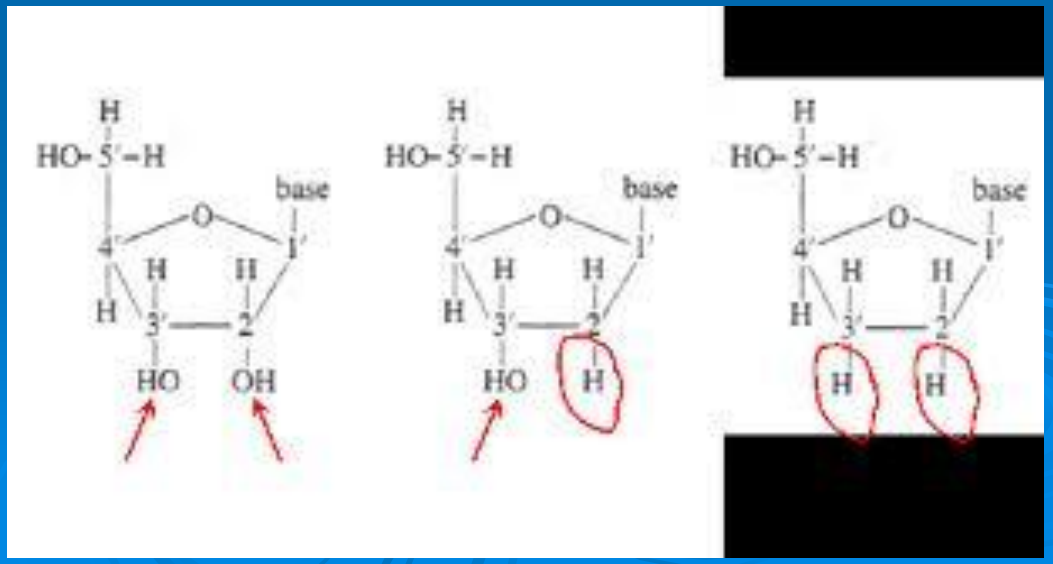
DNA Sequencing

DNA sequencing is the determination of the precise sequence of nucleotides in a sample of DNA.

- The most popular method for doing this is called the **dideoxy method** or Sanger method (named after its inventor, Frederick Sanger, who was awarded the 1980 Nobel prize in chemistry for this achievement).

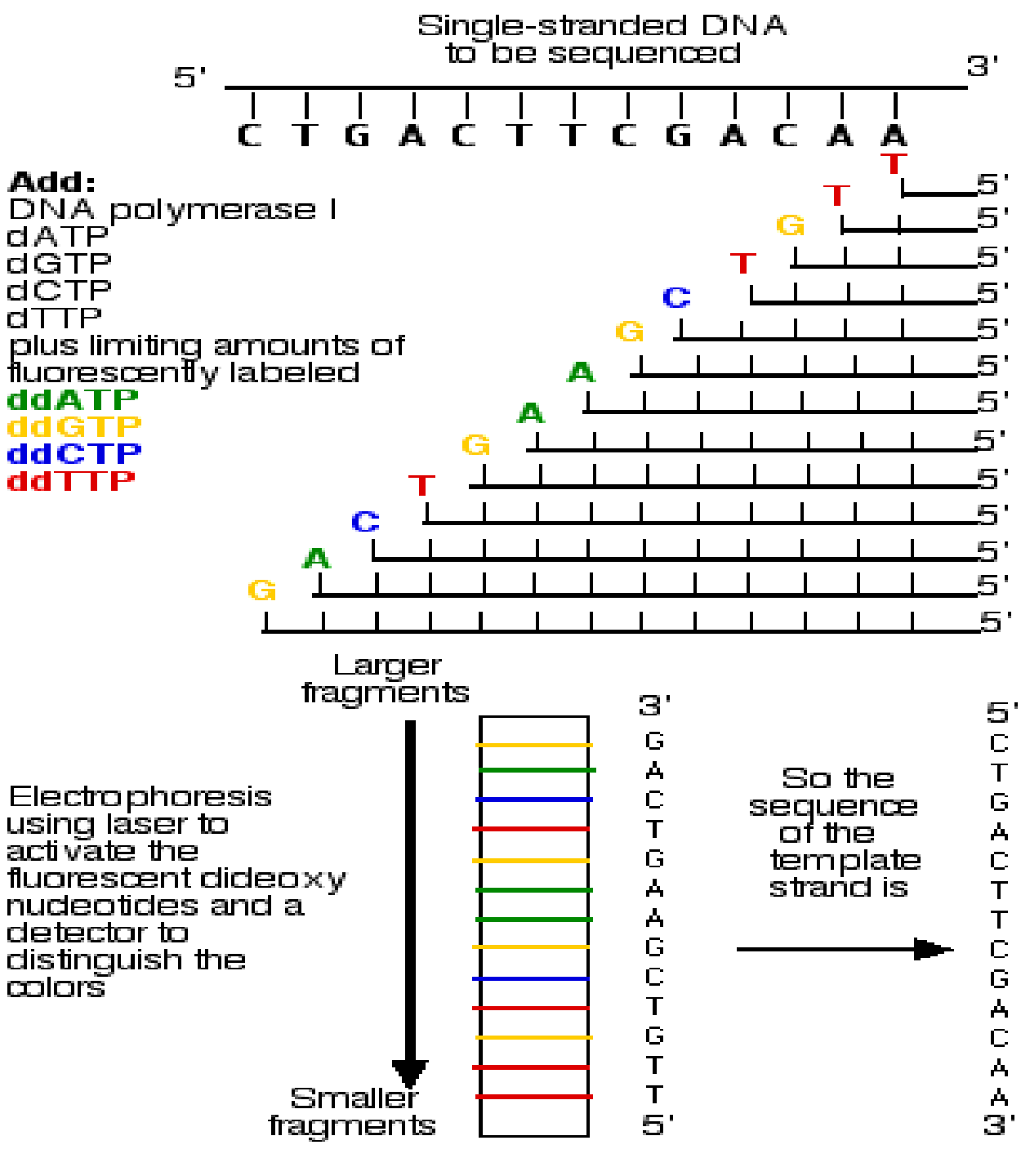
The Procedure

- The DNA to be sequenced is prepared as a single strand.
- This template DNA is supplied with a mixture of all four nucleotides in ample quantities
  - dATP
  - dGTP
  - dCTP
  - dTTP
- a mixture of all four **dideoxy**nucleotides, each present in limiting quantities and each labeled with a "tag" that fluoresces a different color:
  - **ddATP**
  - **ddGTP**
  - **ddCTP**
  - **ddTTP**



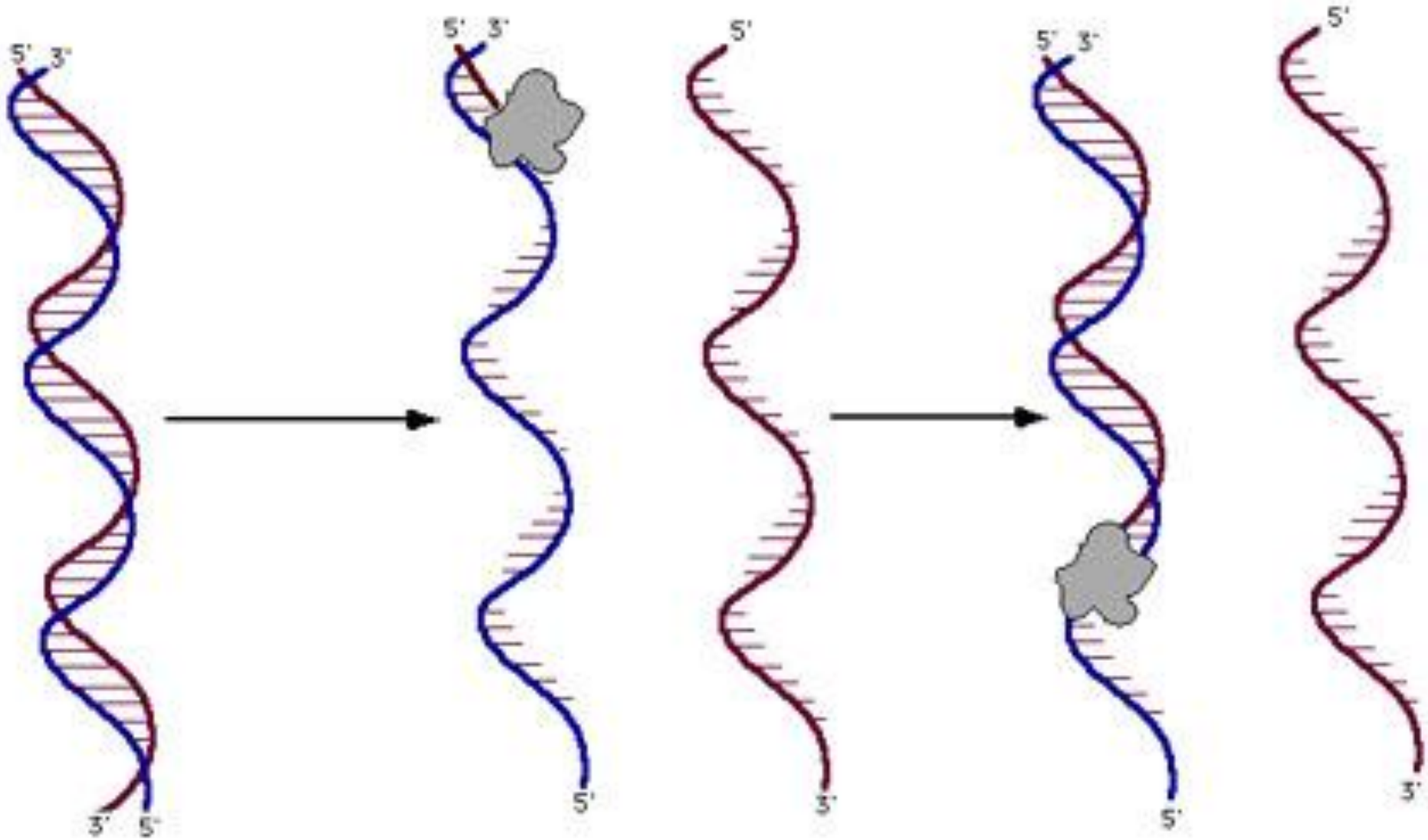
DNA polymerase I

- Because all four normal nucleotides are present, chain elongation proceeds normally until, by chance, DNA polymerase inserts a dideoxy nucleotide instead of the normal deoxynucleotide. If the ratio of normal nucleotide to the dideoxy versions is high enough, some DNA strands will succeed in adding several hundred nucleotides before insertion of the dideoxy version halts the process.
- At the end of the incubation period, the fragments are separated by length from longest to shortest. The resolution is so good that a difference of one nucleotide is enough to separate that strand from the next shorter and next longer strand. Each of the four dideoxynucleotides fluoresces a different color when illuminated by a laser beam and an automatic scanner provides a printout of the sequence.


Single-stranded DNA
to be sequenced
5'
3'
C T G A C T T C G A C A A
Add:
DNA polymerase I
dATP
dGTP
dCTP
dTTP
plus limiting amounts of
fluorescently labeled
ddATP
ddGTP
ddCTP
ddTTP
T 5'
T 5'
G 5'
T 5'
C 5'
G 5'
A 5'
A 5'
G 5'
T 5'
C 5'
A 5'
G 5'
5'
Larger
fragments
Smaller
fragments
3' G A C T G A A G C T G T T 5'
Electrophoresis
using laser to
activate the
fluorescent dideoxy
nucleotides and a
detector to
distinguish the
colors
So the
sequence
of the
template
strand is
5' C T G A C T T C G A C A A 3'

DNA sequencing reactions

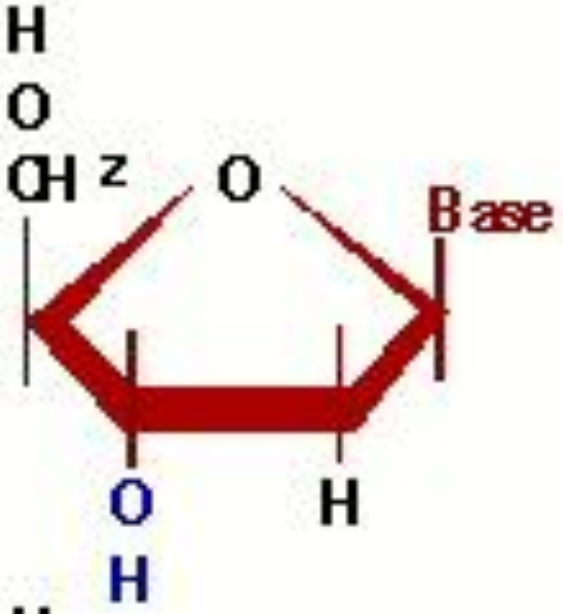
- are just like the PCR reactions for replicating DNA. The reaction mix includes the template DNA, free nucleotides, an enzyme (usually a variant of Taq polymerase) and a 'primer' - a small piece of single-stranded DNA about 20-30 nt long that can hybridize to one strand of the template DNA. The reaction is initiated by heating until the two strands of DNA separate, then the primer sticks to its intended location and DNA polymerase starts elongating the primer. If allowed to go to completion, a new strand of DNA would be the result.


5' 3'
3' 5'
5' 3'
5'
5'
3'
5' 3'
5'

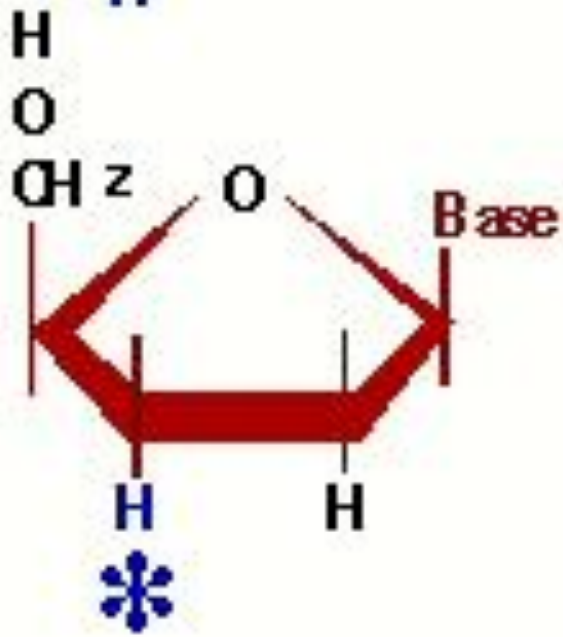
Dideoxynucleotides:

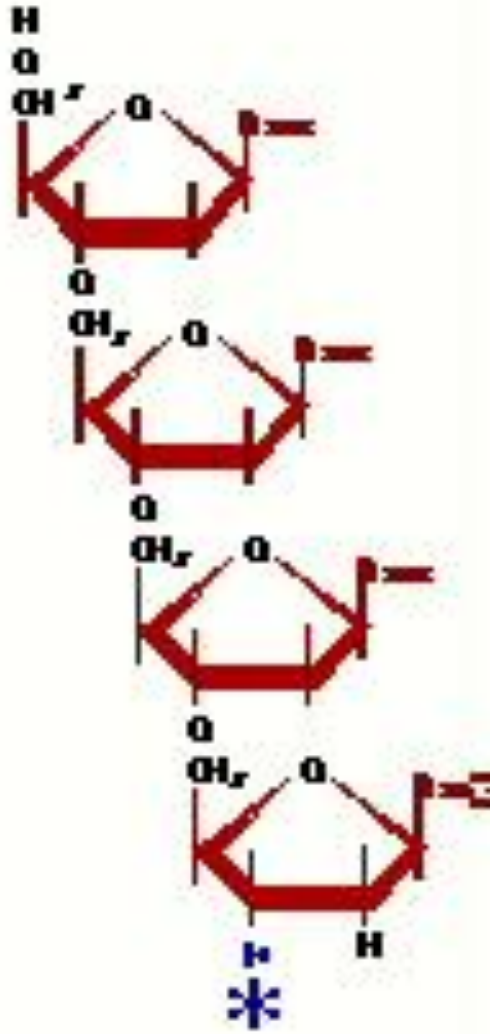
- We run the reactions, however, in the presence of a dideoxyribonucleotide. This is just like regular DNA, except it has no 3' hydroxyl group - once it's added to the end of a DNA strand, there's no way to continue elongating it. Now the key to this is that MOST of the nucleotides are regular ones, and just a fraction of them are dideoxy nucleotides....

Normal nucleotides:



Dideoxy Chain Terminators:





Replicating a DNA strand in the presence of dideoxy-T

- MOST of the time when a 'T' is required to make the new strand, the enzyme will get a good one and there's no problem. MOST of the time after adding a T, the enzyme will go ahead and add more nucleotides. However, 5% of the time, the enzyme will get a dideoxy-T, and that strand can never again be elongated. It eventually breaks away from the enzyme, a dead end product.
- Sooner or later ALL of the copies will get terminated by a T, but each time the enzyme makes a new strand, the place it gets stopped will be random. In millions of starts, there will be strands stopping at every possible T along the way.
- ALL of the strands we make started at one exact position. ALL of them end with a T. There are billions of them ... many millions at each possible T position. To find out where all the T's are in our newly synthesized strand, all we have to do is find out the sizes of all the terminated products!

DNA Polymerase reads the template strand and synthesizes a new second strand to match:

5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCTACCGAT
3' - ATGCGCCATTGCCATACAAGCTGGCAAATCGATGGCTAGAGATCCAA - 5'

IF 5% of the T nucleotides are actually dideoxy T, then each strand will terminate when it gets a ddT on its growing end:

5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCTACCGAT•

5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCT•

5' - TACGCGGTAACGGTATGTTCGACCGTTT•

5' - TACGCGGTAACGGTATGTTCGACCGTT•

5' - TACGCGGTAACGGTATGTTCGACCGT•

5' - TACGCGGTAACGGTATGTT•

5' - TACGCGGTAACGGTATGT•

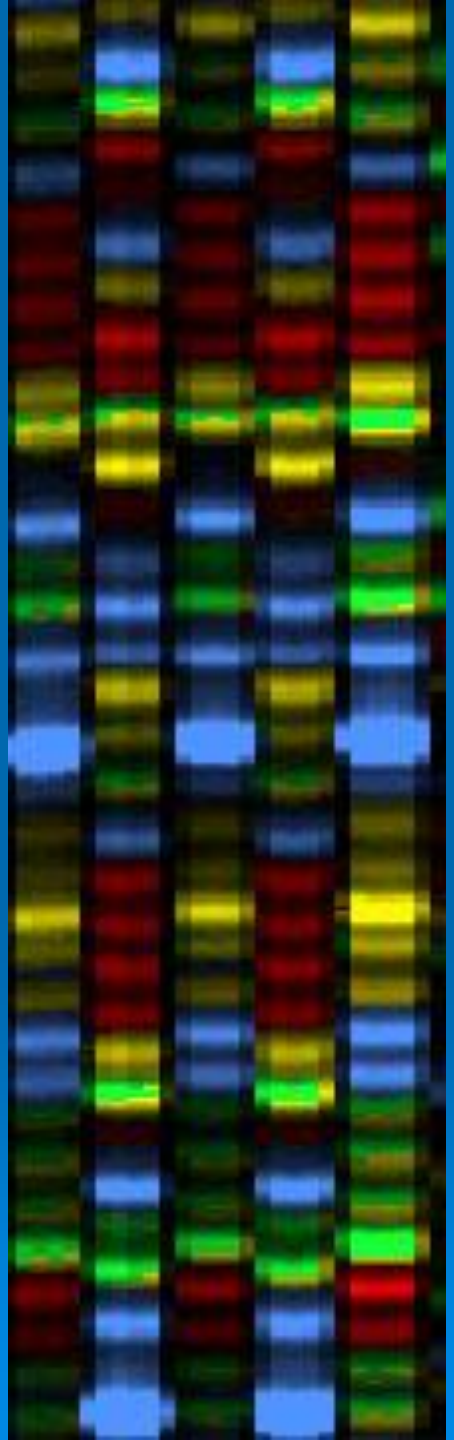
5' - TACGCGGTAACGGTAT•

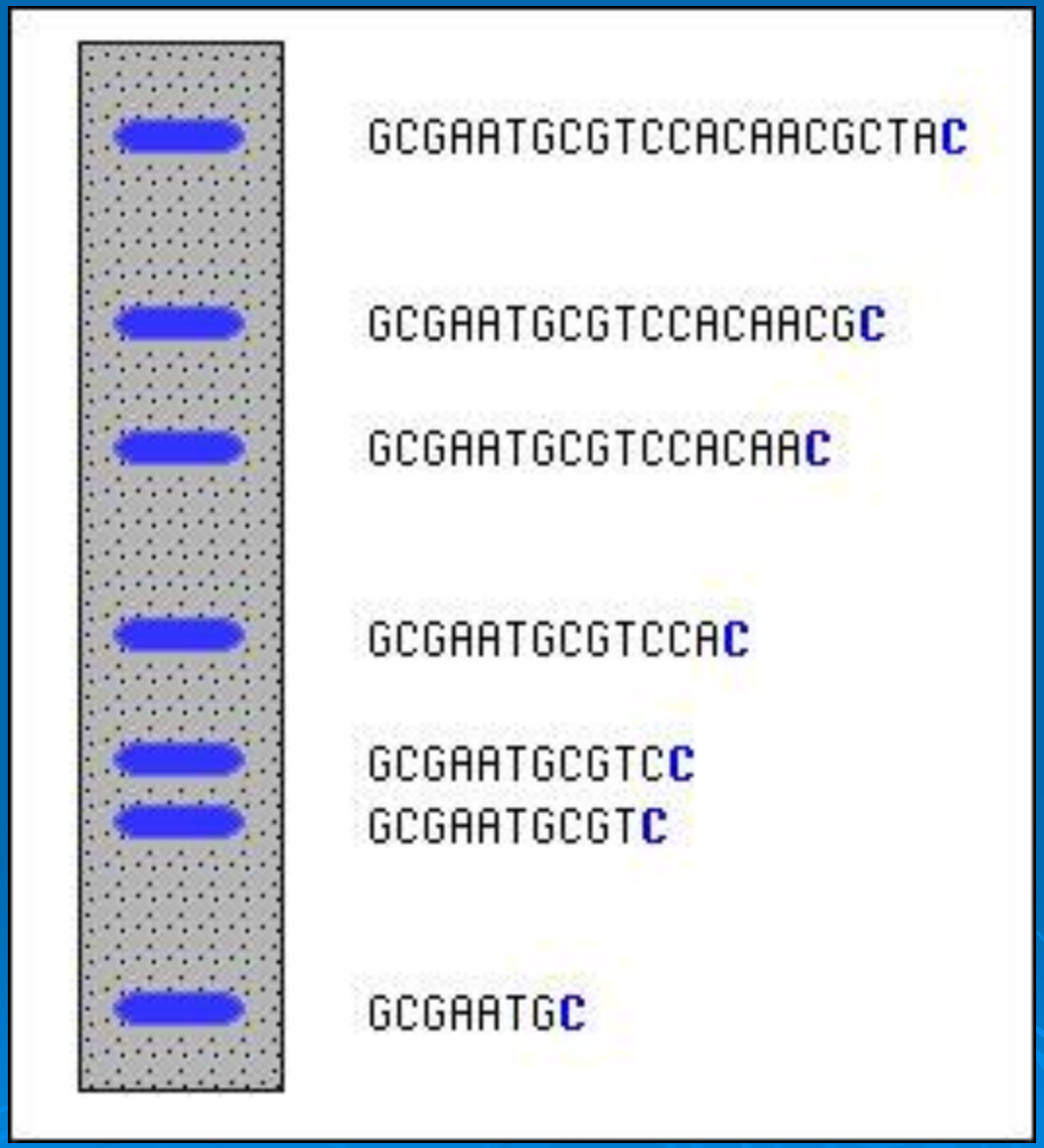
5' - TACGCGGTAACGGT•

5' - TACGCGGT•

Here's how we find out those fragment sizes.

- Gel electrophoresis can be used to separate the fragments by size and measure them.
- First, let's add one fact: the dideoxy nucleotides have been chemically modified to fluoresce under UV light. The dideoxy-C, for example, glows blue. Now put the reaction products onto an 'electrophoresis gel', and you'll see something like the next slide. Smallest fragments are at the bottom, largest at the top. The positions and spacing shows the relative sizes. At the bottom is the smallest fragment that's been terminated by ddC; that's probably the C closest to the end of the primer. Simply by scanning up the gel, we can see that we skip two, and then there's two more C's in a row. Skip another, and there's yet another C. And so on, all the way up. We can see where all the C's are.




GCGAATGCGTCCACAACGCTAC
GCGAATGCGTCCACAACGC
GCGAATGCGTCCACAAC
GCGAATGCGTCCAC
GCGAATGCGTCC
GCGAATGCGTC
GCGAATGC

Putting all four deoxynucleotides into the picture:

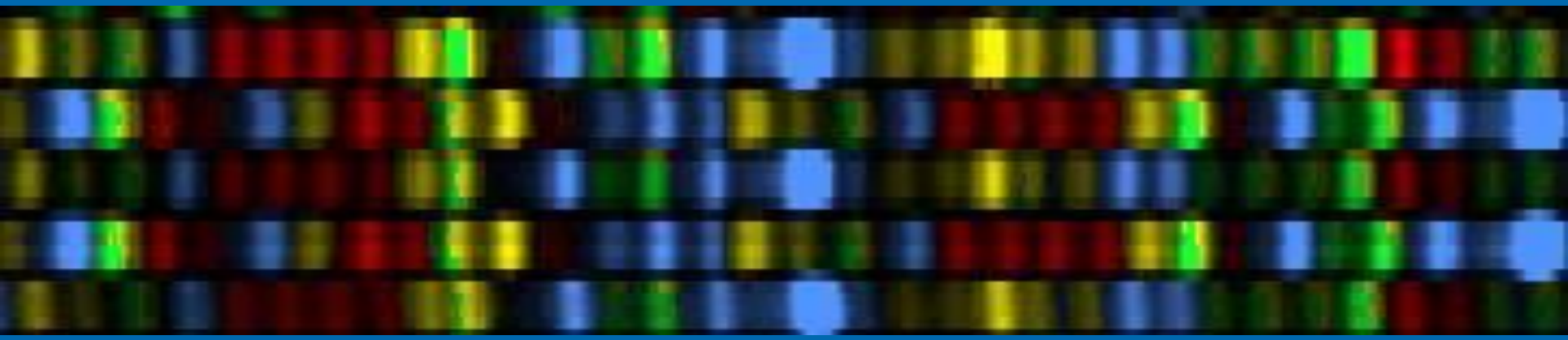
- Well, OK, it's not so easy reading just C's, as you perhaps saw in the last figure. The spacing between the bands isn't all that easy to figure out. Imagine, though, that we ran the reaction with *all four* of the dideoxy nucleotides (A, G, C and T) present, and with *different* fluorescent colors on each. NOW look at the gel we'd get. The sequence of the DNA is rather obvious if you know the color codes ... just read the colors from bottom to top: TGCGTCCA-(etc).

Gel:

G	GCGAATGCGTCCACAACGCTACAGGT**G**
T	GCGAATGCGTCCACAACGCTACAGG**T**
G	GCGAATGCGTCCACAACGCTACAG**G**
G	GCGAATGCGTCCACAACGCTACA**G**
A	GCGAATGCGTCCACAACGCTAC**A**
C	GCGAATGCGTCCACAACGCTA**C**
A	GCGAATGCGTCCACAACGCT**A**
T	GCGAATGCGTCCACAACGC**T**
C	GCGAATGCGTCCACAACG**C**
G	GCGAATGCGTCCACAACG
C	GCGAATGCGTCCACAA**C**
A	GCGAATGCGTCCACA**A**
A	GCGAATGCGTCCAC**A**
C	GCGAATGCGTCCA**C**
A	GCGAATGCGTCC**A**
C	GCGAATGCGTC**C**
C	GCGAATGCGT**C**
T	GCGAATGCG**T**
G	GCGAATGC**G**
C	GCGAATG**C**
G	GCGAAT**G**
T	GCGAA**T**

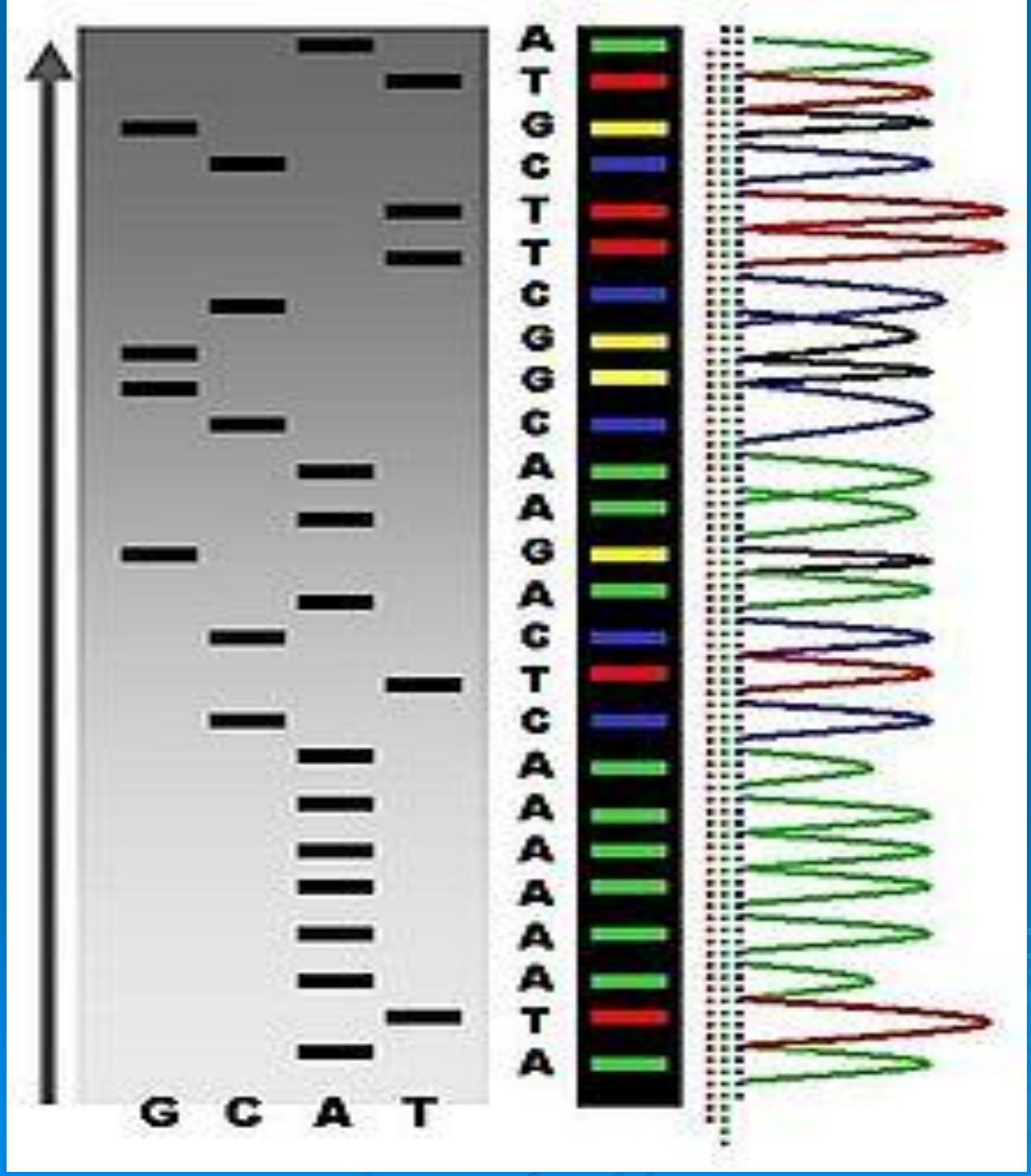
An Automated sequencing gel:

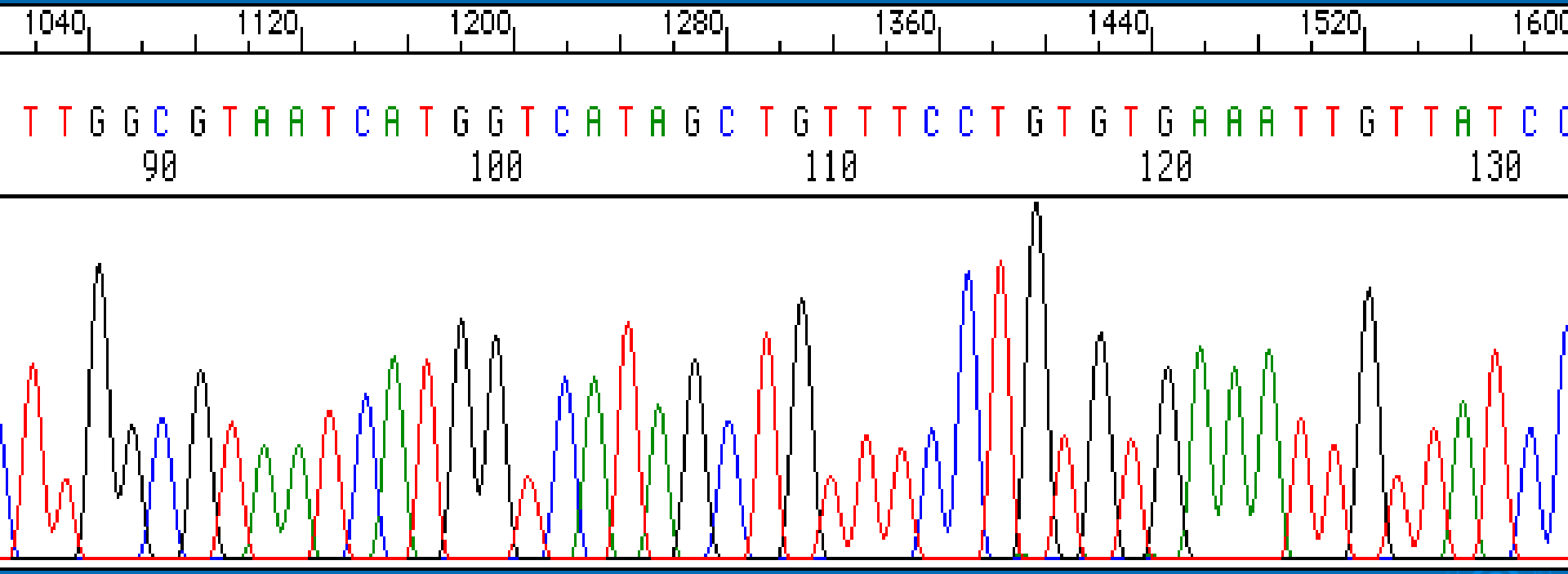
- That's exactly what we do to sequence DNA, then - we run DNA replication reactions in a test tube, but in the presence of trace amounts of *all four* of the dideoxy terminator nucleotides. Electrophoresis is used to separate the resulting fragments by size and we can 'read' the sequence from it, as the colors march past in order.
- Next is a screen shot of a real fragment of sequencing gel. The four colors red, green, blue and yellow each represent one of the four nucleotides.
- The actual gel image, if you could get a monitor large enough to see it all at this magnification, would be perhaps 3 or 4 meters long and 30 or 40 cm wide.



A 'Scan' of one gel lane:

- We don't even have to 'read' the sequence from the gel - the computer does that for us! Below is an example of what the sequencer's computer shows us for one sample. This is a plot of the colors detected in one 'lane' of a gel (one sample), scanned from smallest fragments to largest. The computer even interprets the colors by printing the nucleotide sequence across the top of the plot. **This is just a fragment of the entire file, which would span around 900 or so nucleotides of accurate sequence.**
- The sequencer also gives the operator a text file containing just the nucleotide sequence, without the color traces.

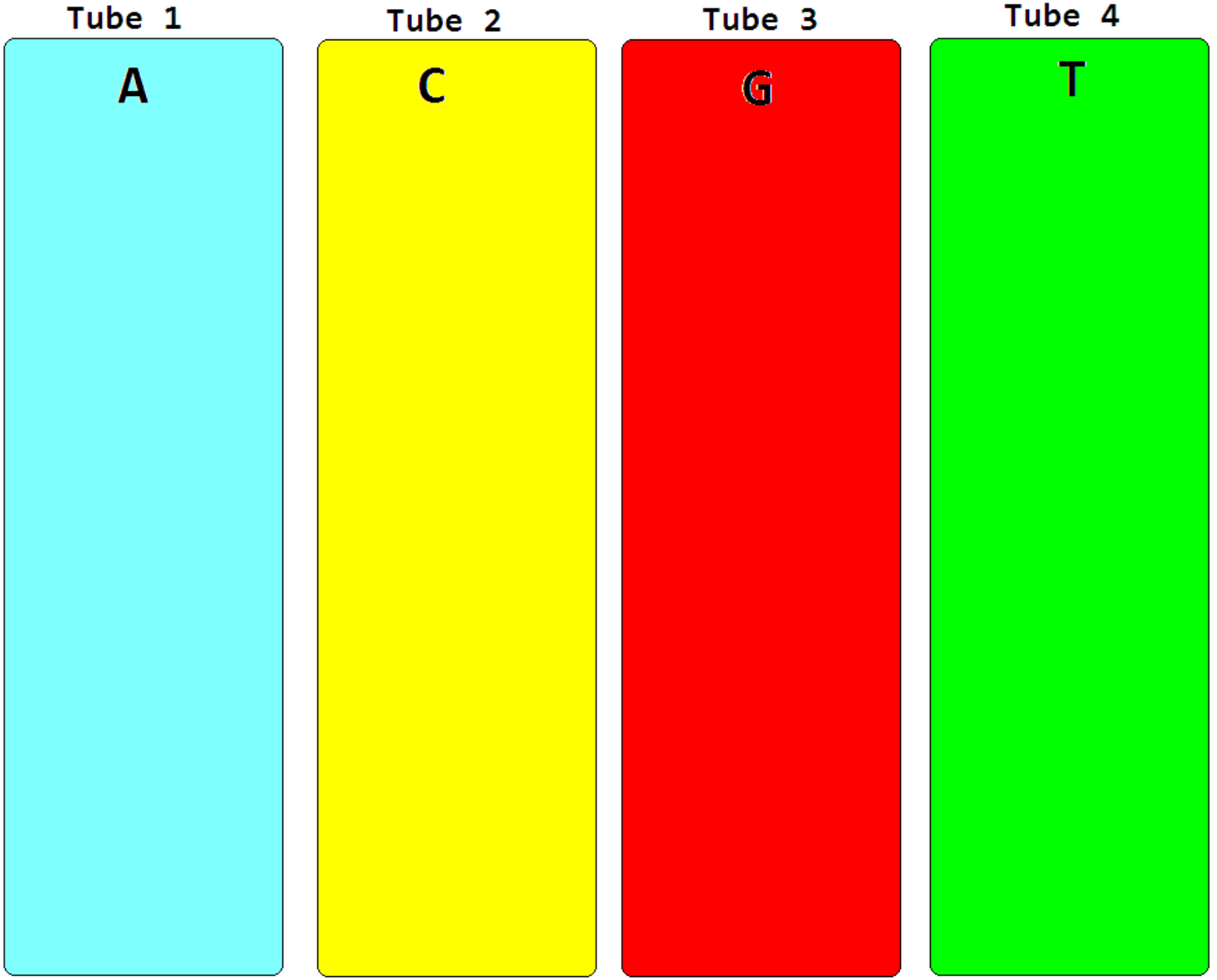

G C A T
A T G C T T C G G C A A G A C T C A A A A A A T A


1040
1120
1200
1280
1360
1440
1520
1600
T T G G C G T A A T C A T G G T C A T A G C T G T T T C C T G T G T G A A A T T G T T A T C C
90
100
110
120
130

- **As you have seen, we can get the sequence of a fragment of DNA as long as 900 or so nucleotides. Great! But what about longer pieces? The human genome is 3 *billion* bases long, arranged on 23 pairs of chromosomes. Our sequencing machine reads just a drop in the bucket compared to what we really need! To do it, we break the entire genome up into manageable pieces and sequence them. There are two approaches currently in use:**
- **The Publically-funded Human Genome Project:** The National Institutes of Health and the National Science Foundation have funded the creation of 'libraries' of BAC clones. Each BAC carries a large piece of human genomic DNA on the order of 100-300 kb.
- **A Privately-Funded Sequencing Project: Celera Genomics** An innovative approach to sequencing the human genome has been pioneered by Celera Genomics. The founders of this company realized that it might be possible to skip the entire step of making libraries of BAC clones. Instead, they blast apart the entire human genome into fragments of 2-10 kb and sequence those. Now the challenge is to assemble those fragments of sequence into the whole genome sequence. Imagine, for example that you have hundreds of 500-piece puzzles, each being assembled by a team of puzzle experts using puzzle-solving computers. Those puzzles are like BACs - smaller puzzles that make a big genome manageable. Now imagine that Celera throws all those puzzles together into one room and scrambles the pieces. They, however, have scanners that scan all the puzzle pieces and huge computers that figure out where they all go.

DNA SEQUENCING LAB instructions:

- Find a partner.
- Take out 2 sheets of paper.
- Get a box of markers from the front of the room.
- On the first paper draw the following:


Tube 1
A
Tube 2
C
Tube 3
G
Tube 4
T

- On the second paper, create a Electrophosesis gel that looks like this:

	G	A	T	C
21				
20				
19				
18				
17				
16				
15				
14				
13				
12				
11				
10				
9				
8				
7				
6				
5				
4				
3				
2				
1				

- There are 24 Single Stranded Sequences of DNA on the page.
- Color Code 6 of these Blue, 6 Yellow, 6 Red, and 6 Green.
- Cut out the 24 strips and place them in the correct colored tube on paper #1.