Sequence Editing, Primer Design, Protein Translation

Shifra Ben-Dor

WEIZMANN INSTITUTE OF SCIENCE

So, you have a DNA sequence. Now what?

Sequence Editing

It is critical to have an accurate copy of the sequence you plan to work with.

Whether you are cloning a known gene, designing a fusion protein, or planning PCR, you should have your ideal sequence in-silico before you start in the lab.

This can save much time, trouble and heartache.

Sequence Editing

There are various programs available for simple sequence editing:

- EMBOSS
- SnapGene
- MacVector
- LaserGene
- Benchling
- ApE (a plasmid editor)
- Word (Microsoft Office)

Sequence Editing

The important things to remember when choosing a program:

- Does it let me jump around the sequence based on coordinates? Sequence?

- How easy is it to combine two existing files?

- File storage?

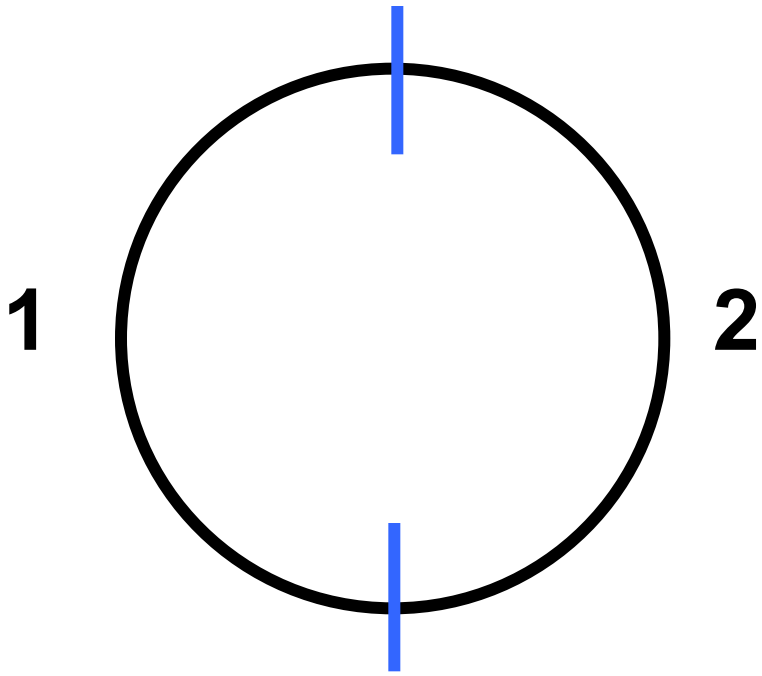
Sequence Editing

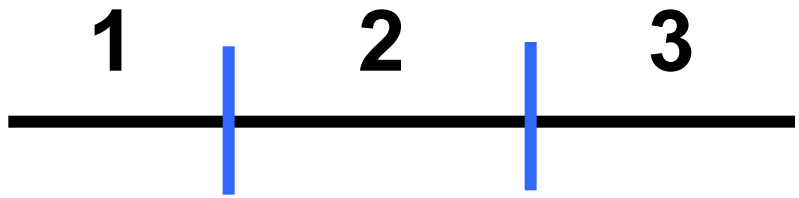
Don’t forget to be very careful with sequence “joins”

- If you are putting sequence into a multiple cloning site, erase what is in-between!

- If you are joining at an enzyme site, be sure you know what each sequence is contributing

Restriction Maps





Attributes of mapping programs

- Choice of enzymes
  - Single cutters
  - x base cutters (6 base)
  - minimum/maximum of sites
- Linear/circular
- Simulation of double digests

Attributes of mapping programs

- Silent mutations
- Output
  - Annotated sequence
  - Table of sites (sorted by enzyme name or position)
  - Table of fragment sizes (sorted by size or position)
  - Restriction site (the actual sequence)
  - Those that do/don’t cut

REVERSE

- Can reverse, complement or reverse and complement a nucleotide sequence

- File remains nucleotide sequence, does not translate

Primer Design

When do we need primers?

- Sequencing (one primer)
- PCR (two, one for each strand)
  - Exact (cloning, add tags, add enzyme sites, site directed mutagenesis, …)
  - Degenerate
- Real time quantitative PCR (qPCR)

Primer Design

- Things to keep in mind:
  - Primer length
  - AT/GC ratio should be around 50%
  - 3’ end should be G/C
  - melting/annealing temperature
  - secondary structure
  - primer dimers

Primer Length

- Primers have to be long enough to be specific, but short enough to detach efficiently from the template
- Ideal lengths are from 18-24 bp long
- For some applications, we use longer ones (adding enzyme sites, tags, changing the end of a sequence…)
- We rarely use shorter ones

GC ratio

- If there are too many Gs and Cs, it will be hard to separate the primer from the template (G and C have 3 hydrogen bonds)
- We generally try to keep the G/C percentage as close to 50% as possible, with a range of 40% - 60%
- If nothing is found, expand the range

3’ clamp

There is a running argument in the literature as to what base is preferable at the 3’ end. Some maintain that an S clamp (G or C) makes for better priming, others say it makes it worse. We generally recommend using an S clamp (unless you’re doing qPCR, in which case an A is recommended)

Melting temperature

- The melting temperature of the primers directly effects the temperature of the annealing step of PCR.
- Currently accepted norms: primer melting temperatures in the 58^{o}C - 60^{o}C range
- The difference in melting temperatures of primers should be as little as possible, but can be up to 5^{o}C

Annealing temperature

- The "rule of thumb" for annealing temperature: it should be 5^{o}C less than the melting temperature
- Optimally, it should be determined for each set of primers on a gradient cycler
- Currently accepted: a minimum of 50^{o}C
- It works down to 37^{o}C, but specificity may become an issue
- If you're working with degenerate primers, you need lower temperatures, though you can use them for a few initial cycles

Secondary structure

Internal complementarity:

There should be no self matching stretches of 3 bases or more, or the primer will bind to itself in a hairpin, and not be able to prime

Other Primer Issues

- Primer Dimers

When the 3’ end of one primer is complementary to the other primer, the primers can anneal to each other and create a new template

- Primer Complementarity

If the primers are complementary anywhere else, it can interfere with hybridization

- Primer/Template:

Avoid stretches of 3 bases or more in a row of the same base - it can lead to mispriming (G, C) or breathing (A, T)

Primer Design

- If you are changing the beginning of a coding region:
  - ATG start codon
  - Kozak sequence
    (GCC) GCC (A/G)CC ATG G
  - signal sequence (secreted, membrane bound)

Reverse (**not complement**)
3’ primer

5’ 3’

GATAAGCTTGATATCGAATTGCCATGTTGAAGCCATCATTACCATT

CTATTCGAACTATAGCTTAACGGTACAACTTCGGTAGTAATGGTAA

5’ 3’

GATAAGC

CTATTCGAACTATAGCTTAACGGTACAACTTCGGTAGTAATGGTAA

Primer = **GATAAGC**

5’ 3’

GATAAGCTTGATATCGAATTGCCATGTTGAAGCCATCATTACCATT

ATGGTAA

3’ 5’

Primer = **AATGGTA**

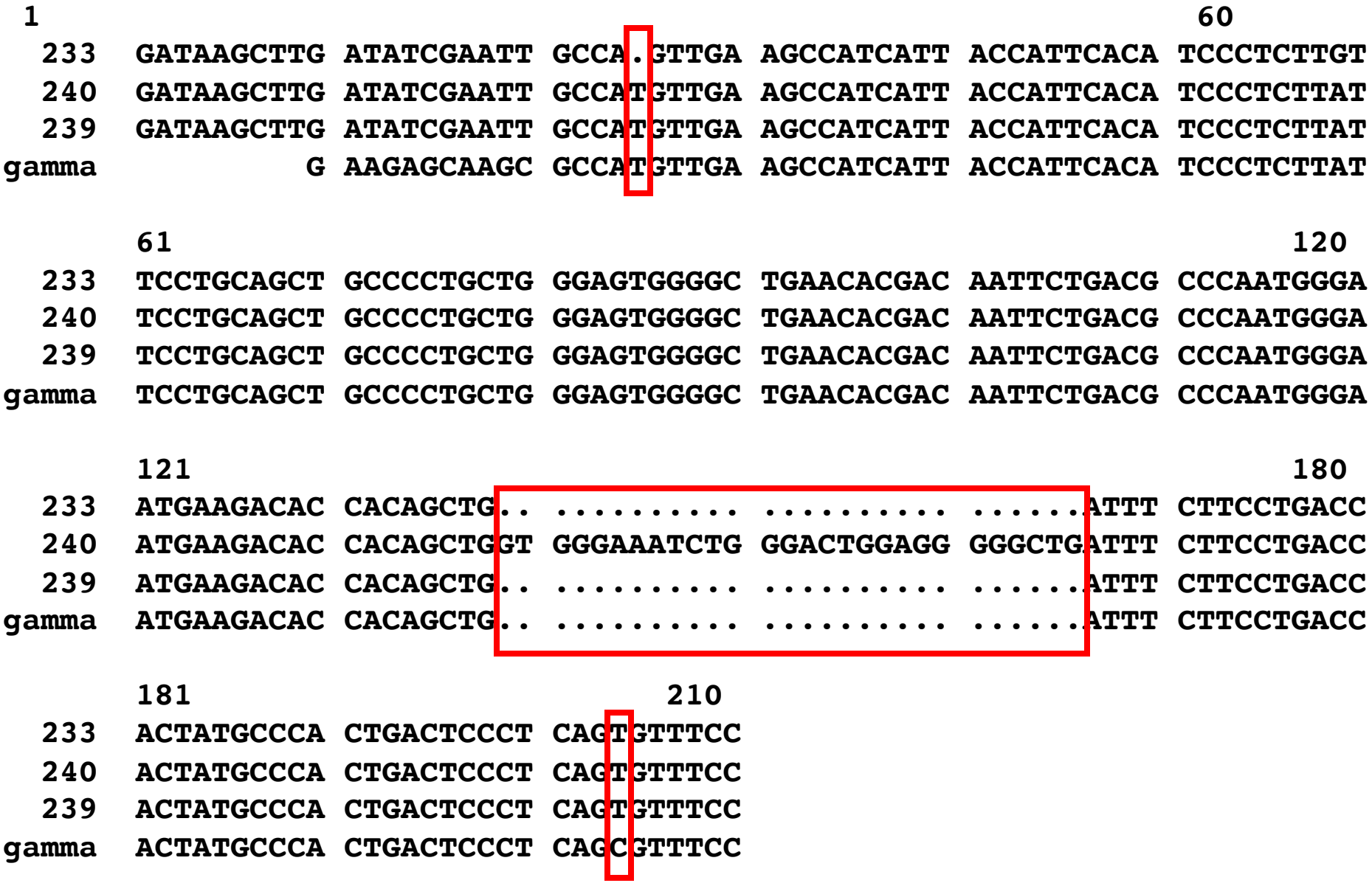
Primer Design

- Always make sure that you are in frame!
- Double check the orientation of the sequence before you submit it for synthesis!

Primer Design

Always sequence PCR products!!!!

(preferably after subcloning, unless you are just checking for presence of product)


1 60
233 GATAAGCTTG ATATCGAATT GCCA.GTTGA AGCCATCATT ACCATTCACA TCCCTCTTGT
240 GATAAGCTTG ATATCGAATT GCCATGTTGA AGCCATCATT ACCATTCACA TCCCTCTTAT
239 GATAAGCTTG ATATCGAATT GCCATGTTGA AGCCATCATT ACCATTCACA TCCCTCTTAT
gamma G AAGAGCAAGC GCCATGTTGA AGCCATCATT ACCATTCACA TCCCTCTTAT
61 120
233 TCCTGCAGCT GCCCCTGCTG GGAGTGGGGC TGAACACGAC AATTCTGACG CCCAATGGGA
240 TCCTGCAGCT GCCCCTGCTG GGAGTGGGGC TGAACACGAC AATTCTGACG CCCAATGGGA
239 TCCTGCAGCT GCCCCTGCTG GGAGTGGGGC TGAACACGAC AATTCTGACG CCCAATGGGA
gamma TCCTGCAGCT GCCCCTGCTG GGAGTGGGGC TGAACACGAC AATTCTGACG CCCAATGGGA
121 180
233 ATGAAGACAC CACAGCTG.. ....ATTT CTTCCTGACC
240 ATGAAGACAC CACAGCTGGT GGGAAATCTG GGACTGGAGG GGGCTGATTT CTTCCTGACC
239 ATGAAGACAC CACAGCTG.. ....ATTT CTTCCTGACC
gamma ATGAAGACAC CACAGCTG.. ....ATTT CTTCCTGACC
181 210
233 ACTATGCCCA CTGACTCCCT CAGTGTTTCC
240 ACTATGCCCA CTGACTCCCT CAGTGTTTCC
239 ACTATGCCCA CTGACTCCCT CAGTGTTTCC
gamma ACTATGCCCA CTGACTCCCT CAGCGTTTCC

Genomic Primers

There are a few additional factors to keep in mind when planning primers to genomic DNA:

- Genomic repeats
- Pseudogenes
- Gene families

Plasmid Design

Things to remember when designing plasmids

- What is your target cell line?
  - Eukaryotic / Prokaryotic
  - Promoter, Origin of replication….
- How are you going to replicate this plasmid?
  - Bacterial origin of replication
  - Copy number control
- What is your target cell “space”
  - Intracellular, extracellular, vesicular
  - Leader sequence

Protein Translation

Shifra Ben-Dor

Translation is relatively simple…

You need:

- A DNA sequence
- A table to convert 3mers to amino acids
- A program to put the two together

Position 2

Position 1

Position 3

TTT	F	Phe	TCT	S	Ser	TAT	Y	Tyr	TGT	C	Cys
TTC	F	Phe	TCC	S	Ser	TAC	Y	Tyr	TGC	C	Cys
TTA	L	Leu	TCA	S	Ser	TAA	*	Ter	TGA	*	Ter
TTG	L	Leu	TCG	S	Ser	TAG	*	Ter	TGG	W	Trp
CTT	L	Leu	CCT	P	Pro	CAT	H	His	CGT	R	Arg
CTC	L	Leu	CCC	P	Pro	CAC	H	His	CGC	R	Arg
CTA	L	Leu	CCA	P	Pro	CAA	Q	Gln	CGA	R	Arg
CTG	L	Leu	CCG	P	Pro	CAG	Q	Gln	CGG	R	Arg
ATT	I	Ile	ACT	T	Thr	AAT	N	Asn	AGT	S	Ser
ATC	I	Ile	ACC	T	Thr	AAC	N	Asn	AGC	S	Ser
ATA	I	Ile	ACA	T	Thr	AAA	K	Lys	AGA	R	Arg
ATG	M	Met	ACG	T	Thr	AAG	K	Lys	AGG	R	Arg
GTT	V	Val	GCT	A	Ala	GAT	D	Asp	GGT	G	Gly
GTC	V	Val	GCC	A	Ala	GAC	D	Asp	GGC	G	Gly
GTA	V	Val	GCA	A	Ala	GAA	E	Glu	GGA	G	Gly
GTG	V	Val	GCG	A	Ala	GAG	E	Glu	GGG	G	Gly

General rules:

- Changing the second position always changes the amino acid
- Changing the last position usually doesn't change the amino acid
- Changing the first amino acid is somewhere in between

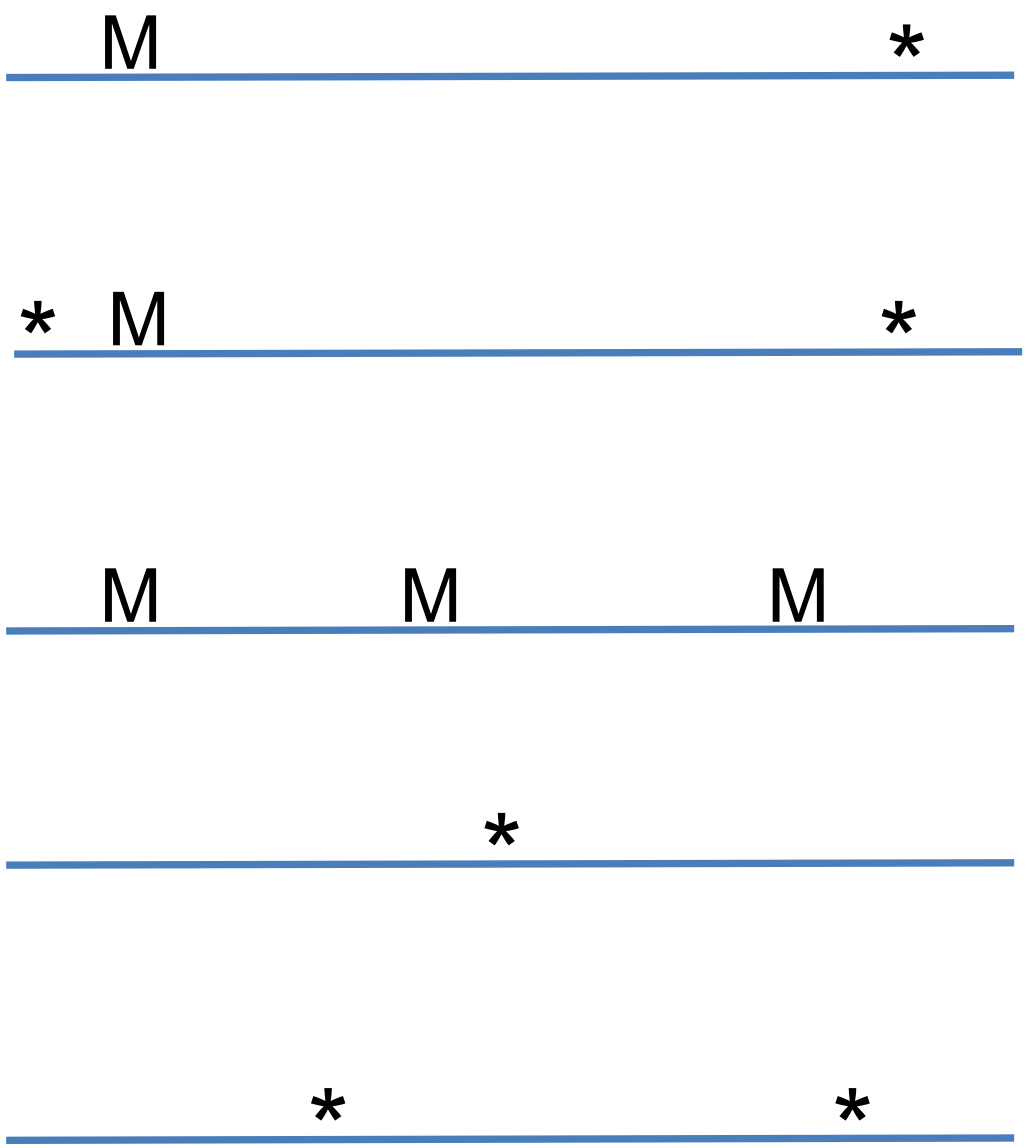
```
AAs   = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts= ---M------**--*----M---------------M----------------------------
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

But there are several genetic codes

https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c

And several definitions of ORF

- Start to Stop : ATG > *
- Stop to Stop : * > *
- Open ended (either requiring ATG or stop or neither)


M
*
* M
*
M
M
M
*
*
*

Defining an ORF

- For example: An ORF is defined as the coding sequence:
  - from the beginning of the transcript to the first stop
  - between two stops
  - from the last stop to the end of the transcript
- minimum length is also defined

Codon Usage

- For amino acids with more than one codon, there is preferential usage of the different codons
- This usage differs by species (and by organelle, for example mitochondria)
- There are databases and programs to calculate codon usage

Codon Optimization

- This is the process of optimizing the codons for maximal translation in a given species

- The same protein sequence can have very different DNA sequences in optimized sequences for different species