy, a specific DNA target from a mixture of DNA molecules. First, two short DNA sequences called primers are designed to bind to the start and end of the DNA target. Then, to perform PCR, the DNA template that contains the target is added to a tube that contains primers, free nucleotides, and an enzyme called DNA polymerase, and the mixture is placed in a PCR machine. The PCR machine increases and decreases the temperature of the sample in automatic, programmed steps. Initially, the mixture is heated to denature, or separate, the double-stranded DNA template into single strands. The mixture is then cooled so that the primers anneal, or bind, to the DNA template. At this point, the DNA polymerase begins to synthesize new strands of DNA starting from the primers. Following synthesis and at the end of the first cycle, each double-stranded DNA molecule consists of one new and one old DNA strand. PCR then continues with additional cycles that repeat the aforementioned steps. The newly synthesized DNA segments serve as templates in later cycles, which allow the DNA target to be exponentially amplified millions of times.

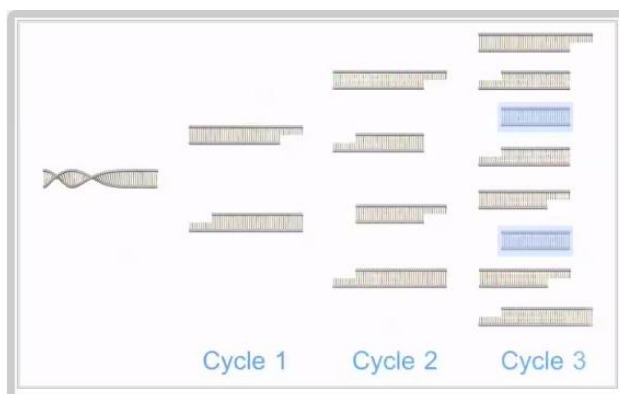


Figure 2.4 PCR process visualization

Source : <http://www.nature.com/scitable/definition/pcr>

2.3 Greedy Algorithm

Greedy algorithm is an algorithm that follows the problem solving heuristic of making the locally optimal choice at each stage, with the hope of finding a global optimum. In many problems, a greedy strategy does not in general produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a global optimal solution in a reasonable time, which is the reason that this algorithm is often used for the cases that require quick solutions such as real-time

systems or gaming.

Greedy algorithms work in phases. In each phase, a decision is made that appears to be good, without regard for future consequences. Generally, this means that some *local optimum* is chosen. This “take what you can get now” strategy is the source of the name for this class of algorithms. When the algorithm terminates, we hope that the local optimum is equal to the *global optimum*. If this is the case, then the algorithm is correct; otherwise, the algorithm has produced a suboptimal solution. Examples of everyday problems using the greedy approach:

- Choosing some type of investment (capital investment)
- Finding the shortest path from Bandung to Surabaya
- Playing remi cards

Greedy algorithm composed by the following elements:

- The candidate set: Contains elements forming solution.
- The set of solutions: Contains candidates selected as solutions to the problems.
- Selection function: Selecting the most likely candidates that are capable of achieving optimal solutions. Candidates that have been selected in one particular step will never be considered again at the next step.
- Feasibility function: Checking whether a candidate has been able to provide a viable solution, namely the candidates together with a set of solutions that have been formed do not violate the existing constraints. Then we add that viable candidates into the solution set, while the candidate which is not feasible get discarded and will never be considered again.
- The objective function: the function to maximize or minimize the value of the solution (eg: length, profits, etc.).

Greedy algorithms produce good solutions on some mathematical problems, but rarely on others. Most problems for which they work, will have two properties:

- Greedy choice property
We can make whatever choice seems best at the moment and then solve the sub-problems that arise later. The choice made by a greedy algorithm may depend on choices made so far but not on future choices or all the solutions to the sub-problem. It iteratively makes one greedy choice after another, reducing each given problem into a smaller one. In other words, a greedy algorithm never reconsiders its choices. This is the main difference from dynamic programming, which is exhaustive and is guaranteed to find the solution. After every stage, dynamic programming makes decisions based on

all the decisions made in the previous stage, and may reconsider the previous stage's algorithmic path to solution.

- Optimal substructure : A problem exhibits optimal substructure if an optimal solution to the problem contains optimal solutions to the sub-problems.

III. BIOLOGICAL CONSTRAINTS

In this paper, the following biological constraints on primers are considered.

- GC content: primers with 40-60% GC content are widely used
- Lengths of primers: In this paper, a primer can match multiple positions in DNA sequences. We can set short length (8-12) for primers
- Complementarity sequence: The set of primers must not contain self-complementary sequence and complementary sequences. (For example, 5'-GCCTAGGC-3' is a self-complementary sequence, 5'-GACAATGC-3' and 5'-GCATTGTC-3' are complementary sequences.)
- Length of amplified regions: PCR products which are between 50 and 500 base pairs are desired. In this paper, the lower bounds for amplified regions are considered.
- Difference of lengths of amplified segments: PCR products should have different lengths with each other to analyze products by electrophoresis. In practical, it is preferable that the difference of the lengths of the PCR products is 5.

IV. GREEDY APPROACH

Given n DNA sequences ($1 \leq j \leq n$). The length of DNA sequence j is denoted by m_j . $(j; p)$ means position p of DNA sequence j . A primer S_i is a DNA sequence of constant length much shorter than m_j . $|S_i|$ denotes the length of primer S_i . Consider a primer set $S = \{S_1, S_2, \dots, S_i, \dots, S_l\}$. $(j; p; r/l) \in S_i$ means primer i match position p of DNA sequence j , with direction r (right) or l (left), that is, for $k = 1$ to $|S_i|$ the $(p + k - 1)^{\text{th}}$ character of the sequence j is identical to the k^{th} character of the primer i in the case of r .

The primer selection problem involves finding the minimum primer subset $S' \in S$ that includes all the desired DNA sequences, and putting a DNA subsequence between sense and antisense primers for each DNA sequence. The length of amplified segment is $(p_2 - p_1 + 1)$. That is,

$$\begin{aligned} \forall j \exists S_i, S_i' \in S' \\ (j, p_1, r) \in S_i, (j, p_2, l) \in S_i' \\ p_2 - p_1 + 1 > 0 \end{aligned}$$

If a DNA sequence is amplified by primers $(j; p_1; r)$ and $(j; p_2; l)$ in PCR experiments, primers which match with position :

$$p(p_1 \leq p \leq p_2)$$

DNA sequence j cannot exist.

In this proposed algorithm, there has to be 2 constraints relating to the selection primer problem, they are selection primer problem with minimum length constraint and selection primer problem with distinguishable length constraint.

Primer selection problem with minimum length constraint is, for a given k , the finding a minimum primer subset $S' \subseteq S$ which covers all DNA sequences, and putting DNA subsequence at least length k between sense and antisense primer for all DNA sequences. ($p_2 - p_1 + 1 \geq k$ instead of $p_2 - p_1 + 1 > 0$)

This problem models a fact that amplified parts by a pair of primers S_i and S_i' have length at least 50.

The primer selection problem with distinguishable length constraint is a primer selection problem and the lengths of DNA subsequences between sense and antisense primers are all different, that is,

$$\forall j \exists S_i, S_i' \in S', (j, p_1, r) \in S_i, (j, p_2, l) \in S_i', p_2 - p_1 > 0$$

This problem models a requirement that may be DNA sequences distinguished with one another by the lengths of their amplified segments using the electrophoresis. If one amplified segment of different length exists for each DNA sequence, DNA sequences can be identified using the electrophoresis.

Basic idea for the proposed algorithm is that the primer selection problem is regarded as the set cover problem. The position p of the DNA sequence j ($j; p$) is regarded as the element of underground set. Primers are regarded as subsets.

All (j, p) ($1 \leq j \leq n, 1 \leq p \leq m_j + 1$) are covered only if amplified sequence exist in all DNA sequences.

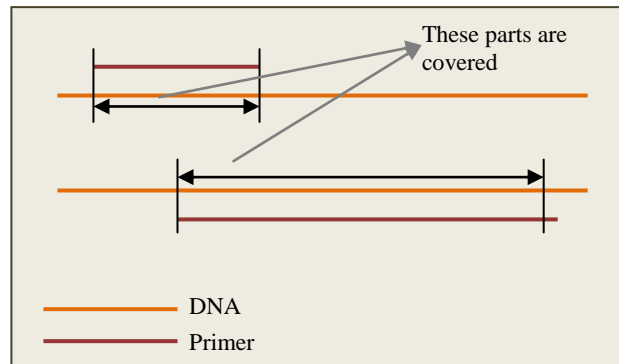


Figure 4.1 Covered Elements of DNA

Steps of greedy algorithm applied to this problem:

- First and the foremost step is scanning DNA sequence. Then pick up primers which satisfy biological conditions we have defined in the previous subsection. Put each of them to the list of candidates of primers

- Select the primer S_i that either covers the largest number of uncovered elements (subsequences of DNA) as in the Figure 4.1 or does not satisfy prohibit conditions. Put S_i to the solution and then delete S_i from the candidate list of primers.
- Repeat 2 until all elements of DNA are covered.

Here is the pseudocodes :

Scanning all bases of single stranded DNA

```
public List<Character> toList
    (String DNA){
    List<Character> NBase = new
        ArrayList<Character>();
    for (int i=0;i<DNA.Length();i++){
        NBase.add(s);
    }
}
```

Put all primers into the list

```
List<List<Character>> prim= new
    ArrayList<List<Character>>();
int i;
List<Character> temp = new
    ArrayList<Character>();
for (string s : primers) {
    for(i=0;i<s.length();i++) {
        temp.clear();
        temp.add(s.charAt(i));
    }
    prim.add(temp);
}
```

Selecting primers that most covering up

```
int max=0, index=0;
List <Character> NBase;
int i, counter;
for (String str : DNACHain {
    NBase=new ArrayList<Character>
        (toList(str));
    for(List<Character> p : prim) {
        counter = 0;
        while(NBase.size()>0) {
            i = 0;
            while(i<p.size()) {
                if(p.get(i)==NBase.get(i) {
                    i++;
                }
                else break;
            }
            if(i==p.size) counter++;
            if (NBase.size()>0) {
                NBase.removeAt(0);
            }
        }
        if (counter > max){
            max = counter;
            index = p.indexOf(p);
        }
    }
}
```

The proposed algorithms above consider the constraints on GC content and lengths of primers only in step 1 where candidates of primers are obtained. Modification to the basic algorithm is necessary for the other constraints.

V. ANALYSIS

This proposed greedy algorithms are applied to randomly selected 300 sequences out of 2000 ORFs (one ORF contains 3 nucleotides) which are randomly generated. The GC content is 50-55% in a primer and length of primer is 8 to 15. Considered lower bounds is 50 and 80. Extra attention is given to amplified segments whose lengths are between 50 and 500.

The analysis might result in an array of arrays, as can be seen on Figure 5.1.

```
Start Generating...

91, 2956, 997, 518, 0
87, 2438, 983, 529, 0
92, 1977, 860, 315, 0
92, 998, 717, 517, 1
90, 1356, 926, 602, 0
95, 2982, 597, 623, 0
92, 3001, 893, 491, 0
93, 2301, 896, 552, 0
91, 1187, 772, 528, 0
88, 981, 540, 503, 1
90, 2555, 919, 498, 0

Finish!
```

Figure 5.1 Result

The format of the result in Figure 5.1, ignoring the header and footer are as following:

```
<#primers>, <#match position>, <#amplified
segments>, <#amplified segments (lengths <
50)>, <number of DNA sequence which have
not amplified segments>
```

The basic algorithm can cover 300 sequences by 100 primers, each has length of 8 nucleotides. By modifying this algorithm by the complementary condition, the number of primers is almost the same. From the result, we can see that the number of primers selected to be used in PCR amplification has reduced by about 10 primers. This can save experimental costs greatly.

As the algorithm is modified by adding biological constraints, better solutions are obtained. This algorithm can also be modified with the existence of string matching algorithm to detect the match point between DNA sequence and primer. Because short primers match many

positions of DNA sequences, the number of primers for covering can be decreased. On the other hand, many short amplified segments are obtained. Considerations of other constraints other than biological constraints are needed to make this algorithm perform at its best.

VI. CONCLUSION

PCR is an efficient and rapid *in vitro* method for enzymatic amplification of specific DNA or RNA sequences from nucleic acids of various sources. One of the most fundamental elements in PCR Amplification is a set of synthetic oligonucleotide primers. The proposed algorithm in this paper has considered minimizing the number of primers for PCR Amplification by selecting suitable set of primers. As the greedy algorithms are mixed with biological constraints, solutions which almost satisfy each constraint are obtained. It is necessary to analyze approximate properties of these algorithms from the theoretical and practical viewpoints. Improving solution could be considered for these algorithms by meta heuristics or other methods. I will also consider take account of the other biological constraints and design the sets of primers for multiple experiments.

VII. ACKNOWLEDGMENT

The author would like to thank her parents for their utmost support. The author would also give her gratitude to Mr. Rinaldi Munir, and Mrs. Nur Ulfa Maulidevi for the knowledge they gave in class and all their support on the course IF2211 Algorithm Strategies (Strategi Algoritma) for the last semester, and also for the chance to write this paper. Last but not least, the author would also like to thank other people who had given their help and support in any way that lead to the finishing of this paper

REFERENCES

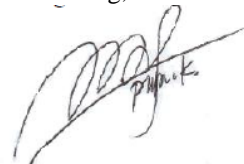
- [1] <http://www.sigmaaldrich.com/life-science/molecular-biology/molecular-biology-products.html>
Access Time : May 2nd 2015, 08:19AM
- [2] <http://www.nature.com/scitable/topicpage/dna-is-a-structure-that-encodes-biological-6493050H>
Access Time : May 2nd 2015, 08:25AM
- [3] R. Munir, Diktat Kuliah Strategi Algoritma, Bandung: Penerbit Sekolah Teknik Elektro dan Informatika, 2009.
- [4] Jane B. Reece, Michal L. Cain, Lisa A. Urry. *Campbell Biology*, 9th Edition. Pearson. 2011
- [5] Rosen, Kenneth H. *Discrete Mathematics and Its Applications*, 7th Edition. The McGraw-Hill Companies. 2012
- [6] W. Bains, G.C. Smith, A novel method for nucleic acid sequence determination, *Journal of Theoretical Biology* 135 (1988) 303–307.
- [7] K.M Konwar, Improved Algorithms For Multiplex PCR Primer Set Selection with Amplification Length Constraints, Department of Computer & Science Engineering, University of Connecticut

- [8] Khoichiro Doi, A Greedy Algorithm for Minimizing The Number of Primers in Multiple PCR Experiments, Department of Information Science, Faculty of Science, University of Tokyo
- [9] M.T.Hajiaghayi, The Minimum k-Colored Subgraph Problem in Haplotyping and DNA Primer Selection

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 Mei 2010



Pipin Kurniawati - 3513089