Lecture 8

DNA Sequencing

In the old days, there were two competing methods of determining DNA sequence:

- Maxam - Gilbert Method, in which a DNA sequence is end-labeled with [P-32] phosphate and chemically cleaved to leave a signature pattern of bands.
- Sanger Method, in which a DNA sequence is annealed to an oligonucleotide primer, which is then extended by DNA polymerase using a mixture of dNTP and ddNTP (chain terminating) substrates.

Since the Maxam - Gilbert method is not frequently used, it will not be described, however one of its advantage is that it permits direct sequencing of small fragments.

Sanger Method (Dideoxynucleotide chain termination) Here's an example of how one goes about sequencing by this method.

First, anneal the primer to the DNA template (usually single stranded):

```
      5'-GAATGTCCTTTCTCTAAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
```

Then split the sample into four aliquots, in tubes labeled "G", "A", "T" and "C" and add the following substrates to the respective tubes:

"G" tube: All four dNTPs, including one that is labeled, **plus ddGTP**
"A" tube: All four dNTPs, including one that is labeled, **plus ddATP**
"T" tube: All four dNTPs, including one that is labeled, **plus ddTTP**
"C" tube: All four dNTPs, including one that is labeled, **plus ddCTP**

When a polymerase (e.g. Klenow fragment) is added to the tubes, the synthetic reaction proceeds until, by chance, a dideoxynucleotide is incorporated instead of a deoxynucleotide. This is a "chain termination" event, because there is a 3' H instead of a 3' OH group. Since the synthesized DNA contains some radiolabeled (or chemically labeled) substrates, the products can be detected and distinguished from the template.

Reminder: The difference between dATP and ddATP
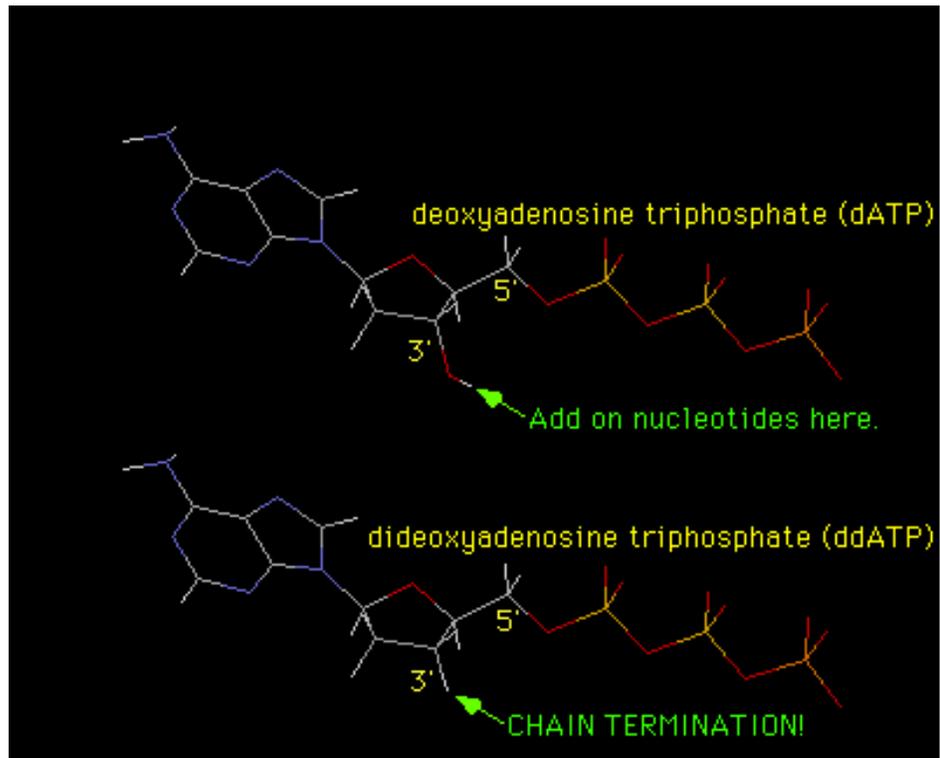
If, for example, we were to look only at the "G" reaction, there would be a mixture of the following products of synthesis:

```
          5'-GAATGTCCTTTCTCTAAGTCCTAAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'


          5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'


          5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'


          5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'


          5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'


          5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTACTTCTAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
```

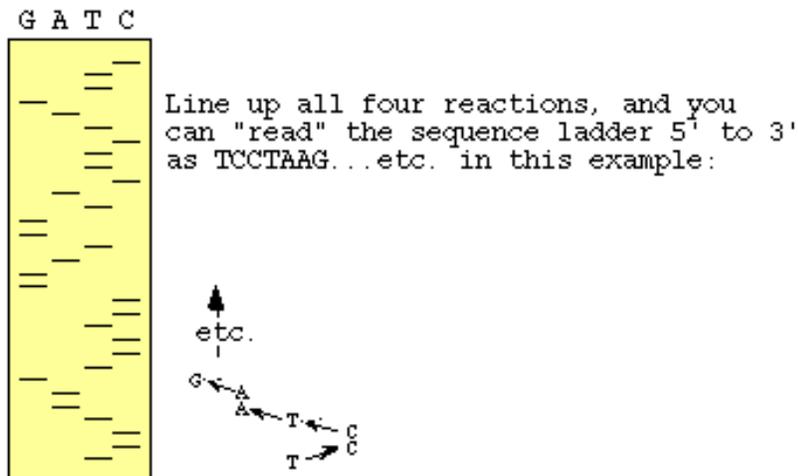(and so on, if the DNA being sequenced continues to the right)

The sequencing gel These products are denatured into single stranded DNA molecules and run on a polyacrylamide/urea gel. The gel is dried onto chromatography paper (to reduce its thickness and keep it from cracking) and exposed to X-ray film. Since the template strand is not radioactively labeled, it does not generate a band on the X-ray film. Only the labeled top strands generate bands, which would look like this:

Polyacrylamide gel electrophoresis of the "G" reaction

what's in the gel bands?

```
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTACTTCTAG

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGG
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATG

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGG
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG

5'-GAATGTCCTTTCTCTAAGTCCTAAG
```

◄——primer——► ◄—new synthesis—► ddG stops synthesis
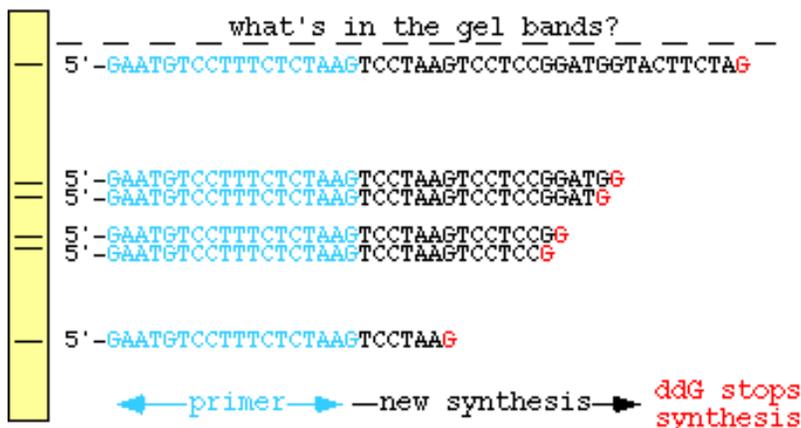
As you can see from this one reaction (the "G" reaction) the chain termination events produce individual bands on a gel. The chain terminations closest to the primer generate the smallest DNA molecules (which migrate further down the gel), and chain terminations further from the primer generate larger DNA molecules (which are slower on the gel and therefore remain nearer to the top).

When similar chain termination reactions are run for each nucleotide, the four reactions can be run next to each other, and the sequence of the DNA can be read off of the "ladder" of bands, 5' to 3' sequence being read from bottom to top:

G A T C

Line up all four reactions, and you can "read" the sequence ladder 5' to 3' as TCCTAAG...etc. in this example:

etc.

That's the way "manual" sequencing is done. We'll have a bit more to say about that before discussing "automated" sequencing methods.

The resolution of the gel electrophoresis is very important in DNA sequencing. Molecules that are 50, 100, or 200 bases in length must be separable from molecules that are 51, 101, or 201 bases in length (respectively). There are several modifications to improve the resolution:

- The gels must be much large (notice the apparatus on the top of the shelf in our classroom?) so that the molecules migrate further and are better resolved.

- The gels must contain a high concentration of urea (7 to 8 molar) to prevent folding of the molecules and formation of secondary structures by hydrogen bonding that would alter the mobility of the molecule. Similarly, the samples are denatured before they are loaded.

- The gels are run at higher temperature (about 50 C), also to prevent H bond formation.

One thing I should confess is that this example has one bit of fiction - you can't really obtain usable sequence information that close to the end of the primer, unless your ddNTP/dNTP ratios are quite high. If you increased that ratio, you would make it more difficult to read sequence 200 to 300 nucleotides further down, because most of the synthetic products would have terminated before that point.

Radioactive labels **It is possible to obtain labeled products in one of two ways:**

Internal labeling of products:
In this case, one of the dNTP substrates is radiolabeled (or chemically labeled) so that the synthetic products are marked internally, and possibly in many places at once. The multiple sites of incorporation mean that the product will have a higher specific activity.

One disadvantage of this is that nonspecific side reactions (perhaps having nothing to do with the oligonucleotide) will also be labeled.

End-labeling of products:
This is used in the Maxim and Gilbert method, of course, but it can also be used in Sanger sequencing methods. An oligonucleotide can be labeled with a P-32 phosphate at its 5' end, for example, by the enzyme polynucleotide kinase and the substrate *gamma-P-32 ATP*. If this radiolabeled oligonucleotide is annealed to a template and extended with a polymerase, the products will be labeled only at their 5' end.

5'-PO$_4$-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTACTTCTAG
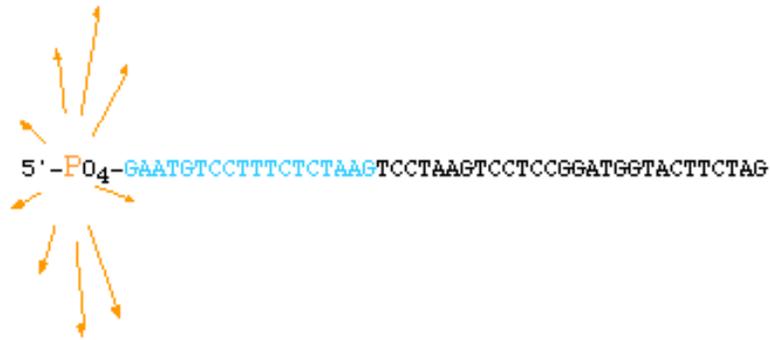
An advantage of this is that only reactions involving the annealing of the oligonucleotide will be labeled. A disadvantage is that at most one labeled atom is incorporated into the synthetic product, so the specific activity is low.

The technology of DNA labeling has changed in the last fifteen years, so that there are many more options.

Radioactivity-based approaches:

P-32 phosphate labeling has many advantages and several disadvantages. The half-life of P-32 is approximately 14 days, and so it is possible to label DNA to a very high specific activity (meaning that a small number of moles of product generate a large number of radioactive disintegrations per minute). One disadvantage of P-32 is that it produces a penetrating burst of beta particles (electrons) that generate a "wider" signal on a piece of film. In the schematic diagram below, you can see how this effect works:
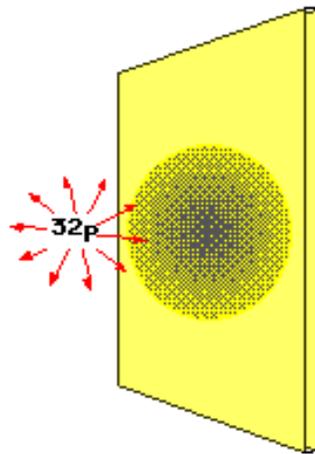
Because of this "spreading out" of the signal, you get bands that are a bit "foggy" looking when you develop the film:
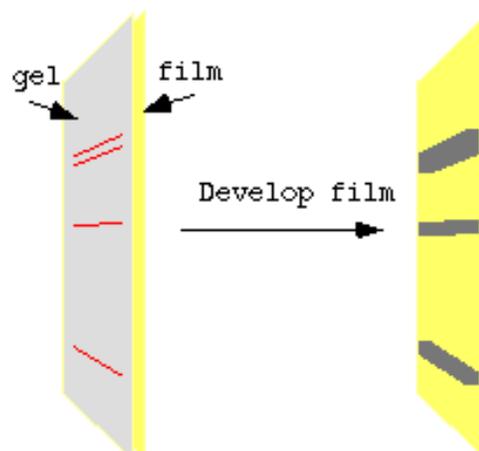
Another disadvantage of P-32 is that radiolabeled substrates not have a long "shelf-life" because the half life of the P-32 nucleus is short.
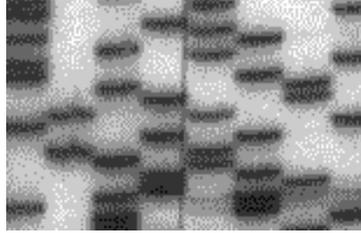
P-33 phosphate is now frequently used because it solves several of the problems associated with P-32. The radioactive decay of P-33 is of lower energy, so a sharper image can be generated on a piece of film. The half-life of P-33 is about twice as long, so dNTP substrates labeled with it have a longer life as well. An unfortunate consequence of a longer half-life however, is that the synthetic products will have a slightly lower specific activity.

S-35 thiophosphate is a second solution to the problems associated with P-32. The radioactive sulfur takes the place of one of the oxygens in the phosphate, which then becomes a "thiophosphate." It has a longer half-life (87 days), meaning that the substrate shelf-life is longer, and a less energetic decay, meaning that the film results are sharper.

A word or two about radioactivity

We measure radioactivity using the "Curie" unit (Ci), named after the Nobel Laureates Marie and Pierre Curie. One Ci of a radioactive substance is equal to approximately 2.2 x 10^12 disintegrations per minute. A micromole of P-32 is approximately 6 Ci, while a micromole of S-35 is approximately 1 Ci. Why the difference? Because while both the P-32 and S-35 are both bent upon self destruction, the P-32 is disintegrating at a 6 times more rapid rate of decay, so more disintegrations per minute will occur per mole of P-32 than per mole of S-35.

Radioactive decay is not a chemical reaction - it is a nuclear reaction. That means you cannot slow it down by putting it into the refrigerator.

Oh no! Here are some quantitative calculations to ponder:

- The amount of P-32 that a laboratory worker would generally use to end label an oligonucleotide is approximately 100 microcuries of gamma-phosphate labeled ATP (an amount that would cost approximately $25). How many moles of ATP is that if the specific activity is 6000 Ci/mmole?

Answer: $(1 \times 10^{-4} \text{ Ci})/(6000 \text{ Ci/mmole}) = 1.6 \times 10^{-8}$ mmole = 16 pmole.

- If this reaction is conducted in a reaction volume of 50 microliters, what is the concentration of ATP that is available for use by the enzyme polynucleotide kinase?

Answer : $(1.6 \times 10^{-11} \text{ moles})/(5 \times 10^{-5} \text{ liters}) = 3.1 \times 10^{-7}$ molar = 0.31 microM

- If the Km of the polynucleotide kinase is approximately 50 micromolar (for the ATP substrate), is this going to be a happy enzyme?

Answer: No, this enzyme is going to get very bored while waiting for substrate!

Let's think ahead to what we might need if we were sequencing using an end-labeled primer: Approximately 0.5 picomole of oligonucleotide primer must be used in a sequencing reaction, to have a reasonable efficiency of annealing to a template.

- If we label 1 pmole of oligonucleotide and the polynucleotide kinase uses approximately half of the P-32 ATP in a 1 hour reaction, what will be the specific activity of the product?

Answer: (50 microCi incorporated/1 pmole oligo) = $(5 \times 10^{-7}$ Ci)/$(1 \times 10^{-9}$ mmole) = 500 Ci/mmole (since the theoretical maximum specific activity is 6000 Ci/mmole, this means that approximately 1 out of every 12 oligonucleotides will have a radioactive phosphate on it. The other 11 will be unlabeled.

That doesn't sound too good!

- Would it help to use more primer? What if we labeled 10 picomole of oligonucleotide instead of 1 picomole?

Answer: That wouldn't help at all! With the same amount of P-32 phosphate applied, you would have 1 out of every 120 oligonucleotides labeled instead of the 1 out of 12 you had before. As a consequence, the synthetic products will be less frequently labeled, and a given number of moles of product will have less radioactivity associated with it (i.e. a lower specific activity).

- Would it help to use more P-32 in the reaction?

Answer: Yes, any additional P-32 that you add will boost the specific activity of your product, but the expense of the experiment will increase and your laboratory safety officer may start to have heart palpatations when he or she sees how much P-32 you're using!

- Let's stick with the original reaction (1 out of 12 labeled) and ask how many disintegrations per minute would be present in each gel band, after annealing an oligonucleotide to 0.1 pmole of plasmid, and extending it to produce 500 different products? To simplify the calculation, we'll assume the 500 products of different length are present in equal molar amounts (though in practice this is not so).

Answer: If all 0.1 pmole of oligonucleotide are used up in this way, then each band is 0.0002 pmole, or 0.2 fmole. Since the specific activity was 500 Ci/mmole and 1 Ci is defined as 2.2×10^{12} dpm, each band will generate
$(2 \times 10^{-13}$ mmole)(500 Ci/mmole)$(2.2 \times 10^{12}$ dpm/Ci)= 220 dpm.
This amount of radioactivity can be detected on X-ray film, after an overnight exposure.

Well! That was fun.

Here's a question for you! Suppose that you annealed an oligonucleotide to a single stranded template (as shown below) and added the enzymes Klenow fragment, and only the following substrates, 32P-dATP, dCTP, dTTP, and ddGTP (note: I said ddGTP, not dGTP).

```
        GATACCATACCGAT
3'-AATCTCTATGGTATGGCTAGGTACTATTAGTCAA-5'
```

How many radioactive phosphates would be incorporated through this reaction?

one
two
five
none

Interesting! Do you think the answer would be different if this were the situation?

```
        GATACCATACCGAT
3'-AATCTCTATGGTATGGCTAGGCACTATTAGTCAA-5'
```
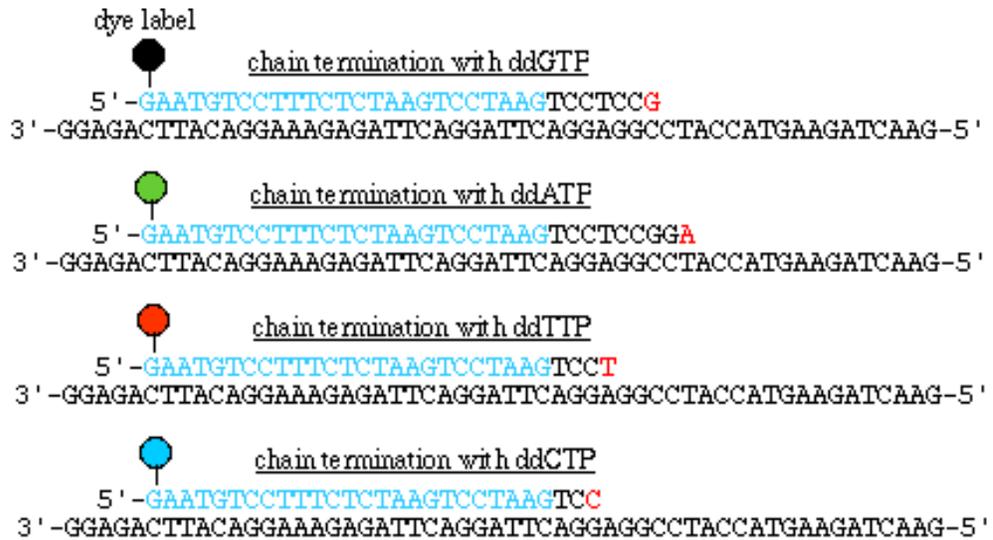
Can you see how this might be useful as a diagnostic test for two alleles?

Approaches based on machine reading How can you avoid having to "read the bases" off of a gel or film yourself? Train a machine to scan the gel with a laser as it is running, and use different fluorescent dyes to indicate the bases. The gel is run using special (and expensive) glass plates that are optically pure, and the laser scans back and forth across the lanes while the DNA fragments migrate past the laser "window."


[Applied Biosystems](#) makes the machine that we use. They have made a program called EditView, which [you can download from their site](#), and which shows the types of chromatograms one obtains from the laser readout.

Dye-labeled primers. In this method, the oligonucleotide primers are end-labeled with four different fluorescent dyes, and four separate synthetic reactions are conducted in the presence of dNTPs and ddNTPs:

dye label

chain termination with ddGTP

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

chain termination with ddATP

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGA
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

chain termination with ddTTP

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCT
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

chain termination with ddCTP

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCC
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

Upon completion, these four reactions can be combined into one lane on a gel, and run on a machine that can scan the lanes with a laser. The wavelength of fluorescence can be interpreted by the machine as an indication of which reaction (ddG, ddA, ddT, or ddC) the product came from.


G A T C
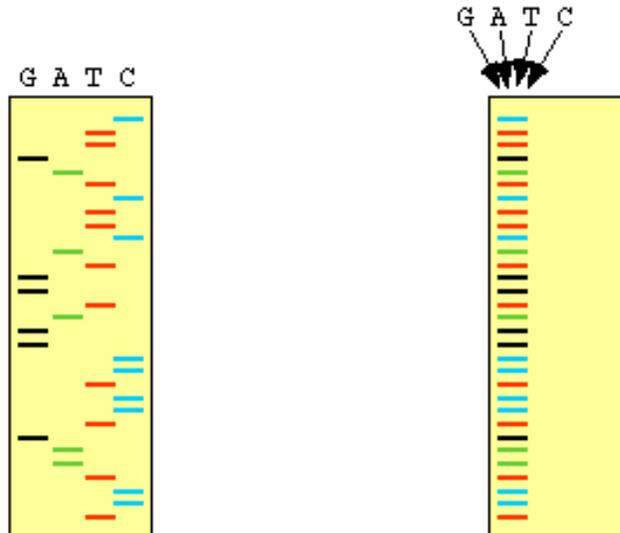
Here's what the products
would look like in
separate gel lanes.

Here's what the products
would look like in
a single gel lanes.

The fluorescence output is stored in the form of chromatograms:

TCAAATTTTGAGTTGCATCCCCTGG
290   300   310
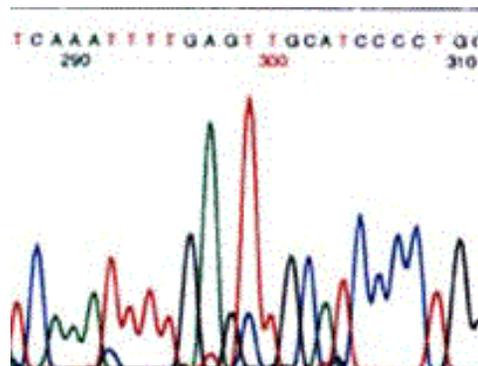
Dye-termination sequencing. This is a much more versatile method of sequencing, because it is not necessary to have a chemically modified oligonucleotide. The fluorescent dyes are conjugated to dideoxynucleotides, so a chain termination event is marked with a unique chemical group. Only one reaction needs to be run in this case, because there is no longer a separation between the label and the terminating group.
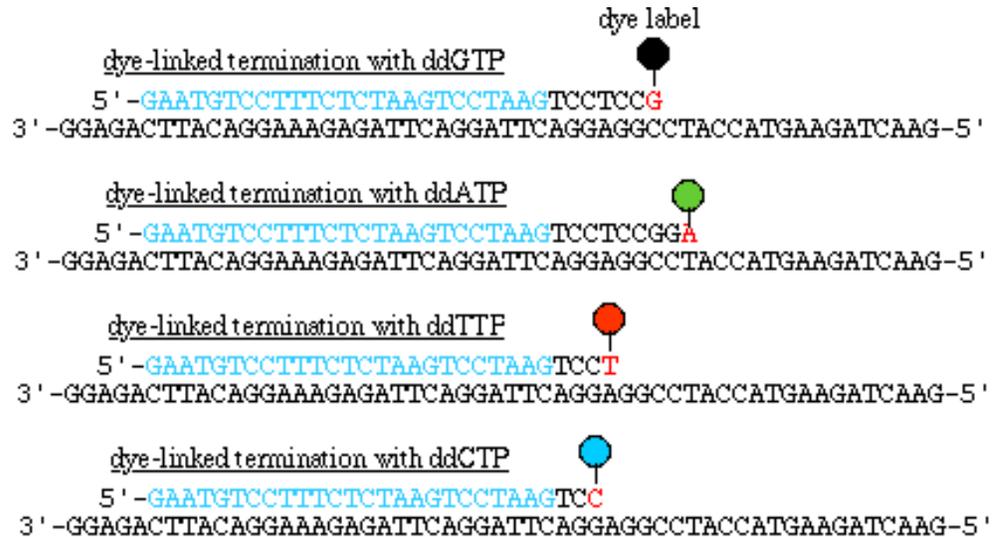
Stan Metzenberg
Department of Biology
California State University Northridge
Northridge CA 91330-8303
stan.metzenberg@csun.edu