Bioinformatics Lab Manual
&
Class Notes

Vivek Rai
12BT30025

April 10, 2015

Primer Design

Primer design has important applications in PCR, DNA sequencing, and hybridization; thus at the core of almost all biological activity today.

Objective: To understand the process of PCR, importance of primers, considerations involved in designing a primer culminating with a hands on session on actual design of regular and degenerate primers.

Primers are short length (12-20 bp) oligonucleotides which serve as the 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.This, however, reveals a very important limitation of primers that we need to have at least some information about the target sequence. In case, the sequence is exactly known we can easily design a primer based on general considerations but otherwise a *degenerate* primer has to be designed with more general target (which in turn reduces the reliability of amplification).

General considerations:

- Melting temperature (T_m) between 55 and 65°C (usually corresponds to 45-55% G+C for a 20-mer)
- Absence of dimerization capability.
- Absence of significant hairpin formation (usually > 3 bp).
- Lack of secondary priming sites in the template.
- Low specific binding at the 3' end, to avoid mispriming.

Degenerate Primers are used to amplify DNA when the target sequence is not accurately known. In this case, we compromise with the target specificity by synthesizing several primers, each having different nucleotide at a particular position. For example, if the protein sequence is known and we want to amplify the corresponding genes, then our primer should include the bases matching to *conserved regions* in protein, and If some of the residues are not completely conserved, then the oligo sequence will need to accommodate all possible codons of all amino acid residues at that site

Laboratory Session

Regular Primer Design

1. Obtain FASTA sequence for a nucleotide (AF103742) from NCBI nucleotide database.

2. We then use Primer3Plus and OligoPerfect tools available online to design our Primer by specifying the downloaded sequence, and parameters taken into concern from previous page (i.e., melting temperature, GC content etc.,)

   Thermodynamics process - SantaLucia 1998
   Salt correction formula - SantaLucia 1998
   Max-PolyX - 3
   Max T_m difference - 5°C

   The optimal values of an annealing temperature, GC content, T_m can be calculated using established formula.

3. We also practiced a the design of a primer for cloning experiment by incorporating suitable restriction sites by referring to vector map of a plasmid and cross checking target sequence to be cloned for restriction sites using NEBCutter.

Degenerate Primer Design

1. We try to design a degenerate primer for a protein called *sericine*, usually found in silk-producing organisms.

2. From the NCBI website we retrieve a couple of protein sequences for the protein of interest and download them in 'sequence.fasta' file.

   We ensure that the protein sequences are within 600-900 aa in length.

3. A clustalW alignment is performed and obtained multiple sequence alignment is saved in file 'clustalw-alignment.clustalw'.

   Now we examine the start and end sequences of obtained alignment file.

   The alignment start for similar nature of organisms is: MRFVLCC
   The alignment end for similar nature of organisms is : LRKNIGV

   As we can see, both of these short sequences contain L, R; the amino acids that we tend to avoid in primer design (because of degeneracy). For this, we consult different online databases that enlist the codon usage pattern in the organism of our interest and try to locate the pattern.

   This information can help us design more specific primers (although primer is intended to be degenerate, we don't want the degeneracy to be too high to reduce the specificity).

4. Corresponding to both these peptide sequences, we design an oligonucleotide, accounting suitably for degenracy.

   For example,

   For forward primer, MRFVLCC

M - ATG
R - (A + C)G(A + T + G + C) = MGN
F - TT(C + T) = TTY
V - GT(A + T + G + C) = GTN
L - (C + T)T(A + T + G + C) = YTN
C - TG(C + T) = TGY

Complete sequence: **ATGMGNTTYGTNYTNTGY**

For reverse primer, LRKNIGV

R - (A + C)G(A + T + G + C) = MGN
K - AA(A + G) = AAR
N - AA(C + T) = AAY
I - AT(A + T + C) = ATH
G - GG(A + T + G + C) = GGN
V - GT(A + T + G + C) = GTN

Complete sequence: **MGNAARAAYATHGGNGTN**

Reverse complement: **NACNCCGATRTTYTTNCK** (Final primer)

Here we can note that R unnecessarily introduces a degeneracy in form of N. Hence, we can drop that amino acid without any harm.

NOTE:

- While designing primers, ensure that you avoid L, R, S amino acids as much as possible. Simply because they are highly degenerate and are coded by 6 codons.
- We try to avoid degeneracy at the 3' end of the sequence because that is where polymerase starts the synthesis. Non degeneracy would ensure correctness and accurate PCR product.
- We also take into account any restriction site, if used, by adding 2-3 bases upstream of primer since many restriction enzymes require a few bases to *sit*.

Sequencing

Question 1. Why it is called the shotgun genome sequencing?

In this sequencing method, a DNA is broken randomly into a large number of small fragments (50-200 Kb) and then sequenced using *chain termination* methods to obtain reads. Since, this appears analogous to shooting down an object and breaking into the pieces, hence the name.

Question 2. Why is it called NGS?

Sanger sequencing was one of the first sequencing methods that was used to obtained a **finished genome** of an organism. The process, however, was very expensive and time consuming. Later, newer technologies were invented which made a major depart from original approach and allowed large scale sequencing at inexpensive rates. The massively parallel nature of these methods along with highly automated method allowed for rapid progress and whole genome sequencing of many organism in succession.

Because of these major changes, the newer technologies were called *next generation sequencing* platforms.

Question 3. What is the function of the variability of insert size?

Insert size refers to the length of DNA sequence between the two primer ends of a template. If the insert size is large and constant, that fragment will never be completely sequenced (or require a large number of attempts) because there will be a gap left in the between. This introduces unreliability in the results which can be avoided by having variable insert sizes. It will ensure that a large number of fragments are sequenced completely and posses overlapping reads, which in turn allow us to be stitch the sequences with more confidence, thereby improving sequence coverage.

Question 4. What is called bridge amplification and why is it required?

Bridge amplification is a solid-phase template amplification technology used by Illumina/Solexa sequencing platform. In this method, universal primers (both forward and reverse) are immobilized on a surface and DNA templates (ligated with similar primers) are allowed to hybridize in the presence of polymerase. These complimentary primers of DNA and the surface anneal partly to immobilize the template as well. The DNA template then forms a bridge with the immediately adjacent primer (using the other free end) and is amplified to form a cluster

of the same template.

Question 5. How Illumina technique is unique compared to other technologies?

Illumina/Solexa sequencing platforms just use DNA polymerase thereby avoiding the need of multiple reagents, use fragmented solid-phase template amplification, and is one of the most widely used and well supported platforms.

Question 6. What is the different between FASTA and FASTQ files? What comes with what function?

FASTA is a file format for storage and transfer of nucleotide and protein bases. FASTQ on the other hand, although similar to FASTA files, is generated by sequencing platforms and contains the raw reads along with their corresponding quality scores.

FASTA format:

```
>gi|39748133|emb|BX571963.1| Rhodopseudomonas palustris CGA009 complete genome
ATCGGTCGAGGCGAAATCTTCACCCTGCCCTCGGAATCATATCCATTGCAGCGGAGGGGCCGTCGTGGTT
TTCATAGTCCACCCGCGACGCCCACGGCTCTTCAGATCAGCGCGGTTTGAGAACCAAGGGCGGACATGCA
```

FASTQ format:

```
@HWUSI-EAS300R_0005_FC62TL2AAXX:8:30:18447:12115#0/1
CGTAGCTGTGTGTACAAGGCCCGGGAACGTATTCACCGTG
+HWUSI-EAS300R_0005_FC62TL2AAXX:8:30:18447:12115#0/1
acdd^aa_Z^d^ddc‘^_Q_aaa‘_ddc\dfdffff\fff
```

Question 7. What is generated after the assembly of the sequence reads?

A contig.

Question 8. How specific function of a stretch of a nucleotide can be predicted?

Predicting the role and function of a stretch of nucleotide is one of the popular problems in bioinformatics called *gene annotation*. It may often require assistance and verification from experimental results.
In general, if nothing is known about the sequence, one can start by BLASTing the sequence and finding if there are already annotated sequences available. Based on the results obtained, on can further investigate the presence of conserved regions, promoter sites etc.,

Question 9. Why adapters are required but primers are not?

Primers are, in general, used when the sequence of the target DNA is already known. In this case, however, we are un aware of the sequence of target DNA - in fact, that itself is the goal of the experiment. Thus, sequencing platforms use *adapters* during the *template amplification* process which is ligated to both of ends of DNA and can anneal to the universal primers. This ensures that complete DNA is amplified and no bases are lost due to primer removal.

Question 10. How can a person do multiple samples at a time? Can it be done?

Yes, multiplexed sequencing can be done and is readily supported by several sequencing platform providers. In this method, DNA libraries are *tagged* with a unique identifier, or index, during the sample preparation. Multiple samples are then pooled into a single sample, analyzed, and segregated based on the identifiers or tags.

QIIME

Question 1. What are the major advantages of QIIME over the other rRNA analysis tools?

QIIME (pronounced *chime*) is one of the recent developed, open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data (Caporaso *et al*, 2010). Compared to other available rRNA analysis tools, it offers distinct advantages such as:

- Ability to read and process large amount of genomic data generated from next generation sequencing methods,
- Ability to run and access multiple software at a time,
- Support for various file formats, demultiplexing and quality filtering, OTU picking, taxonomic assignment, and phylogenetic reconstruction, and diversity analyses and visualizations,
- Cross platform, easy to use, open-source and free of cost

Question 2. What are the pre-processing steps in QIIME?

QIIME has several preprocessing data filters like -

- Primer removal
- Demultiplexing
- Denoising, Quality filter and Chimera Checking

Question 3. What are the different methods of OTU picking?

QIIME provides three ways of OTU picking namely, *de novo*, closed-reference, and open-reference OTU picking.

- *De novo:* reads are clustered against one another without any external reference sequence collection.

- closed-reference: reads are clustered against a reference sequence collection and any reads which do not hit a sequence in the reference sequence collection are excluded from downstream analyses.
- open-reference: reads are clustered against a reference sequence collection and any reads which do not hit the reference sequence collection are subsequently clustered de novo.

Question 4. Compare α and β diversity?

α and β diversity were coined by R. H. Whittaker to describe the total species diversity in a landscape. Any such description according to him, requires describing different species and their interaction at two levels - local or habitat and global.

- α-diversity means the diversity of the community within one site (or one sample), i.e., the number of species and their proportion within one sampling site.
- β-diversity means the dissimilarity between communities of two sites (or two samples) due gain or loss of species from replacement or biotic changes along environmental gradients.

Question 5. What are the different parameters required for the quality filtering?

Quality filering in QIIME is done using USEARCH, which is a sequence analysis tool. The quality of raw reads after filtration depends on many parameters such as -

- Expected Errors (E): Discarding reads based on the total expected error for all bases in the read.
- Truncation Length (L): Truncating sequences at the L^{th} base and discarding if the length is less than L.
- Quality score (N): Discarding all reads having quality score $<= N$.
- Strip Length (L): Number of sequences to strip (delete) from starting of the sequence

Question 6. What are the tools used for aligning the sequence, assigning taxonomy in de novo OTU picking?

QIIME uses MUSCLE for sequence alignment (pyNAST is a suitable alternative) and Uclust Consensus Taxonomy classifier for assigning taxonomy in *de novo* OTU picking.

Question 7. Write down the workflow of QIIME for analyzing the ecology of environmental samples through 16S rRNA amplicon?

The workflow consists of the following steps:

- **Obtaining the data**
- **Validate the mapping file**
- **Demultiplex and quality filer reads**
- **OTU picking**
- **Summarize communities**
- **Diversity indexes**

Question 8. What is the necessity of microbial diversity study?

Microbes play an inevitable role in almost every sphere of life - ranging from the deep trenches of the oceans to freezing glaciers, from hot and extreme conditions of hot springs to the gut of animals. They carry about important processes in the nature and are present at every imaginable place on the Earth. These constant biotic and abiotic challenges have equipped microbes with a very rich set of biochemical activity (by virtue of their genes). Their characterization, study, and analysis of diversity allows us to understand complex and novel processes, and in some ways harness that knowledge for our intellectual and economical use.

Question 9. Why is QIIME needed?

QIIME encompasses diversity analyses tools and techniques from several areas which otherwise are fragmented and difficult to collect. Using QIIME, one can process his entire collection of raw reads and obtain final results with minimal number of steps.

Phylogeny

The study of evolutionary relationship.

All the organisms on earth have descended from a common ancestor, which means that set of species, extant or extinct, is related. This relationship is Phylogeny, and the study of existing organisms' morphological, physiological, and molecular characteristics to create a tree – is called Phylogenetics. **Molecular Phylogenetics** uses the structure and function of the molecule and how they change over the time to infer relationships.

Phylogenetic analysis is the process of testing hypotheses about the descent of species from a common ancestor.The null hypothesis in most cases is that the organisms have a common ancestor (it is because it is natural to accept the fact that any two organism must have been related)

Laboratory session

Objective: 16s rRNA Analysis to determine the relationship of an unknown sequence.

1. We get the unknown sequence and BLAST it against three databases (NCBI,RDP, Ez Taxon).
2. We take the first five results from all three blasts and put it in single file. We also add a sequence from Archeabacteria to act as a negative control.
3. We then use **MEGA** to generate tree.
4. Output files are then saved and submitted to faculty.