UNIVERSITY OF WARSAW

University of Warsaw
Biological and Chemical
Research Centre

GoRNA
STRUCTURAL BIOLOGY GROUP

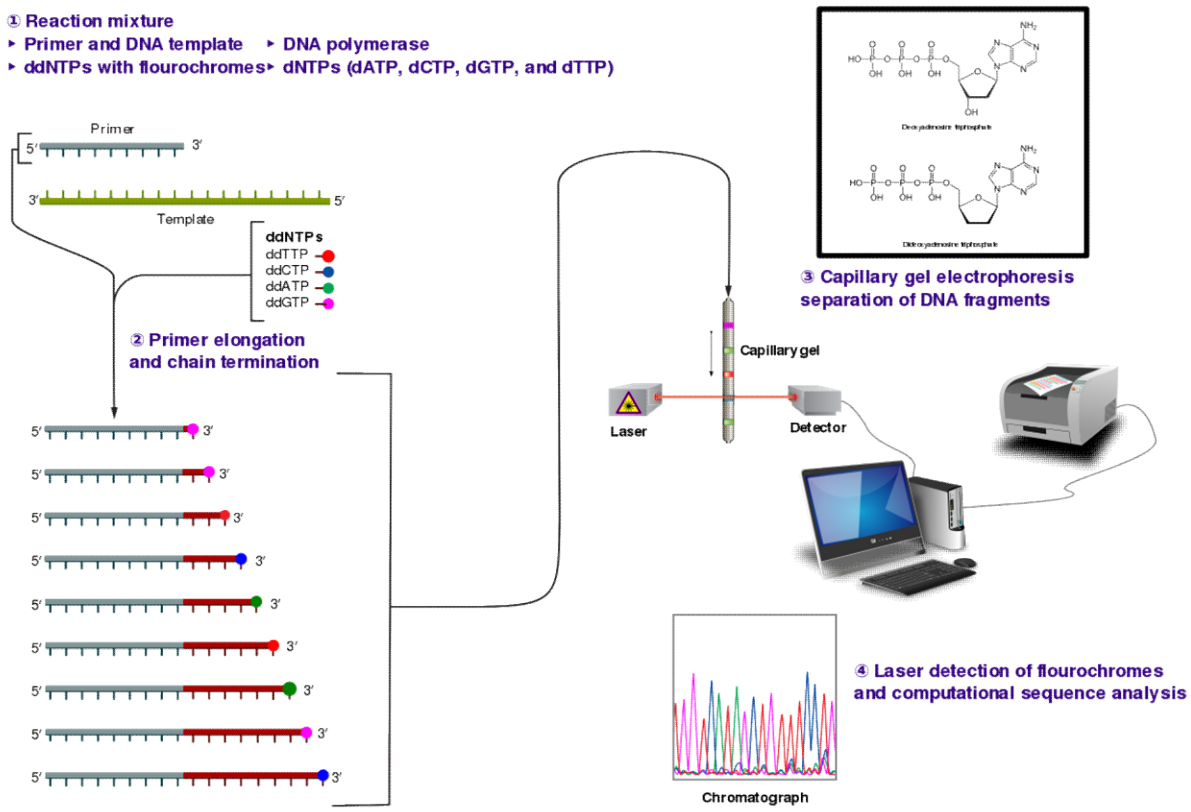
DNA Sequencing

Anna Trzemecka

SANGER Sequencing

- Developed by Frederick Sanger and coworkers in 1977, awarded Nobel Prize in 1980
- The method is based on the chain termination by use of Dideoxynucleotides (ddNTPs)
- Also known as the Enzymatic Method, Dideoxy sequencing or Chain termination method
- Most widely used DNA sequencing method for nearly 40 years, bringing the successful completion of the Human Genome Project (HGP) in 2003

DNA sequencing steps



Variables associated with DNA sequencing that can affect data quality

- **Inadequate template concentration.** (The most common reason) Templates that are too low in concentration will all generate sequences low in signal intensity. the analysis software has difficulty in resolving the base peaks from background noise hence poor base calling and poor data quality. The too highly concentrated template can have a highly detrimental effect on the electrophoresis of sequencing reaction products. They cause the capillary to become blocked, inhibiting the current and causing the reaction product to pass through the capillary slower than normal. Detection of these products begins later, resulting in poor sequencing results.

Variables associated with DNA sequencing that can affect data quality

Template additives: containing impurities that can inhibit the Taq polymerase activity or/and prevent the electrophoresis to be performed successfully. Common contaminants are:

- Salts (EDTA, NaCl, NaAc, Kac, KCl)

  For sequencing, DNA should NOT be dissolved in TE buffer because of EDTA's ability to bind Mg++ which is critical to Taq polymerase activity.

- Phenol quenches fluorescent dyes
- Also proteins, detergents (SDS, Triton X-100), RNA, chromosomal DNA, organic chemicals (ethanol, chloroform, phenol), divalent cations (Mg, Ca,Mn), excess PCR primers, dNTPs, enzyme, and buffer components from PCR

Variables associated with DNA sequencing that can affect data quality

- **Primer concentration.** Primer concentration should be at least 10pmol/ul
- **Primer quality:** Primers should have most of the following characteristics to be able to produce good, consistent sequencing data.

1. High purity
2. No mismatches
3. No potential alternative binding site in the template
4. No secondary structures present, especially at the 3' end
5. 1-8 bases minimum, longer for AT rich primers
6. Avoid runs of more than 4 of the same bases
7. Tm adapted to our annealing temperature (usually 50-55°C)

If primer Tm is much lower than the annealing temperature hybridization to its complementary template will be much less efficient and a lesser number of extending fragments will be generated. Increase primer Tm by adding additional bases to the 5’ or 3’ end to raise the Tm to be within the range of 52°C-58°C. Degenerate primers and those with mismatched bases will also show decreased hybridization efficiency due to a reduction of the stability of primer binding, and if degeneracy or mismatches occur at or near the 3’ end of the primer, it is highly likely that the sequencing attempt will fail.

Variables associated with DNA sequencing that can affect data quality

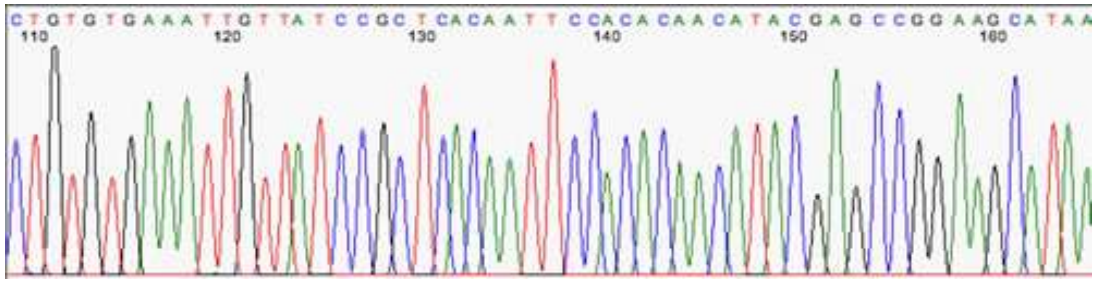
Difficult DNA content: If the DNA template failed to give good sequence data with the standard BigDye Terminator chemistry and the failure is not due to the poor template/primer quality, incorrect quantity or other improper sample preparation steps, then the DNA may be a **Difficult Template** containing one or more of the following:

- AT or GC rich stretches
- Secondary Structure
- Repeats
- Homopolymer regions
- Cosmids, P1 clones

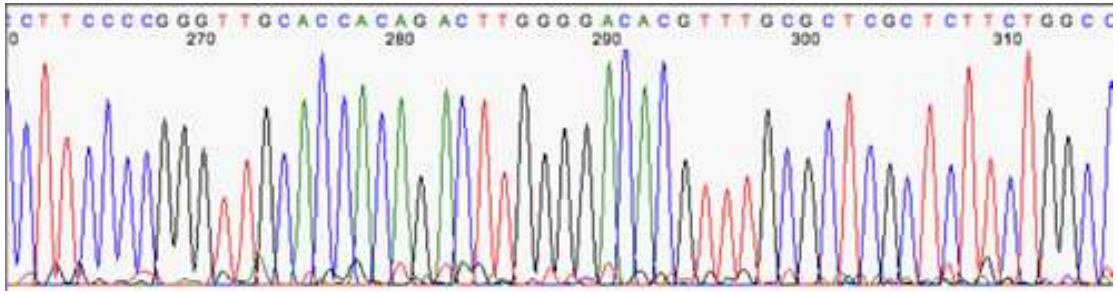
The DNA Sequencing laboratory does have alternative protocols and reagents that might help alleviate some of these issues.

Interpretation of Sequencing Chromatograms

- An example of an excellent sequence: the evenly-spaced peaks and the lack of baseline noise,

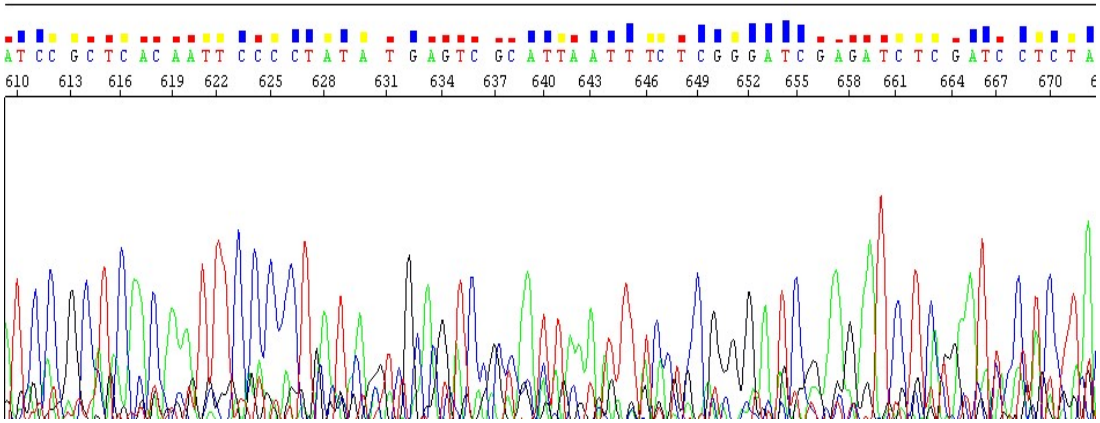


- A little baseline noise, but the 'real' peaks are still easy to call, so there's no problem with this sample:



Interpretation of Sequencing Chromatograms

- Example of mixed signals: peaks are not evenly spaced and overlap



Interpretation of Sequencing Chromatograms

- Example of mixed signals: peaks are not evenly spaced and overlap

Causes of Noisy Sequence

Dirty DNA

Low DNA concentration

Inadequate primers:

Inefficient primer annealing

Multiple priming sites

Noisy data caused by random priming of an N-1 primer synthesis

Possible solutions

Increase template concentration

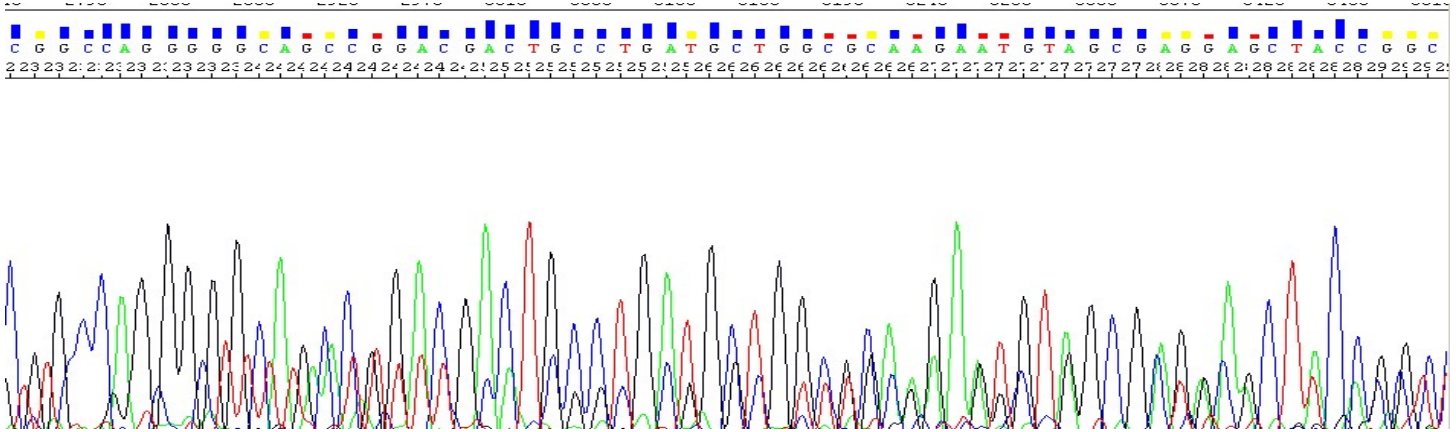
Re-do DNA prep

Request alternative sequencing programs

Choose another primer

Interpretation of Sequencing Chromatograms

- Example of mixed signals: peaks are not evenly spaced and overlap



Interpretation of Sequencing Chromatograms

- Example of mixed signals: peaks are not evenly spaced and overlap

Causes of Mixed signals or mixed sequence

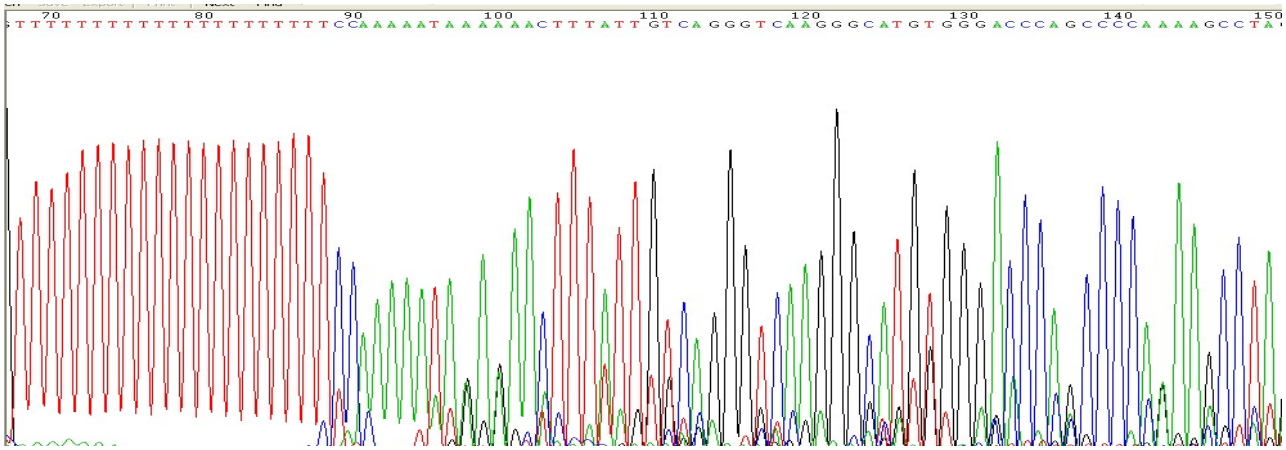
Two or more templates were present in the reaction. More than one clone
or more than one PCR product. This is the most common cause of mixed-signal traces.
Two primers were present in the sequencing reactions
The PCR fragment was not purified of leftover primes before sequencing.
Multiple primer annealing sites are present in the DNA template.
A too-low primer annealing temperature was used in the sequencing reaction.
the primer binding site is within a repeat region on your template
A degraded primer was used.

Possible solutions

Pick a new colony and start over: Prepare a new plasmid prep making sure that only one colony is selected.
Remember that even a relatively low amount of a small PCR product can cause mixed template problems.
Choose another primer: Check the template for possible multiple priming sites. If two sites are present use a
different primer. This can often occur when a fragment containing the priming is sub-cloned into a vector that also
contains the priming site
Request alternative sequencing programs
Run product out on a gel and gel purify or sequence with a nested primer
Ensure that PCR products have been cleaned before receipt so they are free from PCR primers which could
Also, initiate
extension in a sequencing reaction. Even low levels of the PCR primers can cause mixed-signal problems,
especially if they have a high annealing temperature

Interpretation of Sequencing Chromatograms

- Mixed Signals after a homopolymer region



Interpretation of Sequencing Chromatograms

- Mixed Signals after a homopolymer region

Example of mixed signals after a poly T region

A long run of a mono nucleotide base causes the DNA polymerase to "slip" on the template. This occurs by either the template or extension product looping out and rehybridizing and results in the generation of sequencing products of varying size.

These sequencing products then appear as mixed-signal in the trace downstream of the mono nucleotide run.

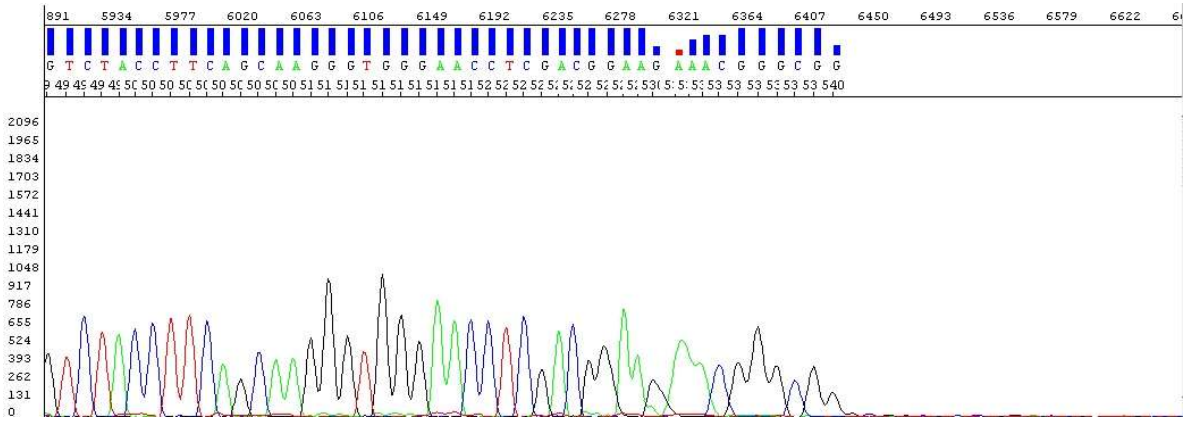
Possible solutions

Sequence the DNA template from both directions.

Use a custom primer designed to hybridize just outside the mononucleotide or dinucleotide run region.

Interpretation of Sequencing Chromatograms

- The signs of secondary structure



Good quality trace signal suddenly stops, or rapidly declines, in regions of the trace that should be well resolved.

The template has a high percentage of G and/or C nucleotides, especially in the region where the stop occurred.

Interpretation of Sequencing Chromatograms

- The signs of secondary structure

Common Causes

Hairpins present in the template block the polymerase's progress. Difficult DNA regions can fold back on themselves and form hairpin structures that the sequencing polymerase can not pass through.

Regions of G and C homonucleotides are problematic to sequence through (GC rich regions). These can not only form hairpins with strong secondary structure, but can also form single-stranded conformational structures that the sequencing the polymerase has difficulty passing through.
Long template regions of guanidine (G runs) are particularly problematic.

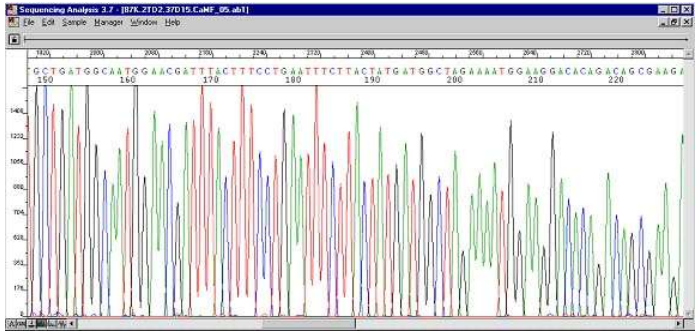
Possible solutions

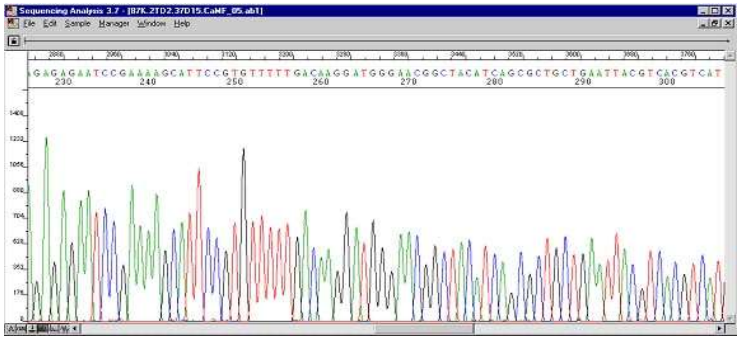
Change the sequencing program allowing for a higher annealing and denaturation step (the Sequencing Lab has these alternate programs).
Custom design a primer to hybridize after the difficult DNA region.
Add a small amount of dGTP nucleotide to the BigDye mix. This can sometimes help the sequencing polymerase pass through long G runs without causing major compression problems. (the Sequencing Lab has stocks of this reagent)

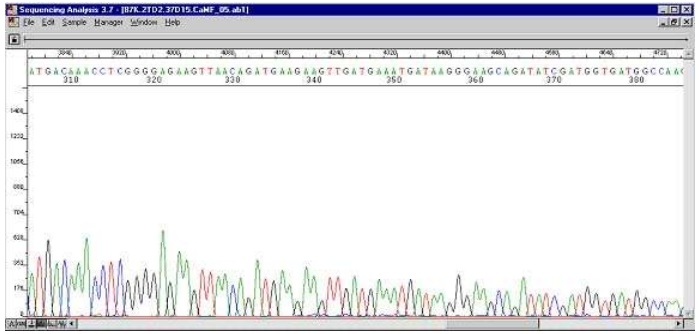
Add either Q Solution or Betaine to the sequencing reaction, both are used as enhancing agents for PCR. (the sequencing Lab have stocks of Q Solution in the house)

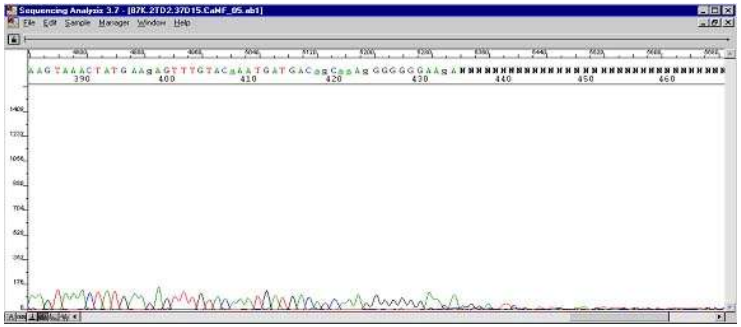
Interpretation of Sequencing Chromatograms

- **Top heavy sequence** The signal intensity is very strong in the beginning and falls off early.
- The signal strength gradually declines as the sequence continues leading to poor read length.









Interpretation of Sequencing Chromatograms

- **Top heavy sequence**

Possible causes of Top Heavy Sequence

Presence of salt within the reaction – this inhibits the enzyme and results in
a shorter read length (**mostly responsible for gradual loss**).
Salt also competes with DNA during injection into the capillaries.

DNA/primer concentration too high

DNA/primer concentration too low

Difficult DNA content

Possible solutions

Dilute to proper concentration and re-sequence

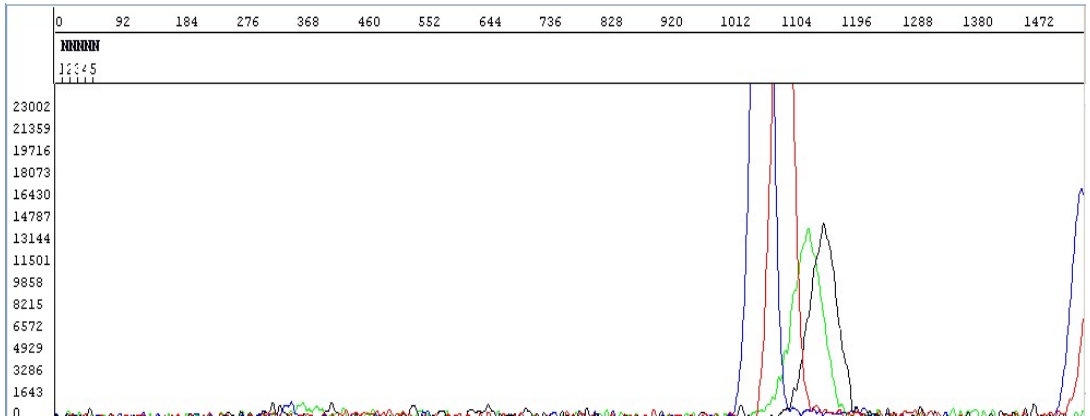
Increase concentration and re-sequence

More cycles for sequencing

Re prep DNA ensuring that is free of any inhibitors.

Interpretation of Sequencing Chromatograms

- **Failed DNA sequencing reaction**



The sequencing reaction has failed to make any extension products that are detectable during the electrophoresis.
Only “dye blobs” (unincorporated dye) and background noise are present

Interpretation of Sequencing Chromatograms

! Failed DNA sequencing reaction

Causes of failed DNA sequencing reactions

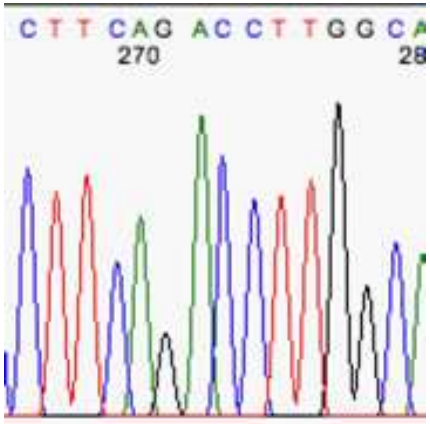
Poor quality DNA.
Low template concentration. A most common cause of reaction failure
Too high template concentration.
The wrong primer was used.
Degraded primer (check age of primer synthesis)
No Priming site (re-check plasmid maps if applicable)
Template is of extremely low quality
DNA could be nicked or degraded by excessive freeze-thawing, nucleases, excessive UV light exposure, or bisulfite treatment.
Presence of contaminants (see below)
Too low primer concentration (must be at least 10pmol/µl)
Inefficient primer annealing

Possible solutions

Check template & primer concentrations
Re-prep DNA or Ethanol precipitation to remove the contaminant(s)
Double-check your plasmid maps/sequences.
Choose another primer—be sure to use a primer design program
If you've designed your own custom primer from previous sequence data, make sure you were using a reliable area of sequence - look for sharp, well-defined peaks with no ambiguity. Avoid areas where the peaks are broader and not well separated.
Request alternative sequencing programs (e.g. new annealing temperature)

Interpretation of Sequencing Chromatograms

- **Mis-spaced peaks**

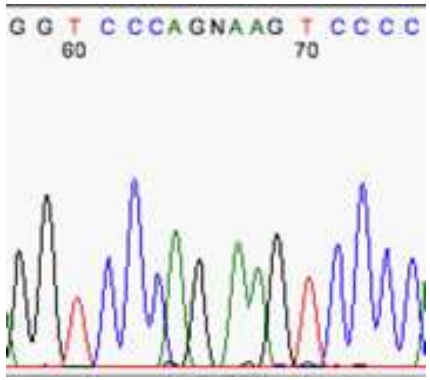


Some sequencers have predictable errors in base spacing. A common one is a G-A dinucleotide, which leaves a little extra space between them. Often, it is ignored by the base caller.

The extra space between the letters G and A (nt's 271 and 272) corresponding to the mis-spaced peaks just below them. No harm done, in this case; the sequence is fine.

Interpretation of Sequencing Chromatograms

- **Mis-spaced peaks**

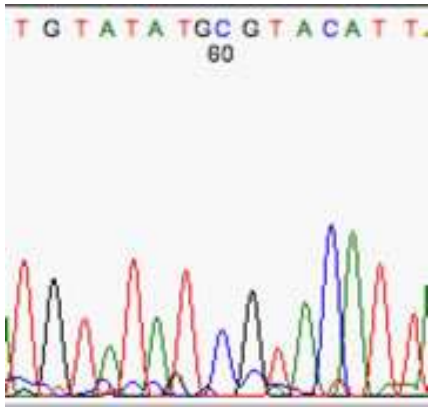


Sometimes, spaces are misinterpreted as missing nucleotides.
In the example, note the 'N' called in the space between the G-A pair.
That is an erroneous call; there is no missing base 'N' at that position.

You can spot this by scanning the text sequence at the top of the window, looking for oddly-spaced letters. Of course, you may also spot this simply by looking for 'N' nucleotides.

Interpretation of Sequencing Chromatograms

- **Mis-spaced peaks**



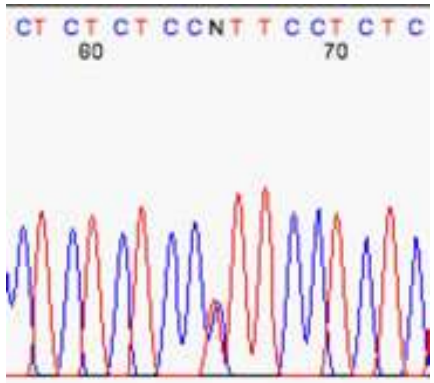
The real problem comes when the basecaller attempts to interpret a gap as a real nucleotide, such as in the example at right. The typical scenario is a sequence with noticeable baseline noise, and a gap is called as if the baseline noise were a real peak. Often it's those aforementioned G-A gaps, but not necessarily, as the example here shows.

The real T peak (nt 58) and the real C peak (nt 60), with the G barely visible between them. Despite its size, the baseline-noise G peak was picked as if it were real. The clues to spot are (i) the oddly-spaced letters, with the G squeezed in, and (ii) the gap in the 'real' peaks, containing a low noise peak.

This is a great example of why a weak sample, with its consequent noisy chromatogram, is untrustworthy.

Interpretation of Sequencing Chromatograms

- **Heterozygous (double) peaks**

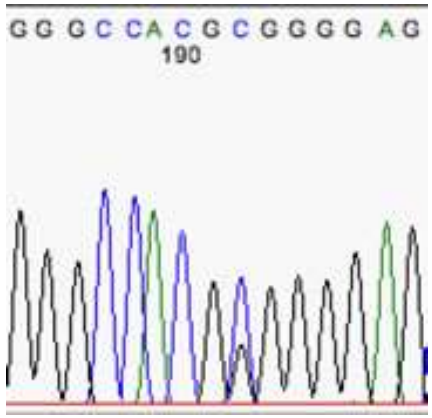


Here is a great example of a PCR amplicon from genomic DNA, with a clear heterozygous single-nucleotide polymorphism (SNP). In this case, one allele carries a C, while the other has a T. Both peaks are present, but at roughly half the height they would show if they were homozygous.

The peak was called an 'N' by the basecaller.
A comparison of text sequences would probably notify you of the presence of a SNP at this particular location

Interpretation of Sequencing Chromatograms

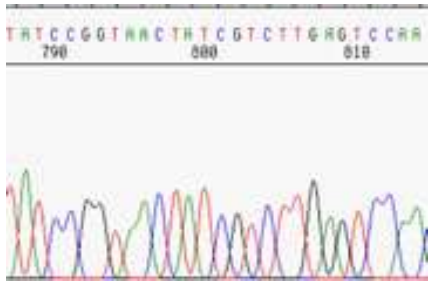
- **Heterozygous (double) peaks**



At right is a heterozygote that was missed by the basecaller. The text sequence simply shows a 'C'. If all of your other sequences also had a 'C' here, you would never realize that you had a het SNP ... unless you scanned your chromatograms.
It can be difficult to go through reams of sequencing chromatograms, looking for heterozygote peaks like this. It's fine for small projects --- just look for the nested multicolor peak. For big SNP-detection projects, though, you should be using a computer program that can detect these (i.e. such as Sequencher by GeneCodes, Inc.).

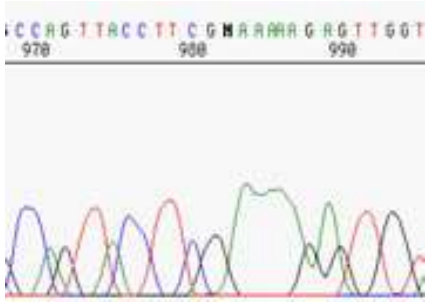
Interpretation of Sequencing Chromatograms

Loss of Resolution Later in the Gel



If you scroll the above chromatogram further to the right (to higher-numbered nucleotides), you see the frame depicted. It is evident that, here at 800 nucleotides, the sequence is still quite reliable. The peaks are broader and clearly less well-resolved, but there still is an evident separation between them, and basecalls can still be made accurately.

The spacing between the basecall letters at top is regular, which is often a good indication of the reliability of the data. When that spacing becomes irregular, be careful!



The very limit of resolution is around 900-1000 nt. You get only a general sense of the sequence here. There are only a few basecalls that can be considered reliable. The G at 981 may in fact be two G's, the N could be a G or an A, and who knows how many A's there are afterward.

Late in the chromatogram, watch for multiple bases of any one nucleotide where there really should be only one. Watch, too, for wide peaks miscounted by the program as two nucleotides, when it should have been just one nucleotide. Wide peaks may also obscure smaller adjacent peaks

Thank you