- **METHODS IN MOLECULAR DIAGNOSIS**

SDS-PAGE (**P**oly**A**crylamide **G**el **E**lectrophoresis)

The purpose of SDS-PAGE is to separate proteins according to their size, and no other physical feature. In order to understand how this works, we have to understand the two halves of the name: SDS and PAGE.

SDS

Since we are trying to separate many different protein molecules of different shapes and sizes, we first want to denatured so that the proteins no longer have any secondary, tertiary or quaternary structure (i.e. we want them to retain only their primary amino acid structure). We use SDS to denature all proteins to the same linear shape.

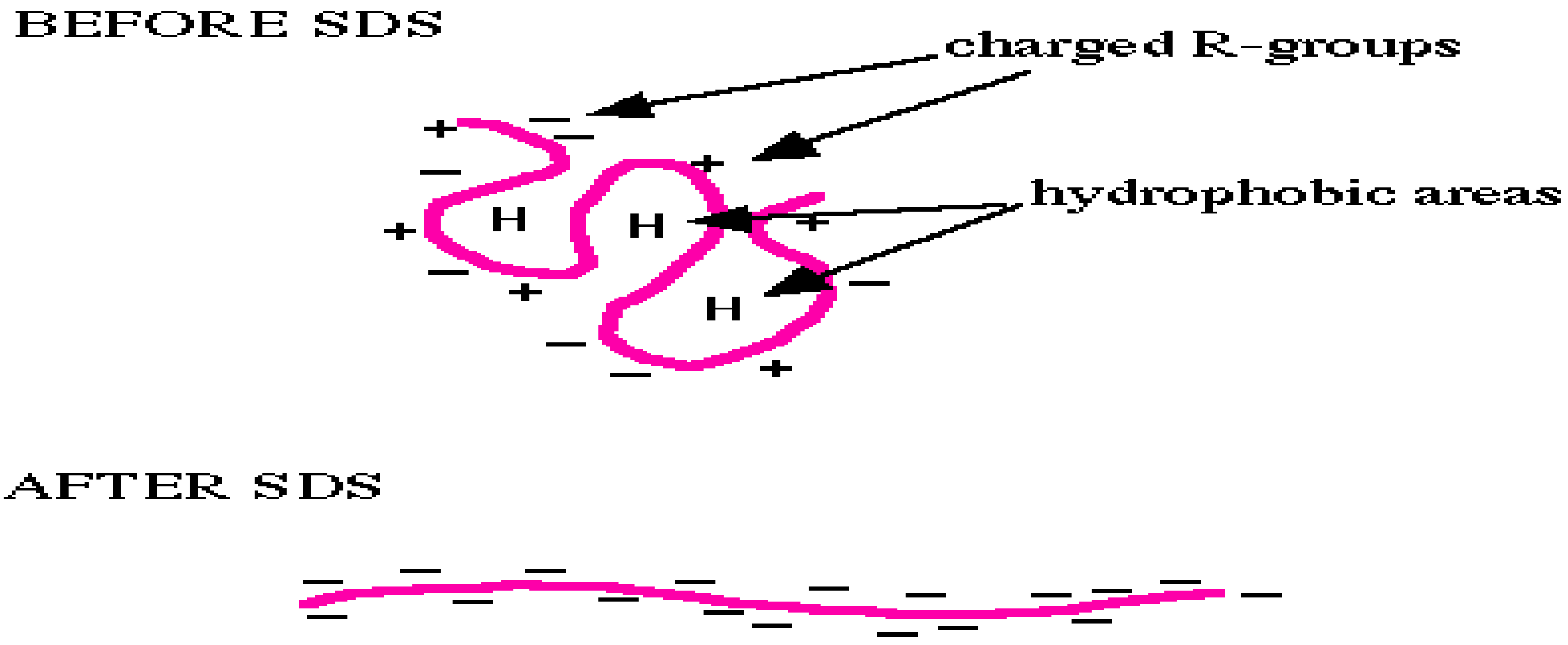


Figure 1. This cartoon depicts what happens to a protein (pink line) when it is incubated with the denaturing detergent SDS. The top portion of the figure shows a protein with negative and positive charges due to the charged R-groups in the protein. The large H's represent hydrophobic domains where nonpolar R-groups have collected in an attempt to get away from the polar water that surrounds the protein. The lower diagram shows that SDS can disrupt hydrophobic areas and coat proteins with many negative charges which overwhelms any positive charges the protein had due to positively charged R-groups. The resulting protein has been denatured by SDS (reduced to its primary structure) and as a result has been linearized.

SDS (sodium dodecyl sulfate) is a detergent (soap) that can dissolve hydrophobic molecules but also has a negative charge (sulfate) attached to it. Therefore, if a cell is incubated with SDS, the membranes will be dissolved, all the proteins will be solubilized by the detergent, plus all the proteins will be covered with many negative charges. So a protein that started out like the one shown in the top part of figure 1 will be converted into the one shown in the bottom part of figure 1. The end result has two important features: 1) all proteins retain only their primary structure and 2) all proteins have a large negative charge which means they will all migrate towards the positve pole when placed in an electric field. Now we are ready to focus on the second half - PAGE.

PAGE

If the proteins are denatured and put into an electric field, they will all move towards the positive pole at the same rate, with no separation by size. So we need to put the proteins into an environment that will allow different sized proteins to move at different rates. The environment of choice is polyacrylamide, which is a polymer of acrylamide monomers. When this polymer is formed, it turns into a gel and we will use electricity to pull the proteins through the gel so the entire process is called polyacrylamide gel electrophoresis (PAGE). A polyacrylamide gel is not solid but is made of a laberynth of tunnels through a meshwork of fibers (figure 2 and figure 3).

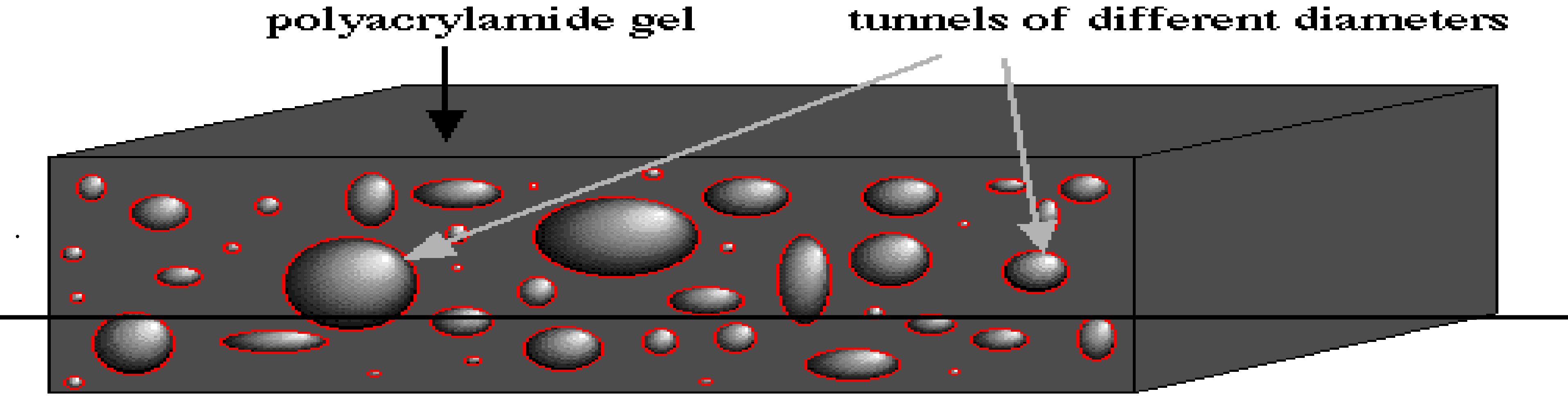


Figure 2. This cartoon shows a slab of polyacrylamide (dark gray) with tunnels (different sized red rings with shading to depict depth) exposed on the edge. Notice that there are many different sizes of tunnels scattered randomly throughout the gel.

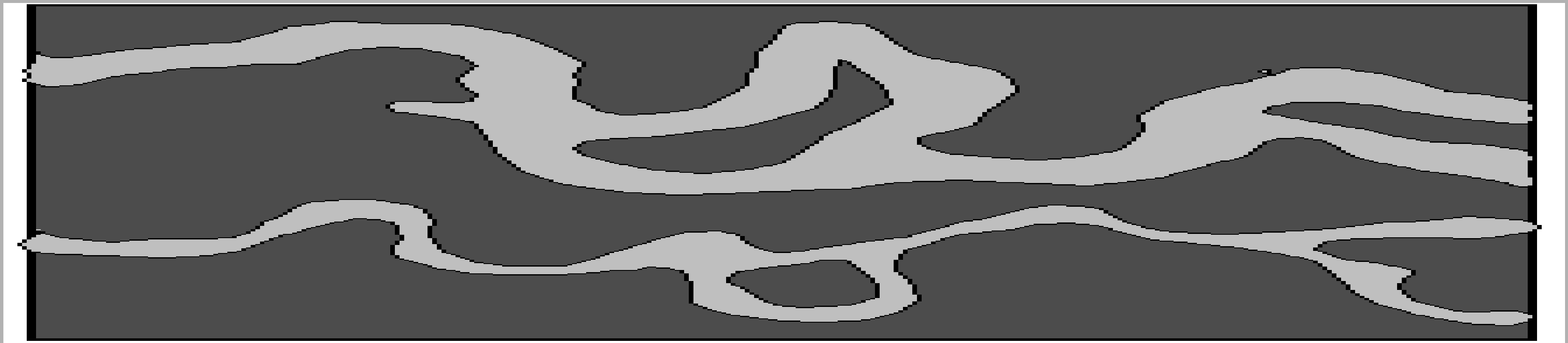


Figure 3. This is a top view of two selected tunnels (only two are shown for clarity of the diagram). Notice the difference in diameter of the two tunnels.

Now we are ready to apply the mixture of denatured proteins to the gel and apply the current (figure 4). If all the proteins enter the gel at the same time and have the same force pulling them towards the other end, which ones will be able to move through the gel faster? Small molecules can circulate through the polyacrylamide "forest" faster than big molecules.

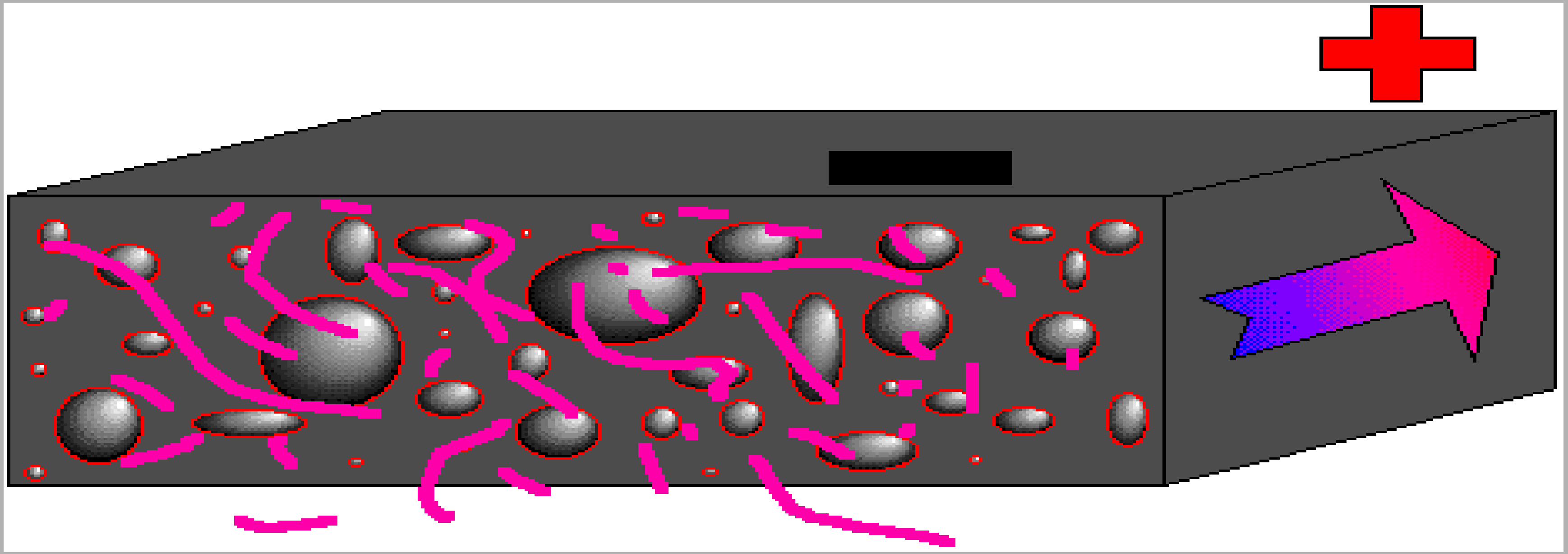


Figure 4. Cartoon showing a mixutre of denatured proteins (pink lines of differen lengths) beginning their journey through a polyacrylamide gel (gray slab with tunnels). An electric filed is established with the positive pole (red plus) at the far end and the negative pole (black minus) at the closer end. Since all the proteins have strong negative charges, they will all move in the direction the arrow is pointing (run to red).

When running an SDS-PAGE, we never let the proteins electrophorese (run) so long that they actually reach the other side of the gel. We turn off the current and then stain the proteins and see how far they moved through the gel (until we stain them, they are colorless and thus invisible). Figure 5 shows a cartoon gel and figure 6 shows a real one. Notice that the actual bands are equal in size, but the proteins within each band are of different sizes.

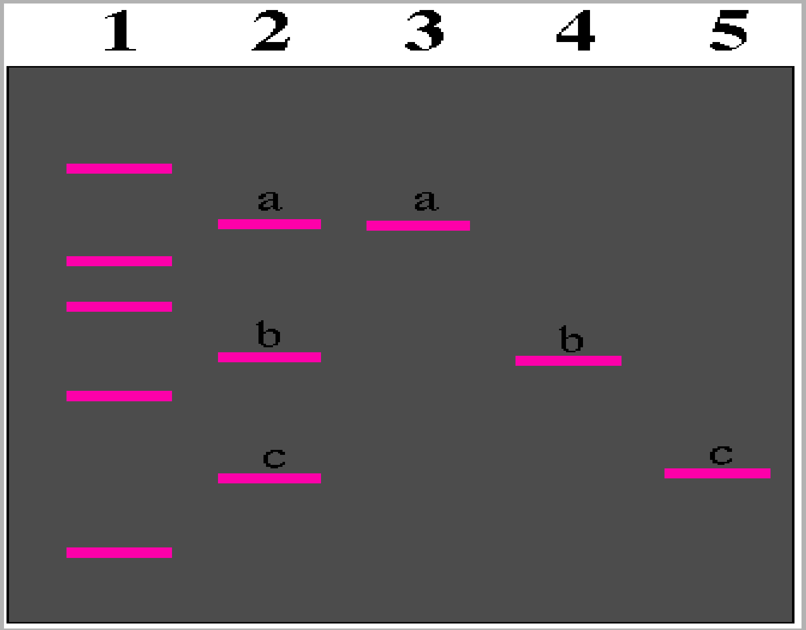
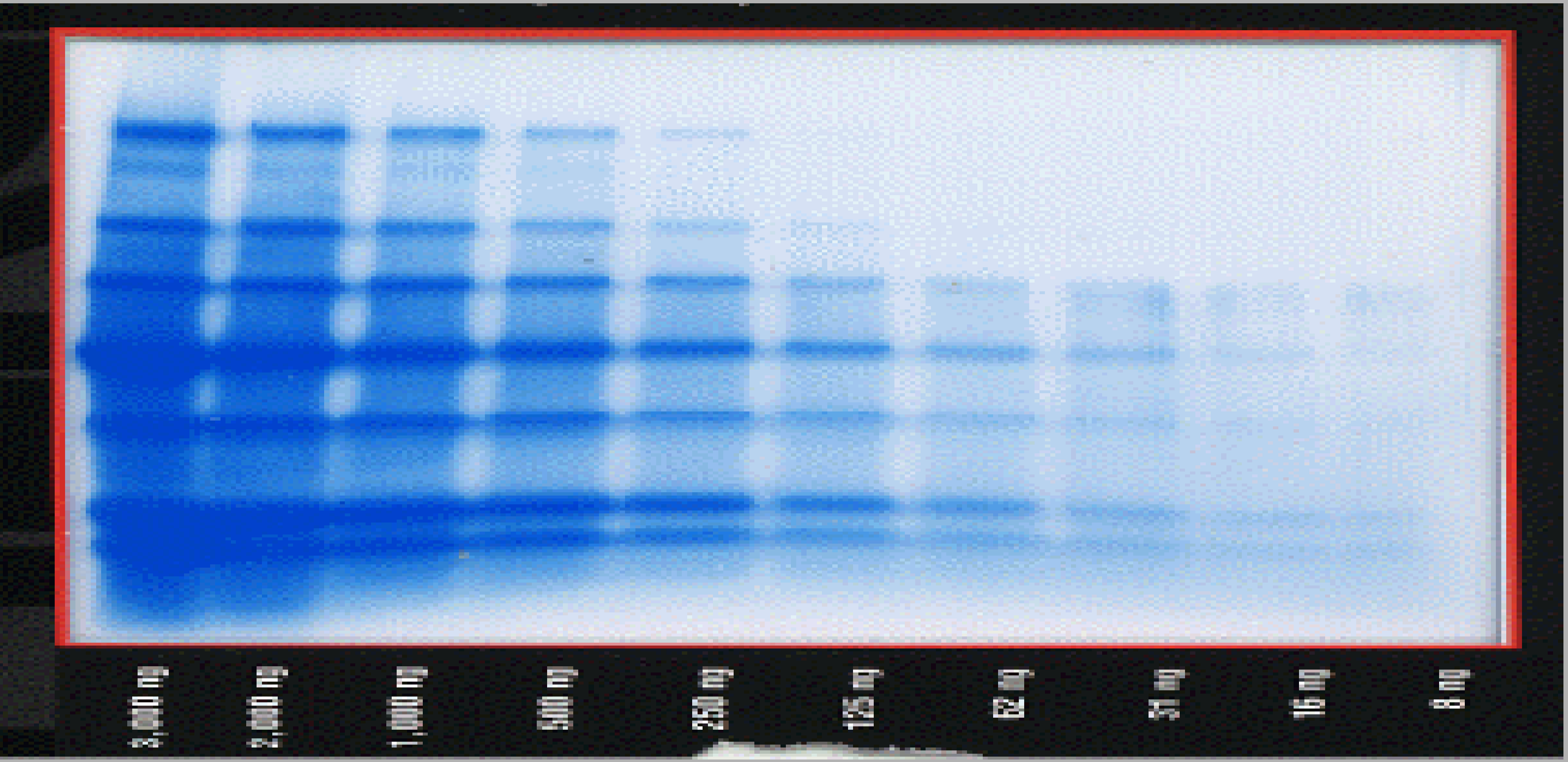


Figure 5. Lane 1, molecular weight standards of known sizes; Lane 2, a mixture of three proteins of different sizes with a being the largest and c being the smallest protein; Lane 3, protein a by itself; Lane 4, protein b by itself; Lane 5 protein c by itself. Notice that each group of the three proteins migrated the same distance in the gel whether they were with other proteins (lane 2) or not (lanes 3-5). The molecular weight standards are used to measure the relative sizes of the unknow proteins (a, b, and c).

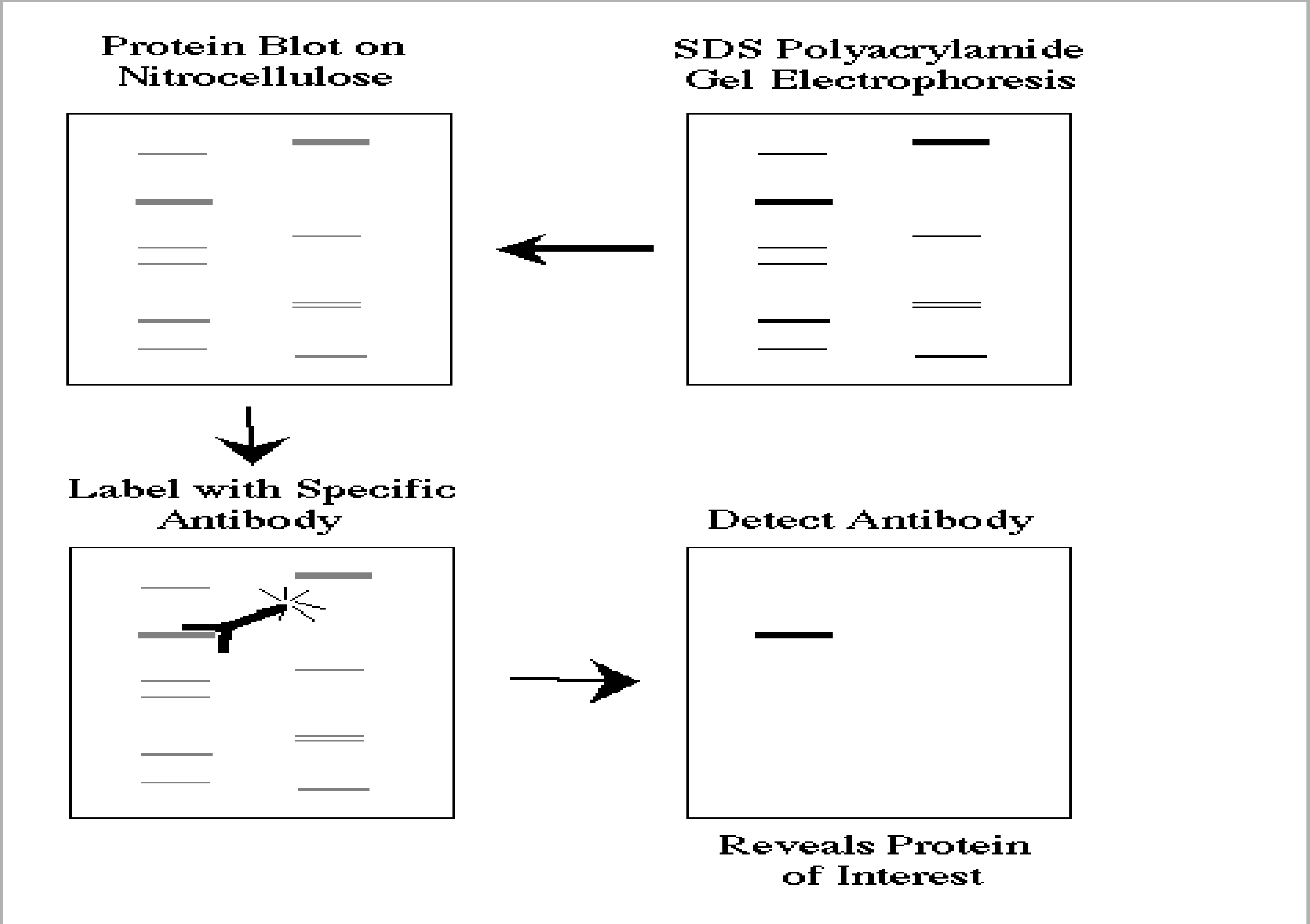
SDS-PAGE separates proteins based on their primary structure or size but not amino acid sequence. Therefore, if we had many copies of two different proteins that were both 500 amino acids long, they would travel together through the gel in a mixed band. As a result, we would not be able to use SDS-PAGE to separate these two proteins of the same molecular weight from each other.



This figure has from 3,000 ng (far left lane) to 8 ng (far right lane) of total protein loaded in the lanes. The proteins have been stained with coomassie blue.

Western Blot Procedure

This is a brief overview of how a western blot (more formally called a protein immunoblot) is performed and what type of data you can obtain from .


Protein Blot on Nitrocellulose
SDS Polyacrylamide Gel Electrophoresis
Label with Specific Antibody
Detect Antibody
Reveals Protein of Interest

Western blots allow investigators to determine the molecular weight of a protein and to measure relative amounts of the protein present in different samples.

1) Proteins are separated by gel electrophoresis, usually SDS-PAGE.

2) The proteins are transfered to a sheet of special blotting paper called nitrocellulose, though other types of paper, or membranes, can be used. The proteins retain the same pattern of separation they had on the gel.

3) The blot is incubated with a generic protein (such as milk proteins) to bind to any remaining sticky places on the nitrocellulose. An antibody is then added to the solution which is able to bind to its specific protein. The antibody has an enzyme (e.g. alkaline phosphatase or horseradish peroxidase) or dye attached to it which cannot be seen at this time.

4) The location of the antibody is revealed by incubating it with a colorless substrate that the attached enzyme converts to a colored product that can be seen and photographed

Epitope Tags for Antibody Binding

An epitope is a portion of a molecule to which an antibody binds. Epitopes can be composed of sugars, lipids or amino acids. In most cases, epitope tags are constructed of amino acids. Epitope tags are added to a molecule (usually proteins) which an investigator wants to visualize. Visualization can take place in a gel, a western blot or labeling via immunofluorescence.

If you wanted to follow a particular protein but did not have an antibody that would bind your protein, you might consider adding an epitope tag onto your protein. Epitope tags range from 10 to 15 amino acids long and are designed to create a molecular handle for your protein. An epitope tag could be placed anywhere within your protein, but typically they are placed on either the amino or carboxyl terminus to minimize any potential disruption in tertiary structure and thus function of your protein.

Although any short stretch of amino acids known to bind an antibody could become an epitope tag, there are a few that are especially popular. Three examples include:

c-myc - a 10 amino acid segment of the human protooncogene myc (EQKLISEEDL)

HA - haemaglutinin protein from human influenza haemagglutinin protein (YPYDVPDYA)
His$_6$ - If six histidines are placed in a row, they form a structure that binds the element Nickle. This is especially useful for affinity chromostography but can also be used as an epitope tag.

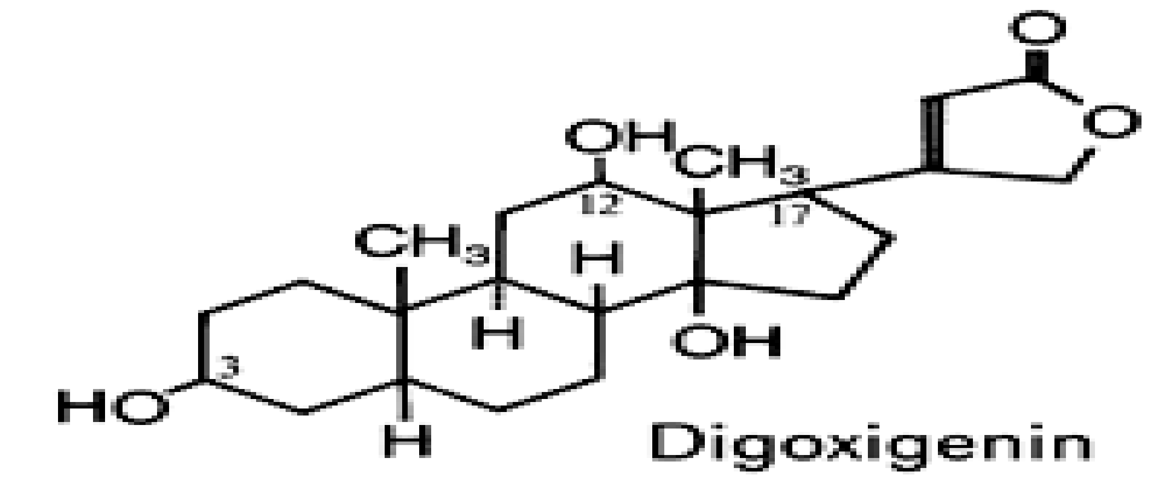


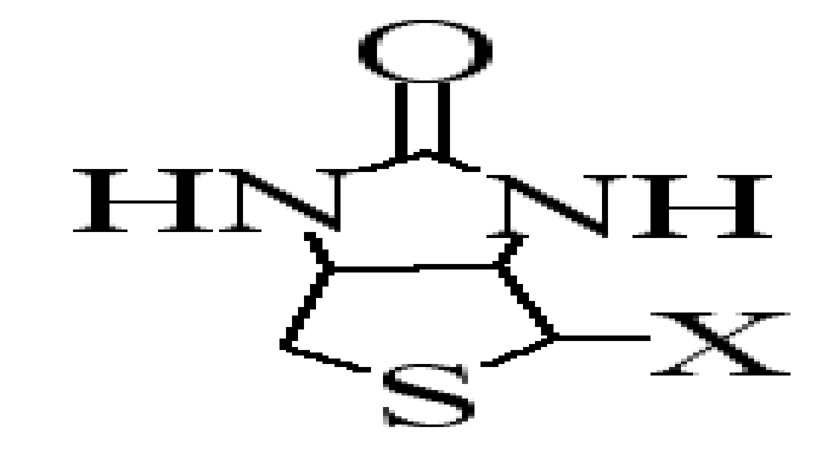
Figure 1. Structure of digoxigenin.

DIG - digoxigenin is a small organic molecule rather than amino acids (figure 1). DIG can be covalently added to proteins or nucleic acids which makes it very handy in its diverse applications.

·

biotin - biotin is a small molecule that can be covalently linked to proteins after they have been translated. Therefore, unlike most other protein epitope tags, it can be added at any point in time and is often used to label proteins located in particular sites such as on the extracellular surface of cells.

Biotin (also known as vitamin H) is a small organic molecule found in every cell (figure 1). Avidin (also called strepavidin) is a much larger protein that binds biotin with a very high affinity . When these two molecules are in the same solution, they will bind with such high affinity that the binding is essentially irreversible.



Fiugre 1. Structure of biotin. Because biotin can be modified differently in different organisms, the X is used to denote a variable side chain.

flourescent dyes (e.g. FITC, Cy3, Cy5, etc.) - since antibodies that bind to any antigen, epitope tags can be generated from any molecule. Investigators have taken full advantage of this and used fluorecent dyes as eiptope tags.

Affinity Chromatography Method

Affinity chromatography is designed to purify a particular protein from a mixed sample.

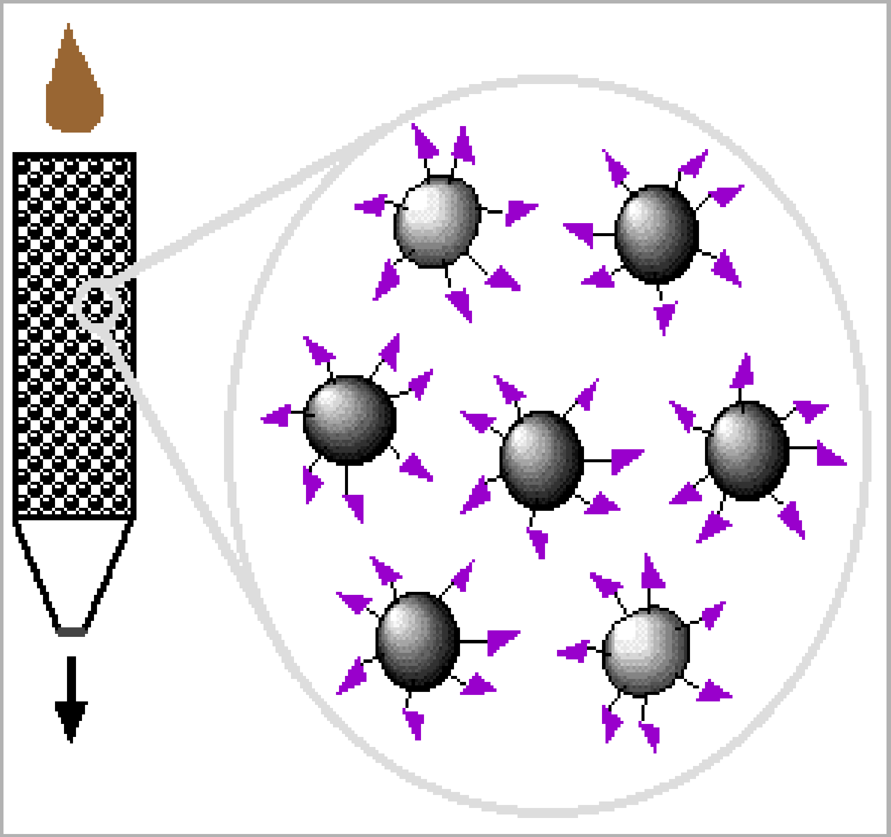


Figure 1. Loading affinity column.

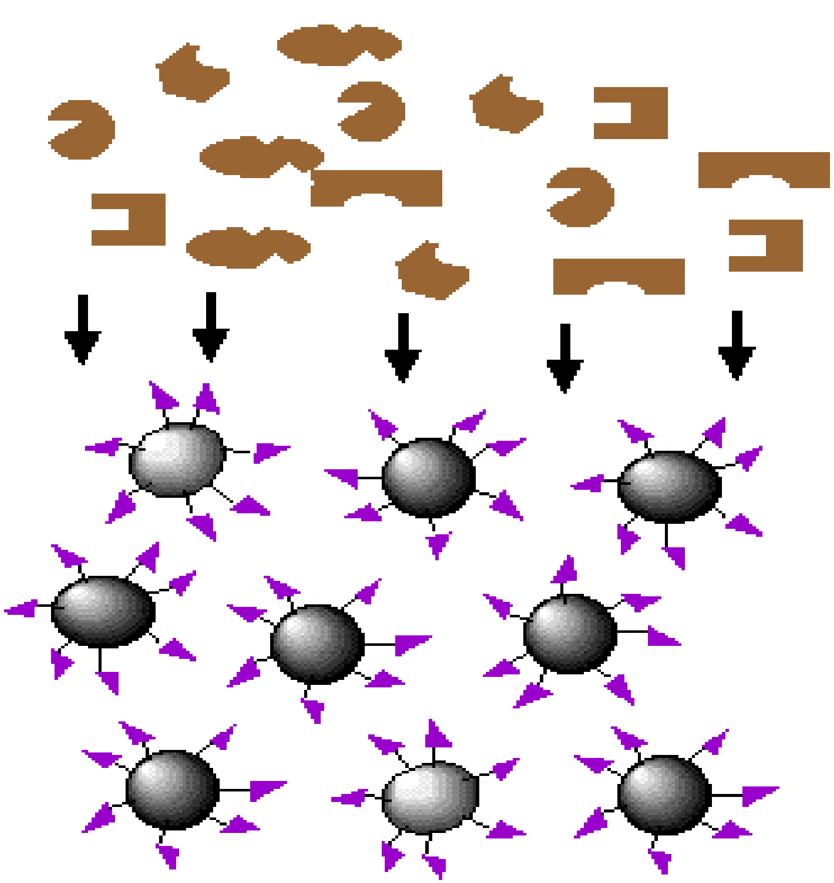


Figure 2. Proteins sieve through matrix of affinity beads.

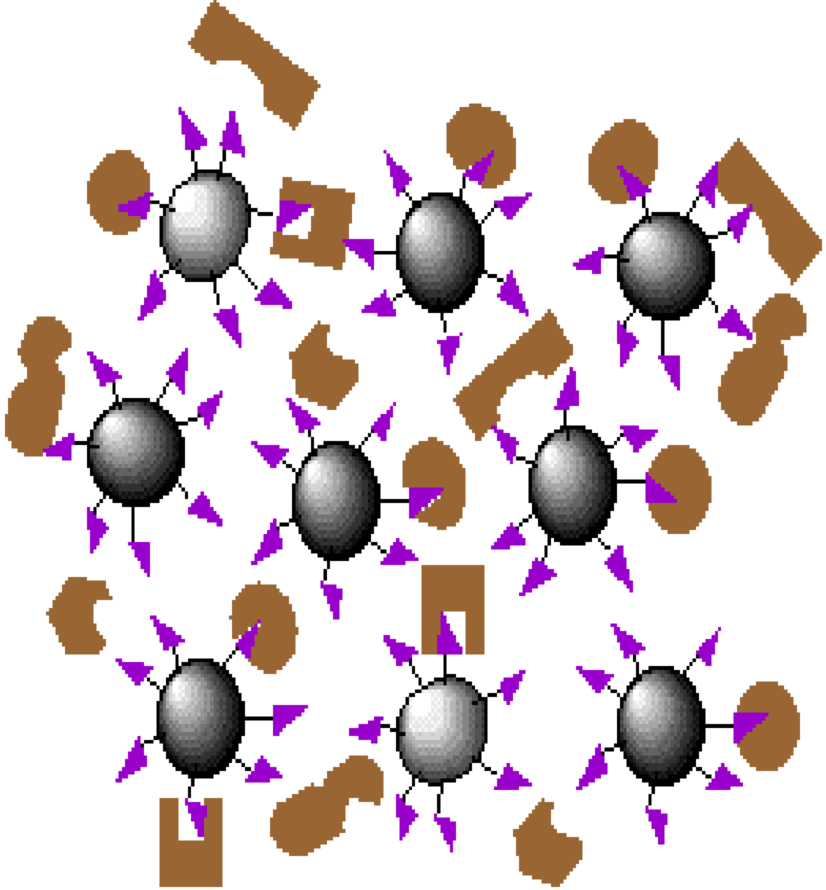


Figure 3. Proteins interact with affinity ligand with some binding loosely and others tightly.

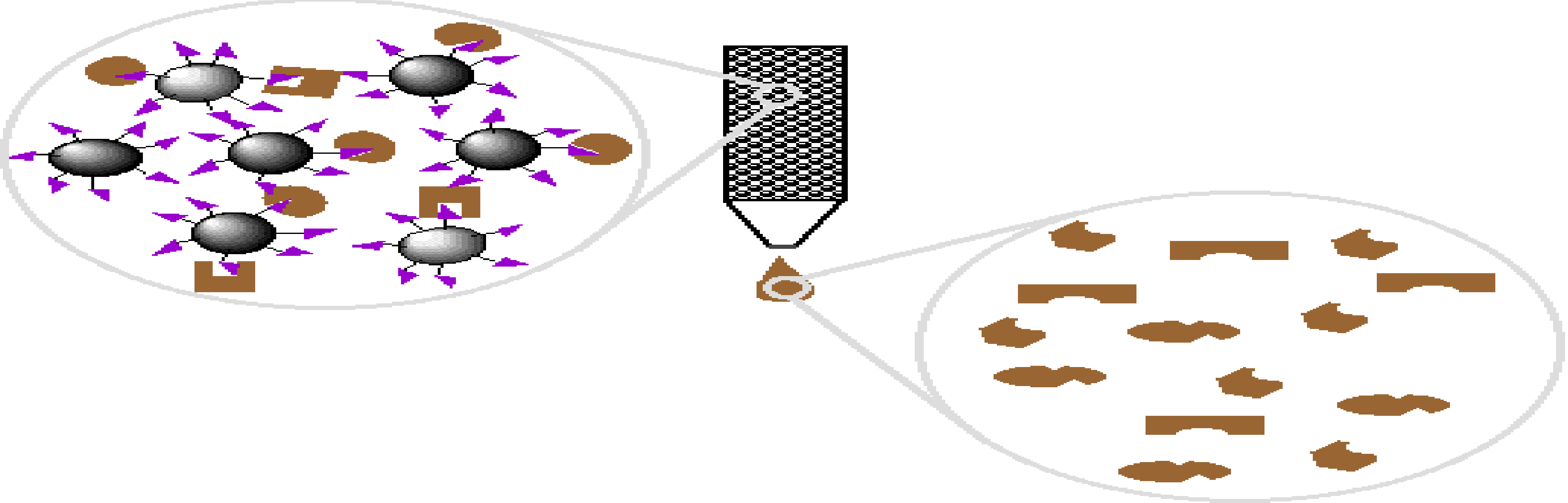


Figure 4. Wash off proteins that do not bind.

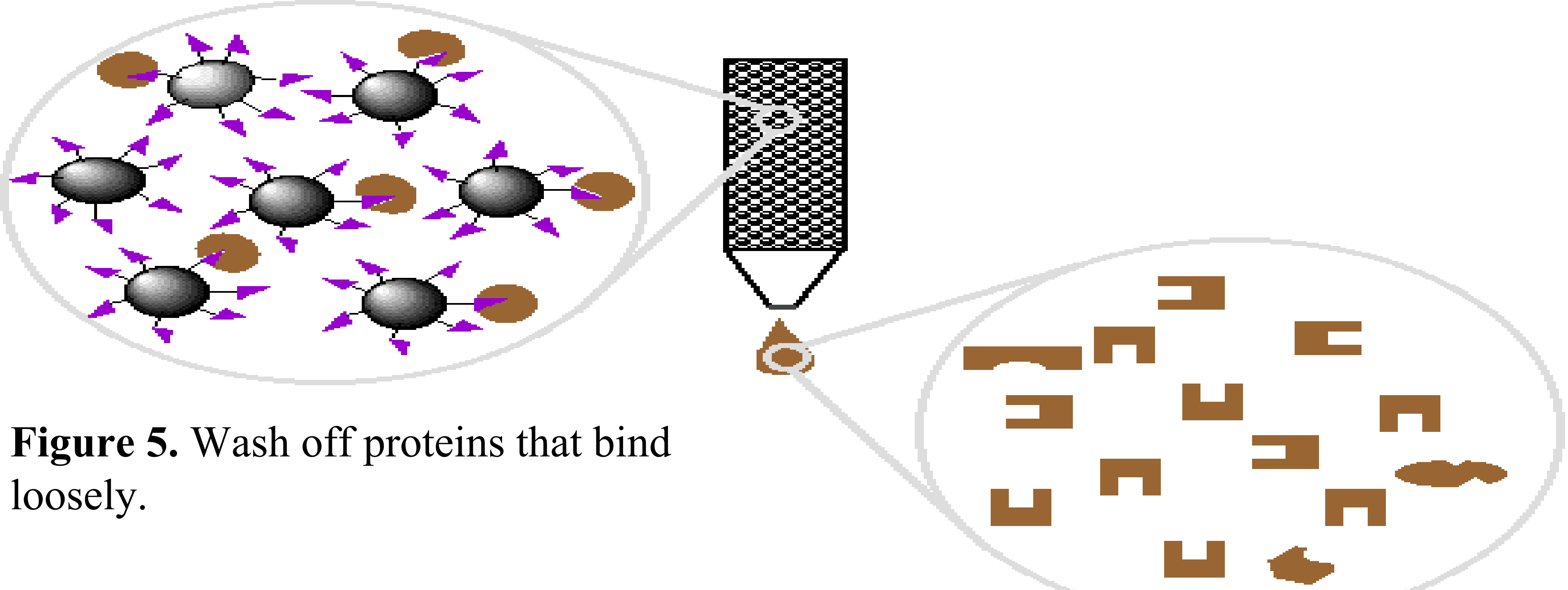


Figure 5. Wash off proteins that bind loosely.

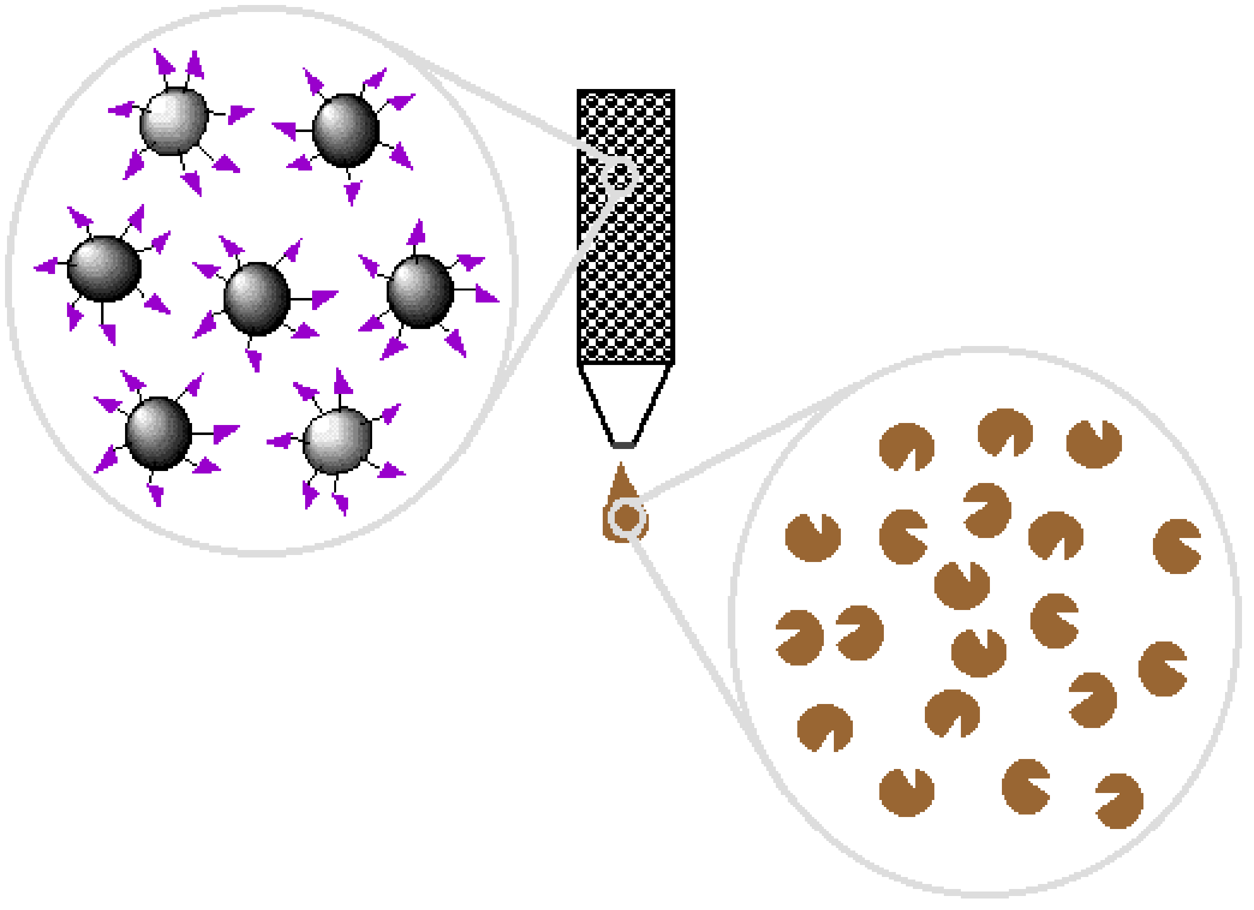


Figure 6. Elute proteins that bind tightly to ligand and collect purified protein of interest.

Immunofluorescence Method

The purpose of immunofluorescence is to detect the location and relative abundance of any protein for which you have an antibody. Once you have antibodies to your favorite protein, you can use them to indicate where the protein is located. In this example, we will use antibodies for the calcium ATPase, or pump, that is located in the endoplasmic reticulum (ER) of every cell. The antibody used here only recognized the chicken calcium ATPase but immunofluorescence can be used on any protein.

The key to this entire process is the ability to visualize the antibody when looking through a microscope. Therefore, you have to use a fluorescent dye that is covalently attached to the antibody. When a light illuminates the fluorescent dye, it absorbs the light and emits a different color light which is visible to the investigator and can be photographed.

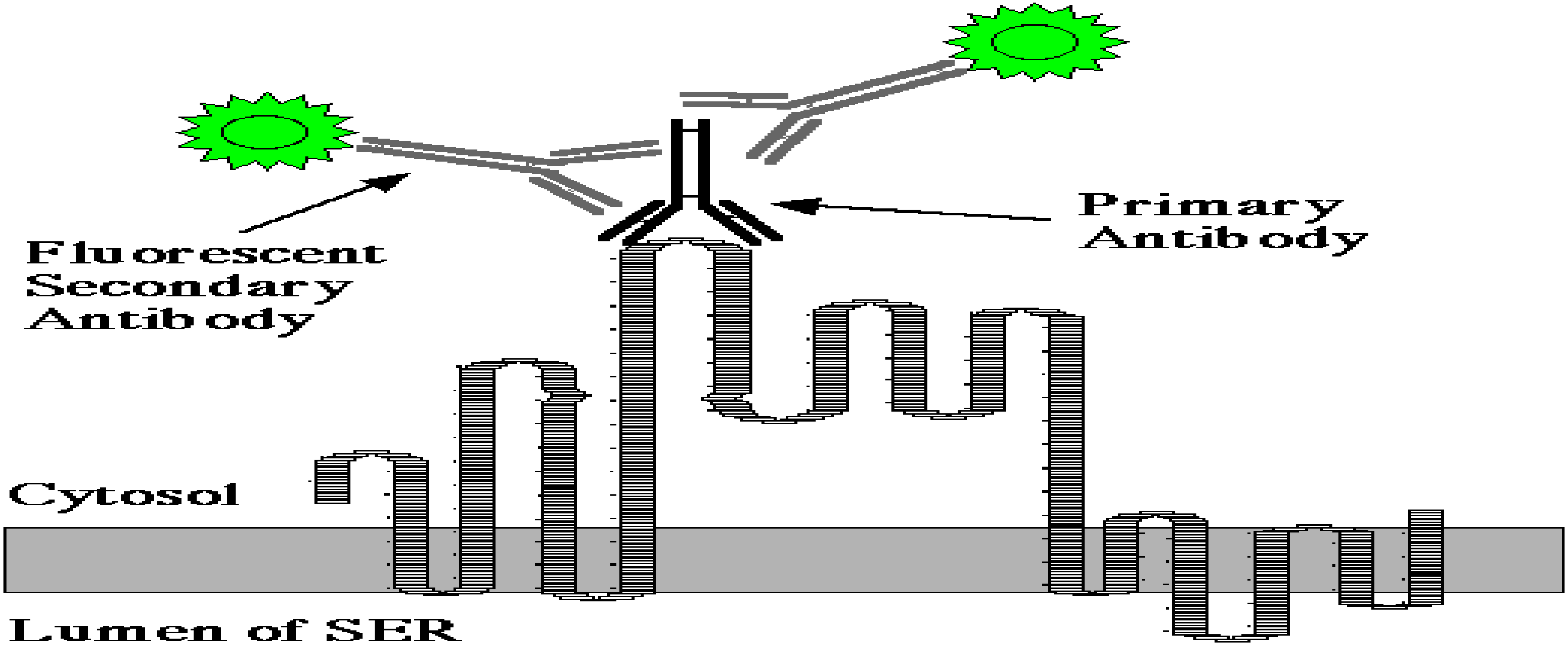


Figure 1. In most immunofluorescence experiments, two antibodies are employed. The first one, called the primary antibody, is typically generated in a mouse and binds to your favorite protein, which in this case is the chicken calcium ATPase . The secondary antibody was purchased from a company that sells antibodies that bind to mouse antibodies and have a fluorescent dye covalently attached to it. As illustrated here, the secondary antibodies can bind to multiple sites on the primary antibody and thus produce a brighter signal since more dyes are brought to a single location.

The first step is to choose your cells of interest. In this case, we will look at a chicken fibroblast, or skin cell. It was grown in tissue culture and so it appears as an isolated cell with no visible neighbors.

The cell was fixed with formaldehyde to retain the shape and location of all cellular proteins. The cell was treated with a mild detergent to disolve small holes in the membranes so the antibodies could have access to the cytoplasm. Because the calcium ATPase is located in the ER, the antibodies must have access to the cytoplasm or they could not bind to the target protein.

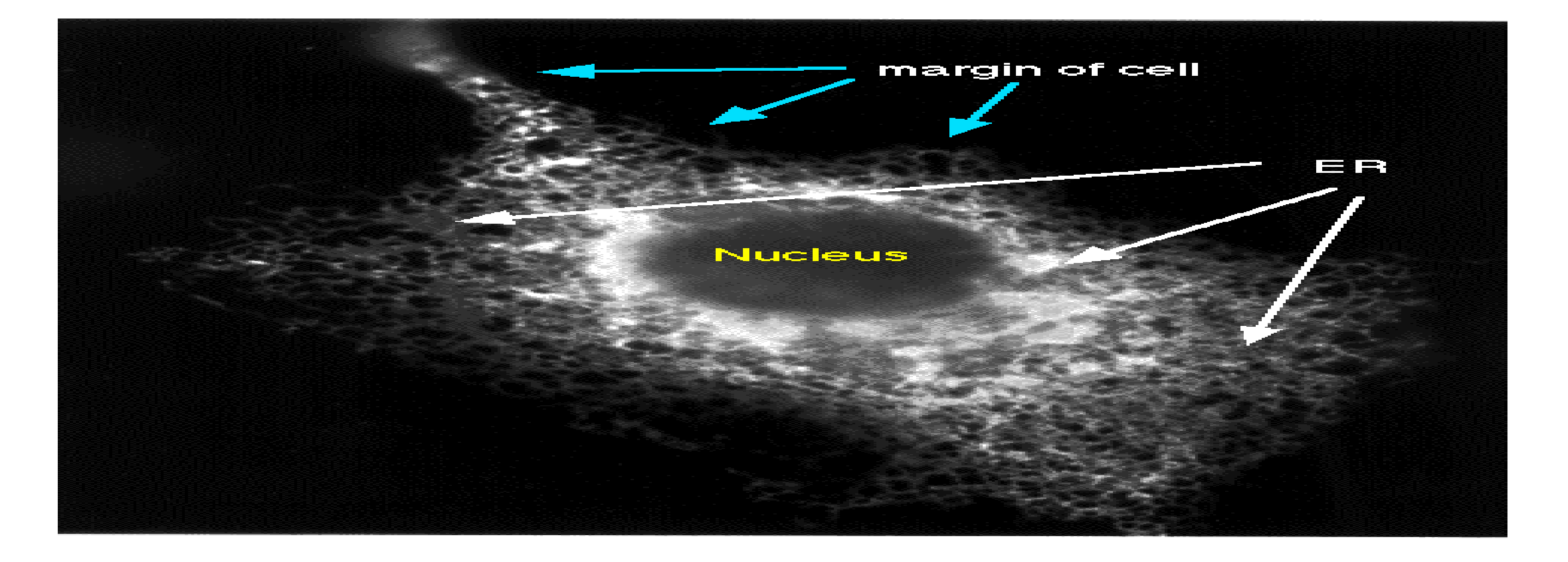
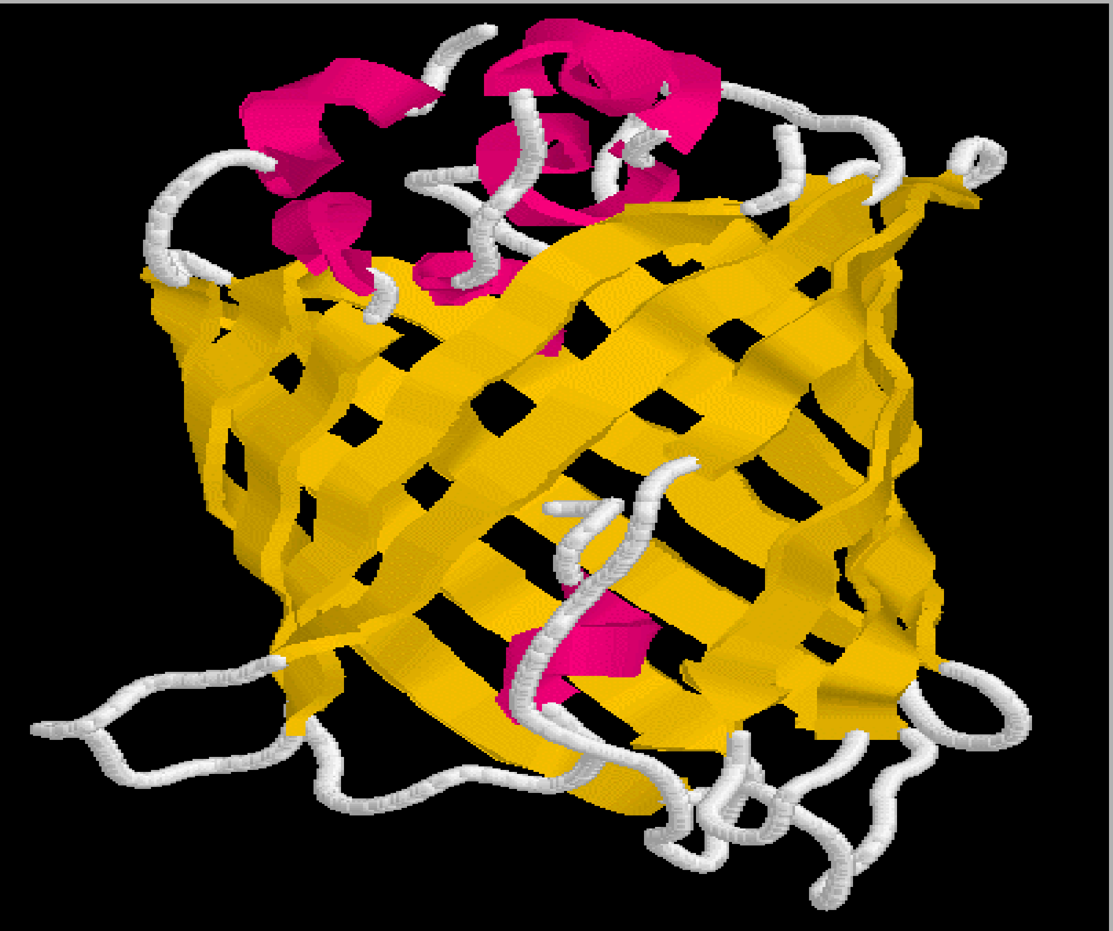


Figure 2. This immunofluorescence micrograph shows the ER being labeled with a monoclonal antibody against the chicken calcium ATPase. This chicken cell was fixed, permeabilized, and processed for immunofluorescence. White indicates the location of the fluorescent antibody and thus the calcium ATPase to which the antibody was bound. Using immunofluorescence, investigators can see when, where and how much of their favorite protein is expressed in any cell or tissue.

GFP - Green Fluorescent Protein as a Reporter

If you wanted to know whether a particular promoter were activated or not you'd like to have an easy way to see that a promoter is activated. Or perhaps you would like to be able to watch your favorite protein as it traveled around a cell , performing its cellular role. One protein has been very successful at both of these molecular methods - Green Flourescent Protein (GFP; figure 1).

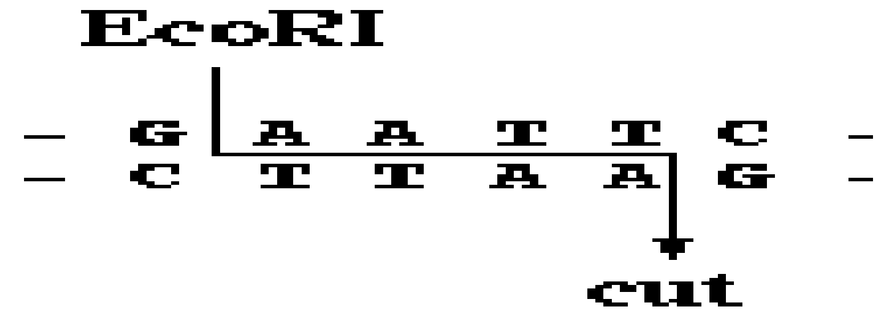


GFP was the first of many flourescent proteins to be used as a reporter protein. When excited with a blue light, it will flouresce green. It is very stable, can function when added to either end of a protein of interest, and does not fade easily

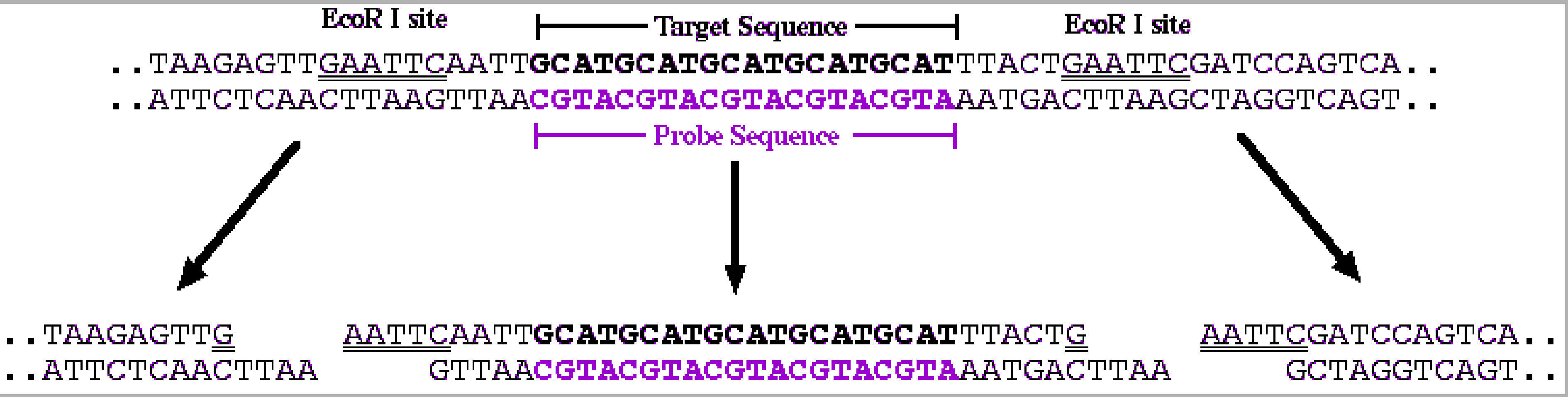
RFLP Method - Restriction Fragment Length Polymorphism

RFLP is a method used by molecular biologists to follow a particular sequence of DNA. RFLPs can be used in many different settings to accomplish different objectives. RFLPs can be used in paternity cases or criminal cases to determine the source of a DNA sample.

RFLP Production
Each organism inherits its DNA from its parents. Since DNA is replicated with each generation, any given sequence can be passed on to the next generation. An RFLP is a sequence of DNA that has a restriction site on each end with a "target" sequence in between. A target sequence is any segment of DNA that bind to a probe by forming complementary base pairs.
A probe is a sequence of single-stranded DNA that has been tagged with radioactivity or an enzyme so that the probe can be detected. When a probe base pairs to its target, the investigator can detect this binding and know where the target sequence is since the probe is detectable. For example, let's follow a particular RFLP that is defined by the restriction enzyme EcoR I and the target sequence of 20 bases GCATGCATGCATGCATGCAT. EcoR I binds to its recognition seuqence GAATTC and cuts the double-stranded DNA as shown:



In the segment of DNA shown below, you can see the elements of an RFLP; a target sequence flanked by a pair of restriction sites. When this segment of DNA is cut by EcoR I, three restriction fragments are produced, but only one contains the target sequence which can be bound by the complementary probe sequence (purple).



EcoR I site | Target Sequence | EcoR I site

..TAAGAGTTGAATTCAATT**GCATGCATGCATGCATGCAT**TTACTGAATTCGATCCAGTCA..
..ATTCTCAACTTAAGTTAA**CGTACGTACGTACGTACGTA**AATGACTTAAGCTAGGTCAGT..

Probe Sequence

..TAAGAGTTG
..ATTCTCAACTTAA

AATTCAATT**GCATGCATGCATGCATGCAT**TTACTG
GTTAA**CGTACGTACGTACGTACGTA**AATGACTTAA

AATTCGATCCAGTCA..
GCTAGGTCAGT..

Let's look at two people and the segments of DNA they carry that contain this RFLP (for clarity, we will only see one of the two stands of DNA). Since Jack and Jill are both diploid organisms, they have two copies of this RFLP. When we examine one copy from Jack and one copy from Jill, we see that they are identical:

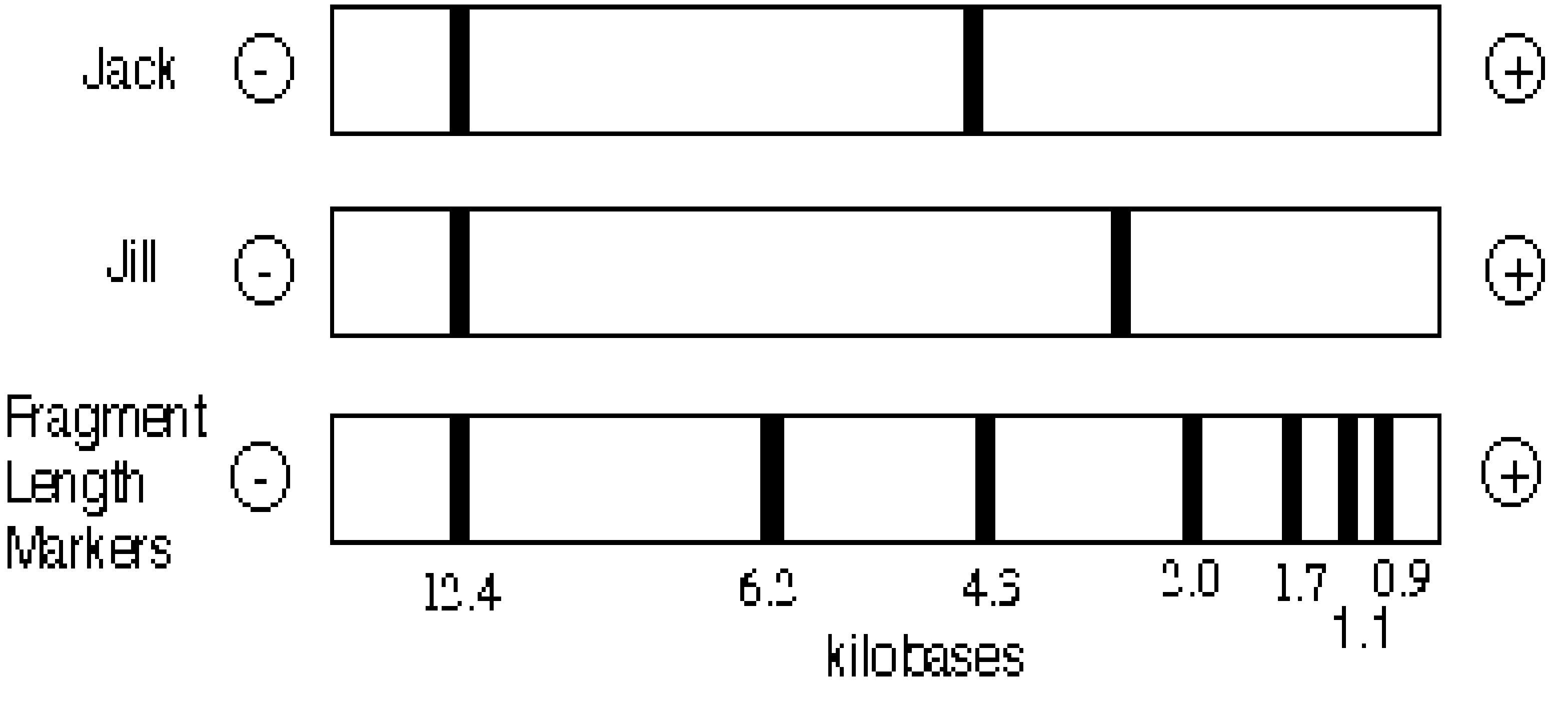
Jack 1: -GAATTC---(8.2 kb)---**GCATGCATGCATGCATGCAT**---(4.2 kb)---GAATTC-
Jill 1: -GAATTC---(8.2 kb)---**GCATGCATGCATGCATGCAT**---(4.2 kb)---GAATTC-

When we examine their second copies of this RFLP, we see that they are not identical. Jack 2 lacks an EcoR I restriction site that Jill has 1.2 kb upstream of the target sequence (difference in italics).

Jack 2: -GAATTC--(1.8 kb)-*CCCTTT*--(1.2 kb)--**GCATGCATGCATGCATGCAT**--(1.3 kb)-GAATTC-

Jill 2: -GAATTC--(1.8 kb)-*GAATTC*--(1.2 kb)--**GCATGCATGCATGCATGCAT**--(1.3 kb)-GAATTC-

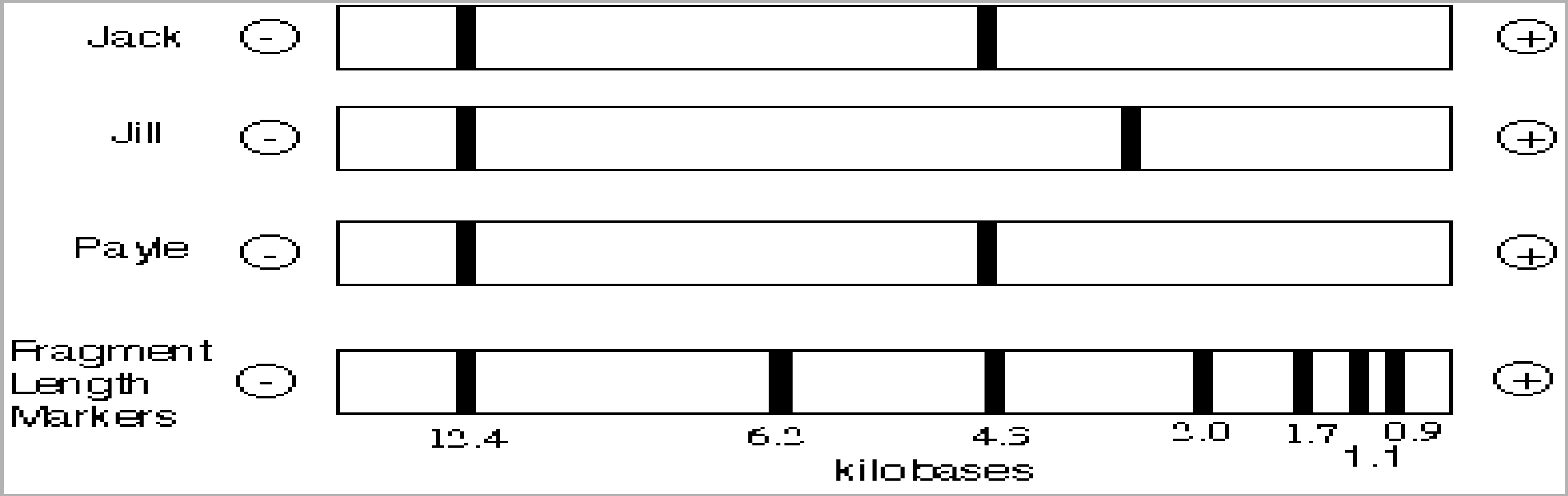
Therefore, when Jack and Jill have their DNA subject to RFLP analysis, they will have one band in common and one band that does not match the other's in molecular weight:



paternity Case

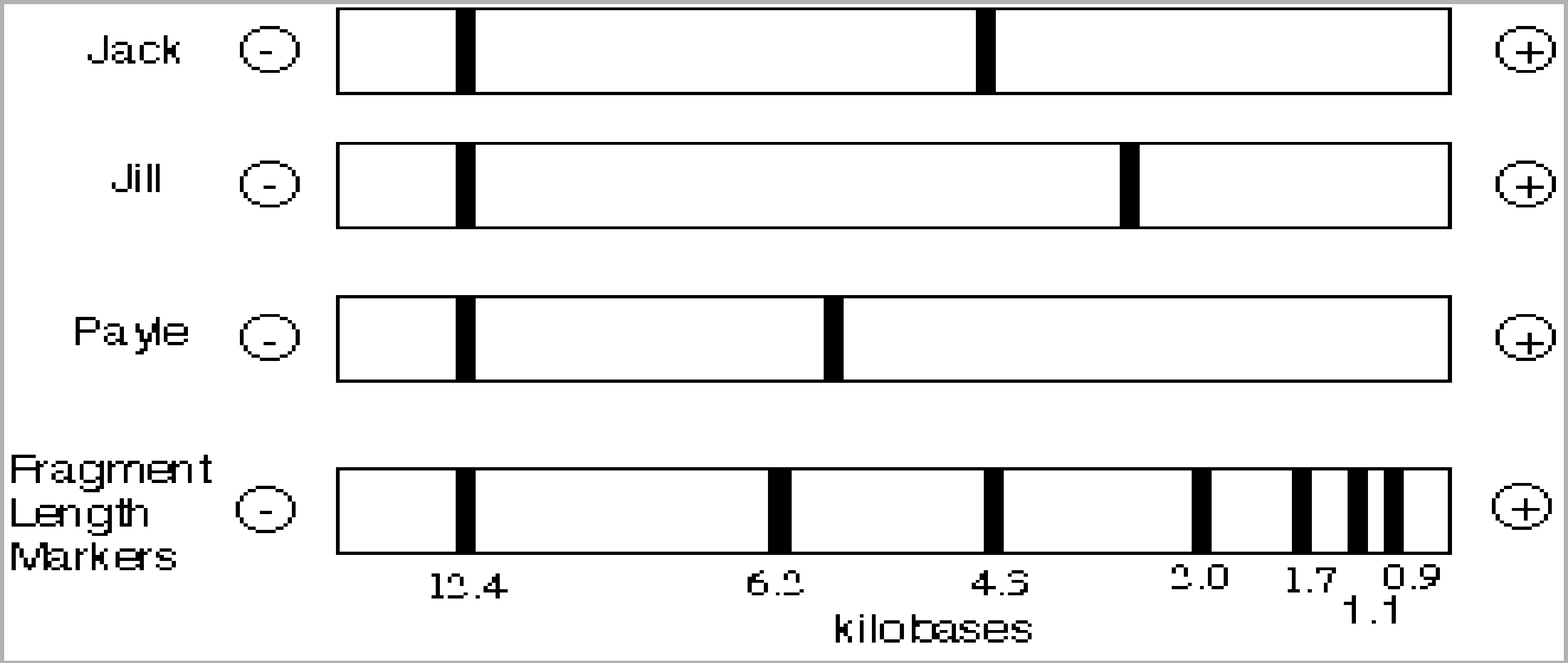
Let's use RFLP technology to determine if Jack is the father of Jill's child named Payle.

In this scenario, DNA was extracted from white blood cells from all three individuals and subjected to RFLP analysis. The results are shown below:



In this case, it appears that Jack could be the father, since Payle inherited the 12.4 kb fragment from Jill and the 4.3 fragment from Jack. However, it is possible that another man with similar RFLP pattern could be as well. <u>To be certain, several more RFLP loci would be tested.</u> It would be highly unlikely that two men (other than identical twins) would share multiple RFLP patterns and so paternity could be confirmed.

In a different scenario, DNA was extracted from white blood cells from all three individuals and subjected to RFLP analysis. The results are shown below:

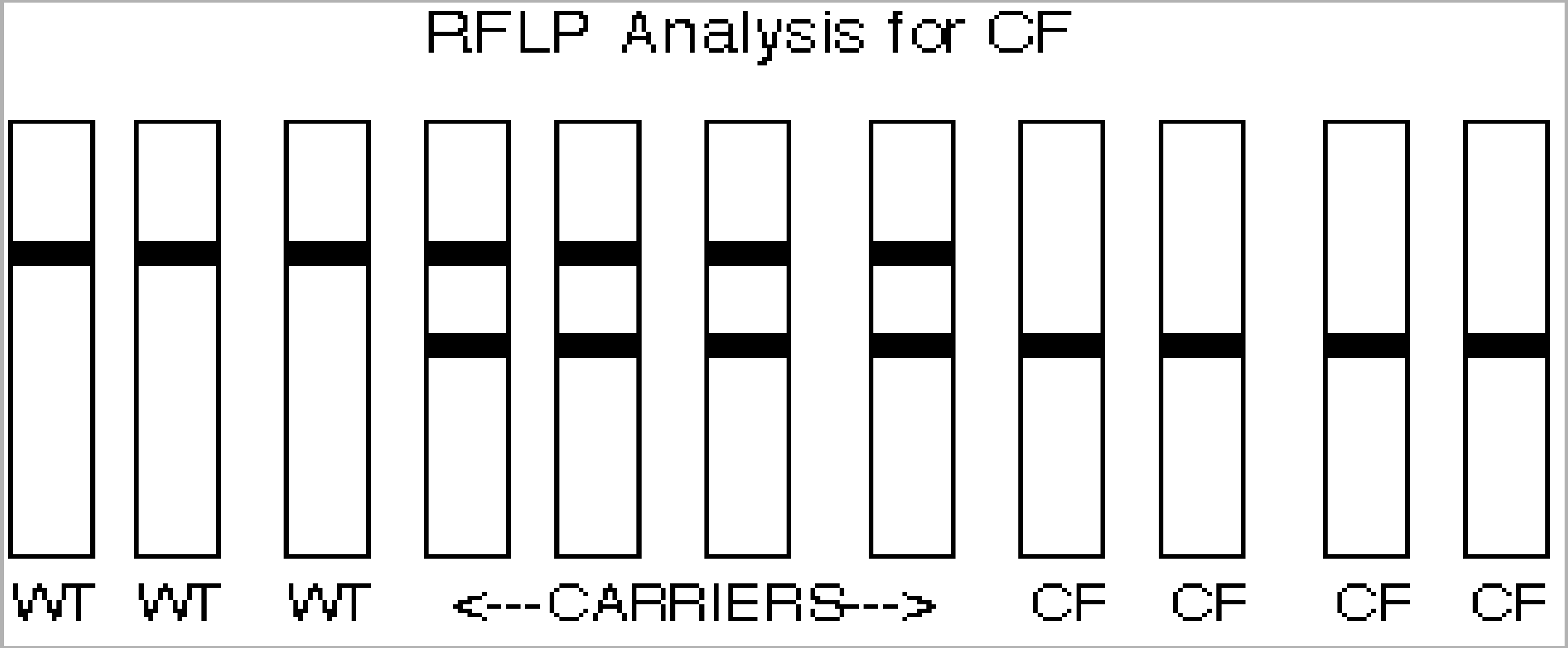


This time, it can be determined that Jack is NOT the father of Payle since Payle has a band of about 6 kb and Jack does not. Therefore, it is very probable that Payle's father is not Jack, though it is possible that Payle carries a new mutation at this locus and a different sized band was produced.

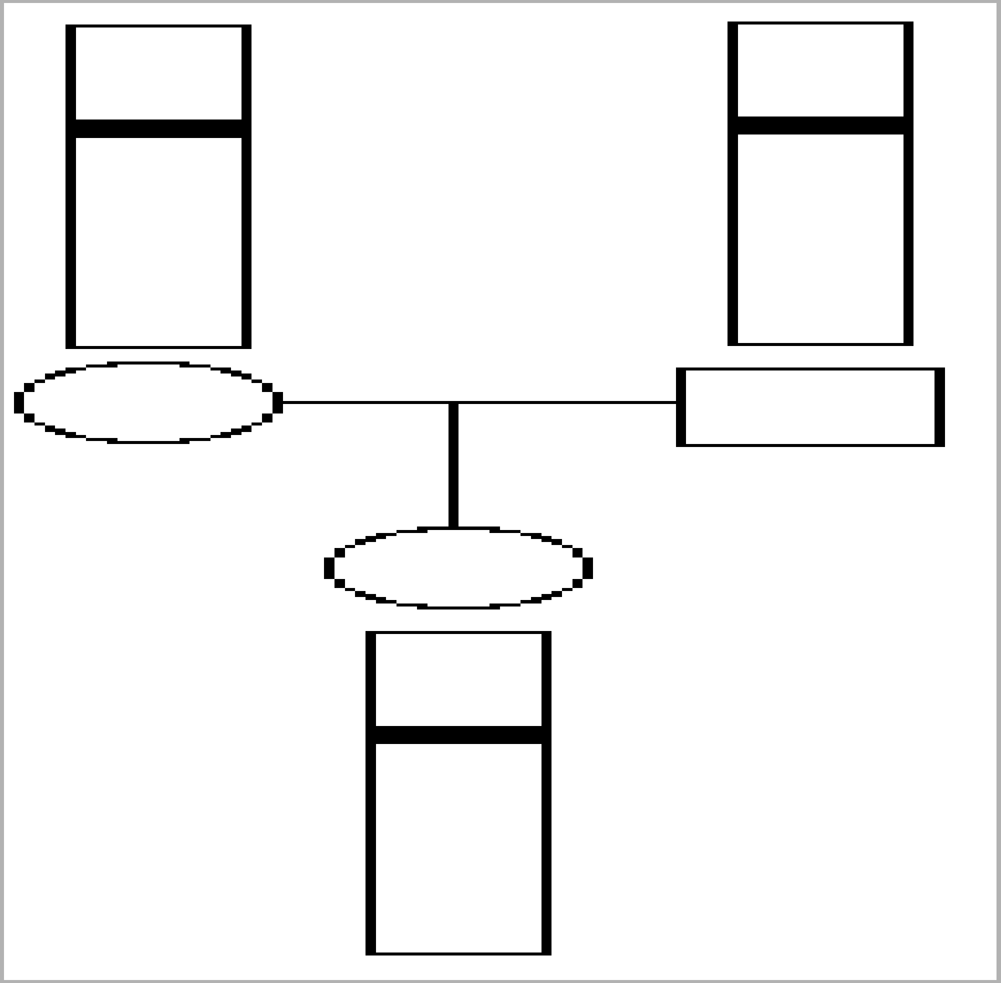
Disease Status

In this example, we want to know if a person carries any cystic fibrosis (CF) alleles and if so, how many. Because CF is a recessive disease, anyonne with CF must be homozygous for disease alleles. From pedigree information, we can often determine who in this family is a carrier. However, if a couple comes to a genetic counselor, often an RFLP analysis is performed on the couple's DNA.

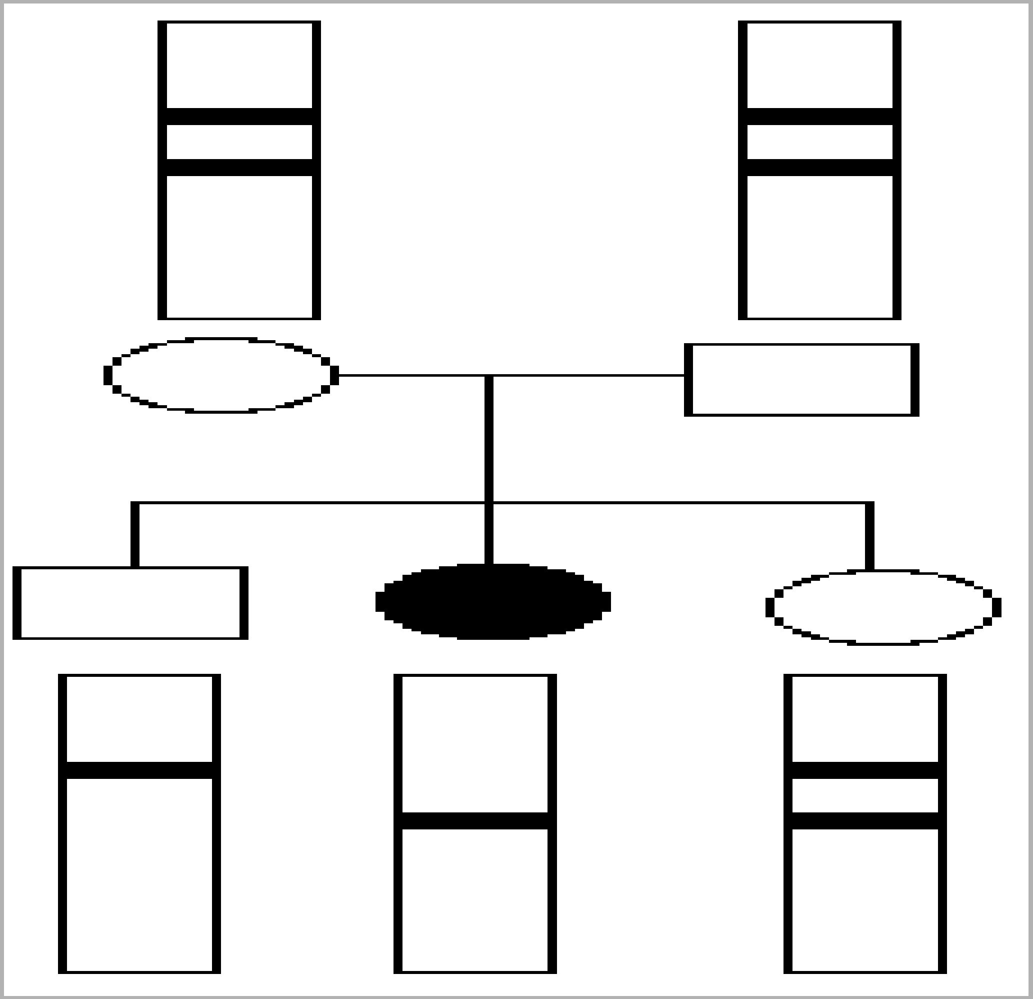
RFLPs are known for CF and so it would be easy to determine if a person were homozygous wild-type (*wt*), heterozygous "carrier", or homozygous disease alleles and thus have CF.



For couples expecting a child, it would be simple to test both parents and make a prediction about the eventual disease status of their fetus. For example, if both parents were homozygous *wt*, then all of their children would also be homozygous *wt*:



However, if both parents were heterozygous, they could have children with any of the three genotypes, though heterozygous children would be twice as likely as either of the homozygous genotypes.

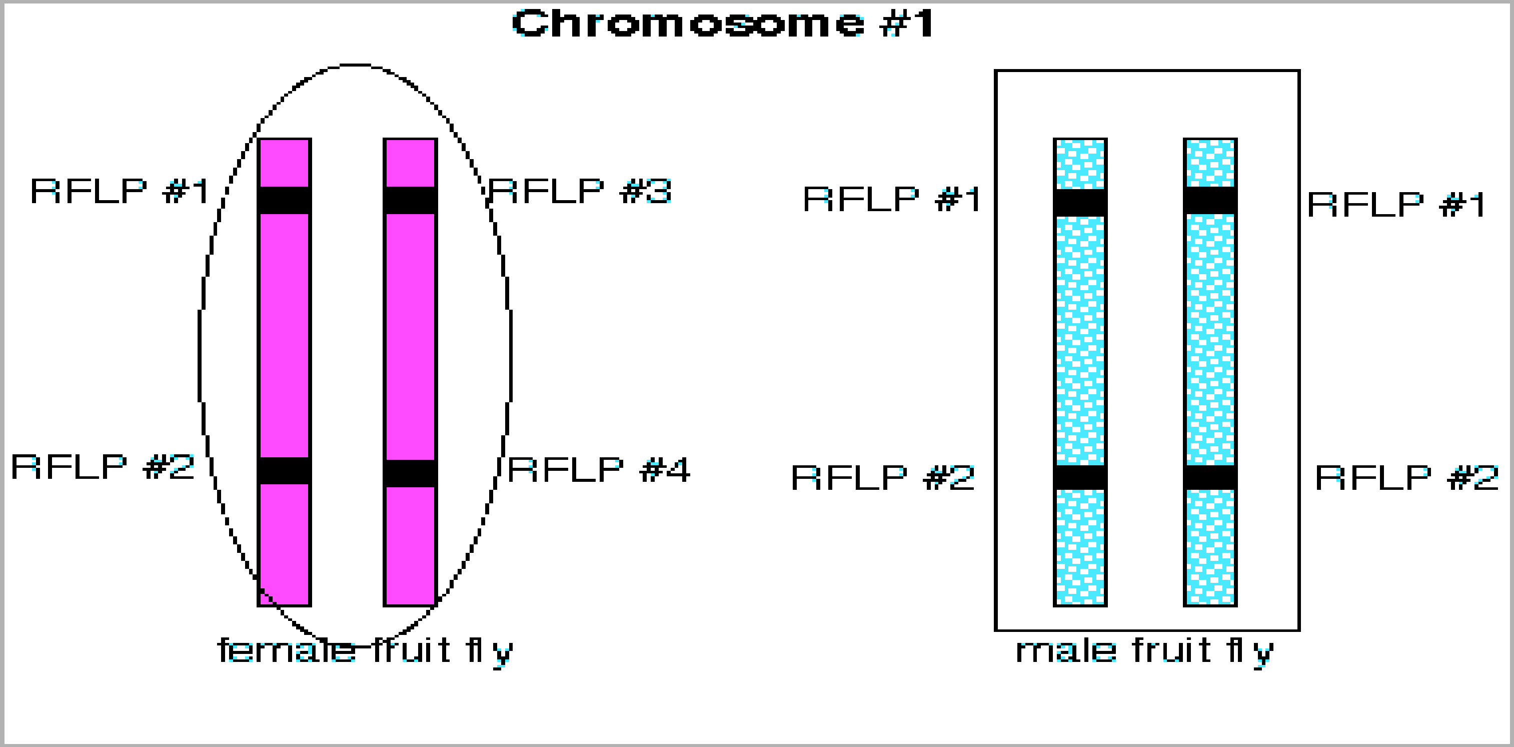


With increasing genomic sequence information, increasing numbers of genetic disease can be predicted from RFLP analyses.

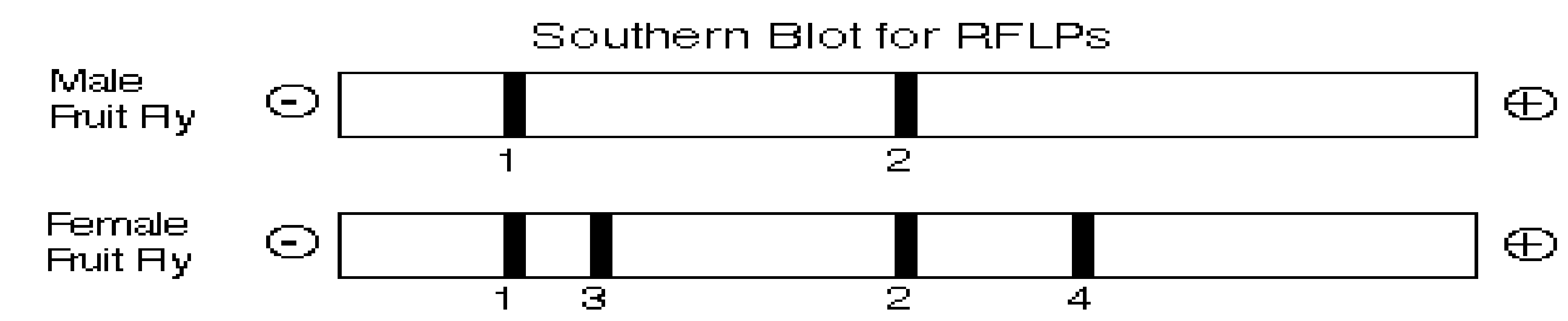
Genetic Mapping

To calculate the genetic distance between to loci, you need to be able to observe recombination. Traditionally, this was performed by observing phenotypes but with RFLP analysis, it is possible to measure the genetic distance between two RFLP loci whether they are a part of genes or not.

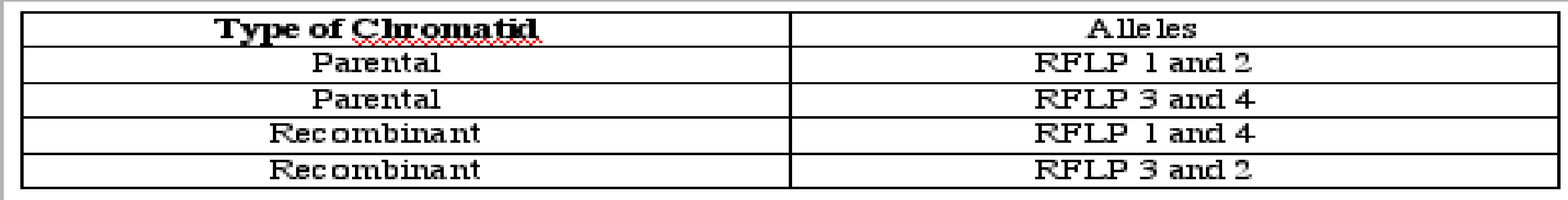
Let's look at a simple example in fruit flies. Two RFLP loci with two RFLP bands possible at each locus:



These loci are located on the same chromosome for the female (left) and the male (right). The upper locus can produce two different bands called 1 and 3. The lower locus can produce bands called 2 or 4. The male is homozygous for band 1 at the upper locus and 2 for the lower locus. The female is heterozygous at both loci. Their RFLP banding patterns can be seen on the Southern blot below:

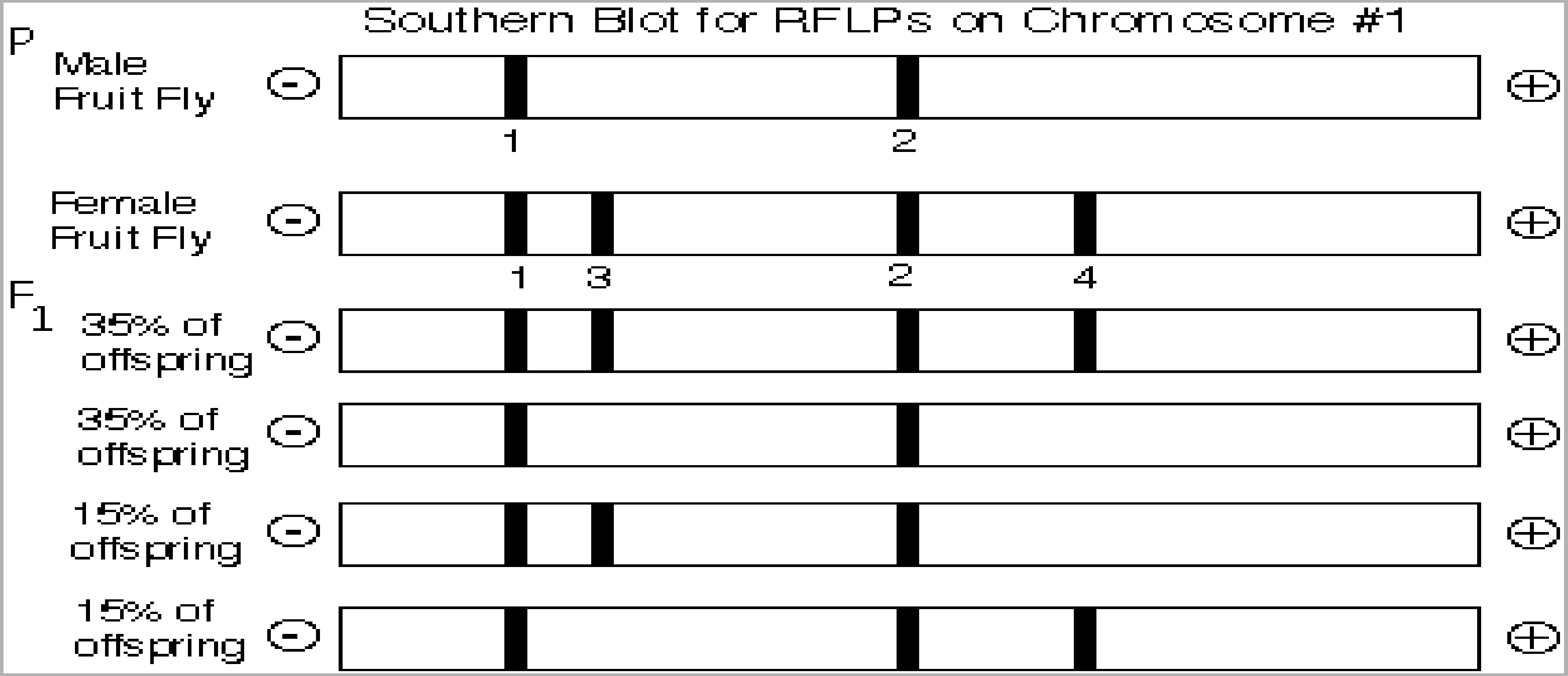


The male can only produce one type of gamete (1 and 2) but the female can produce four different gametes. Two of the possible four are called parental because they carry both RFLP bands from the same chromosome; 1 and 2 from the left chromosome or 3 and 4 from the right chromosome. The other two chromosomes are recombinant because recombination has occurred between the two loci and thus the RFLP bands are mixed so that 1 is now linked to 4 and 3 is linked to 2.



Type of Chromatid	Alleles
Parental	RFLP 1 and 2
Parental	RFLP 3 and 4
Recombinant	RFLP 1 and 4
Recombinant	RFLP 3 and 2

When these two flies mate, the frequency of the four possible progeny can be measured and from this information, the genetic distance between the two RFLP loci (upper and lower) can be determined.



In this example, 70% of the progeny were produce from parental genotype eggs and 30% were produced by recombinant genotype eggs. Therefore, these two RFLP loci are 30 centiMorgans apart from each other.

Southern Blot Method

This is a brief overview of how a Southern blot (more formally called an DNA blot) is performed and what type of data you can obtain.

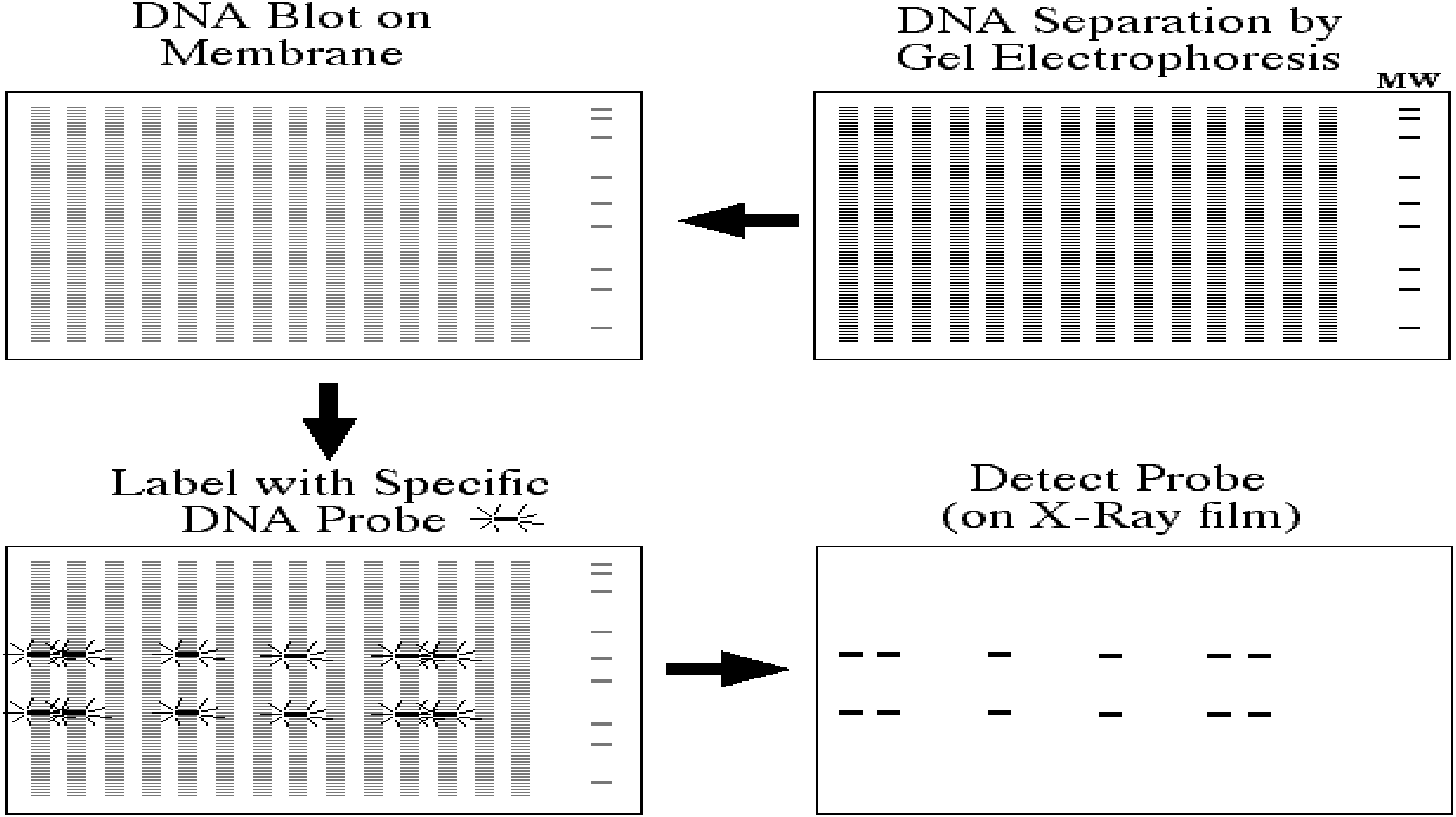

DNA Blot on Membrane
DNA Separation by Gel Electrophoresis
MW
Label with Specific DNA Probe
Detect Probe (on X-Ray film)

Figure 1. Southern blots allow investigators to determine the molecular weight of a restriction fragment and to measure relative amounts in different samples.

Procedure

1) DNA (genomic or other source) is digested with a restriction enzyme and separated by gel electrophoresis, usually an agarose gel. Because there are so many different restriction fragments on the gel, it usually appears as a smear rather than discrete bands. The DNA is denature into single strands by incubation with NaOH.

2) The DNA is transfered to a membrane which is a sheet of special blotting paper. The DNA fragments retain the same pattern of separation they had on the gel.

3) The blot is incubated with many copies of a probe which is single-stranded DNA. This probe will form base pairs with its complementary DNA sequence and bind to form a double-stranded DNA molecule. The probe cannot be seen but it is either radioactive or has an enzyme bound to it (e.g. alkaline phosphatase or horseradish peroxidase).

4) The location of the probe is revealed by incubating it with a colorless substrate that the attached enzyme converts to a colored product that can be seen or gives off light which will expose X-ray film. If the probe was labeled with radioactivity, it can expose X-ray film directly.

Below is an example of Southern blot used to detect the presence of a gene .

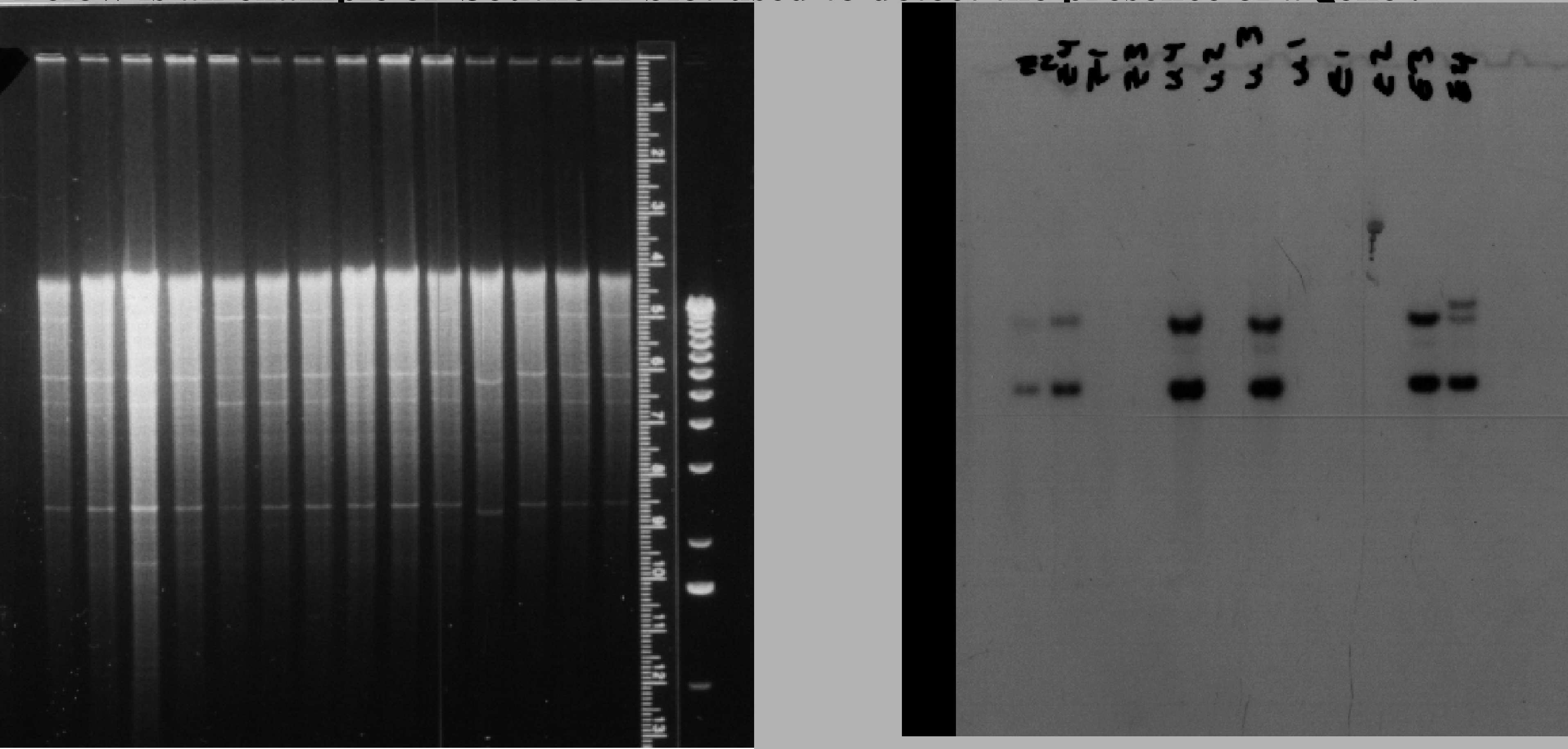
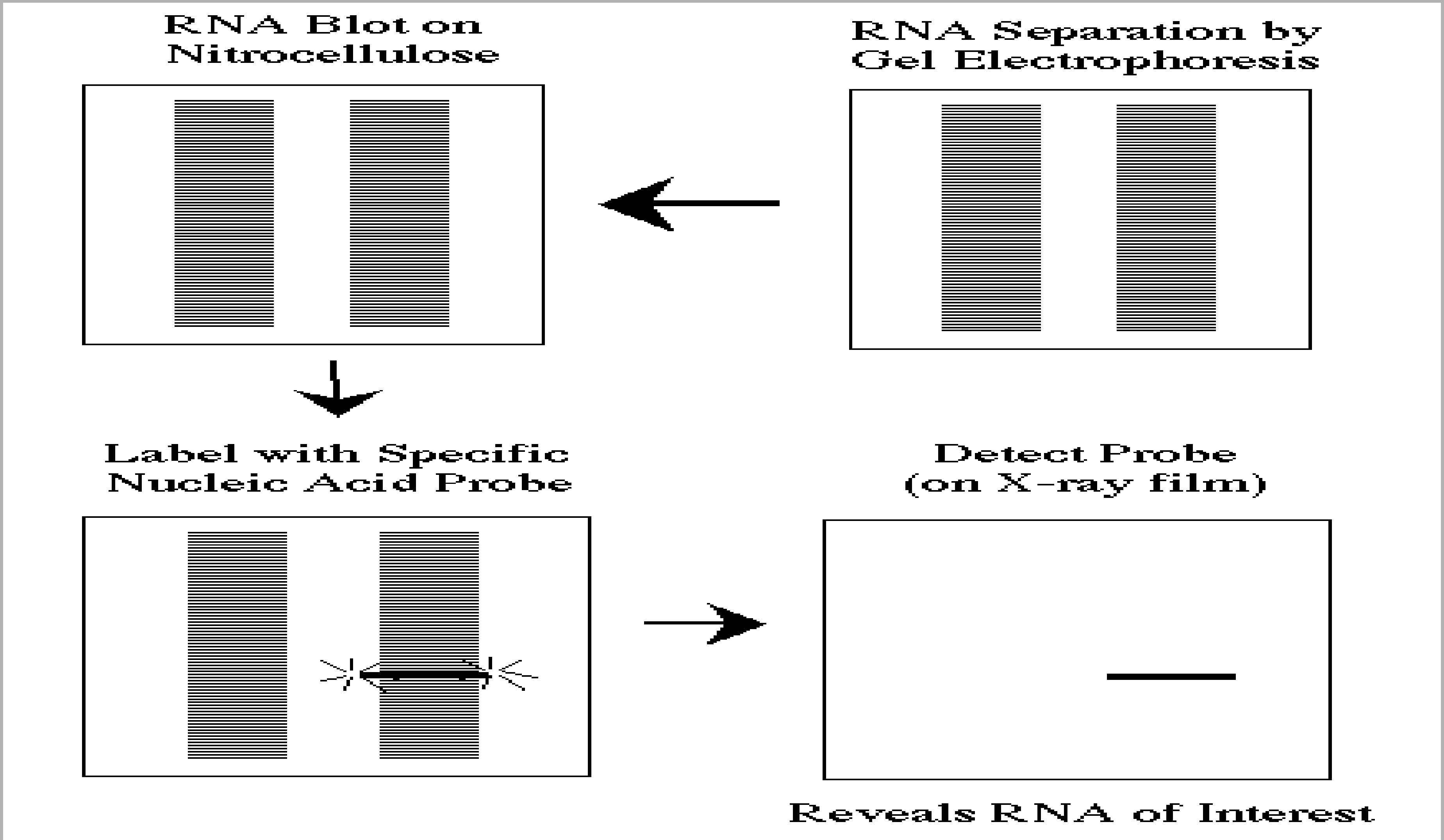


Figure 2. The figure on the left shows a photograph of a 0.7% agarose gel that has 14 different samples loaded on it (plus molecular weight marker in the far right lane). Each sample of DNA has been digested with the same restriction enzyme (EcoRI). Notice that the DNA does not appear as a series of discrete bands but rather as a smear. The DNA was transferred to nitrocellulose and then probed with a radioactive fragment of DNA that was derived from the transformed gene. The figure on the right is a copy of the X-ray film and reveals which strains contain the target DNA and which ones do not.

Northern Blot Procedure

This is a brief overview of how a Northern blot (more formally called an RNA blot) is performed and what type of data you can obtain .


RNA Blot on Nitrocellulose
RNA Separation by Gel Electrophoresis
Label with Specific Nucleic Acid Probe
Detect Probe (on X-ray film)
Reveals RNA of Interest

Northern blots allow investigators to determine the molecular weight of an mRNA and to measure relative amounts of the mRNA present in different samples.

1) RNA (either total RNA or just mRNA) is separated by gel electrophoresis, usually an agarose gel. Because there are so many different RNA molecules on the gel, it usually appears as a smear rather than discrete bands.

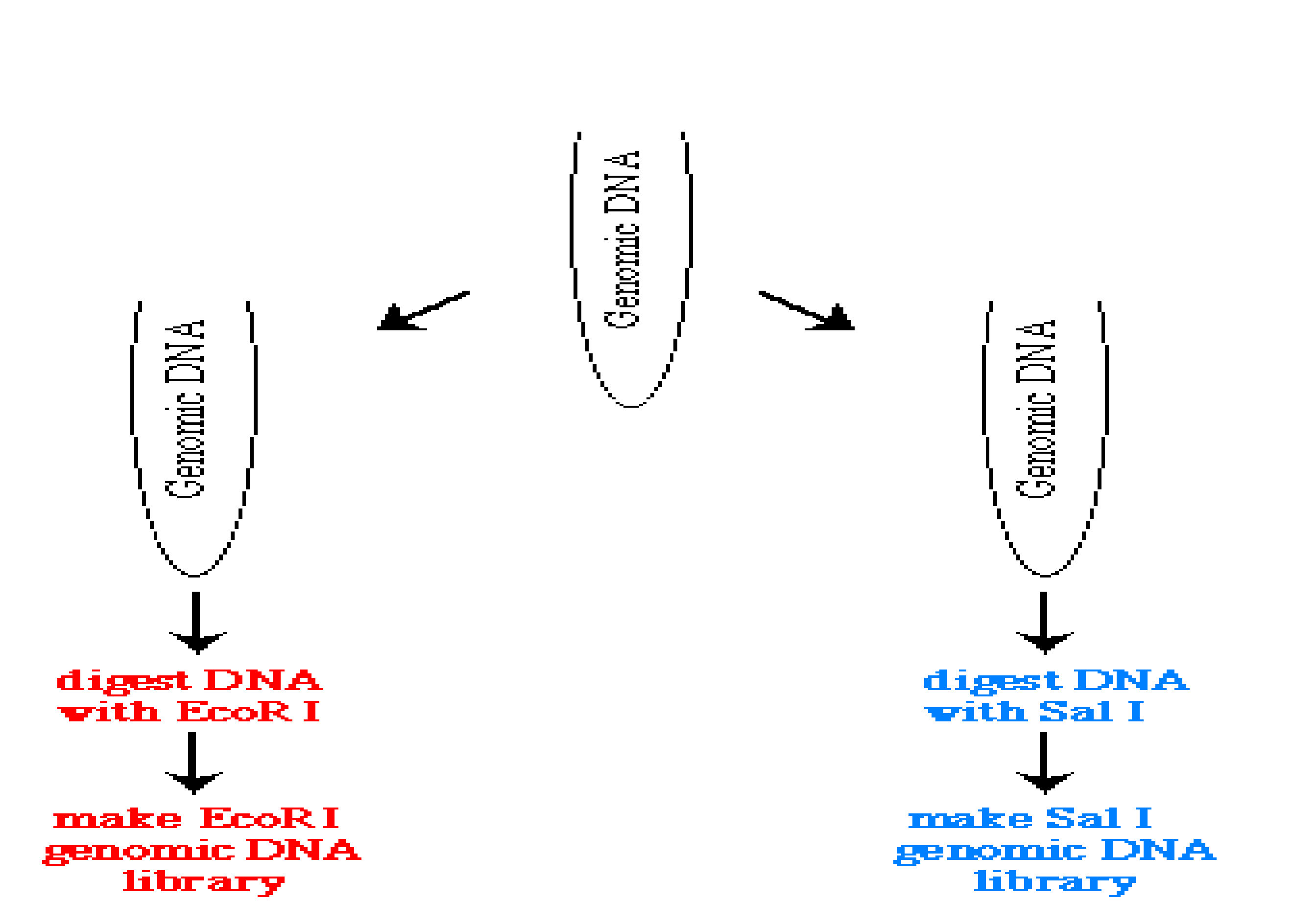
2) The RNA is transfered to a sheet of special blotting paper called nitrocellulose, though other types of paper, or membranes, can be used. The RNA molecules retain the same pattern of separation they had on the gel.

3) The blot is incubated with a probe which is single-stranded DNA. This probe will form base pairs with its complementary RNA sequence and bind to form a double-stranded RNA-DNA molecule. The probe cannot be seen but it is either radioactive or has an enzyme bound to it (e.g. alkaline phosphatase or horseradish peroxidase).

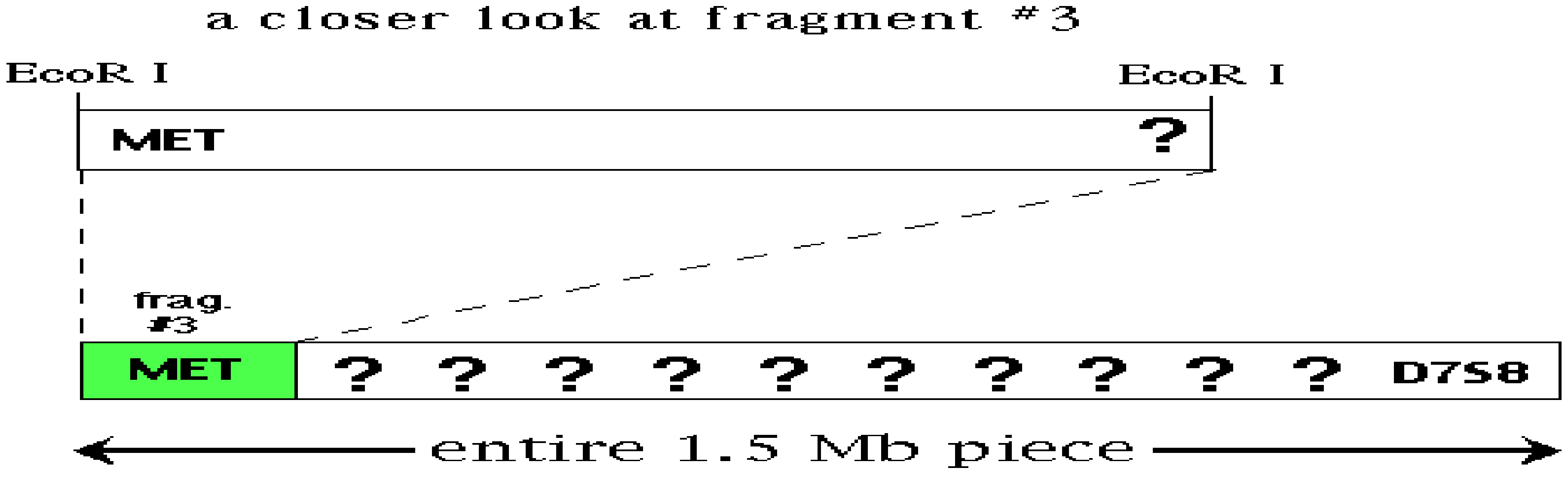
4) The location of the probe is revealed by incubating it with a colorless substrate that the attached enzyme converts to a colored product that can be seen or gives off light which will expose X-ray film. If the probe was labeled with radioactivity, it can expose X-ray film directly.

Chromosomal Walking to Clone the Cystic Fibrosis Gene

The first thing you need to do is create two genomic libraries of the same DNA but each library used a different restriction enzyme, such as EcoR I and Sal I.



You would screen a genomic library with this probe and isolate a piece of DNA that binds to your probe. If you cloned an EcoR I restriction fragment from the genomic library (let's say it is fragment #3 of many possible fragments) that binds to a particular probe (called MET) that is linked to CF. You want to slide down the chromosome from MET towards CF so you can clone and sequence CF. D7S8 is another RFLP marker located on the other side of CF so CF is located between MET and D7S8.



To clone CF, you will employ **chromosomal walking** to take baby steps towards CF, starting with the EcoR I restriction fragment #3 you just cloned. Now, you need to generate a restriction map of EcoR I fragment #3. You must **digest the Eco RI restriction fragment with multiple** restriction enzymes and analyse the results on an agarose gel as shown here:

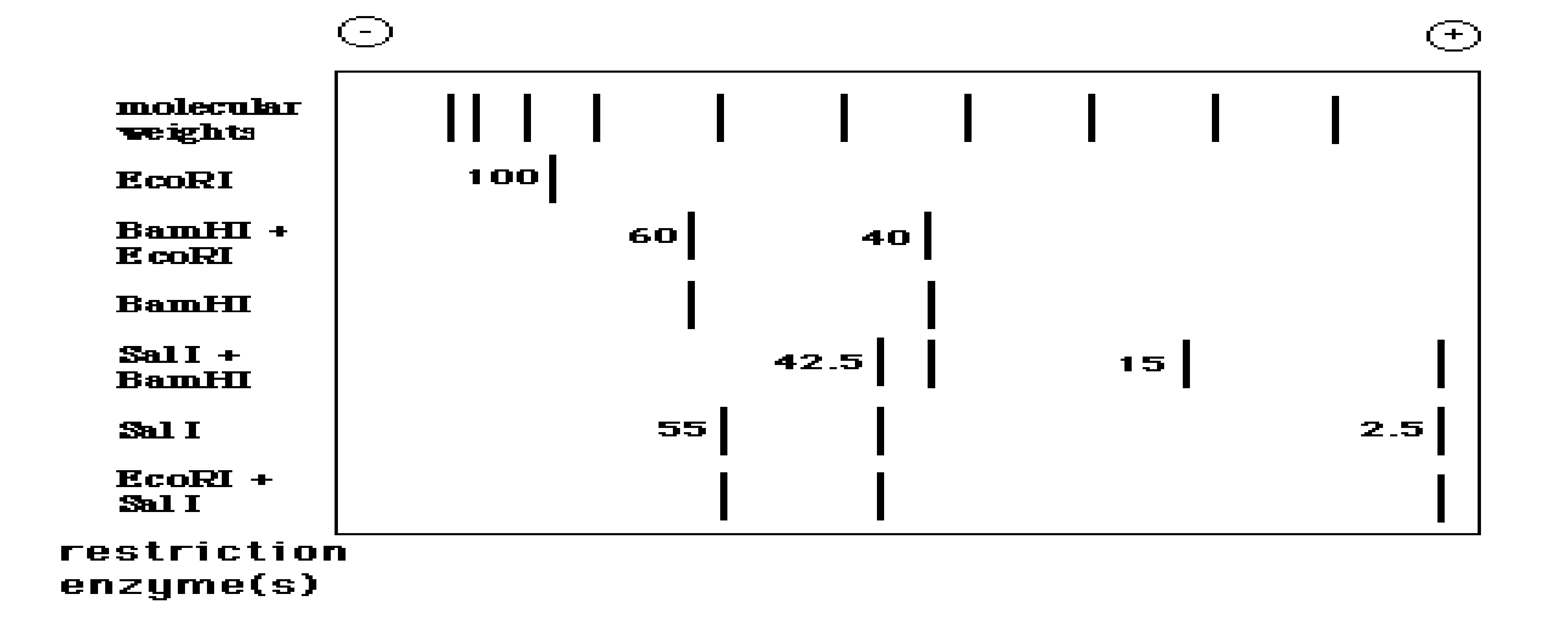


Figure 3. Cartoon of an agarose gel that contains the restricted DNA as described above.

If you then performed a Southern blot with this gel and used the original **MET probe that allowed you to isolate the EcoR I restricion fragment,** you might see the following resutls on an X-ray film:

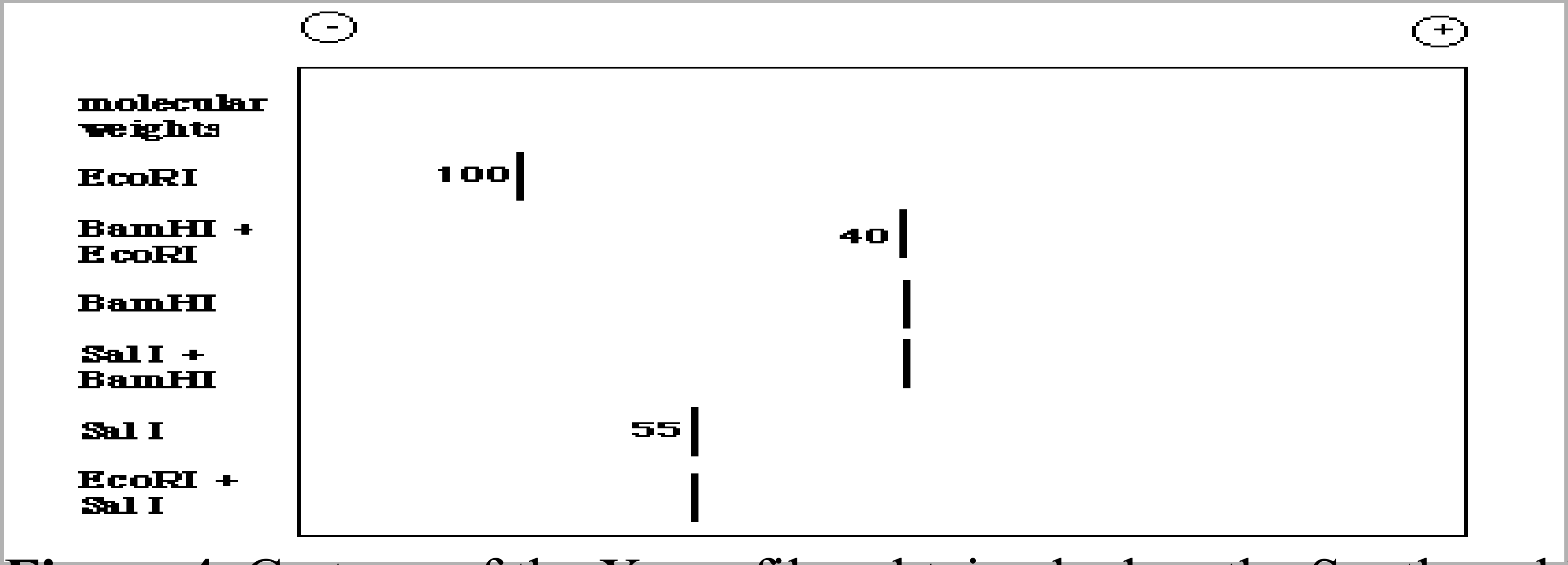


Figure 4. Cartoon of the X-ray film obtained when the Southern blot is probed with MET.

From these data, you could construct the following restriction map that also indicates where the probe binds (note: two slightly different restriction maps could be generated from these data but this one is fine for our purposes):

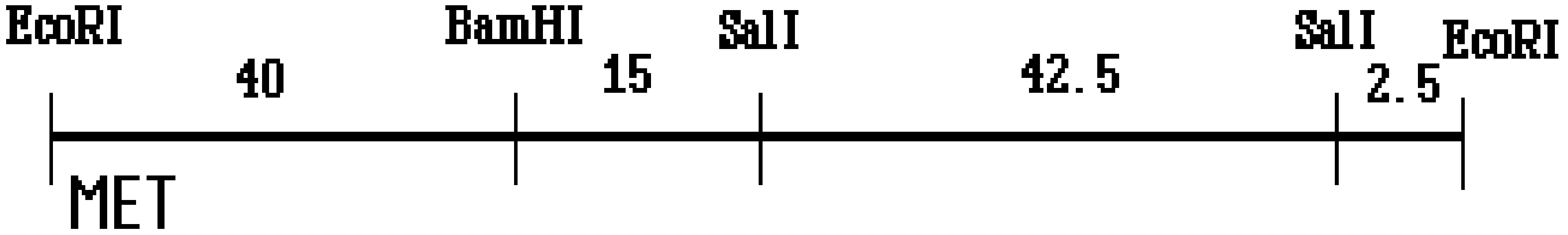


Figure 5. One possible restriction map, given the data from figures 1 and 2 above.

The next task at hand is to isolate the 2.5 kb Sal I - EcoR I fragment and use it as your second probe on the Sal I genomic library because this 2.5 kb piece is the DNA fragment furthest from the MET marker and therefore must be closer to CF. You are ready to screen the Sal I genomic library that used identical DNA but was digested with the restriction enzyme Sal I instead of EcoR I. Because probe #2 is flanked on the left by a Sal I site, you know any new fragment that has Sal sites on both ends and binds to the second probe will extend towards the right (in the direction of CF) as shown:

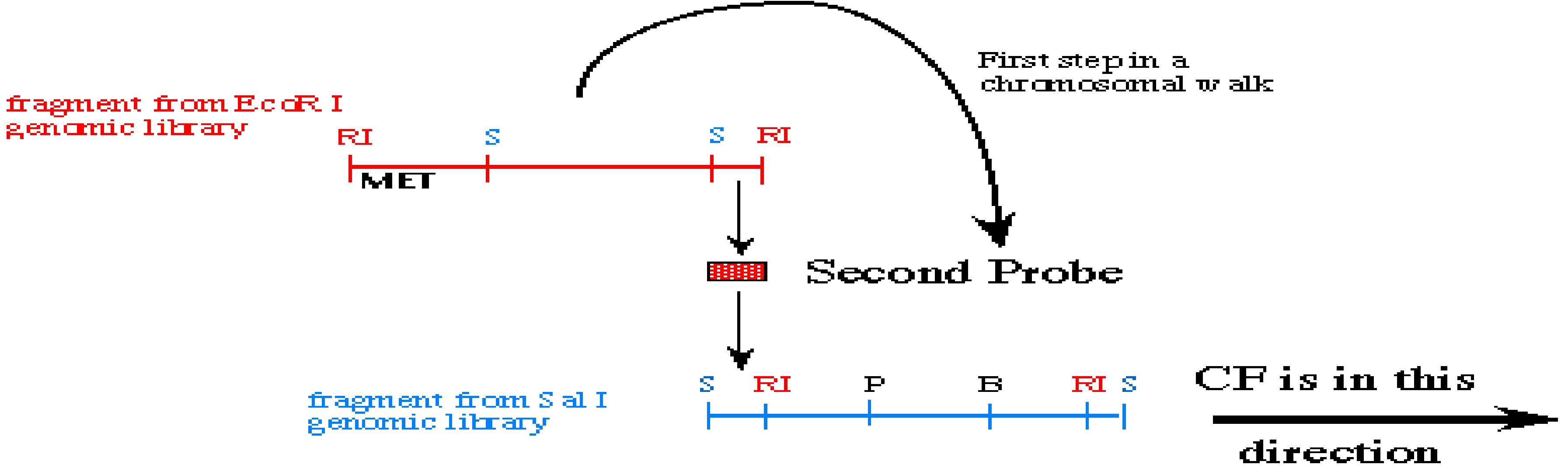

First step in a chromosomal walk
fragment from EcoR I genomic library
RI
S
S
RI
MET
Second Probe
fragment from Sal I genomic library
S
RI
P
B
RI
S
CF is in this direction

Figure 6. This cartoon illustrates how you use the first genomic DNA clone to generate a second probe that takes one step in the direction of CF.

When you have cloned a Sal I fragment that binds to probe #2, you need to figure out its restriction map the same way we did for the EcoRI fragment #3 above. This process continues until you reach D7S8. The final product for a chromosomal walk is a series of overlapping restriction maps starting at your original probe (MET) and extending to D7S8. The final combined restriction map, and the overlapping fragments, might look like this:

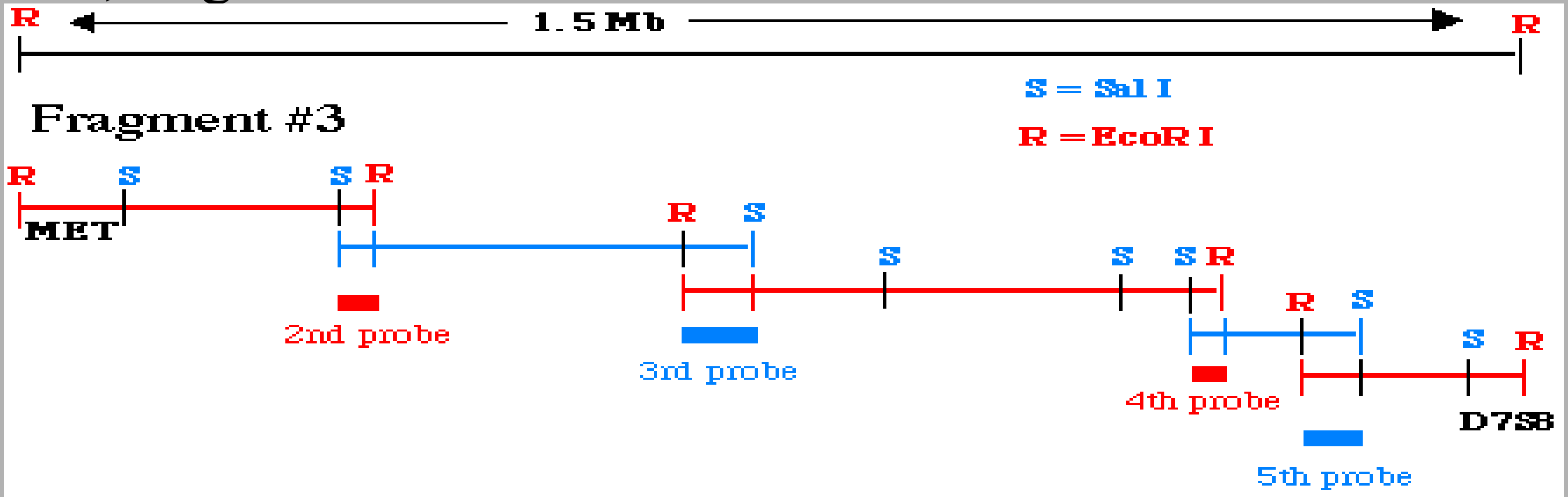


Figure 7. This cartoon illustrates how a series of overlapping pieces of genomic DNA has been isolated from alternate genomic libraries.

This is a simplistic example involving only 5 steps for the walk. Now that you have a restriction map spanning the area of interest, you can use a number of different methods to determine which fragment contains the CF gene. Once you know this, it is very simple to sequence the CF gene and continue your analysis.

Karyotypes

Karyotypes are images of chromosomes to display their banding patterns. When a nucleus is in during metaphase of mitosis, its chromosomes are condensed and the banding of the chromosomes can be visualized when certain dyes (e.g. Giemsa dye) are added to the chromosomes.

For those of us who are unaccustomed to seeing real chromosomes, often they are drawn in a cartoon fashion called an ideogram. Below is an ideogram of the X chromosome. The short arm of any chromosome is called the "p" arm which stands for the French word for small - *petite*. The long arm is called the "q" arm.

Many years ago, histologist numbered the bands for each arm so we can refer to particular bands as genomic locations and everyone will be looking at the same band. The two telomeres are refered to as "ter" for termini.

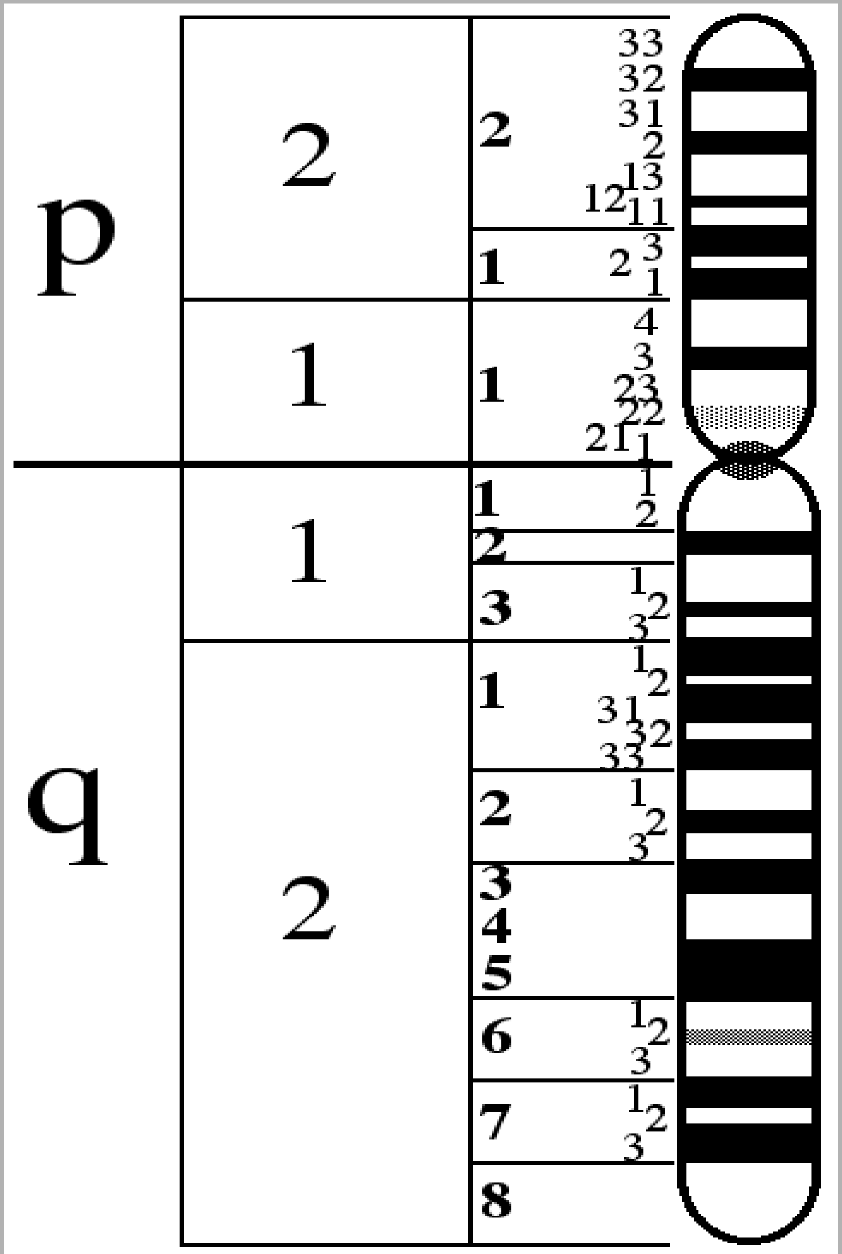


Figure 1. Ideogram of a human X chromosome.

FISH

Principle

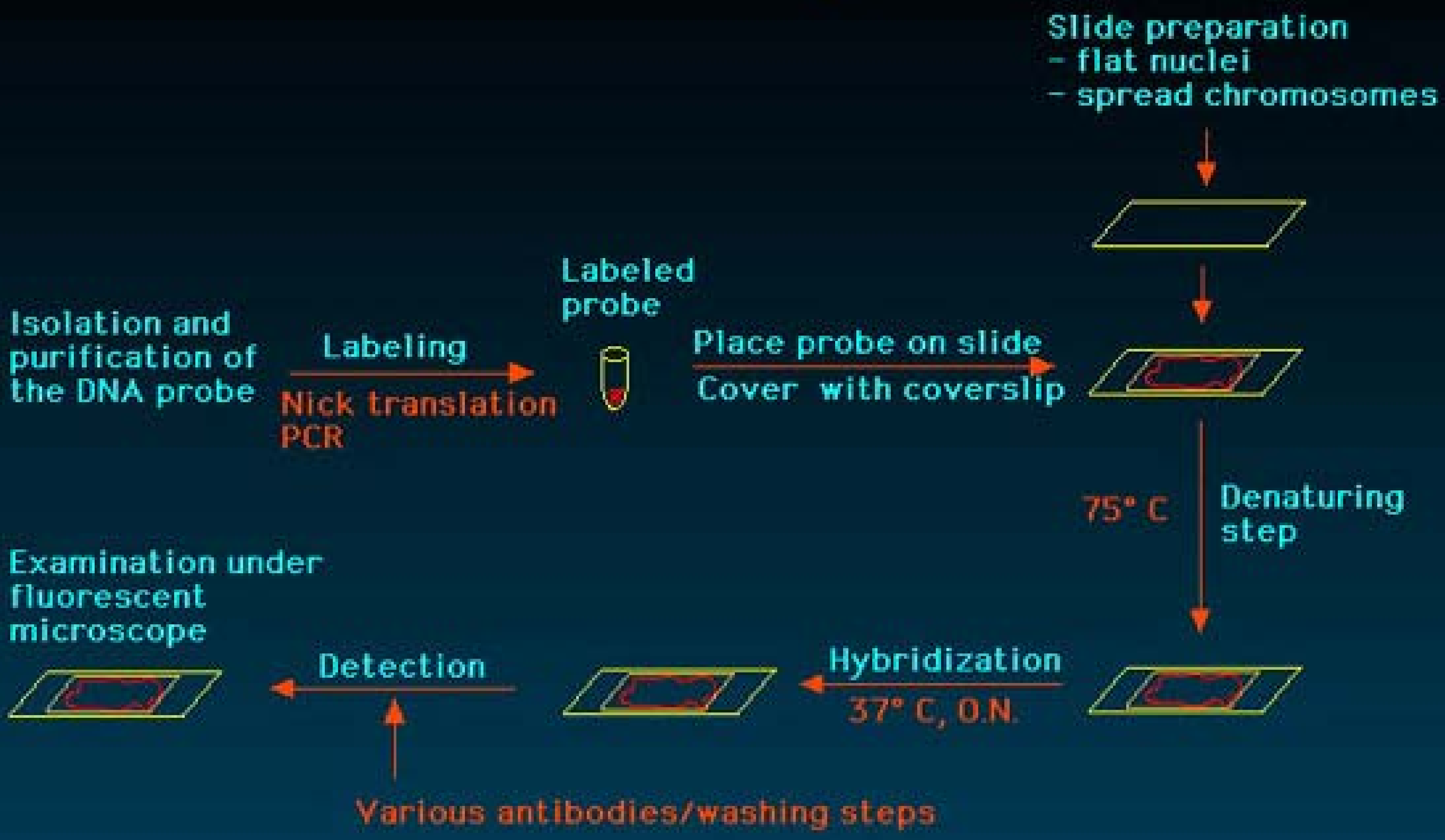
In **fluorescence in situ hybridization (FISH), a DNA probe is labeled with a fluorescent dye or a haptene (usually in the form of fluor-dUTP or haptene-dUTP, that is incorporated into the DNA using enzymatic reactions, such as nick translation or PCR).**

The labeled DNA is purified, concentrated, resuspended in hybridization buffer (containing formamide) and is hybridized onto chromosomes and nuclei on slides (cytogenetic preparations).

After overnight hybridization, the slides are rinsed in washing solutions and, if needed, one or several layers of fluorescent-labeled antibodies are added to detect haptene-labeled DNA. The slides is mounted with antifade solution and is visualized at the fluorescent microscope, using appropriate filters

This procedure is illustrated in the figure below.

General FISH protocol



Slide preparation
- flat nuclei
- spread chromosomes

Isolation and purification of the DNA probe

Labeling

Nick translation
PCR

Labeled probe

Place probe on slide

Cover with coverslip

75° C

Denaturing step

Hybridization

37° C, O.N.

Detection

Various antibodies/washing steps

Examination under fluorescent microscope

Cell type considerations (for FISH)

Due to the ease in culturing and to the high metaphase index, lymphocytes from peripheral blood cultures are preferred for use in most laboratories.

•A 30-40 minutes colcemid treatment of 48-72 hour cultures can yield a high metaphase index, with remarkably similar chromosome lengths.

•Fibroblasts can also be easily used for general FISH purposes (probe mapping, interphase FISH). The metaphase index is, in general, lower than for fresh lymphocyte preparations. There is a larger variability in chromosome sizes among metaphases of the same pellet, due primarily to longer colcemid times. 4-8 hours colcemid treatment on fibroblast cultures can significantly increase the metaphase index, but many of the metaphases will have very short chromosomes. In general, the chromosomes of fibroblast cultures overspread very easily on the slide, indicating an increased fragility of the cell membranes with usual fixative treatment.

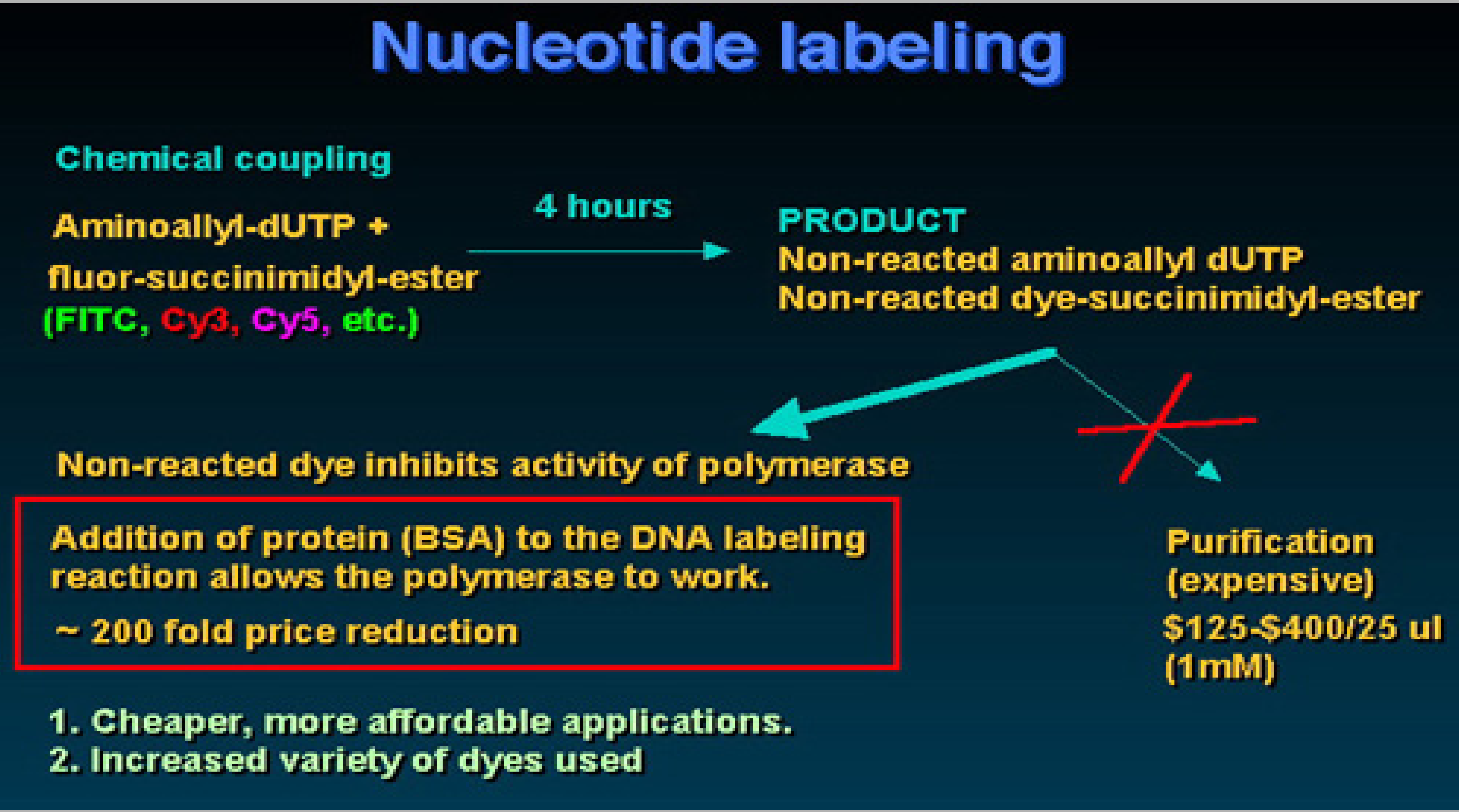
• Lymphoblastoid and tumor cell lines are more difficult to use in FISH analyses, as the metaphase index is lower and the shape/spread of the chromosomes is more difficult to control.

Long-term storage

For good FISH results on older cytogenetic slides, the slides should not be stored dry at room temperature. Instead, slides should be stored either in 100% ethanol at -20° C, or should be placed in a plastic box wrapped up in Saran Wrap, and stored at -20° or -80° C. With any of these procedures, slides can be stored for years and can be used in FISH without major problems

Briefly, reactive aminoallyl-dUTP and succinimidyl-ester derivatives of many dyes can be chemically coupled in a simple reaction, in the presence of bicarbonate (alkaline pH). After the reaction, the nucleotides can be used directly for enzymatic DNA labeling reactions, if BSA (bovine serum albumin) is added to the mixture (probably to block free radicals). After labeling, the DNA MUST be purified, in order to remove especially the residual free dye. Using custom-made fluor-dUTPs decreases significantly the costs of the FISH analysis.

A schematic description of the process is depicted in the **figure** below.



Nucleotide labeling

Chemical coupling

Aminoallyl-dUTP +
fluor-succinimidyl-ester
(FITC, Cy3, Cy5, etc.)

4 hours

PRODUCT
Non-reacted aminoallyl dUTP
Non-reacted dye-succinimidyl-ester

Non-reacted dye inhibits activity of polymerase

Addition of protein (BSA) to the DNA labeling reaction allows the polymerase to work.

~ 200 fold price reduction

Purification
(expensive)
$125-$400/25 ul
(1mM)

1. Cheaper, more affordable applications.
2. Increased variety of dyes used

All antibodies used were IgG molecules raised against other whole IgG molecules. Detection schemes were chosen carefully, to prevent unwanted antibody-antibody interactions

Table 1. DNA probe labeling procedures.

NICK TRANSLATION	***PCR LABELING***
•1 ug DNA (20 ng/µl final conc.) •5 µl (10x) nick translation buffer * •5 µl (100mM=10x)-mercaptoethanol •2.5µl (1mM=20x) each dACG •5µl (0.325/0.175 mM=10x) dTTP/dUTP •5µl (1µg/µl=10x) DNase (Sigma), •1µl (10U/µl) E. Coli polymerase •ultrapure sterilized water to 50 µl.	•Template DNA (various amounts) •2.5-3.5µl (10x) PCR buffer •0.15µl (33.3mM) each dACG •0.7µl (5mM) dTTP •1.6µl (1mM) dUTP •0.1-1µl (50µM) primer(s) •0.2-0.4µl (5U/µl) Taq polymerase •ultrapure sterilized water to 25 µl

Table 2. Commercially-labeled nucleotides used.

FLUORESCENT labeled dUTP	***HAPTENE labeled dUTP***
AMCA-6-dUTP	Biotin(BIO)-11-dUTP
CascadeBlue-4-dUTP	Digoxygenin(DIG)-11-dUTP
Fluorescein-12-dUTP	Dinitrophenyl (DNP)-11-dUTP
Rhodamine-6-dUTP	
TexasRed-6-dUTP	
Cy3-6-dUTP	
Cy5-dUTP	

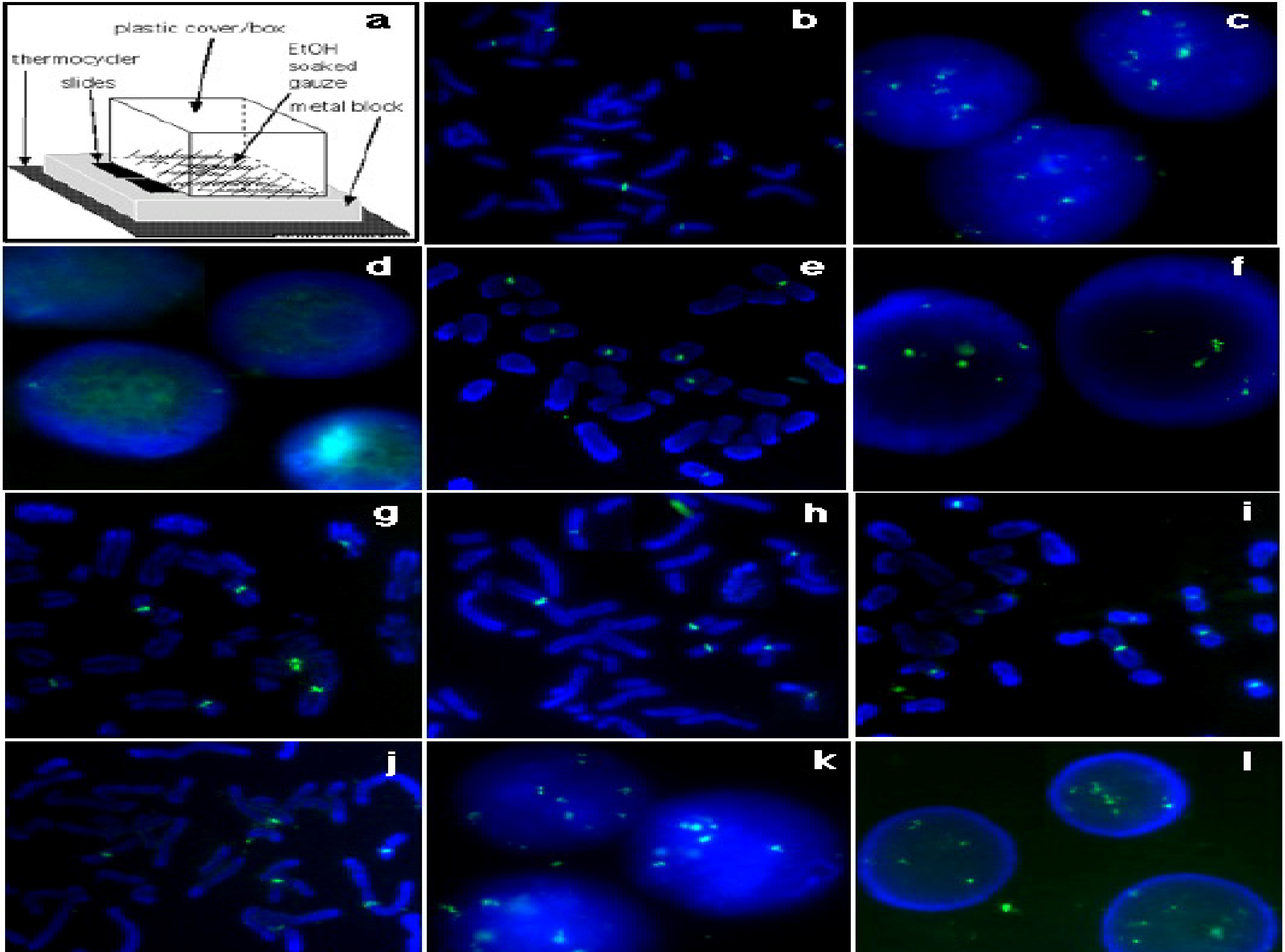

a
plastic cover/box
thermocycler
slides
EtOH
soaked
gauze
metal block
b
c
d
e
f
g
h
i
j
k
l

Fig.1. a) Chemical aging setting, using the metal block of a thermocycler: slides are placed on the metal block, 150-200ul ethanol pipetted onto them and covered with coverslip. A plastic cover, containing gauze or paper soaked in ethanol is placed so as to cover the slides and maintain an ethanol-saturated atmosphere during the heating cycle. b-f) Non-aged slides. Slides were kept 1-2 hours at RT, then subjected to pretreatment in: pepsin (c,b), 2xSSC (d) and trypsin (e,f). A BIO-labeled commercial chromosome 1 centromere was hybridized onto slides for 30 minutes, then detected with avidin-FITC. g-l) Aged slides: pepsin vs. trypsin pretreatment. Same chromosome 1 centromere was used in all six images. (g,h) Examples of pepsin pretreatment onto non-aged (g) and chemically aged (h) slides; (i,j) trypsin pretreatment on non-aged (i) and chemically aged (j) slides; k,l) nuclei hybridizations on chemically aged slides pretreated with pepsin (k) and trypsin (l).

Fig 1 results indicate that, chemically aged chromosomes and nuclei preserve better their architecture compared to non-aged slides, without losing in hybridization quality.
Pepsin pretreatment preserves better the architecture of nuclei and chromosomes than trypsin.
Hybridization was more efficient on pretreated compared to non-pretreated slides.
All posthybridization washes were done at 42 C.

Sequencing Whole Genomes

Hierarchical Shotgun Sequencing v. Shotgun Sequencing

How do you sequence a whole genome?
There are two general strategies for sequencing a complete genome. The method preferred by the Human Genome Project is the **hierarchical shotgun sequencing** method.

In this approach, genomic DNA is cut into pieces of about 150 Kb and inserted into BAC vectors, transformed into *E. coli* where they are replicated and stored. The BAC inserts are isolated and mapped to determine the order of each cloned 150 Kb fragment. Each BAC fragment in the Golden Path is fragmented randomly into smaller pieces and each piece is cloned into a plasmid and sequenced on both strands.
These sequences are aligned so that identical sequences are overlapping. These contiguous pieces are then assembled into finished sequence once each strand has been sequenced about 4 times to produce 8X coverage of high quality data.

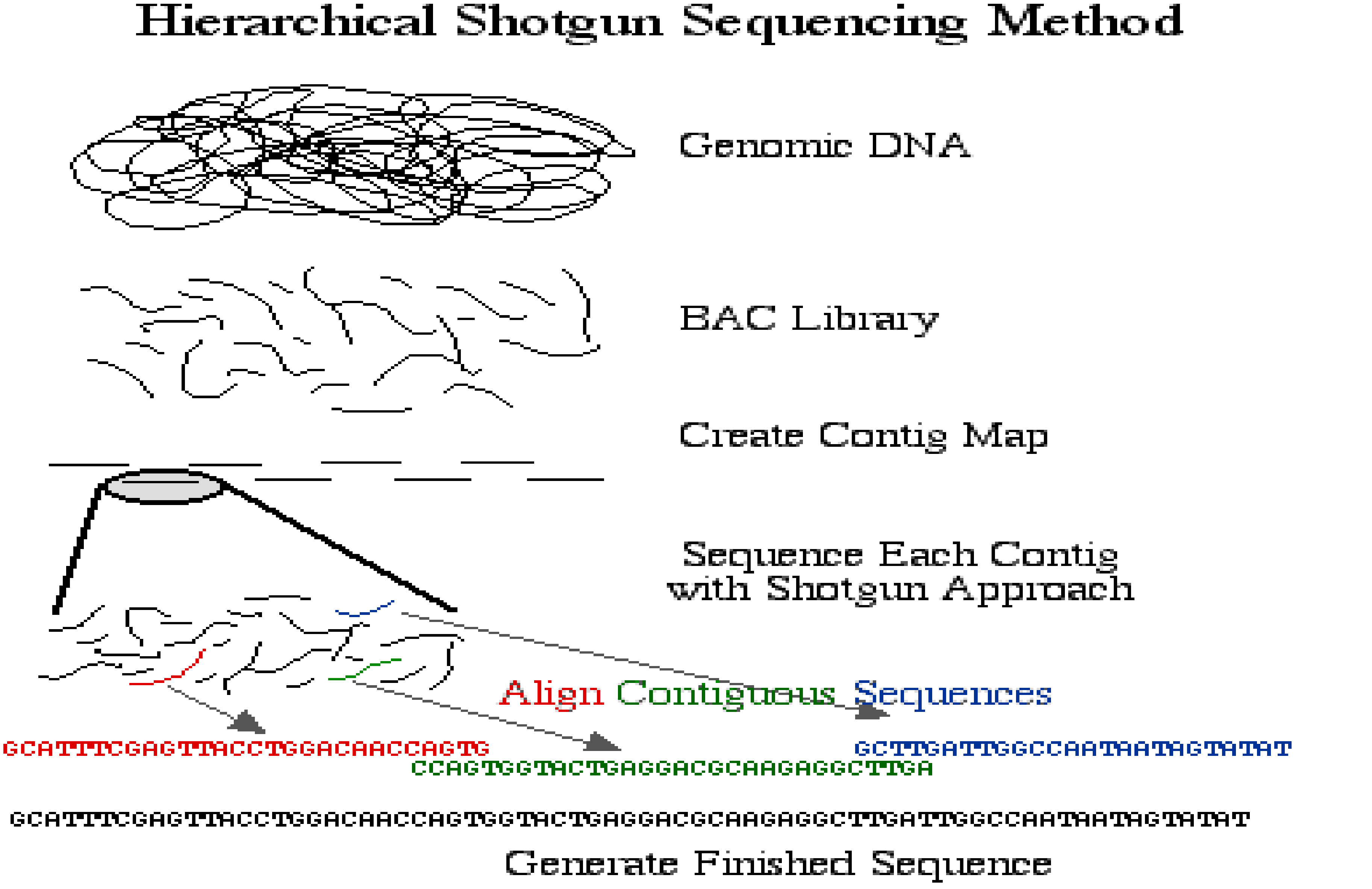

Hierarchical Shotgun Sequencing Method
Genomic DNA
BAC Library
Create Contig Map
Sequence Each Contig with Shotgun Approach
Align Contiguous Sequences
GCATTTCGAGTTACCTGGACAACCAGTG
CCAGTGGTACTGAGGACGCAAGAGGCTTGA
GCTTGATTGGCCAATAATAGTATAT
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT
Generate Finished Sequence

Figure 1. Schematic diagram of sequencing strategy used by the publicly funded Human Genome Project. The DNA was cut into 150 kb fragments and arranged into overlapping contiguous fragments. These contigs were cut into smaller pieces and sequenced completely.

The method developed and preferred by Celera is simply called shotgun sequencing. This approach was developed and perfected on prokaryotic genomes which are smaller in size and contain less repetitive DNA. Shotgun sequencing randomly shears genomic DNA into small pieces which are cloned into plasmids and sequenced on both strands, thus eliminating the BAC step from the HGP's approach. Once the sequences are obtained, they are aligned and assembled into finished sequence.

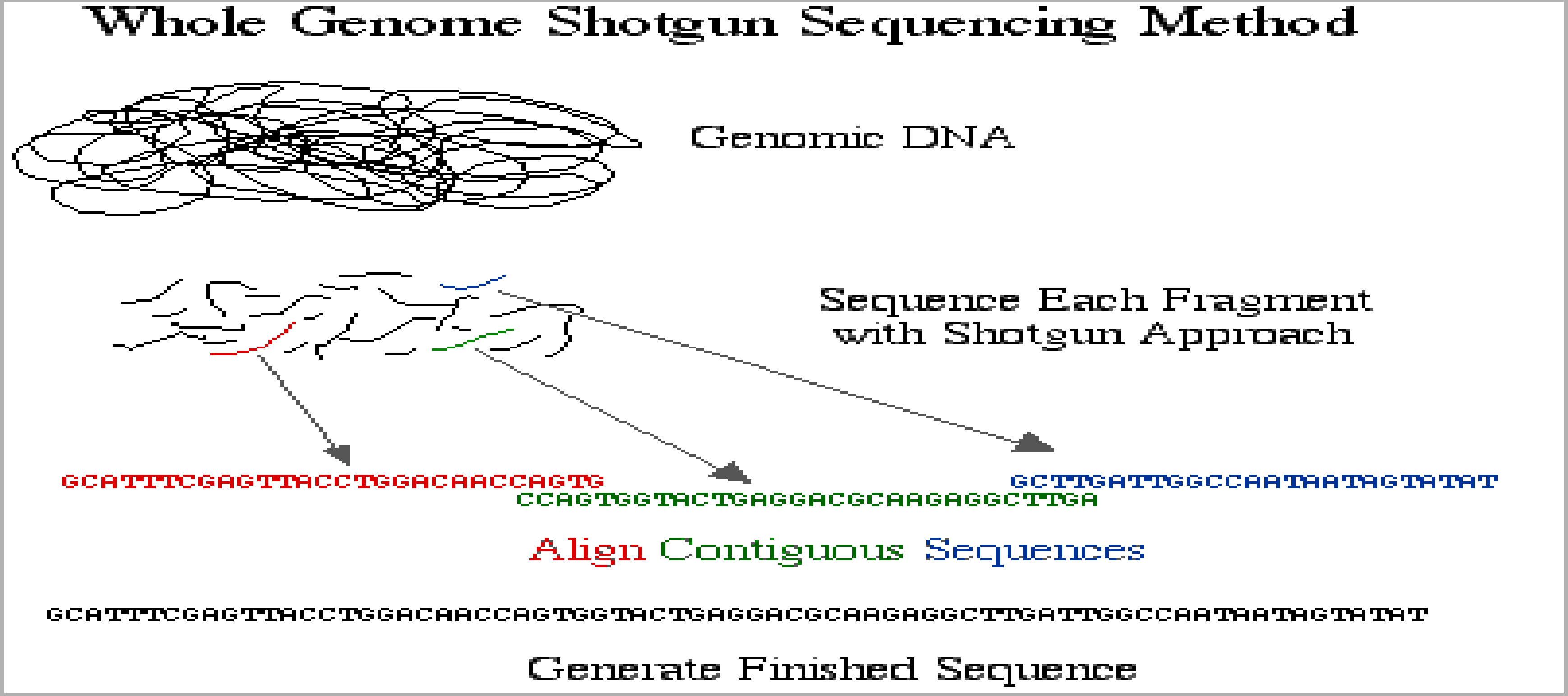

Whole Genome Shotgun Sequencing Method
Genomic DNA
Sequence Each Fragment
with Shotgun Approach
GCATTTCGAGTTACCTGGACAACCAGTG
CCAGTGGTACTGAGGACGCAAGAGGCTTGA
GCTTGATTGGCCAATAATAGTATAT
Align Contiguous Sequences
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT
Generate Finished Sequence

Figure 2. Schematic diagram of sequencing strategy used by Celera. The DNA was cut into small pieces and sequenced completely. These fragments were organized into contigs based on overlapping sequences.

The advantage to the hierarchical approach is sequencers are less likely to make mistakes when assembling the shotgun fragments into contigs as long as full chromosomes. The reason is that the chromosomal location for each BAC is known, and there are fewer random pieces to assemble. The disadvantage to this method is time and expense.

The shotgun method is faster and less expensive, but it is more prone to errors due to incorrect assembly of finished sequence. For example, if a 500 kb portion of a chromosome is duplicated and each duplication is cut into 2kb fragments, then it would be difficult to determine where a particular 2 kb piece should be located in the finished sequence since it occurs twice. But duplications seldom retain their original sequences; they tend to drift over time. So a small region may be retained while other parts may mutate. This might create overlapping sequences for small pieces that are located several hundred kb apart on the chromosome.

Which method is better? It depends on the size and complexity of the genome. With the human genome, each group believes its approach to be superior to the other. We only have draft sequences and each has gaps and unfinished regions so it is not possible to say for sure. It is worth mentioning that Celera had access to the HGP data but the HGP did not have access to the Celera data. Furthermore, since the Celera data is not freely available, most investigators will use the HGP sequence for further research.

Therefore, we may never know which method "won".

The Human Genome Project

We should think the human genome as a database of critical information that serves as a tool for exploring the workings of the cell and, ultimately, understanding how a complex living organism functions.

Sequencing a Genome

Sequencing a genome is an enormous task. It requires not only finding the nucleotide sequence of small pieces of the genome, but also ordering those small pieces together into the whole genome.

A useful analogy is a puzzle, where you must first put together the pieces of a smaller puzzle and then assemble those pieces into a much larger picture. Two general strategies have been used in the sequencing of large genomes: clone-based sequencing and whole genome sequencing (Fig. 1).

In clone-based sequencing (also known as hierarchical shotgun sequencing) the first step is mapping. One first constructs a map of the chromosomes, marking them at regular intervals of about 100 kilobases (kb). Then, known segments of the marked chromosomes (which can contain very small fragments of DNA) are cloned in plasmids. One special type of plasmid used for genome sequencing is a BAC (bacterial artificial chromosome), which can contain DNA fragments of about 150 kb. The plasmid's fragments are then further broken into small, random, overlapping fragments of about 0.5 to 1.0 kb. Finally, automated sequencing machines determine the order of each nucleotide of the many small fragments.

Data management and analysis are critical parts of the process, as these sequencing machines generate vast amounts of data. As the data are generated, computer programs align and join the sequences of thousands of small fragments. By repeating this process with the thousands of clones that span each chromosome, researchers can determine the sequences of all the larger clones.

Once they know the order of all the larger clones, the researchers can join the clones and determine the sequence of each chromosome.

The challenge is assembling all the pieces. The National Human Genome Research Institute (the public consortium headed by Francis Collins) used clone-based sequencing for the human genome. In doing so, they relied heavily on the work of computer scientists to assemble the final sequence.

Whole genome shotgun sequencing clones millions of the genome's small fragments in plasmids, sequences all of these small overlapping fragments, and then uses computers to find matches and join them together.

Celera Genomics, a private company used this approach to clone the human genome. Although they started much later than the public consortium, Celera completed its draft sequence at about the same time as the consortium; however, it had the advantage of having access to all the consortium's maps.

The technology developed for sequencing the human genome - both in terms of sequencing DNA and in the software and hardware used to assemble the sequences into a genome - has resulted in the rapid sequencing of many other genomes.

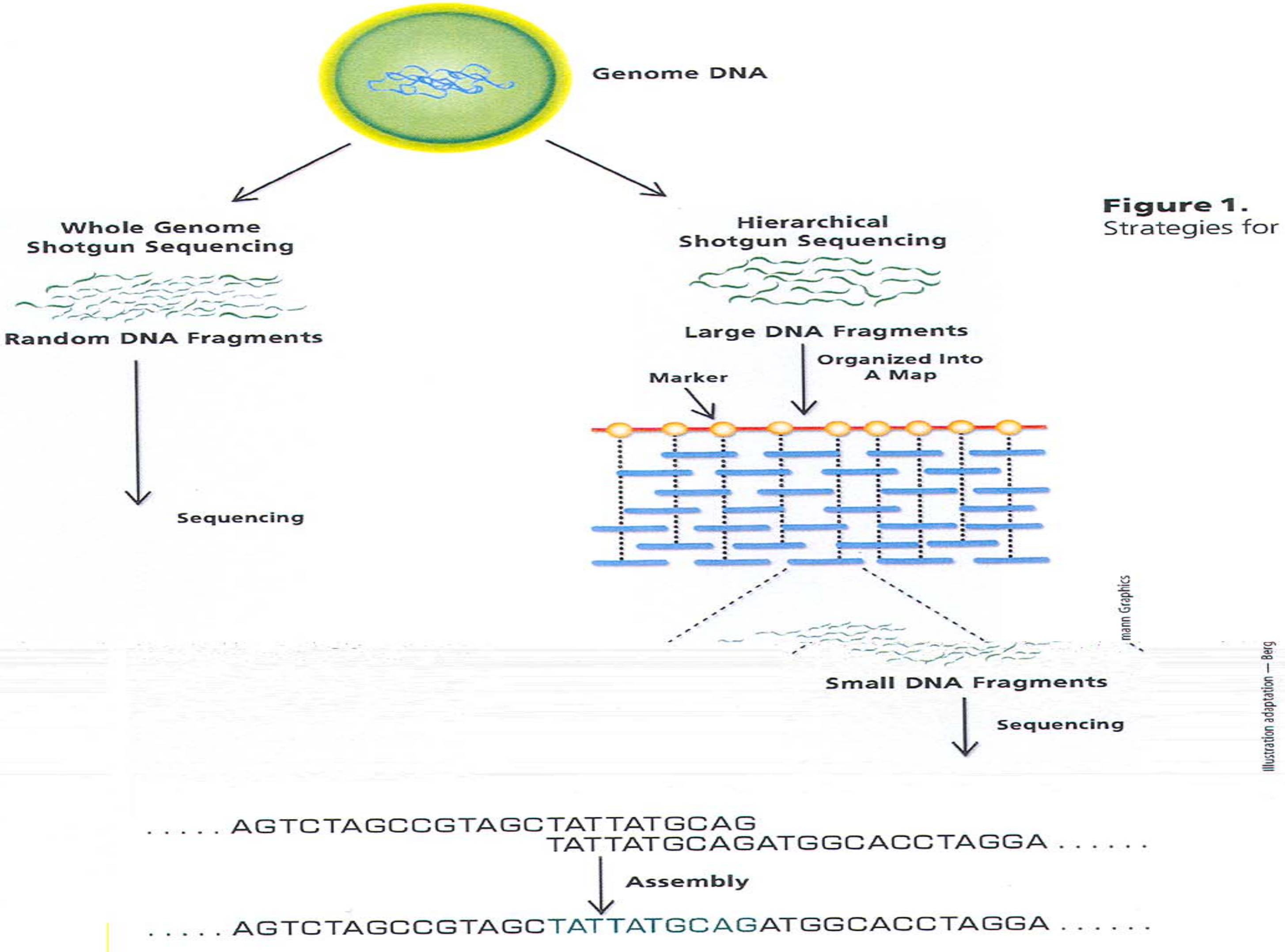

Genome DNA
Whole Genome Shotgun Sequencing
Random DNA Fragments
Sequencing
Hierarchical Shotgun Sequencing
Large DNA Fragments
Organized Into A Map
Marker
Small DNA Fragments
Sequencing
. . . . . AGTCTAGCCGTAGCTATTATGCAG
TATTATGCAGATGGCACCTAGGA . . . . . .
Assembly
. . . . . AGTCTAGCCGTAGCTATTATGCAGATGGCACCTAGGA . . . . . .

Figure 1.
Strategies for cloning whole genomes.

Illustration adaptation — Berg
mann Graphics

Finding Genes

Imagine the genome as an encyclopedia with a volume for each chromosome. If you were to open a volume, you would find page after page containing only four letters - A, T, G, and C - without spaces or punctuation. How could you read such a book, or even identify possible words and sentences? The genome sequence itself does not provide direct information on the location of a gene, but there are clues embedded in the sequence that computer programs can find.

Most simple gene prediction programs use several pieces of sequence information to identify a potential gene in a DNA sequence. The programs look for sequences in the DNA that have the potential to encode a protein. These sequences are called open reading frames (ORFs). An ORF usually begins with a codon of AUG, and then contains a long sequence of codons that specify the protein's amino acids. The ORF then ends with a stop codon of UAA, UAG, or UGA (Fig. 2).

Using overlapping frames of three nucleotides each, the computer program searches the database until it identifies an ORF region. For example, the sequence "abcdefghijk" could be read in three-letter "words" of "abc-def-ghi'" "bcd-efg-hij," or "cde-fgh-ijk." Computer programs can scan DNA sequences quickly, using these overlapping reading frames on both the original strand and on the complementary strand, producing a total of six different reading frames for any sequence.

Using these programs to find ORFs in bacterial genomes is relatively easy. Here, the DNA sequence matches the mRNA.

The situation is more complicated for eukaryotic genes, which often contain one or more noncoding regions (introns). To find ORFs in these genes, the introns are removed in a process called splicing (Fig. 3). The final spliced mRNA, which encodes the protein product of the gene, is smaller than the original RNA transcript that matches the genome. The introns are removed, leading to the splicing of the coding regions of a gene (exons) together into the final mRNA.

The problem is that a simple ORF-finding program cannot be used with genomic DNA that has introns because those genes do not match the mRNA. While computer programs can identify eukaryotic genes with introns, they are not always accurate.

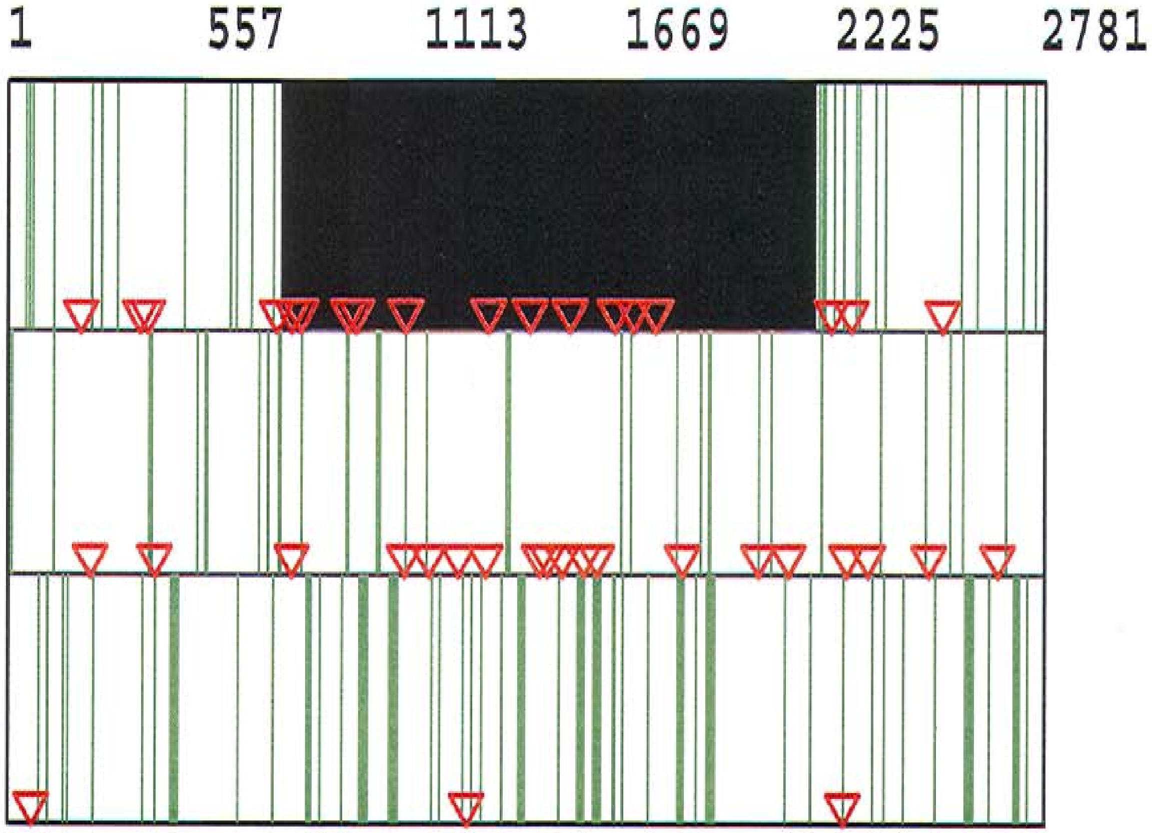


Figure 2. To find an open reading frame (ORF), a computer program identifies start codons (red arrows) and stop codons (green lines) in all three reading frames (represented by the three stacked rows). The black box is the largest ORF found in this sequence.

An alternate approach to characterize genes in eukaryotes is to first make a DNA copy of the mRNA encoded by the gene. To do this, one uses an enzyme called reverse transcriptase.

The copy, called cDNA or complementary DNA, has the same sequence as the mRNA, except that the U is replaced by a T. Because the cDNA lacks introns, the sequence of the cloned cDNA can be used to find an ORF. While eukaryotes generally have more genes than bacteria, the difference in gene content is not as great as the difference in DNA content: there is much more noncoding DNA in eukaryotes.

In fact, gene-coding regions comprise only about two percent of the human genome.

Most eukaryotic genes are interrupted by large introns.

In eukaryotes, repeated sequences characterize great amounts of noncoding DNA. Some of this repetitive DNA is dispersed more or less randomly throughout the genome. There are also millions of copies of other,

shorter repeats, but they are typically found in larger blocks.

Some trinucleotide (3 bp) repeats are associated with diseases such as fragile X and Huntington's disease, which result from extra copies of the repeat sequence.

Most of these repeat sequences are <u>transposable elements</u>, that can replicate and insert a copy in a new location in the genome. The result is the amplification of these repetitive elements over time. Transposable elements can be harmful because they can cause mutation when they move into a gene.
Are these elements useful components of the genome? We don't really know, but there are some suggestions of functions for some of these elements. About one million copies of the repetitive DNA element called Alu repeats lurk in the genomes of each one of us. What are they doing? One study found that these bind to proteins used to reshape chromatin during cell division.

Perhaps this apparent junk DNA is actually helping provide structure to the chromosome and regulate the production of proteins in different cell types.

Genomes differ in size, in part because they have different proportions of repetitive DNA .

More than half the human genome is repetitive DNA.

More than ninety-nine percent of human genes have a related copy in the mouse.

As one examines animals that are more distantly related, the proportion of the genes they share decreases; This remarkable conservation of gene structure is striking considering how much these animals differ in morphology, physiology, and behavior.

If they share so many of the same genes, why are different animals so different?

Differences among species result largely from differences in the time and location of the genes' expression. Let us consider our closest relative, the chimpanzee. Not only do chimpanzees and humans share nearly all of the same genes, but the DNA sequences of those genes also are very similar between the two species.

Svante Pããbo sequenced three million bases of the chimp genome and found that chimps and humans differ overall by less than two percent at the sequence level. Based on the low sequence divergence, Pããbo hypothesized that the difference between humans and chimpanzees was due mainly to how the genes were expressed in the different species.

To test this hypothesis, Pããbo compared the expression pattern of 20,000 human genes in humans and chimps. He found that while expression levels were similar in liver cells and blood, there were larger differences in brain cells.

This suggests that the human brain has increased the use of certain genes compared to those same genes in a chimp.

So, it not so much the sequence of the genes that is important, but how they are expressed to make the cell's proteins that determines the unique characteristics of each organism.

The Difference May Lie Not in the Sequence but in the Expression

Determining Gene Function from Sequence Information

Researchers have produced an enormous number of genome sequences from a variety of organisms. Publicly available databases, such as GenBank at the NCBI (National Center for Biotechnology Information), store many of these sequences. The databases have been a tremendous boon for comparative biology. The NCBI database stores not only the genome sequences, but also information about the function (if it is known) of the genes.

The NCBI can also identify unknown genes by comparing them with known genes in the database. One program commonly used for this purpose is BLAST (Basic Local Alignment Search Tool).

Sequence similarity searching algorithms like BLAST are based on the premise that if two sequences are similar then they are likely to be homologous (that is, they share a common evolutionary ancestor). Using this database, one can infer the function of an unknown gene by finding similar sequences of known genes and proteins. For example, suppose you were to use BLAST to search for sequences similar to a new gene. Upon viewing your results, you noticed that all the sequences with a high degree of similarity to the new gene belonged to a family of genes known to break down hydrogen peroxide. You could logically conclude, then, that this new gene encoded a protein with a similar function.

BLAST searches can be done at the nucleotide level; however, comparisons at the amino acid level provide much greater sensitivity. Therefore, unless one is particularly interested in the DNA sequence itself, it is better to search for genes using protein. If you have only raw nucleotide sequence data, computer programs can automatically translate the DNA into amino acids using all six

reading frames (three frames from one strand and three frames from the complementary strand) before searching the protein database.

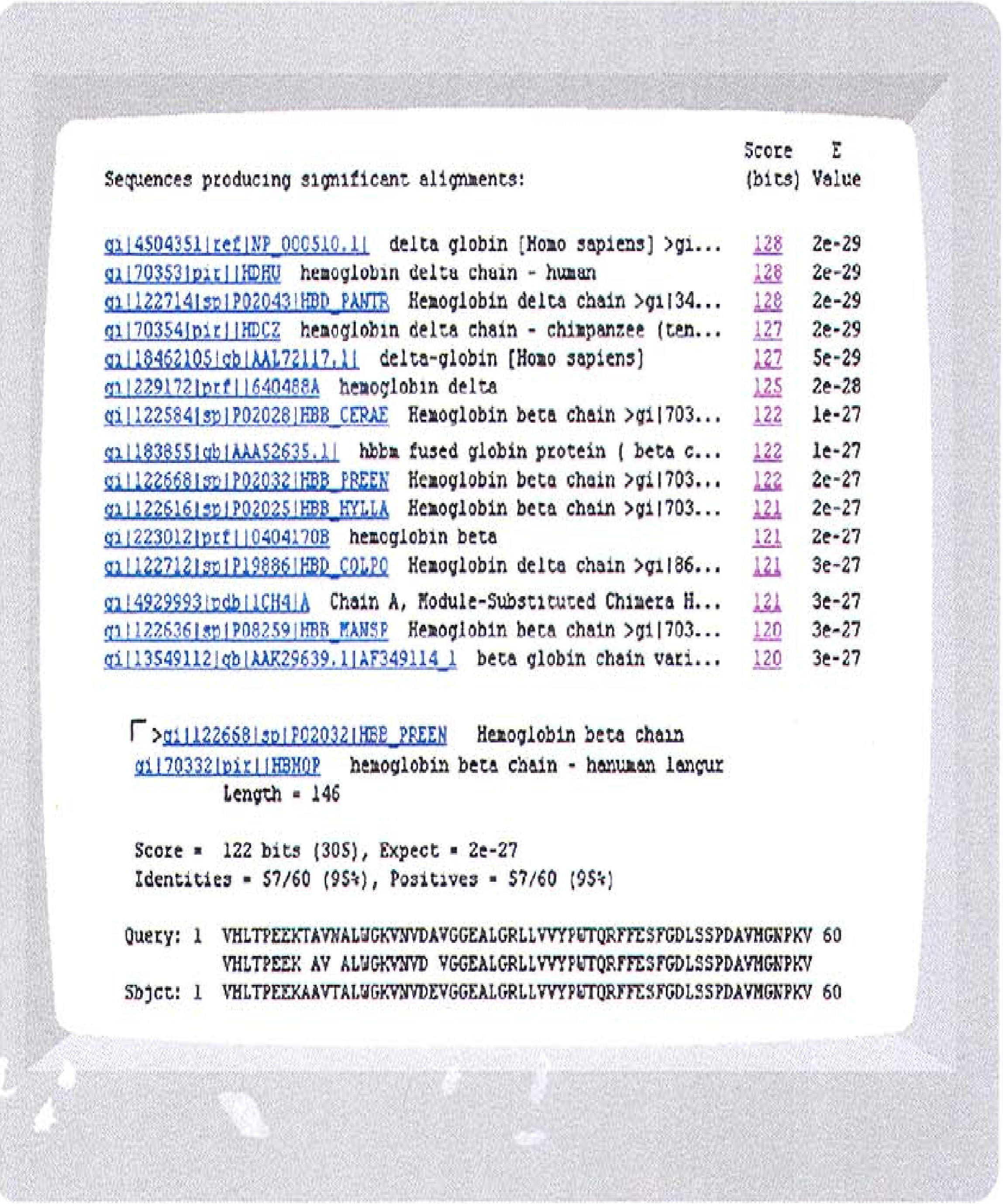
In addition to whole proteins, similarity searches can identify protein motifs. A motif is a distinctive pattern of amino acids, conserved across many proteins, which gives a particular function to the protein. For example, the presence of one particular motif in a protein indicates that this protein probably binds ATP and may therefore require ATP for its action.

The result of a database search is a list of matches, ranked from highest to lowest, based on the probability of a significant match (Fig. 4). The reported alignment scores are given "expectation values" (E), which represent the probability that a match with the reported score would be expected to occur by random chance.

The smaller the E- value, the higher the assigned score and the less likely that the match was coincidence. Some of the easiest results to interpret are very high scores (small E-values, low-probability), which usually result from two very similar proteins.

Other easily identifiable results are very low scores, which indicate that the outcome is probably the result of chance similarity.

Search results also provide links (in blue) to a database page with information on each sequence similar to the query sequence. This page gives extensive information on the match sequence, including the organism it came from, the function of the gene product - if it is known - and references to journal articles concerning the sequence. BLAST results also provide the actual alignment results for nucleotides or amino acids between the query sequence and the match sequences.



```
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

gi|4504351|ref|NP_000510.1|  delta globin [Homo sapiens] >gi...    128   2e-29
gi|70353|pir||HDHU  hemoglobin delta chain - human                 128   2e-29
gi|122714|sp|P02043|HBD_PANTR  Hemoglobin delta chain >gi|34...    128   2e-29
gi|70354|pir||HDCZ  hemoglobin delta chain - chimpanzee (ten...    127   2e-29
gi|18462105|gb|AAL72117.1|  delta-globin [Homo sapiens]            127   5e-29
gi|229172|prf||640488A  hemoglobin delta                           125   2e-28
gi|122584|sp|P02028|HBB_CERAE  Hemoglobin beta chain >gi|703...    122   1e-27
gi|183855|gb|AAA52635.1|  hbbm fused globin protein ( beta c...    122   1e-27
gi|122668|sp|P02032|HBB_PREEN  Hemoglobin beta chain >gi|703...    122   2e-27
gi|122616|sp|P02025|HBB_HYLLA  Hemoglobin beta chain >gi|703...    121   2e-27
gi|223012|prf||0404170B  hemoglobin beta                           121   2e-27
gi|122712|sp|P19886|HBD_COLPO  Hemoglobin delta chain >gi|86...    121   3e-27
gi|4929993|pdb|1CH4|A  Chain A, Module-Substituted Chimera H...    121   3e-27
gi|122636|sp|P08259|HBB_MANSP  Hemoglobin beta chain >gi|703...    120   3e-27
gi|13549112|gb|AAK29639.1|AF349114_1  beta globin chain vari...    120   3e-27


   >gi|122668|sp|P02032|HBB_PREEN  Hemoglobin beta chain
   gi|70332|pir||HBMQP  hemoglobin beta chain - hanuman langur
           Length = 146

   Score =  122 bits (305), Expect = 2e-27
   Identities = 57/60 (95%), Positives = 57/60 (95%)

  Query: 1  VHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV 60
            VHLTPEEK AV ALWGKVNVD VGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV
  Sbjct: 1  VHLTPEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV 60
```

Figure 4. The results of a BLAST search using the delta chain of hemoglobin as the query.

The Virtues of Knockouts

Gene prediction programs have been valuable in the preliminary identification of genes; however, they have limitations. Unless the gene of interest is homologous to a gene of known function, the function is generally still not known. A biological approach to determining the function of a gene is to create a mutation and then observe the effect of the mutation on the organism. This is called a knockout study. While it is not ethical to create knockout mutants in humans, many such mutants are already known, especially those that cause disease. One advantage of having a genome sequence is that it greatly facilitates the identification of genes in which mutations lead to a particular disease.

The mouse, where one can make and characterize knockout mutants, is an excellent model system for studying genetic diseases of humans; its genome is remarkably similar to a human's.

Nearly all human genes have homologs in mice, and large regions of the chromosomes are very well conserved between the two species.. Thus, it is possible to create mutants in mice to determine the probable function of the same genes in humans. One goal of the mouse genome project is to make and characterize mutations in order to determine the function of every mouse gene. After a particular gene mutation has been linked to a particular disorder, the normal function of the gene may be determined.

Genetic Variation Within Species and SNPs

A polymorphism, the existence of two or more forms of sequence between different individuals of the same species, can arise from a change in a single nucleotide. These single nucleotide polymorphisms (SNPs) account for ninety percent of all polymorphisms in humans. The number of SNPs between two genomes provides a measure of sequence variation; however, the variation is not uniform over the genome. About two-thirds of SNPs are in noncoding DNA and tend to be concentrated in certain locations in the chromosome. In addition, sex chromsomes have a lower concentration of SNPs than autosomes.

There are about three million SNPs in the human genome, or about 1 per 1000 nucleotides. SNPs are ideal genetic markers for many applications because they are stable, widespread, and can often be linked to particular characteristics (phenotypes) of interest. They are proving to be among the most useful human markers for studies of evolutionary genetics and medicine.

Not all SNPs, even when they are present in coding genes, lead to visible or phenotypic differences among individuals.

Changes in the DNA sequence don't always change the amino acid sequence of the protein.

For example, a change from GGG to GGC results in no change in the protein because both codons result in a glycine in the protein. This is called a synonymous mutation or silent mutation; non-synonymous substitutions do cause a change in the amino acid.

About half of all SNPs in genes are non-synonymous and therefore can account for diversity between individuals or populations.

Depending on the particular change in an amino acid caused by a nonsynonymous mutation, the resulting protein may be an active, inactive, or partially active. It may also be active in a different way.

One well-characterized SNP exists in a gene in chromosome 6.

Individuals with cysteine at amino acid position 282 are healthy; however, about 1 in 200-400 Caucasians of Northern European descent possess two copies of that gene where the amino acid is tyrosine instead of cysteine.

Due to this one change, these individuals have a disease called hereditary hemochromatosis.

People afflicted with this disease accumulate high levels of iron, which causes permanent damage to the organs, especially the liver.

About ten percent of these individuals carry only one copy of this mutation; they are heterozygous and are carriers of the disease.

A genetic test for hereditary hemochromatosis is available.

Identifying and Using SNPs

In order to identify SNPs, nucleotide sequences of two or more genomic regions must be aligned so that the polymorphisms are apparent. Sequence alignments are easy when the sequences are similar, but can be very difficult when there are many polymorphisms. The alignment of two sequences is determined by a program that compares the two sequences, nucleotide by nucleotide. For multiple sequences, the program continues the same type of pairwise alignment for all possible pairs. The result is a pairwise distance matrix based on all possible alignments of any two sequences. This matrix is then used to construct a phylogenetic tree that predicts how closely related two sequences are, based on their similarity. The program then uses this information to align the sequences, again in order of their relatedness.

This is the method used in a program called CLUSTAL. A typical output from CLUSTAL is shown in Figure 5.

SNPs appear to cluster in blocks called haplotypes. Grouping individuals that share a particular haplotype is called *haplotyping*. Because these particular sequences of SNPs on a chromosome are inherited together as blocks, they can be used to distinguish individuals and populations.

One can determine what specific diseases or other traits are associated with different haplotypes. In most cases, there are much fewer haplotypes than SNPs. Although it is the SNPs that actually cause disease, looking for changes in one SNP out of millions in the genome is not practical; looking for a particular haplotype is much easier.

Crohn's disease is a chronic inflammatory disease of the digestive tract that tends to cluster in families. Researchers identified a haplotype on chromosome 5 that correlates with the disease. This region of the chromosome contains genes involved in immunity; these genes then may be important in other inflammatory diseases, such as lupus or asthma.

at the bottom.

```
        R  P  P  G  K  S  G  K  Y  Y  Y  Q  L  N  S  K  K  H  H    159
  Human CGGCCGCCGGGCAAGAGCGGCAAGTACTACTACCAGCTCAACAGCAAGAAGCACCAC  642
  Mouse CCCCCGCCAGG-AAGAGCGGCAAGTATTATTATCAGCTAAATAGCAAAAAGCACCAC  614
Chicken CAGTCCCACAGCAAG---GGCAAGTACTACTACCAGCTCAACAGCAAGAAGCACCAC  583
   Frog CATTCCAGTAACAAG---AAAAAATACTATTATCAGCTCAATAGCAAAAAACATCAT  500
        *   *       ***      ** ** ** ** ***** ** ***** ** ** **
```

Figure 5. A CLUSTAL alignment of a segment of a gene from four species. The red letters show the amino acid sequence (R=arginine, P=proline, G=glycine, etc.). The nucleotides that are conserved in all four species are shown in the columns with an asterisk at the bottom.

Phylogenetic Trees and Their Parts

Rooted Tree

Phylogenetic trees are designed to reveal evolutionary relationships among DNA or protein sequences. The use of the term "tree" has given rise to arborial terminology to describe the different parts of the overall tree. In order to depict related sequences on a rooted tree, the investigator must know one sequence is very distantly related to all the others. This most distant sequence is used to root the tree.

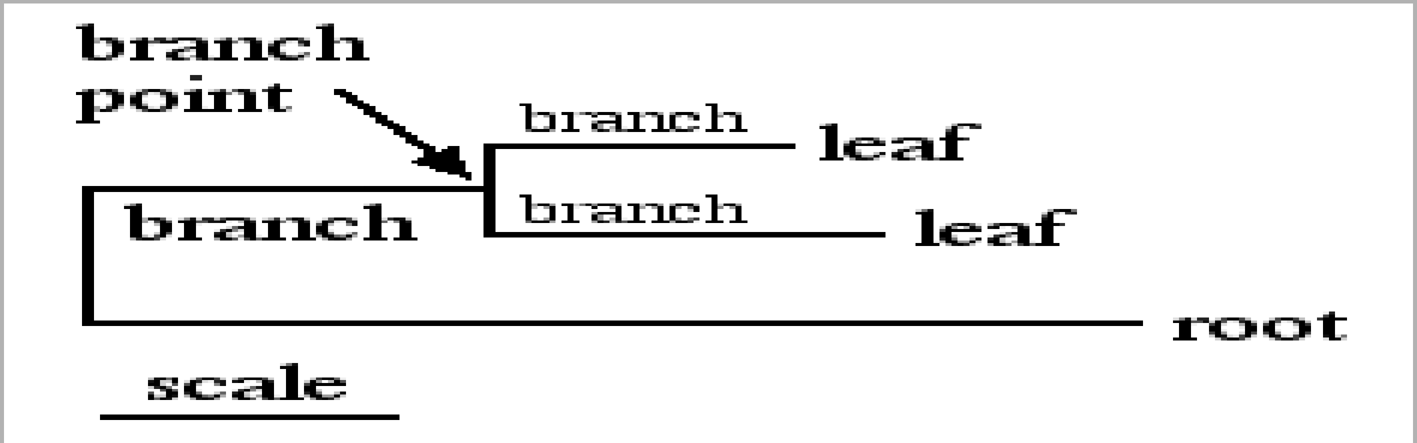


Figure 1. This figure illustrates the most common terminology for phylogenetic trees: root, branch, branch point and leaf. Though there are variations of these terms which you may encounter in other readings, the ones illustrated here are widely used..

Unrooted Tree

When the investigator has not included one distantly related sequence for comparison, then an unrooted tree is required. Since the visual display of the unrooted tree is very different, it might be beneficial to choose an unrooted tree to depcit clusters of related sequences.

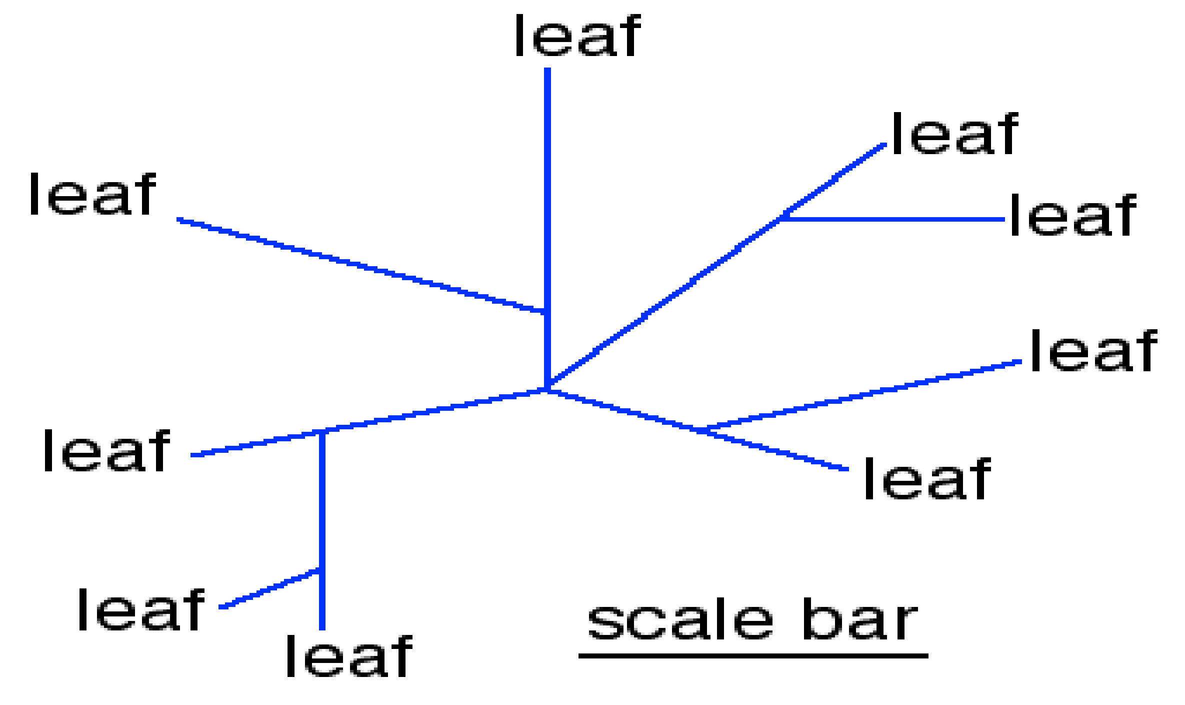


Figure 2. In this unrooted tree, the branches are blue and the leaves are labeled as such.

Sequence Analysis of Functional Domains

About 40% of all genes have no known function. To reduce the number of "unknown genes", it is possible to perform a sequence analysis looking for functional domains. To perform this sequence analysis, it is necessary to compare the genomes of different species. In species A, a protein exists that has two functional domains (labeled 1 and 2 in figure 1). In species B, there may be two orthologs of the two domains but in species B, the orthologs are found in separate genes (labeled 1' and 2' in figure 1).

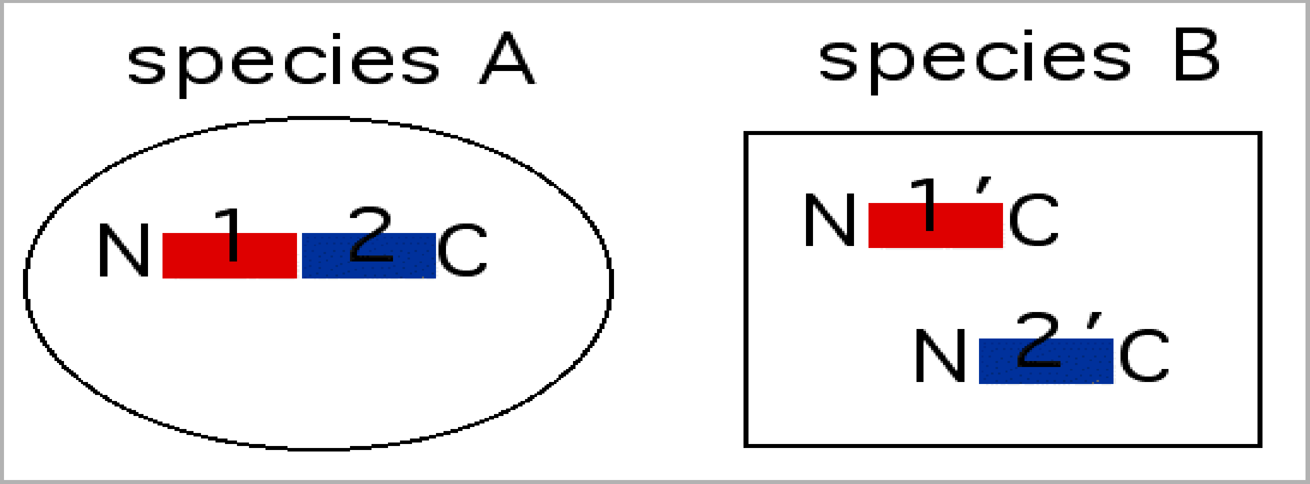


Figure 1. Two species with conserved functions. Species A performs a particular task using two domains in a single protein. Species B performs the same task but utilizes two genes encoding similar domains.

Let's consider an hypothetical example.

Kinases must have at least two domains to perform their task of phosphorylating a substrate. One domain binds the substrate and the other binds ATP and transfers the terminal phosphate onto the substrate. It is easy to image that in species A, this task could be performed by a single protein while in species B, two genes form a heterodimer to accomplish the same task.

Using this type of conservation of sequence but divergence of gene number and size, investigators have been able to mine genome sequences to determine the functional roles of previously "unknown" genes.

Kinase (Enzyme) Assay

A **kinase** is protein enzyme that adds a phosphate onto a molecule, though typically only proteins. The phosphorylated molecule can be another protein, the kinase itself (autophosphorylation) or any other molecule. The source for the phosphate is terminal phosphate which is called the gamma (γ) phosphate from ATP (figure 1).

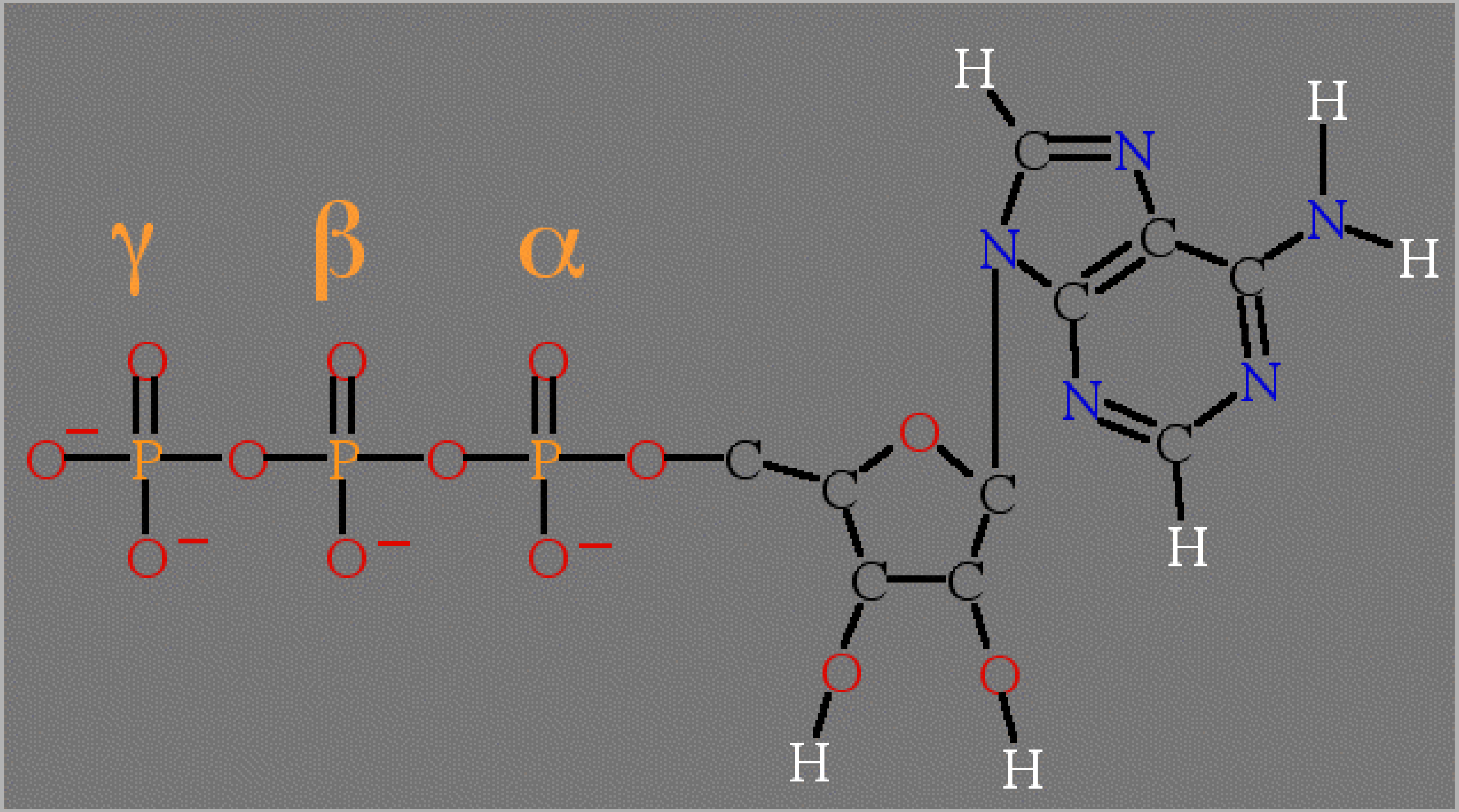


Figure 1. Structure of ATP

The addition of a phosphate covalently modulates the substrate protein and typically alters the substrate conformation sufficiently to either activate or inactivate it. Some kinases can phosphorylate more than one protein and sometimes one substrate will be activated while another will be inactivate .

Any enzyme can be measured in an assay where either the production of product or the consumption of substrate can be measured. When the accumulation of product or loss of substrate is measured over time and there are no limiting factors (such as very little substrate), then we expect to see the rate of enzyme activity to be linear over time (figure 2).

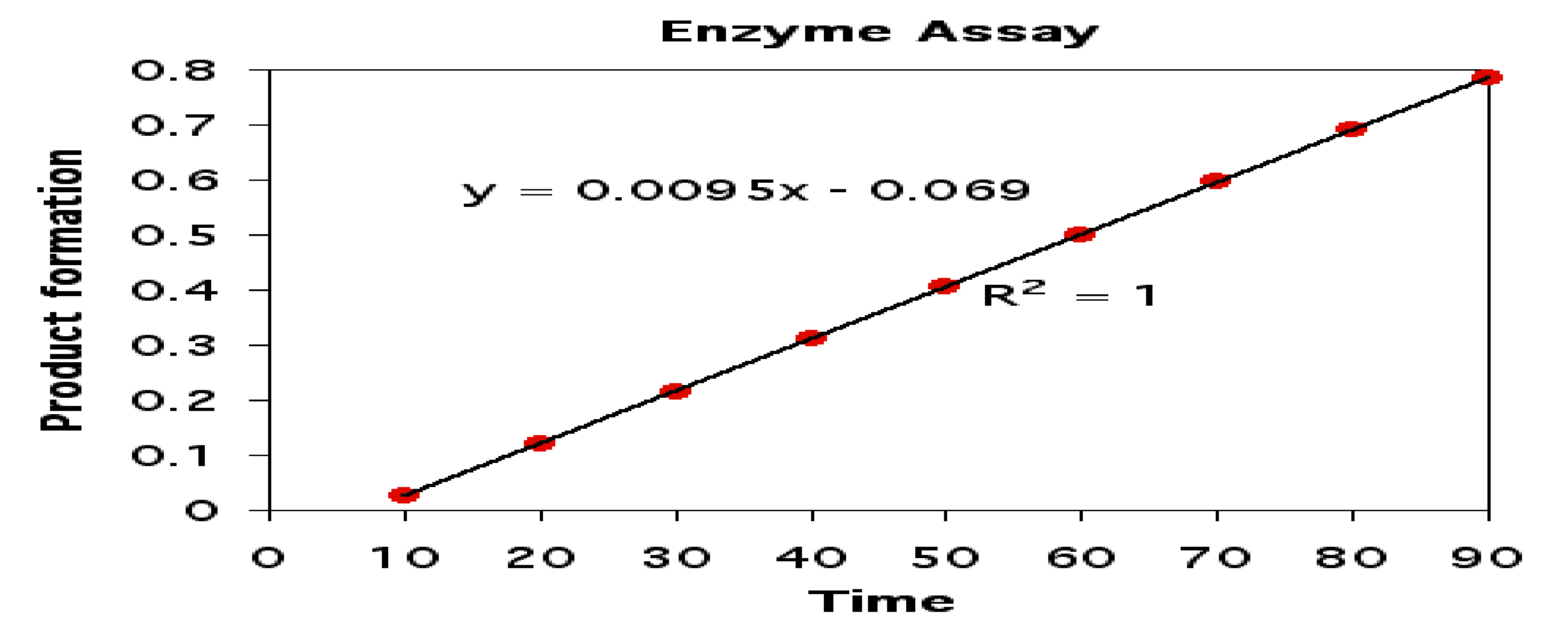


Figure 2. Rate of product formation is linear over time when substrate is plentiful and product accumulation does not inhibit the enzyme.

Practical Applications of Genomics

DNA is an invaluable tool in forensics because - aside from identical twins - every individual has a uniquely different DNA sequence. Repeated DNA sequences in the human genome are sufficiently variable among individuals that they can be used in human identity testing.

The FBI uses a set of thirteen short tandem repeat (STR) DNA sequences for the Combined DNA Index System (CODIS) database, which contains the DNA fingerprint or profile of convicted criminals. Investigators of a crime scene can use this information in an attempt to match the DNA profile of an unknown sample to a convicted criminal. DNA fingerprinting can also identify victims of crime or catastrophes, as well as many family relationships, such as paternity.

While we think of forensics in terms of identifying people, it can also be used to match donors and recipients for organ transplants, identify species, establish pedigree.

The basis of many diseases is the alteration of one or more genes. Testing for such diseases requires the examination of DNA from an individual for some change that is known to be associated with the disease. Sometimes the change is easy to detect, such as a large addition or deletion of DNA, or even a whole chromosome. Many changes are very small, such as those caused by SNPs. Other changes can affect the regulation of a gene and result in too much or too little of the gene product.

In most cases if a person inherits only one mutant copy of a gene from a parent, then the normal copy is dominant and the person does not have the disease; however, that person is a carrier and can pass the disease on to offspring.

If two carriers produce a child and each passes the mutant allele to the child (a one-in-four probability), that individual will have the disease.

Normal adults may also be tested to determine whether they are carriers of alleles for cystic fibrosis, or sickle cell anemia. This can help them determine their risk of transmitting the disease to children. These tests as well as others (such as for Down's syndrome) are also available for prenatal diagnosis of diseases. As new genes are discovered that are associated with disease, they can be used for the early detection or diagnosis of diseases such as familial adenomatous polyposis (associated with colon cancer) or p53 tumor-suppressor gene (associated with aggressive cancers).

The ultimate value of gene testing will come with the ability to predict more diseases, especially if such knowledge can lead to the disease's prevention.

Gene therapy is a more ambitious endeavor:
its goal is to treat or cure a disease by providing a normal copy of the individual's mutated gene. The first step in gene therapy is the introduction of the new gene into the cells of the individual. This must be done using a vector (a gene carrier molecule), which can be engineered in a test tube to contain the gene of interest. Viruses are the most common vectors because they are naturally able to invade the human host cells. These viral vectors are modified so that they can no longer cause a viral disease.

Patients often experience negative side effects and expression of the desired gene introduced by viral vectors is not always sufficiently effective.
Gene therapy is the long-term goal for the treatment of genetic diseases for which there is currently no treatment or cure.

Examining Gene Expression

Understanding the functions of genes depends on knowing when and in what cells they are each expressed. How can one measure the amount of mRNA transcribed from a gene in a particular cell type?

The standard method uses a probe - a DNA sequence unique for that gene - which binds to the mRNA that has the complementary sequence. The more mRNA particular cell produces, the more mRNA that is bound to the probe, giving the probe an increased signal. Because cDNA is complementary in sequence to mRNA, it can also be used to measure the expression of a particular gene.

Organisms have so many genes in their genomes that studying the expression of all of these genes had been exceedingly difficult. Going from studying gene expression one gene at a time to examining expression patterns of a multitude of genes required new technology.

In the late 1990s the development of microarray chips allowed researchers to examine the expression of thousands of genes simultaneously. This allowed for a much broader perspective of gene expression than was possible when genes were analyzed singularly. Microarray chips are glass slides spotted with many rows containing tiny amounts of probe DNA, one for each of thousands of genes (Fig. 5). The target sample of interest, usually made from mRNA of a specific type of cell, is labeled with a fluorescent dye and added to the chip. If there is a match between the sample of interest and the DNA probe on the chip, the two molecules will bind to each other. Then, when exposed to a laser, the spot will produce a signal that will fluoresce. (Figure 5 describes this process in more detail.)

Using microarrays, one can measure expression patterns of large numbers of genes in different cell types (such as cancer cells versus normal cells, or liver cells versus kidney cells). It can also be used to examine the changes in gene expression over time (for example, as an embryo develops), or changes in a given cell type under different environmental conditions (various temperatures, for instance).

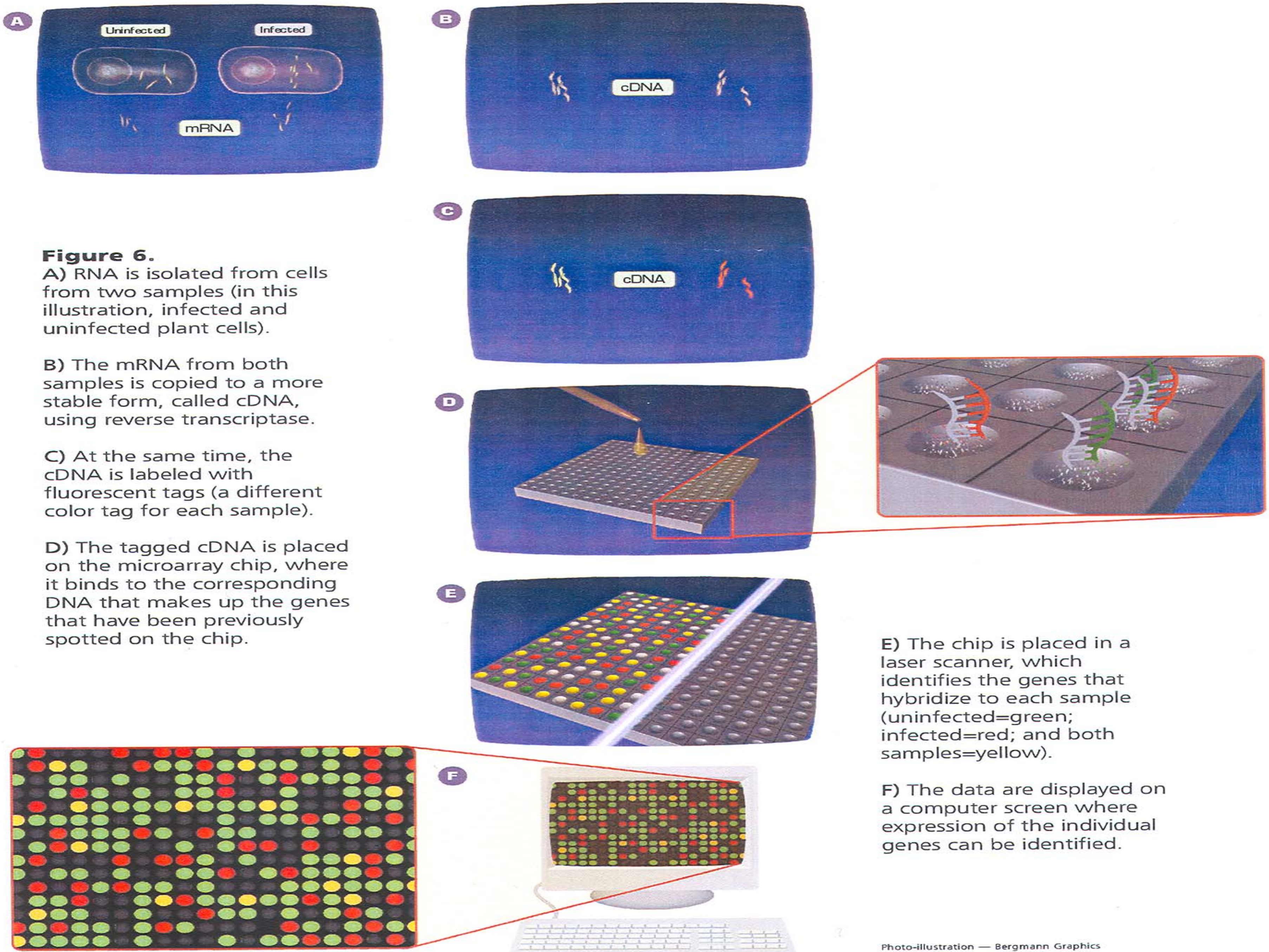


Figure 6.
A) RNA is isolated from cells from two samples (in this illustration, infected and uninfected plant cells).

B) The mRNA from both samples is copied to a more stable form, called cDNA, using reverse transcriptase.

C) At the same time, the cDNA is labeled with fluorescent tags (a different color tag for each sample).

D) The tagged cDNA is placed on the microarray chip, where it binds to the corresponding DNA that makes up the genes that have been previously spotted on the chip.

E) The chip is placed in a laser scanner, which identifies the genes that hybridize to each sample (uninfected=green; infected=red; and both samples=yellow).

F) The data are displayed on a computer screen where expression of the individual genes can be identified.

Photo-illustration — Bergmann Graphics
3D Graphics — KStar Productions

This is a robot used to create microarrays

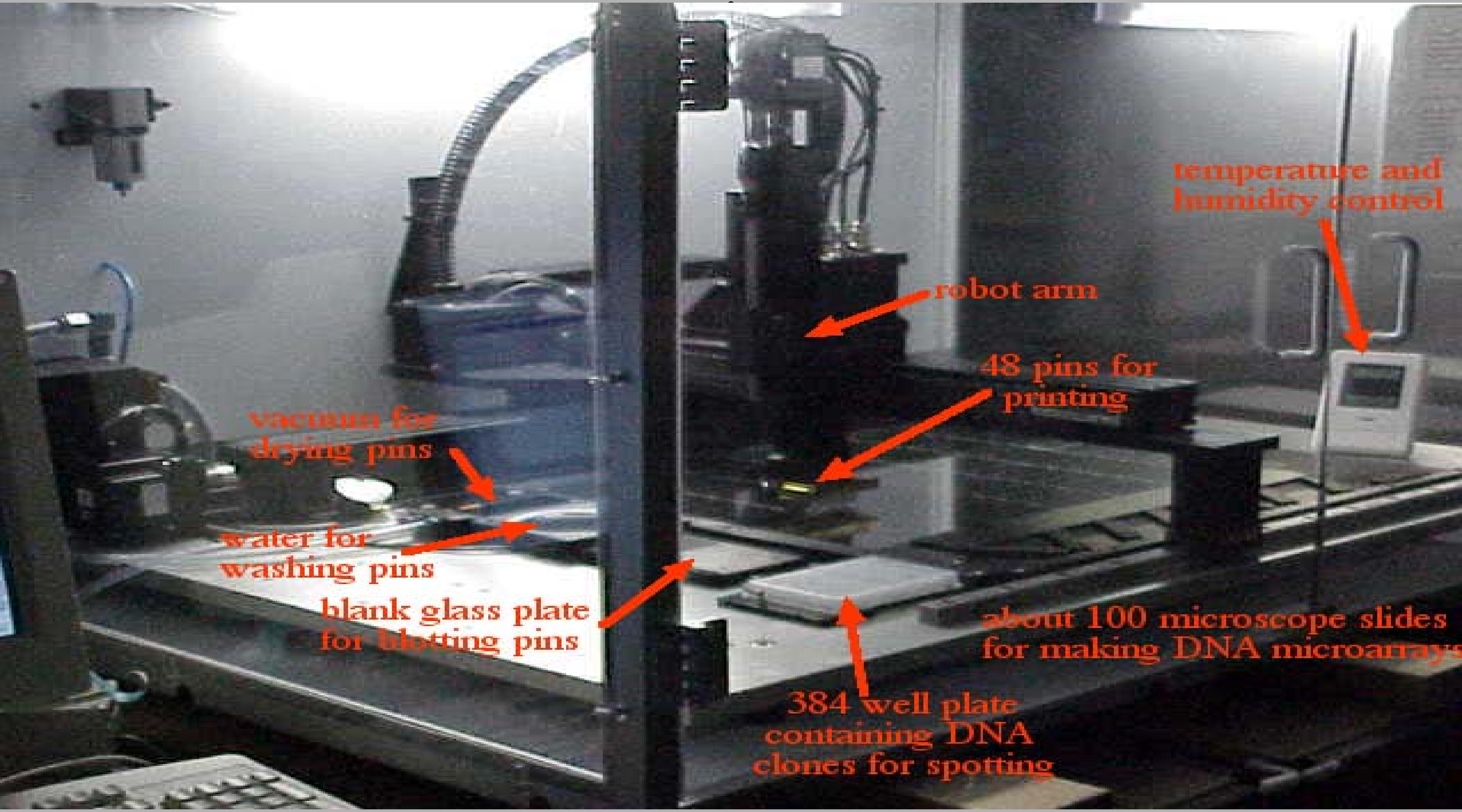


Figure 1. DNA micrarray robot enclosed in a temperature and humidity controlled environment. The temperature and humidity must be closely regulated or the spot sizes will vary or the printing pins will dry out too quickly.

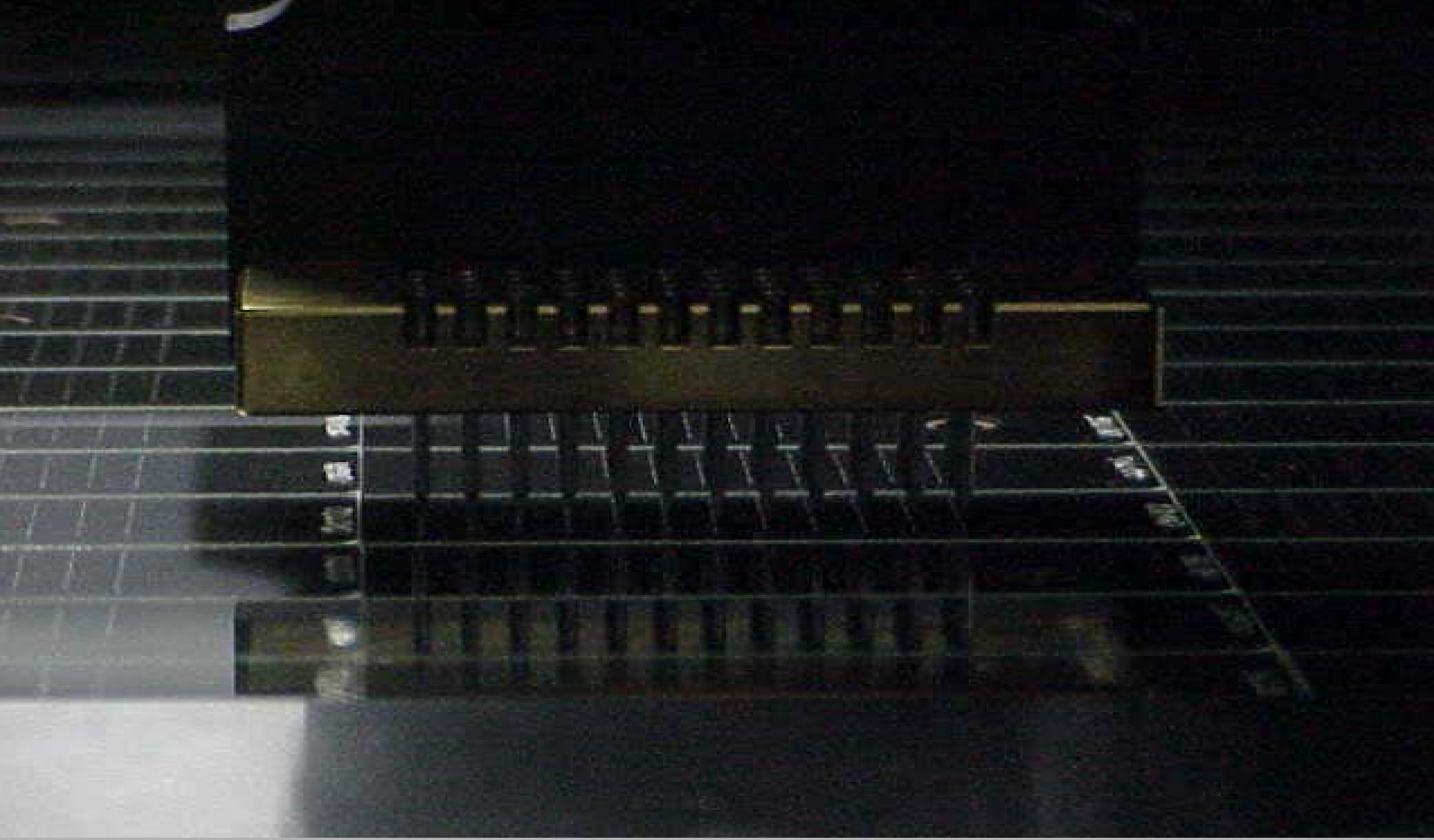
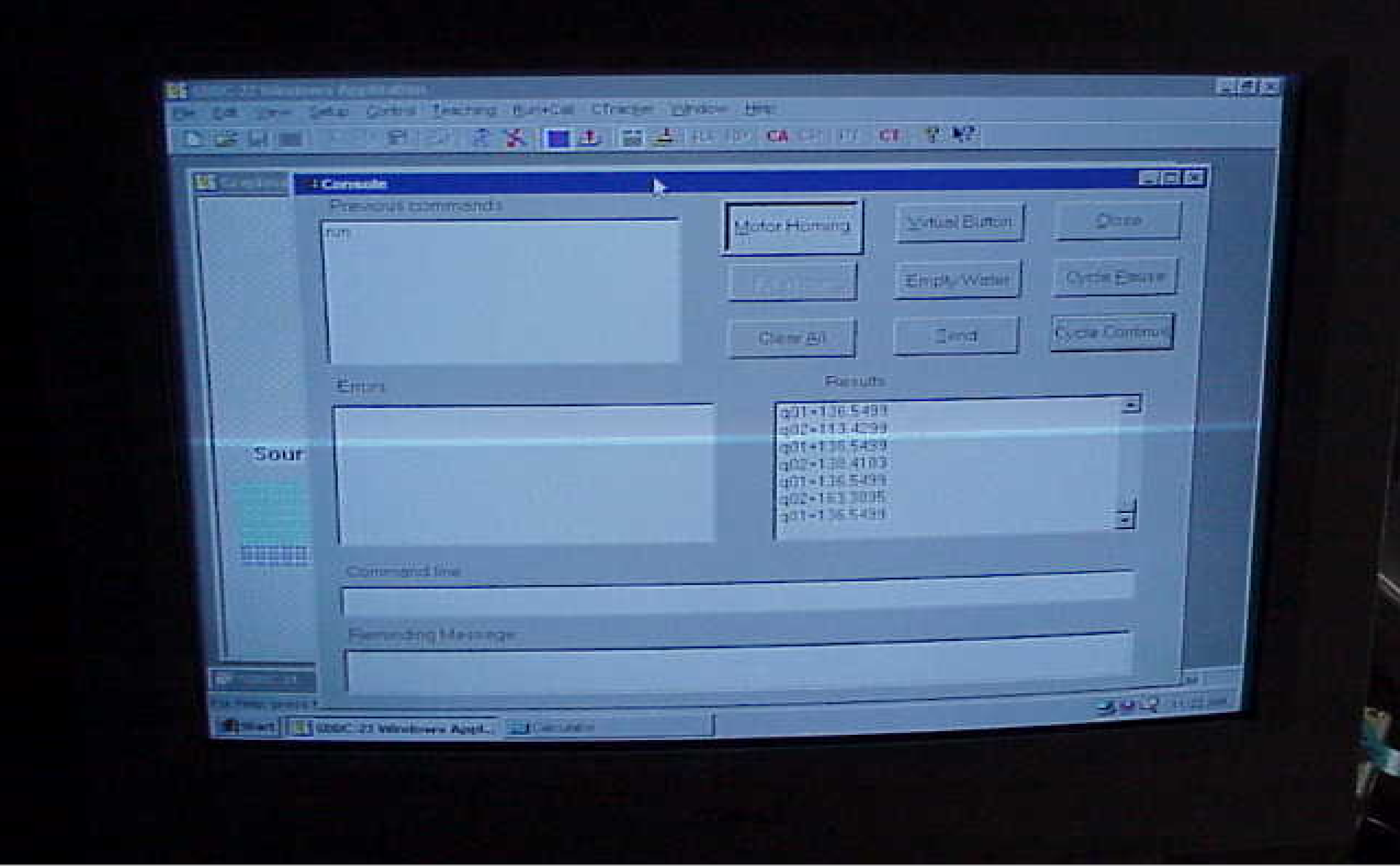


Figure 2. This shows the print head with 48 pins. Each pin is spring loaded and has a small slit in it which draws up the DNA solution by capillary action. You can see the rows of spotted DNA on the glass slides to the left of the print head; they appear as small grey lines. Each slide is numbered for identification.



```
q01=136.5499
q02=113.4299
q01=136.5499
q02=138.4103
q01=136.5499
q02=163.3095
q01=136.5499
```

Figure 3. The robot is controlled by a computer with a graphic user interface (GUI) similar to a web page.

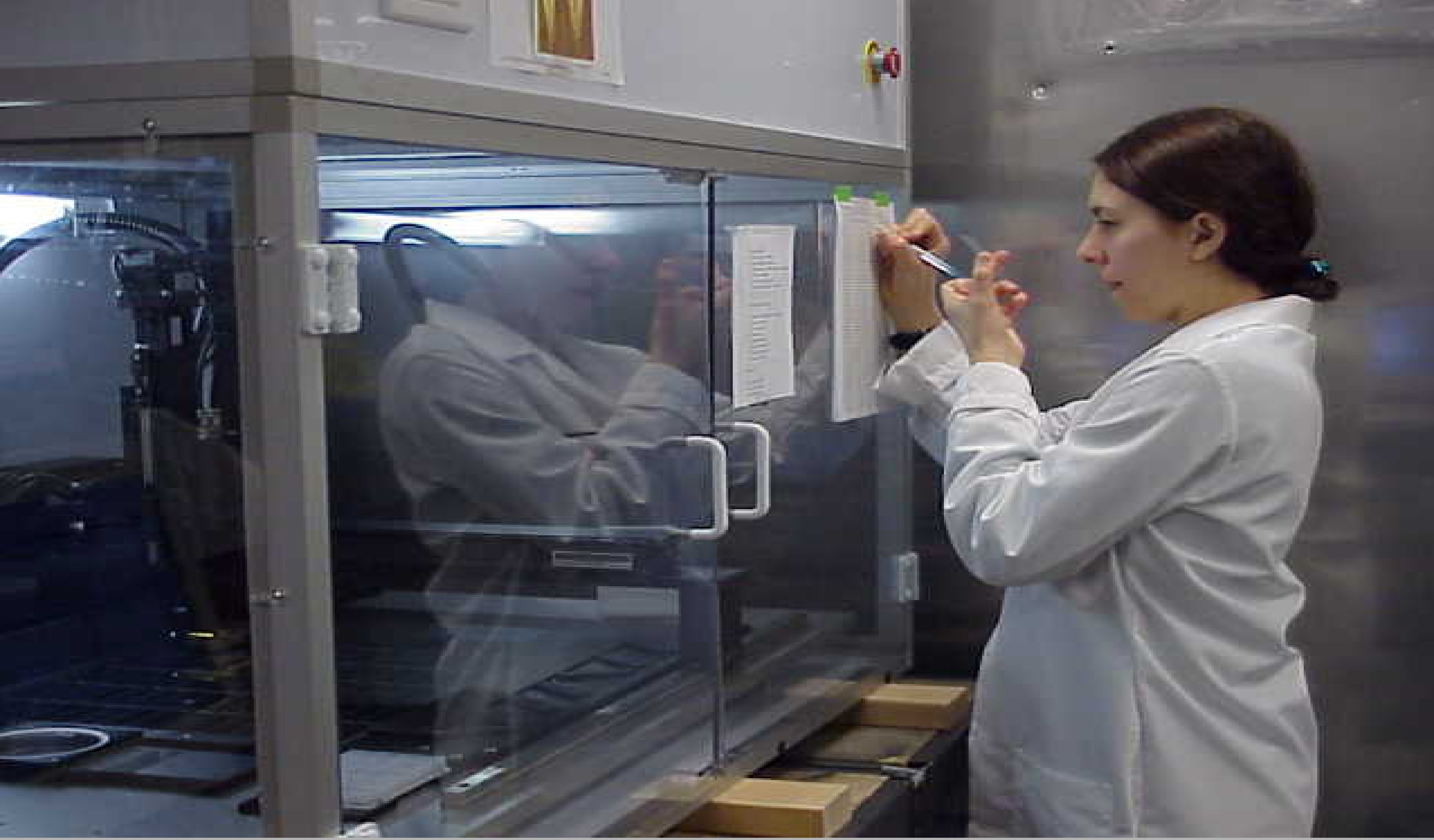


Figure 4. The production of DNA microarrays is automated but it takes a lot of personal attention to details to make sure the robot is peforming properly the progress of the robot as it is producing about 100 human DNA chips with 40,000 spots on each chip.

Ethics

Possessing detailed knowledge about the genetic makeup of individuals raises several complex ethical quandaries. How confidential should genetic information be? How should privacy concerns be weighed against other interests? If genetic information related to disease genes should be as confidential as any other health-related information, should there be databases of detailed genomic information on individuals?

Even without detailed genomic databases, thirteen genetic markers are sufficient for the FBI to identify every person except identical twins. Should this type of genetic information be stored on all convicted criminals; everyone arrested for a crime; or on every individual, regardless of his or her past?

Who should have access to detailed genetic information if it becomes available? Should it be accessible to law enforcement officers, physicians, research scientists, employers and potential employers, or insurance providers?

The NIH-DOE Working Group on the Ethical, Legal, and Social Implications (ELSI) has recommended that employers can request and use genetic information, but only to protect the health and safety of workers; such information must remain confidential. They also recommend that insurers cannot use genetic information to deny or limit health insurance coverage or to charge different fees based on this information. Overall, the focus of legislation should be to prevent discrimination of individuals based on genetic information.

In 1993, long before the human genome was completed, a committee of the Institute of Medicine of the National Academy of Sciences developed recommendations to prevent involuntary genetic testing and protect confidentiality. They concluded that the responsible use of genetic testing requires that individuals understand the tests, their significance, and their implications. Testing for diseases should be done only when individuals are capable of providing informed consent. This means not only that individuals must be informed, but that they also should understand the implications of that consent.

Patenting of human genes is another ethical concern emerging from the human genome project. In order to be patentable under the U.S. Federal Patent Act, an invention must be "novel, nonobvious, and have utility." In applying for a patent on a human gene, applicants generally claim that the patent's holders will add to the utility of the natural gene by developing tests and therapies to fight diseases associated with that gene. Opponents of gene patenting think that patents will limit the ability of other scientists to do additional research on these genes.

Most patents are filed by private companies that plan to develop and market diagnostic tests and treatments that come from their research on a particular gene. These companies feel that, without a patent, they cannot afford to do the research that will lead to useful products. They argue they need the protection of a patent before they can invest millions of dollars in the development of new tests, drugs, and therapies. Some scientists counter that companies tend to patent genes even before they know what the gene does, so it is hard to understand how they can claim that they will increase the utility

of such a gene.

Epilogue

The explosion of information coming from the sequencing of genomes has changed the landscape of biology. We now have tools to better understand the basis of disease and its prevention and control. These tools also allow us to design more effective drugs, and even understand the genetic relationships among all living things that make the universal tree of life. Acquiring the sequence was only the beginning.

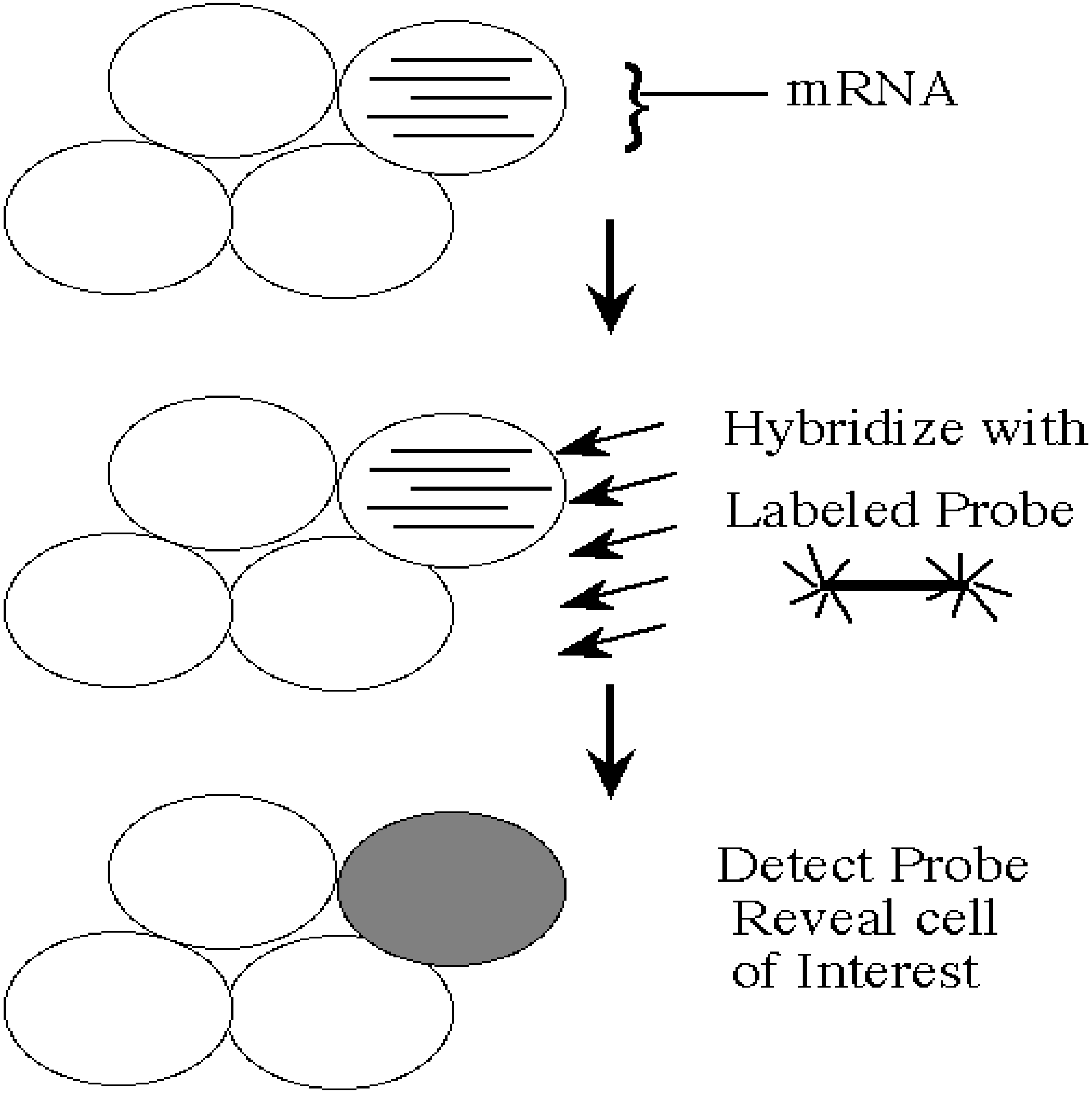
In Situ Hybridization Method

In situ hybridization allows you to visualize where a particular mRNA is located inside a tissue .
In situ **means in the location, so this method allows the investigator to see the normal location for the mRNA of interest.**

In top panel of the diagram, one of four cells in a tissue is transcribing the gene of interest, as shown by the lines.
A single-stranded DNA probe (middle panel) is linked to an enzyme and allowed to base pair with the mRNA. After a series of washes, only probes that are correctly base paired with the target mRNA remains in the tissue.

A colorless substrate is incubated with the probe and the substrate is enzymatically converted to a colored product that is easily visualized. Cells that contain the colored product indicate where your mRNA of interest was expressed.

Four cells, one expressing mRNA of interest.



mRNA

Hybridize with
Labeled Probe

Detect Probe
Reveal cell
of Interest

In Situ Polymerase Chain Reaction

Principles and Methods

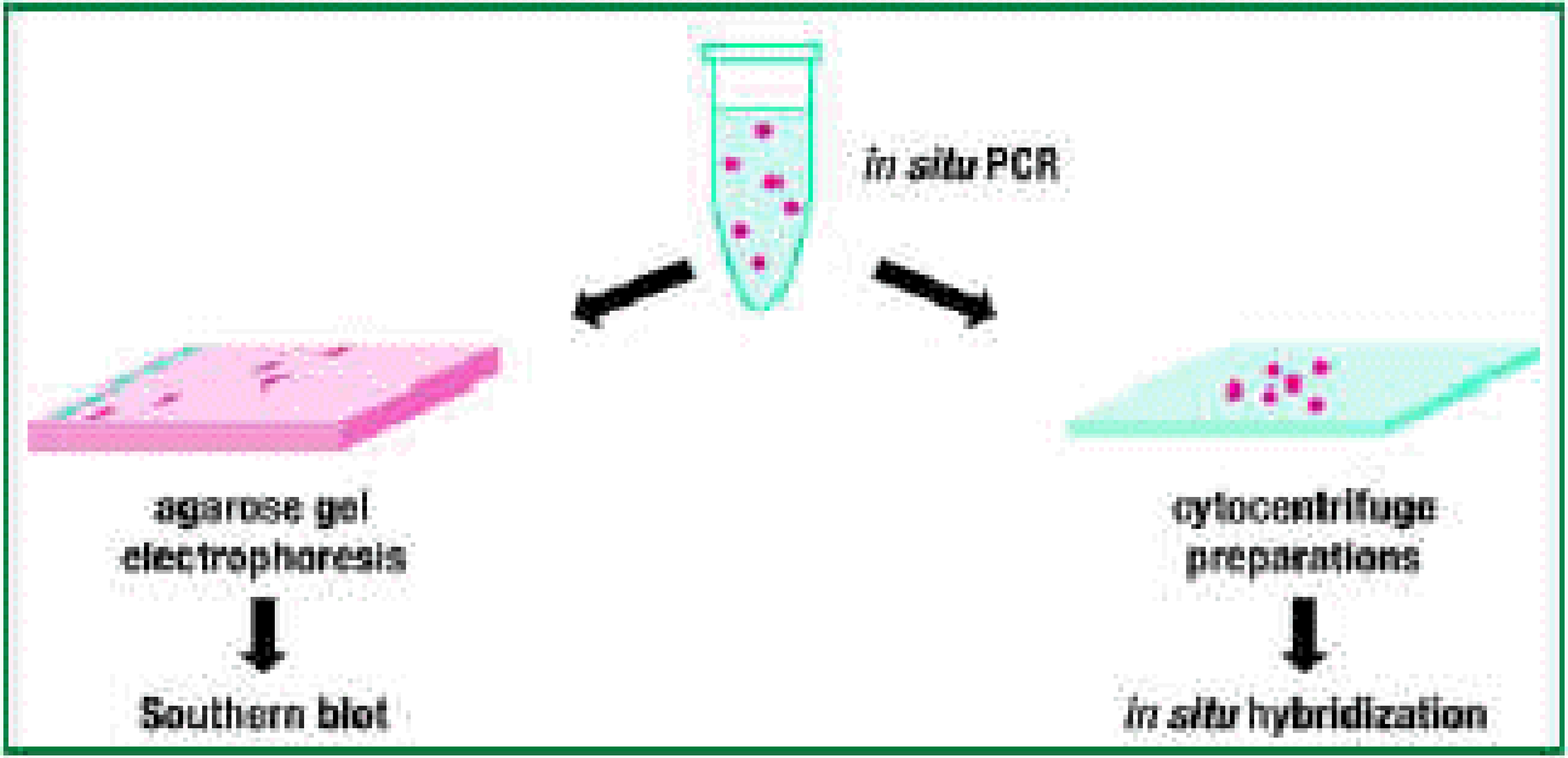
Experimental protocols for successful *in situ* PCR include fixation and permeabilization during sample preparation, a mechanism for thermal cycling cellular material in solution or on glass slides and a means to detect the amplificants .

Prior to *in situ* PCR, cells or tissue samples are fixed and permeabilized to preserve morphology and permit access of the PCR reagents to the intracellular sequences to be amplified. PCR amplification of target sequences is next performed either in intact **cells held in suspension** in micro-Eppendorf tubes or directly in cytocentrifuge preparations or tissue sections **on glass slides** . In the former approach, fixed cells suspended in the PCR reaction mixture are thermal cycled in micro-Eppendorf tubes using conventional block cyclers.

After PCR the cells are cytocentrifugated onto glass slides with visualization of intracellular PCR products by ISH or immunohistochemistry.

In situ PCR on glass slides is performed by overlaying the samples with the PCR mixture under a coverslip which is then sealed with nail polishor mineral oil to prevent evaporation of the reaction mixture. Thermal cycling is achieved by placing the glass slides directly on top of the heating block of a specially designed thermal cycler .

Detection of intracellular PCR-products is achieved by one of two entirely different techniques: (i) indirectly by ISH with PCR-product specific probes **(indirect *in situ* PCR)**, or (ii) without ISH through direct detection of labeled nucleotides (digoxigenin-11-dUTP, fluorescein-dUTP, 3H-CTP or biotin-16-dUTP) which have been incorporated into PCR products during thermal cycling **(direct *in situ* PCR)** (Fig. 92)



Principles of *in situ* PCR performed in cells in suspension.

Applications

In situ PCR has a number of potential research and diagnostic applications. To date several different groups have reported successful *in situ* PCR detection of single copy nucleic acid sequences in cell preparations and also of low copy DNA or RNA sequences in tissue sections . Most of the studies to date have focused on the detection of viral or proviral DNA, but *in situ* PCR has also been applied to the study of endogenous DNA sequences including human single copy genes, rearranged cellular genes and chromosomal translocations and to map low copy number genomic sequences in metaphase chromosomes .

Pitfalls and Limitations

Direct *in situ* PCR, a rapid alternative to indirect *in situ* PCR for DNA and RNA detection has proved unreliable in the hands of most workers about the specificity of the results obtained. Even when controlled fixation, controlled protease digestion and ”hot-start procedures” are used, the direct detection approach yields very significant false positive results, especially when working with tissue sections .

The false positive signals result mainly from non-specific incorporation of labeled nucleotides into fragmented endogenous DNA undergoing ”repair” by the DNA polymerase **(”DNA-repair artifacts”)** or by priming of non-specific PCR products by cDNA or DNA fragments **(”endogenous priming”)** .

Future developments may allow for a modification of this position but for now caution and adequate use of appropriate controls are urged in the interpretation of data generated by the direct detection approach .

The indirect method, by using probes which recognize the amplified sequences, provides maximum specificity in detection of intracellular PCR products and is the approach that most investigators now use. At a theoretical level, probes targeted to regions in between the primers represent the ideal for reasons of specificity. The overall efficiency of *in situ* DNA amplification appears to be low.

It has been estimated to be approx. 50 fold after 30 cycles in suspended cell and probably even lower in cytospin preparations and tissue sections. Many factors may be contributing to this low amplification efficiency including certain deleterious effects of the aldehyde based cross-linking fixatives such as histone cross-linking to DNA and single stranded breaks in DNA as well as sequestration of DNA polymerases and other reagents in the surface of silanized slides .

Diffusion artifacts” represent a significant problem of both indirect and direct *in situ* PCR performed in cells in suspension . PCR products and DNA leak out of template positive cells and serve as templates for extra-cellular amplification which is probably far more efficient than intracellular amplification. These extracellularly amplified DNA sequences have the potential to adhere to the surface of adjacent template negative cells or perhaps to even diffuse into them resulting in false positive signals .

Rigorous and logical controls must be employed to avoid the many dangerous pitfalls that have been identified . Whenever possible, new data obtained by this technology should be confirmed by another method or at least corroborated by parallel investigations.

Obstacles with *in situ* PCR have been low amplification efficiency, poor reproducibility and difficulty in quantitation of results .
Quantitation remains a limitation and *in situ* PCR remains at best comparative or semi-quantitative .

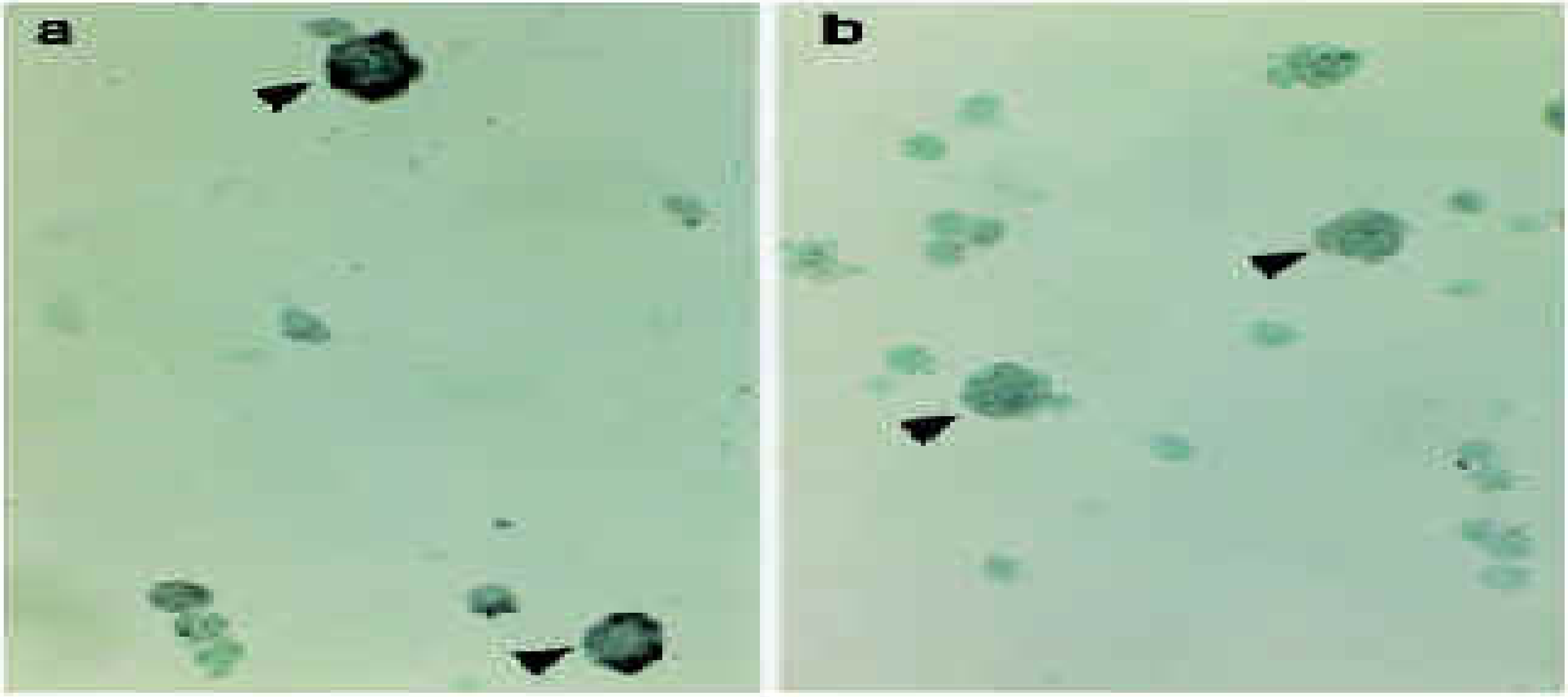
The range of potential applications is large. Our ability to detect latent viruses in low copy number represents a huge step forward in leading to our understanding of viral disease pathogenesis . We will now have the ability to study the time course and determinants of viral activation as well as the morphologic correlates of that activation.

Genetic determinants of tumorigenesis including DNA mutations or chromosomal translocations will be studied to understand cause and effect and the latency period between DNA alterations and morphologic characteristics of malignancy.

Furthermore, the detection of intracellular mRNAs by *in situ* RT-PCR has applications in any situation where the level of gene expression is below that detectable by ISH.

In conclusion, *in situ* PCR is an exciting technique that is already providing a mechanism to gain insight into disease pathogenesis.

Its true strengths and weaknesses are being increasingly understood and further refinements of the methodology
to render it more reliable are steadily emerging.



Detection of VH$_3$-rearrangement-specific messenger RNA (mRNA) in clonal B cells using *in situ* reverse-transcriptase (RT) followed by *in situ* PCR amplification of cDNA.

Digoxigenin-labeled dUTP was incorporated into PCR products (direct *in situ* RT-PCR). A positive signal in the cytoplasm after *in situ* amplification is only observed in the larger clonal cells (arrow heads) but not in smaller negative cells **(a).** The signal is absent (arrow heads) in the control where RNase digestion was performed prior to *in situ* RT-PCR **(b)**.

Single-Strand Conformation Polymorphism (SSCP)

DEFINITION

- SSCP is the electrophoretic separation of single-stranded nucleic acids based on subtle differences in sequence (often a single base pair) which results in a different secondary structure and a measurable difference in mobility through a gel.

BACKGROUND

- **The mobility of double-stranded DNA in gel electrophoresis is dependent on strand size and length but is relatively independent of the particular nucleotide sequence. The mobility of single strands, however, is noticeably affected by very small changes in sequence, possibly one changed nucleotide out of several hundred. Small changes are noticeable because of the relatively unstable nature of single-stranded DNA; in the absence of a complementary strand, the single strand may experience intrastrand base pairing, resulting in loops and folds that give the single strand a unique 3D structure, regardless of its length. A single nucleotide change could dramatically affect the strand's mobility through a gel by altering the intrastrand base pairing and its resulting 3D conformation**

•Like restriction fragment length polymorphisms (RFLPs), SSCPs are allelic variants of inherited, genetic traits that can be used as genetic markers. Unlike RFLP analysis, however, SSCP analysis can detect DNA polymorphisms and mutations at multiple places in DNA fragments.

• As a mutation scanning technique, though, SSCP is more often used to analyze the polymorphisms at single loci, especially when used for medical diagnoses .

Most experiments involving SSCP are designed to evaluate polymorphisms at single loci and compare the results from different individuals.

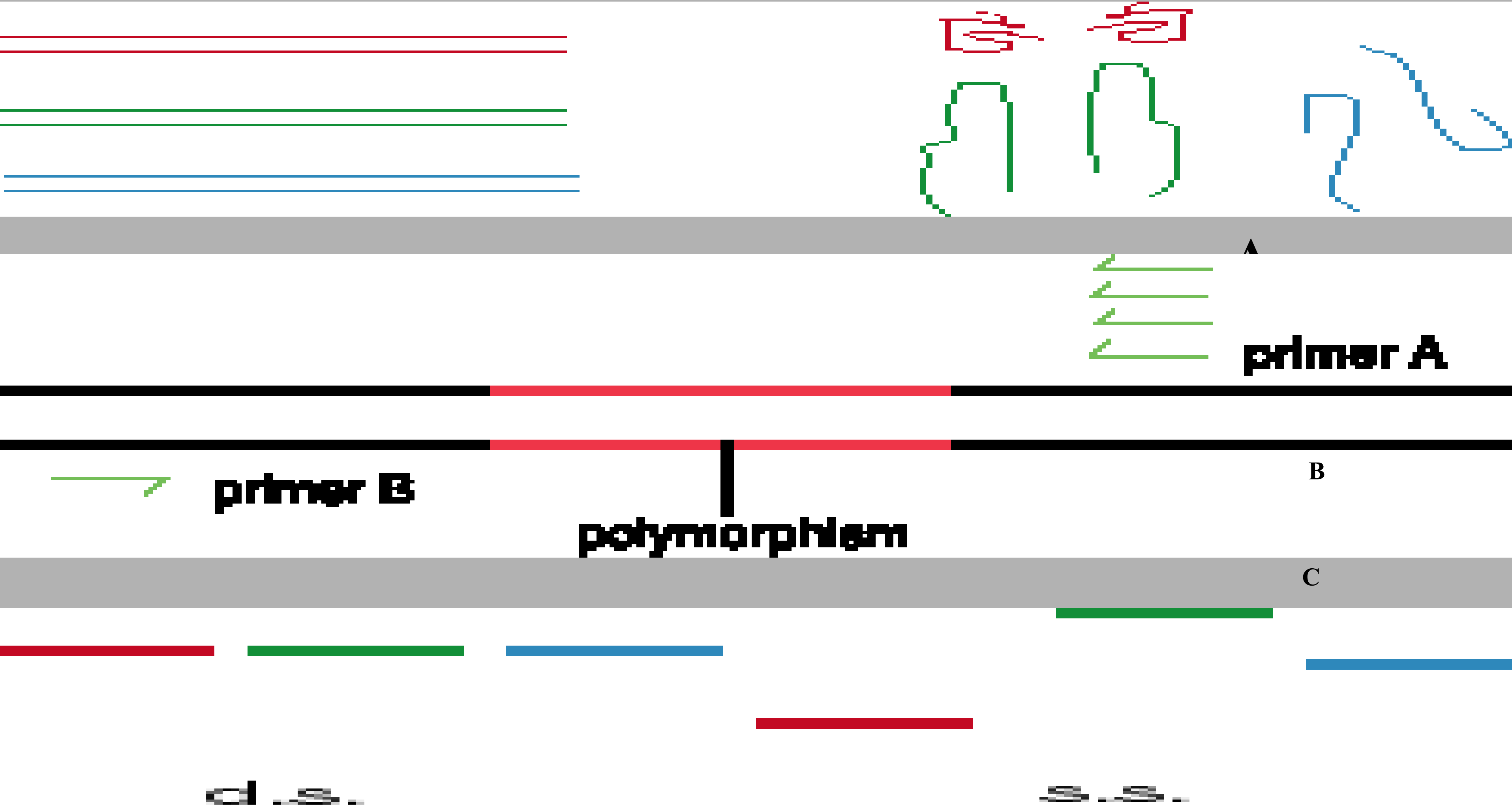


Figure 1: SSCP Procedure.

The three equal-length double-stranded DNA fragments are shown with the corresponding single-stranded structures, the red fragment folding into the smallest molecule and the green the largest (Panel A). The desired polymorphism is selected with PCR primers; primer A is in excess to amplify only a single strand (Panel B). Both the double-stranded and single-stranded fragments are run through gel electrophoresis (Panel C). If not for the color labels, there would be no distinction between the double-stranded fragments. The single-stranded fragments, however, show considerable variation in mobility; the small red molecule migrates more quickly through the gel than either the blue or the large green molecule. Using this SSCP result, it becomes clear that the different lanes (red, blue, or green) contain strands with different sequences; the more far apart the bands, the less similar the nucleotide sequences

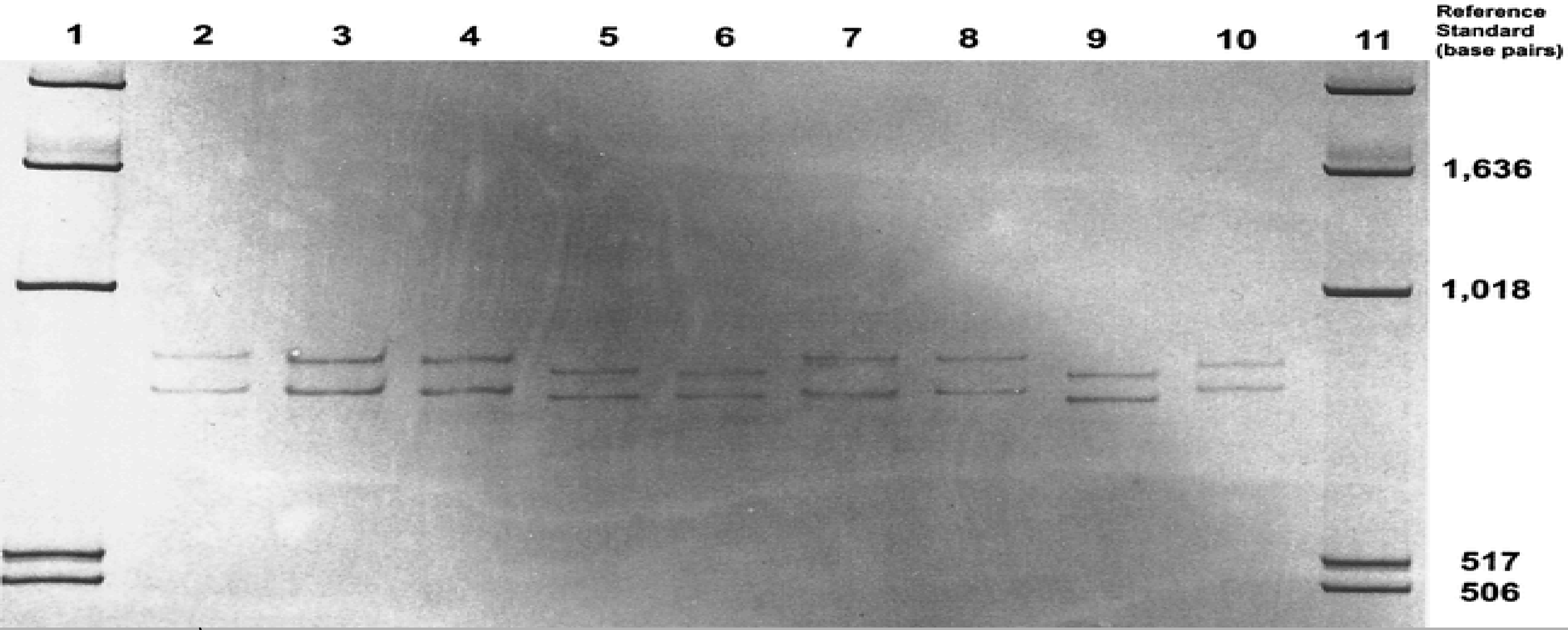
- **Procedure as illustrated in Figure 1:**

1. A specific pair of PCR primers (forward and reverse) is used to amplify the desired DNA fragments from individuals.

2. Single-stranded DNA is produced by asymmetric PCR: the primer on one side of the fragment is greatly in excess over the other primer. After the low-concentration primer supply is exhausted, continued PCR produces only the target single strand.

3. The mobilities of the single stranded fragments are compared by electrophoresis on a neutral polyacrylamide gel.

4. Bands are detected by radioactive labeling or (more often) silver staining, and the pattern is interpreted



).

Figure 2: Sample SSCP Gel Result and Interpretation. **DNA was isolated and amplified from sand flies (*Lutzomyia longipalpis*). SCCP analysis of the DNA shows multiple haplotypes, or sets of alleles usually inherited as a unit. Lanes 3 and 4 were identical haplotypes from two individuals. The difference in band migration in adjacent lanes is associated with the number of nucleotide differences (in parentheses): lanes 2-3 (2), lanes 3-4 (0), lanes 4-5 (3), lanes 5-6 (1), lanes 6-7 (3), lanes 7-8 (1), lanes 8-9 (1), and lanes 9-10 (4).**

SSCP LIMITATIONS AND CONSIDERATIONS

- Single-stranded DNA mobilities are dependent on temperature. For best results, gel electrophoresis must be run in a constant temperature.
- Sensitivity of SSCP is affected by pH. Double-stranded DNA fragments are usually denatured by exposure to basic conditions: a high pH.
- Kukita et al. found that adding glycerol to the polyacrylamide gel lowers the pH of the electrophoresis buffer--more specifically, the Tris-borate buffer--and the result is increased SSCP sensitivity and clearer data.

- **Fragment length also affects SSCP analysis. For optimal results, DNA fragment size should fall within the range of 150 to 300 bp.**

- **The presence of glycerol in the gel may also allow a larger DNA fragment size at acceptable sensitivity .**

- **Under optimal conditions, approximately 80 to 90% of the potential base exchanges are detectable by SSCP.**

- **If the specific nucleotide responsible for the mobility difference is known, a similar technique called Single Nucleotide Polymorphism (SNP) may be applied**.

Pulsed Field Electrophoresis for Separation of Large DNA

In 1984, Schwartz and Cantor described pulsed field gel electrophoresis (PFGE), introducing a new way to separate DNA.

In particular, PFGE resolved extremely large DNA for the first time, raising the upper size limit of DNA separation in agarose from **30-50 kb** . **DNA above 30-50** kb migrates with the same mobility regardless of size. This is seen in a gel as a single large diffuse band.

If, however, the DNA is forced to change direction during electrophoresis, different sized fragments within this diffuse band begin to separate from each other.
With each reorientation of the electric field relative to the gel, smaller sized DNA will begin moving in the new direction more quickly than the larger DNA. Thus, the larger DNA lags behind, providing a separation from the smaller DNA.

Instrumentation

Although many types of PFGE instrumentation are available (fig. 1), they generally fall into two categories.

The simplest equipment is designed for field inversion gel electrophoresis **(FIGE)**. FIGE works by periodically inverting the polarity of the electrodes during electrophoresis. Because FIGE subjects DNA to **a 180ψ** reorientation, the DNA spends a certain amount of time moving backwards. Only an electrical field switching module is needed; any standard vertical or horizontal gel box that has temperature control can be used to run the gel.

The other category contains instruments that reorient the DNA at smaller oblique angle, **generally between 96 and 120ψ.** This causes DNA to always move forward in **a zigzag** pattern down the gel. For a similar size range under optimal conditions, these separations are faster, resolve a wider size range compared to FIGE. Although extremely useful for separating relatively small DNA, 4- 1,000 kb (fig. 2), FIGE's reorientation angle of 180ψ results in a separation range most useful under **2,000 kb.**

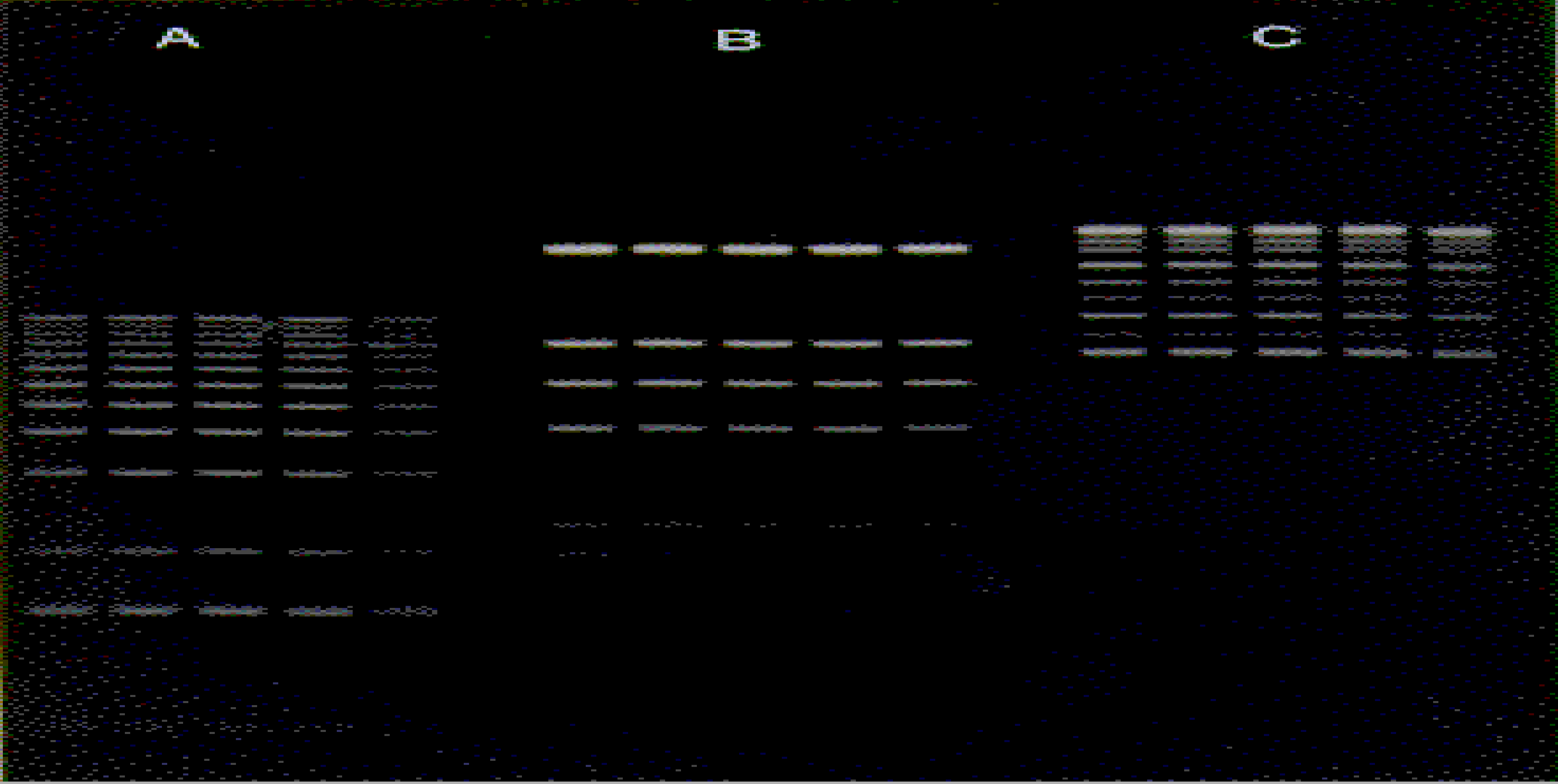
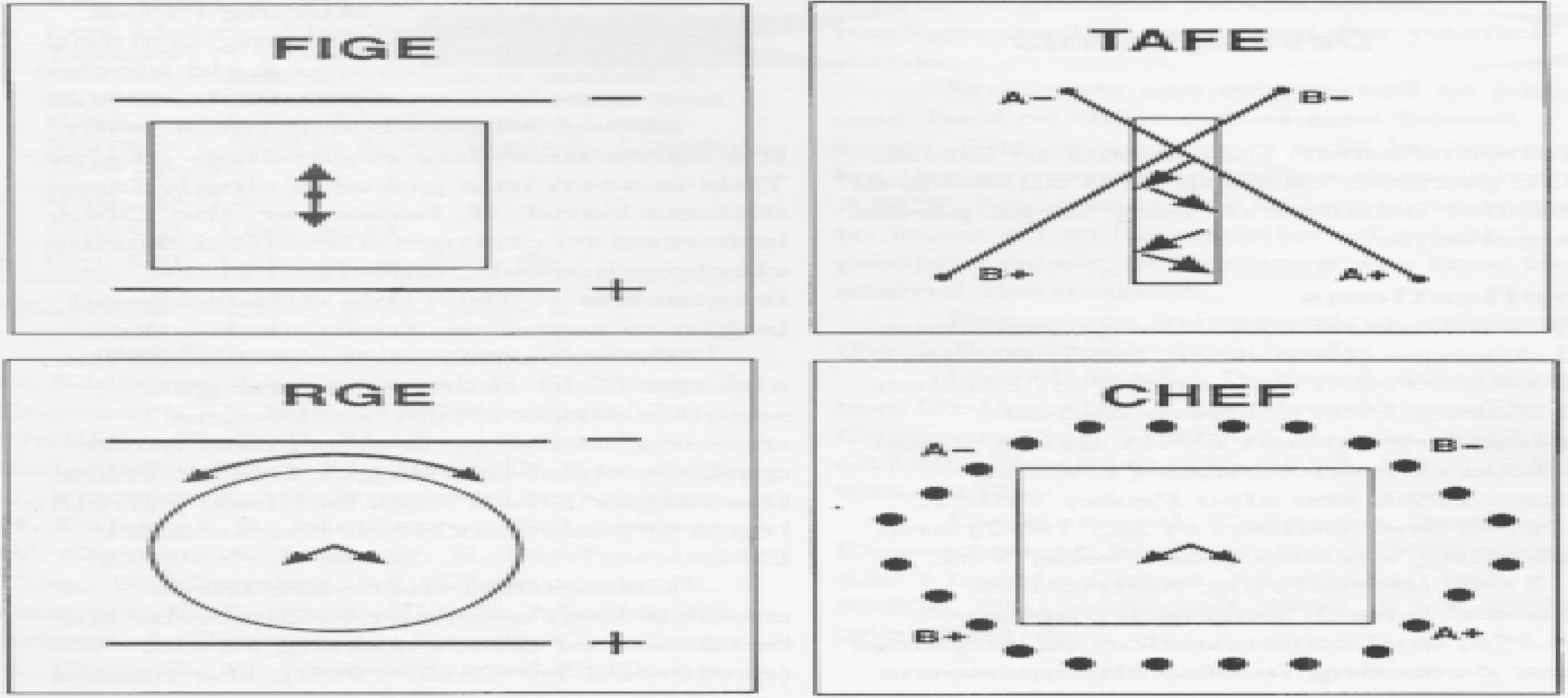


Figure 2. *Increased separation of the 20-50 kb range with field inversion gel electrophoresis (FIGE). Run conditions: 230 V, 7.9 V/cm, 16 hrs., 50 msec. pulse, forward:reverse pulse ratio = 2.5:1, 1% GTG agarose, 0.5X TBE, 10 C.a) 1 kb ladder, 0.5-12 kb; b) Lambda/Hind III, 0.5-23 kb; and c) High molecular weight markers, 8.3-48.5 kb.*



TAFE use a complicated geometry between the electrodes and a vertically placed gel to get straight lanes. **CHEF and RGE** maintain a homogeneous electric field in combination with a **horizontal gel**. **CHEF changes the direction of the electric field electronically** to reorient the DNA by changing the polarity of an electrode array. With **RGE** the electric field is fixed; to move the DNA in a new direction, the **gel simply rotates.**

Rotating Gel Electrophoresis

RGE is one of the most recent commercial introductions of pulsed field equipment and combines variable angles with a homogeneous electric field (figs. 3 and 4) . The electrodes are positioned along opposite sides of the buffer chamber with their polarity fixed. Briefly, the gel is cast on a circular running plate and then placed in the buffer chamber.

Separation Parameters

The minimum amount of information needed to repeat a separation should include a short description of the pulsed field instrumentation used; **applied voltage and field strength (e.g., 180 V at 5.3 V/cm); pulse length (e.g., 87 seconds); reorientation angle (e.g., 120ψ); the buffer (0.5X TBE); the agarose type and concentration (SeaKem Gold, 1.1%); the buffer chamber temperature** ; **the type of standards (Clontech *S. cerevisiae*); and, if possible, the amount of DNA loaded**

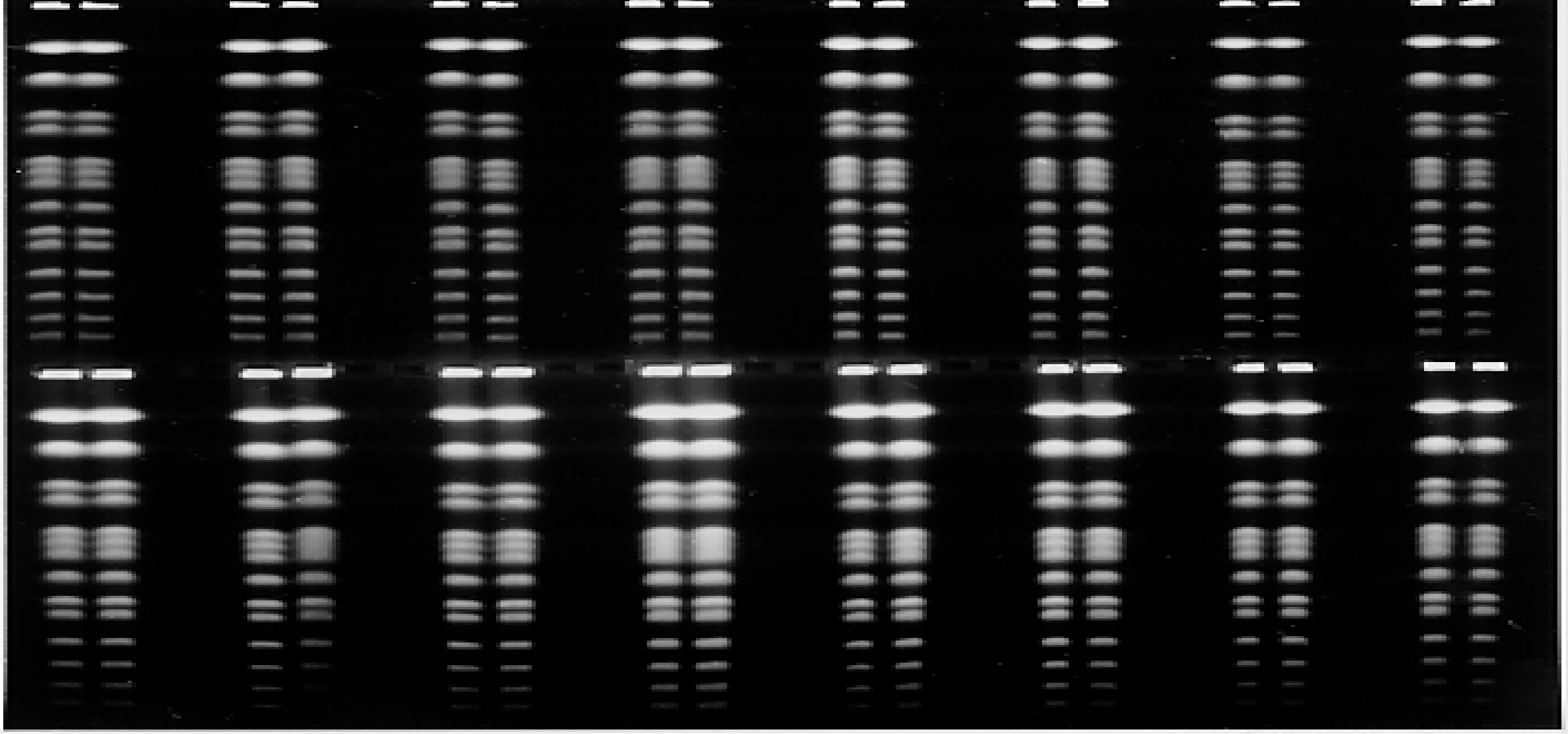


Figure 3. *Rotating gel electrophoresis (RGE) separation Saccharomyces cercevisiae chromosomes* ***(245-2190 kb).*** *Run conditions: 180 V, 5.1 V/cm, 34 hrs., 120 angle, 60-120 sec. pulse ramp, 0.5X TBE, 1.2% GTG agarose, 10 C. Two combs were used on the same gel to load 32 samples, a maximum of 72 are possible*

Separation Area

Most PFGE systems separate DNA over a relatively small area, limiting the resolution of complex samples. **RGE** is an exception to this, with a useful separation distance up to **20 cm** and a maximum gel size of 18 x 20 cm.

Field Strength

The field strength has a profound effect on pulsed field separations and is a compromise between separation time and resolution of a particular size class. **Four to six volts/cm is generally required for resolving DNA up to 2000 kb (e.g., *S. cerevisiae* chromosomes) in a reasonable period of time (e.g., 1-2 days).**

Pulse Time

Pulse time primarily changes the size range of separation. **Longer pulse times lead to separation of larger DNA. For example, at 5.4 V/cm, the 1.6 Mb and 2.2 Mb chromosomes from *S. cerevisiae* separate as a single band with 90-second pulse length. Increasing the pulse length to 120 seconds resolves these into two bands**

Reorientation Angle

The smaller the angle, however, the faster the DNA mobility. And for separating extremely **large DNA, 96 to 105ψ is almost a requirement to get a good separation in the shortest possible time.**

Buffers

Two buffers are commonly employed for PFGE--**TAE and TBE** (1x TAE is 40 mM Tris acetate, 1 mM EDTA, pH 8.0; 1x TBE is 89 mM Tris, 89 mM boric acid, 2 mM EDTA, pH 8.0). Both are used at a relatively **low ionic strength to prevent heating and carry** the designations of either **0.25 and 0.5x** to indicate the dilution relative to the standard concentration. An added benefit to low ionic strength buffers is an increase in DNA mobility. For example, while using RGE to compare various buffers and agaroses, lowering both **TAE and TBE to 0.25 x gave the maximum mobility (40-50% faster than 1x).**

Below 0.25x, the DNA mobility dropped off.

Agarose

The type of agarose also affects DNA separation, with the fastest mobilities and best resolution achieved in gels **made of low electroendosmosis (EEO)** agarose. Although most standard electrophoresis grades of agarose are suitable for PFGE (e.g., SeaKem GTG), agarose with minimal EEO will provide a faster separation. Several low EEO "pulsed field grades" are available, including FastLane and Gold (FMC BioProducts), and Megarose (Clontech).

The concentration of agarose affects both the resolution and mobility of the DNA . Higher concentrations of agarose yield sharper, but slower moving bands. And the typical concentrations used **(0.8-1.2%)** represent a tradeoff between speed and resolution.

Temperature

Because DNA mobility also depends on the separation temperature, the **temperature must be constant** both during and between runs.

Although higher temperatures increase DNA mobility, it does so at the expense of resolution

Conclusion

Since its introduction , PFGE has evolved into a routine technique for molecular biologists.

What does the future hold? Possibilities include using a new or improved separation material, and going beyond the current size limit of up 10 Mb.

DENATURING GRADIENT GEL ELECTROPHORESIS (DGGE)

THEORY

The theory behind DGGE is very simple, the two strands of a DNA molecule separate, or melt, when heat or a chemical denaturant is applied. The temperature at which a DNA duplex melts is influenced by two factors:

1. The hydrogen bonds formed between complimentary base pairs, GC rich regions melt at higher temperatures than regions that are AT rich.
2. The attraction between neighbouring bases of the same strand or "stacking"

Consequently, **a DNA molecule may have several melting domains with characteristic melting temperatures (Tm) determined by the nucleotide sequence.**

DGGE exploits the fact that otherwise identical DNA molecules, which differ by only one nucleotide within a low melting domain will have different melting temperatures. When separated by electrophoresis through a gradient of increasing **chemical denaturant (usually formamide and urea),** the mobility of the molecule is retarded at the concentration at which the DNA strands of low melt domain dissociate. The branched structure of the single stranded moiety of the molecule becomes entangled in the gel matrix and no further movement occurs. Complete strand separation is prevented by the presence of a high melting domain, which is usually artificially created at one end of the molecule by incorporation of **a GC clamp**. This is accomplished during PCR amplification using a **PCR primer with a 5' tail consisting of a sequence of 40 GC.**

PRELIMINARY PREPARATION: PRIMER DESIGN AND OPTIMISATION OF GEL RUNNING CONDITIONS.

DGGE is usually performed on PCR products, primers must carefully be chosen so that the region to be **screened for mutations has one or at the most two discrete melting domains** (excluding the GC clamp).

The GC clamp is usually positioned adjacent to the highest melting domain. Thus, full sequence data must be available so that a melt map of the molecule can be constructed, and primers can be designed to amplify a region of unit melting domain. The optimal gradient and gel running conditions must also be established.

Computer Generation of Melt Maps and Primer Design

The programs MELT87, MELT95 and MACMELT are used to generate a melt map from a known DNA sequence. The programs identify primer pairs that will amplify short segments of unit melting domain. Ideally the PCR products should be between **100 and 400bp in size**. The program predicts the effect on the melting temperature of the PCR product when a GC clamp is positioned at one of the four ends of the molecule.

The program SQHTX calculates the difference in melting temperature (TC) for the wild type molecule and any possible mutation and is used to determine the gel denaturing gradient concentration and run times for the PCR products.

Alternatively, optimal gradient concentrations can be determined empirically by performing perpendicular gel electrophoresis.

In such an experiment the denaturing gradient is perpendicular to the direction of electrophoresis.

Optimisation of Gel Running Conditions

The computer programs described above reduce the number of preliminary experiments required for optimisation of the gel running conditions. However, it is still necessary to run some preliminary gels to determine the optimal voltages and running times and to confirm the optimal denaturing gradient that has been chosen.

The aim is to have well separated bands (normal and mutation positive control are simultaneously loaded on the gels) which are "focused" by the gradient.

PCR products with two low melting domains require different gel conditions for the analysis of each domain.

When optimised gel running conditions have been established the method can be used for mutation screening.

Detection Rate and Sensitivity

The detection rate is very high, in many cases it approaches 100%. It has been possible to identify a mutation present in a 100bp sequence at the level of 0.5%. Thus, DGGE is an ideal choice for mutation detection in the diagnostic laboratory

ADVANTAGES of DGGE

1. High detection rate and sensitivity.
2. The methodology is simple and a non-radioactive detection method is used.
3. PCR fragments may be isolated from the gel and used in sequencing reactions.

DISADVANTAGES of DGGE

1. Computer analysis and preliminary experiments are essential when setting up DGGE .
2. Purchase of DGGE equipment may be required.
3. Primers are more expensive because of the 40 bases of GC clamp.

4.Additional primers may be required for sequencing?

5.Analysis of PCR fragments over 400bp is less successful.

6.Genes which are exceptionally GC rich are not easily analysed by DGGE.

VARIATIONS OF DGGE

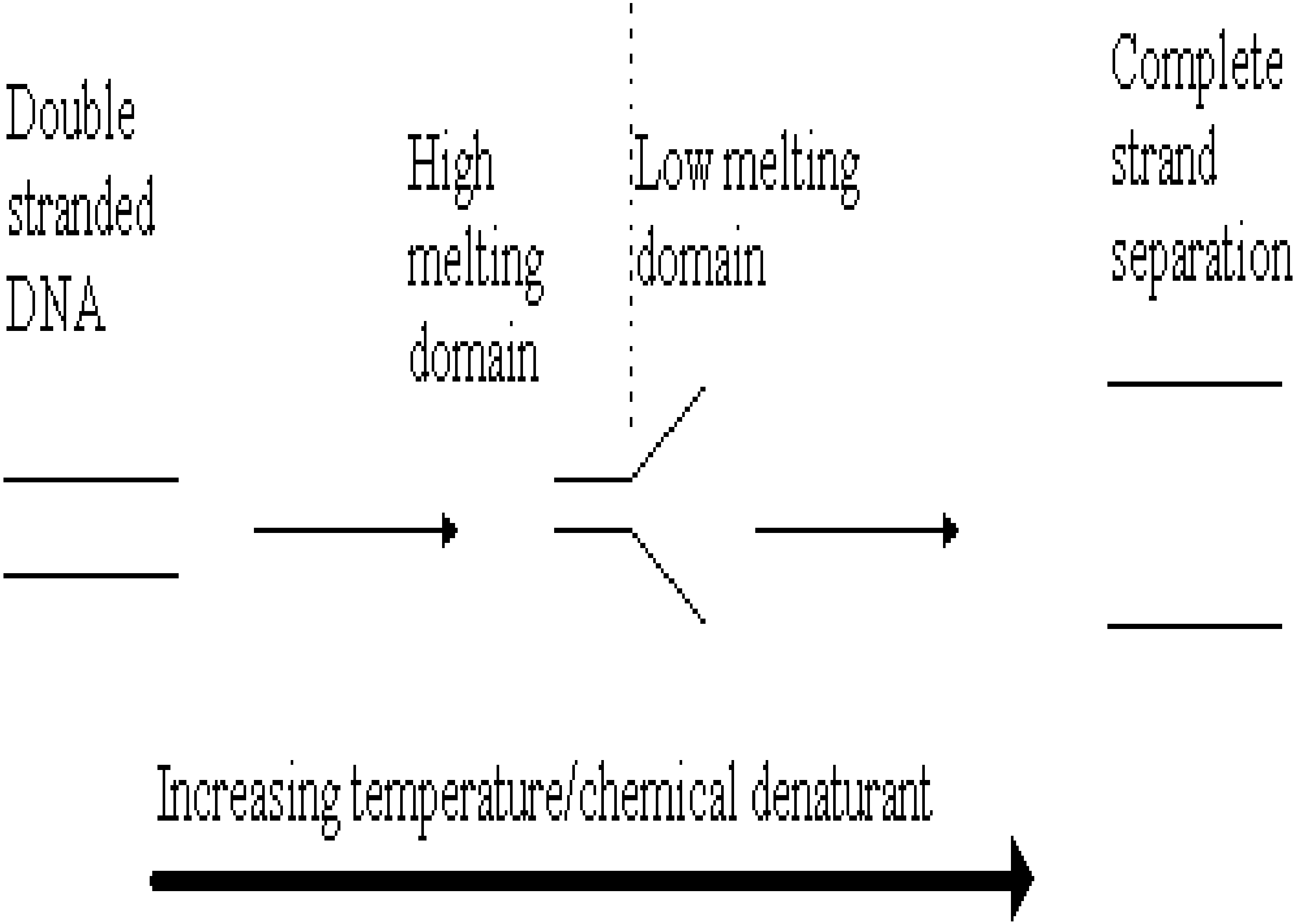
TEMPERATURE GRADIENT GEL ELECTROPHORESIS (TGGE)

The chemical denaturant gradient is replaced by a gradient of increasing temperature down the gel.

CONSTANT DENATURANT GEL ELECTROPHORESIS (CDGE)

The chemical denaturant is at a constant concentration throughout the gel, equivalent to the melting temperature of the low melting domain. This approach requires different gel conditions for each PCR fragment to be analysed. The main application of CDGE is limited to the identification of known mutations.

DNA MOLECULES ANNEAL IN DISCRETE DOMAINS



Double stranded DNA

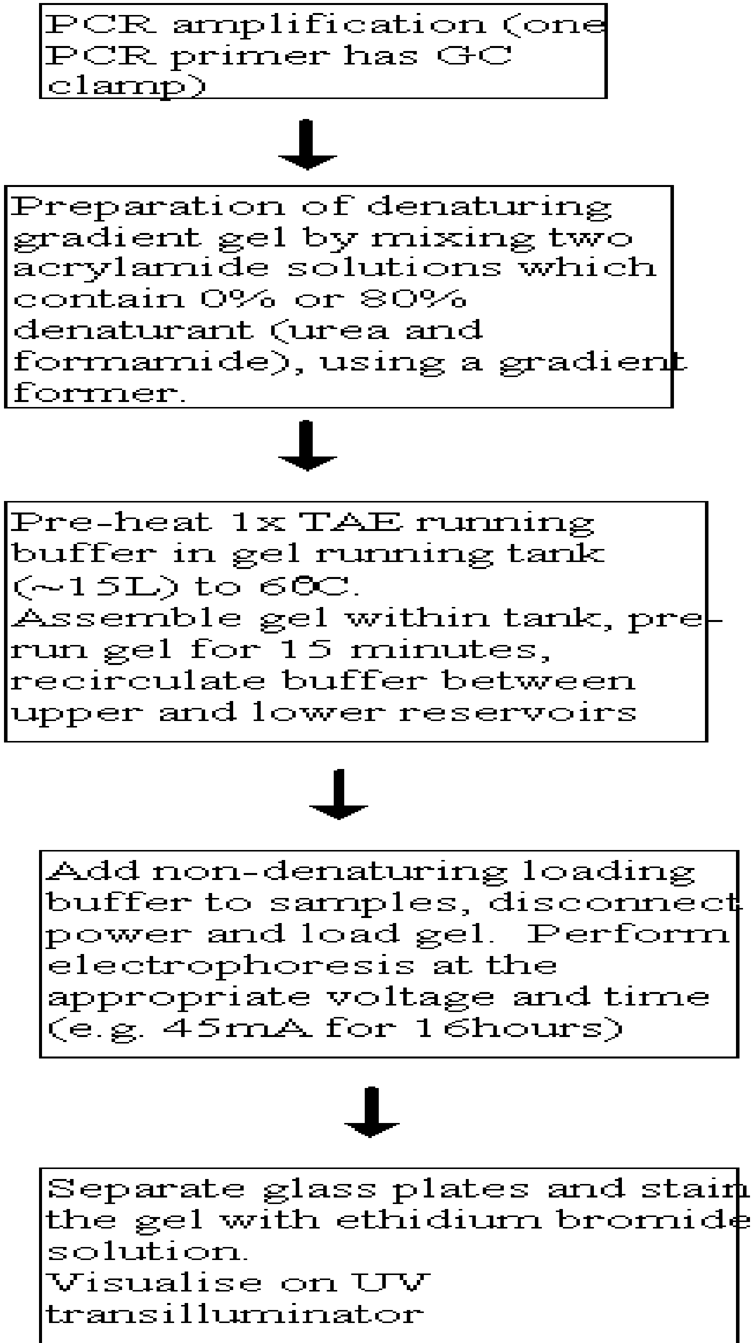
High melting domain

Low melting domain

Complete strand separation

Increasing temperature/chemical denaturant

A SIMPLE GUIDE TO THE STEPS INVOLVED IN DGGE



PCR amplification (one PCR primer has GC clamp)

Preparation of denaturing gradient gel by mixing two acrylamide solutions which contain 0% or 80% denaturant (urea and formamide), using a gradient former.

Pre-heat 1x TAE running buffer in gel running tank (~15L) to 60C.
Assemble gel within tank, pre-run gel for 15 minutes, recirculate buffer between upper and lower reservoirs

Add non-denaturing loading buffer to samples, disconnect power and load gel. Perform electrophoresis at the appropriate voltage and time (e.g. 45mA for 16hours)

Separate glass plates and stain the gel with ethidium bromide solution.
Visualise on UV transilluminator

Heteroduplex Analysis

Mutations are detected by *heteroduplex analysis* based on the retardation of the heteroduplex compared with the corresponding homoduplex on a non-denaturing polyacrylamide gel. Heteroduplexes migrate more slowly than their corresponding homoduplexes due to a more open double-stranded configuration surrounding the mismatched bases.

Basic Protocol

Heteroduplexes are formed by mixing wild-type and mutant DNA amplified by PCR. The samples are denatured and 're-annealed' (usually by heating and cooling).

Four distinct species are generated by this reassortment:

wild-type homoduplex,
mutant homoduplex,
and two heteroduplexes.

The formation of the heteroduplexes and their stability depend primarily on the type of mutation in the fragment. Single-base changes are more sensitive to temperature, solvents, and ionic strength of the buffer. There is no way to predict the influence of these parameters on the stability of the heteroduplex, and thus electrophoretic conditions must be optimized ,(solvents, and ionic strength of the buffer).

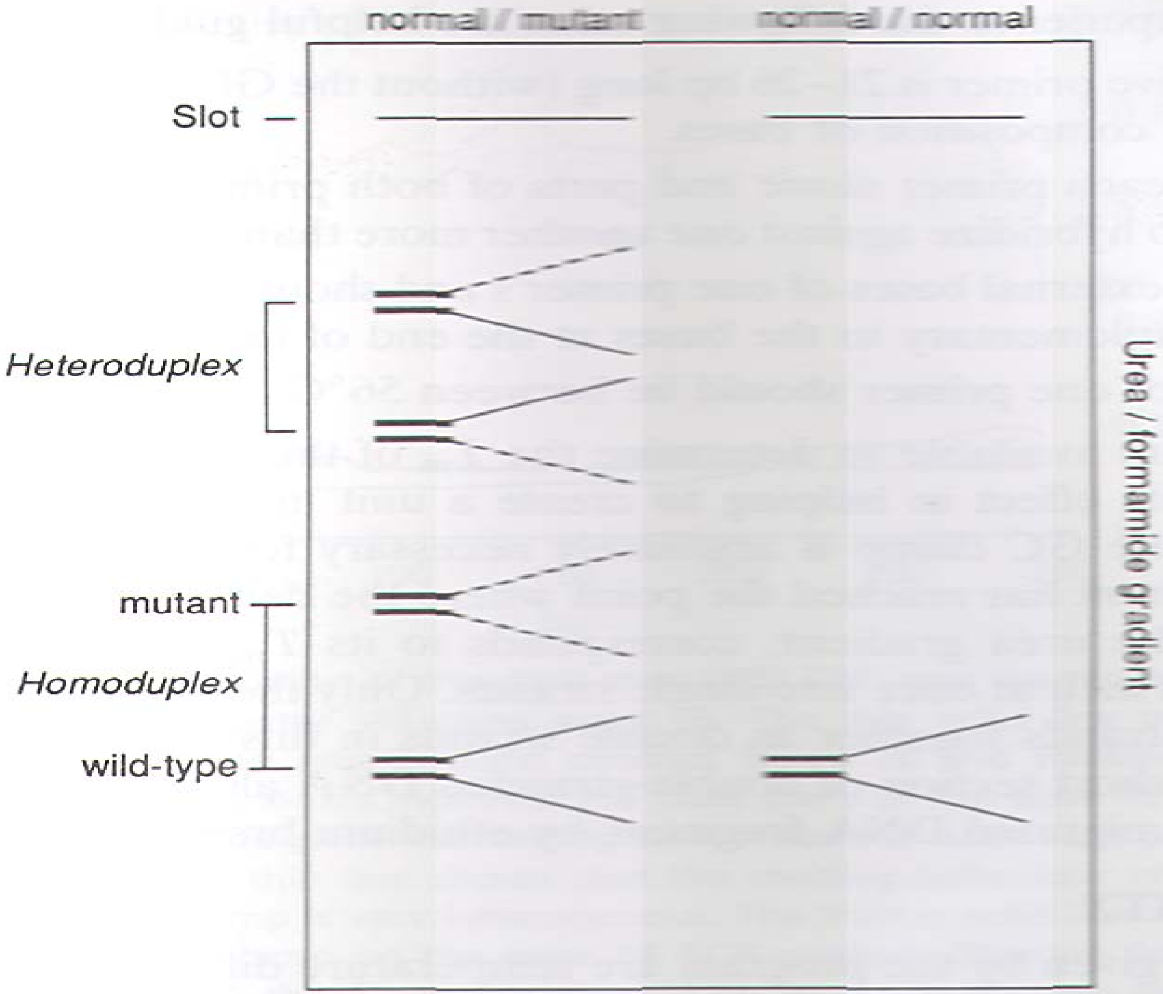


Figure 7. Diagram representing a denaturing gradient gel showing the separation of two mixed wild-type DNAs (homoduplexes) and a wild-type DNA mixed with a mutant DNA (heteroduplexes). While the wild-type homoduplexes melt at one position, the heteroduplexes show four bands (two different heteroduplexes and the original homoduplexes) which differ in their melting behaviour. The position of the two heteroduplexes always lies above the melting point of the homoduplexes. The mutated original homoduplex may be positioned above or below the wild-type homoduplexes. At their melting point the single strands are still held together by the GC clamp, shown as the thick line.

Fluorescence Activated Cell Sorting (FACS)

It would be very useful if we could separate cells that are phenotypically different from each other. In addition, it would be great to know how many cells expressed proteins of interest, and how much of this protein they expressed. Fluorescence Activated Cell Sorting (FACS) is a method that can accomplish all these goals.

The process begins by placing the cells into a flask and forcing the cells to enter a small nozzle one at a time (figure 1). The cells travel down the nozzle which is vibrated at an optimal frequency to produce drops at fixed distance from the nozzle. As the cells flow down the stream of liquid, they are scanned by a laser (blue light in figure 1). Some of the laser light is scattered (red cone emanating from the red cell) by the cells and this is used to count the cells. This scattered light can also be used to measure the size of the cells.

If you wanted to separate a subpopulation of cells, you could do so by tagging those of interest with an antibody linked to a fluorescent dye. The antibody is bound to a protein that is uniquely expressed in the cells you want to separate. The laser light excites the dye which emits a color of light that is detected by the photomultiplier tube, or light detector. By collecting the information from the light (scatter and fluorescence) a computer can determine which cells are to be separated and collected.

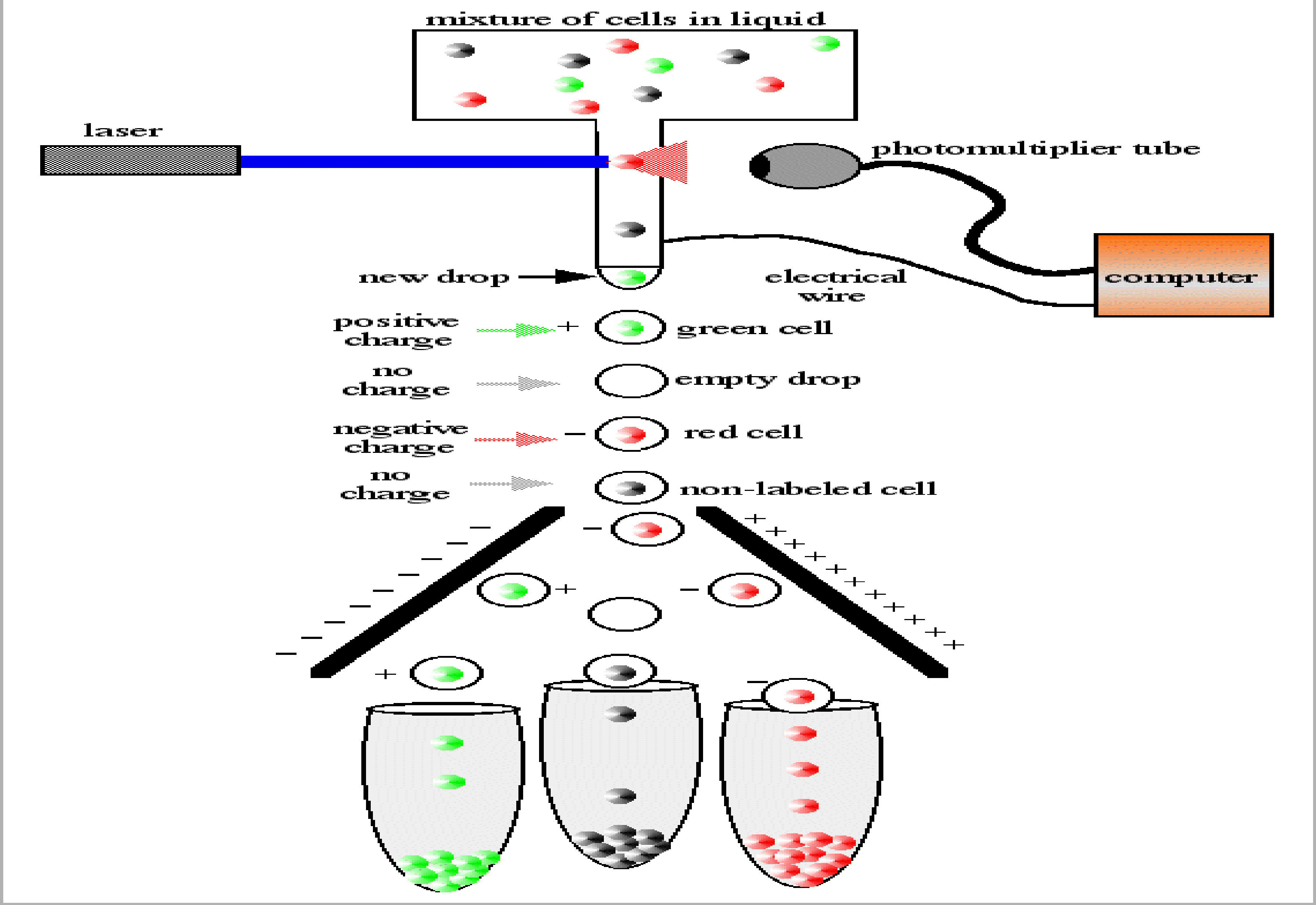


Figure 1. Diagram of FACS machine. Cells have been fluorescently tagged with either red or green antibodies, though not every cell expresses the epitope and therefore some are not tagged either color.

The final step is sorting the cells which is accomplished by electrical charge. The computer determines how the cells will be sorted before the drop forms at the end of the stream. As the drop forms, an electrical charge is applied to the stream and the newly formed drop will form with a charge. This charged drop is then deflected left or right by charged electrodes and into waiting sample tubes. Drops that contain no cells are sent into the waste tube. The end result is three tubes with pure subpopulations of cells. The number of cells in each tube is known and the level of fluorescence is also recorded for each cell.

Quantifying FACS Data

FACS data collected by the computer can be displayed in two different ways. What we want to know is how many cells of each color were sorted. In the first example (figure 2), we see the intensity of the green or red fluorescence is plotted on the X-axis and the number of cells with each level of flourescence is plotted on the Y-axis. In this example, there were twice as many red cells sorted as green or unlabeled cells, but the level of light was greater from the green cells than the red cells. This method is best if all cells are either green, red or unlabeled and no cells are labeled both colors.

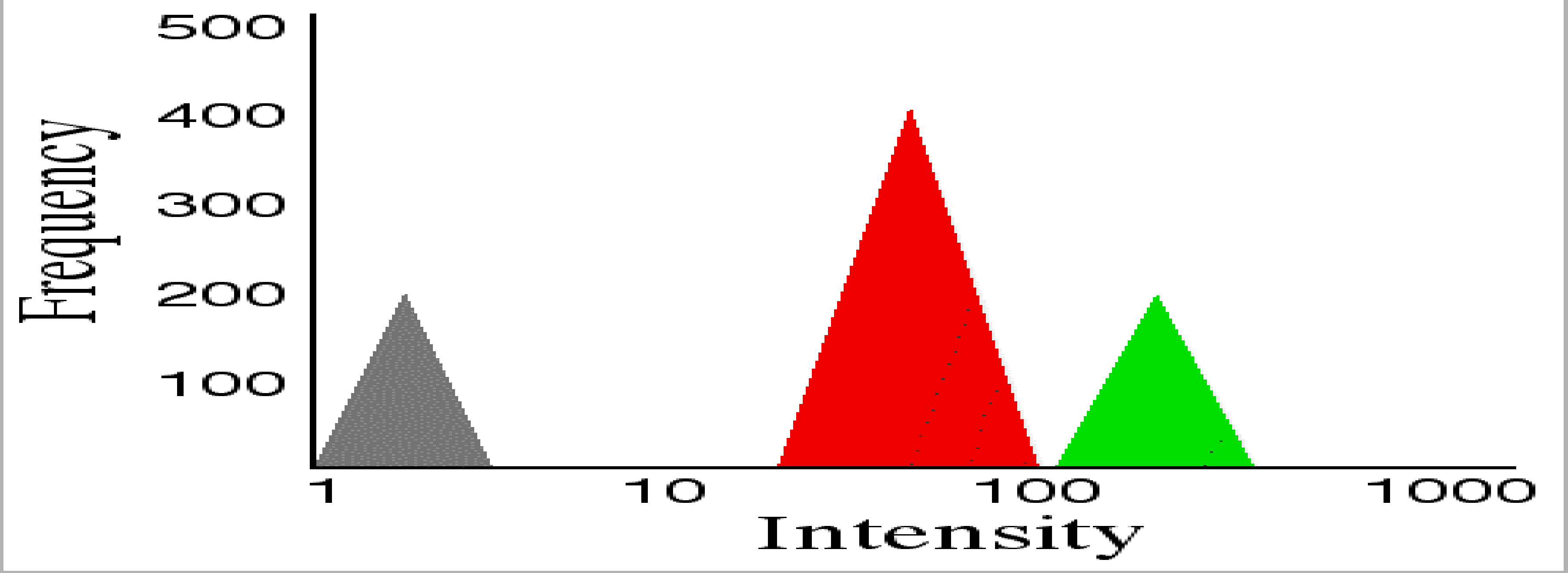


Figure 2. Quantifying FACS data. This graph shows the number of cells (X-axis) and the level of fluorescence emitted (Y-axis) by the labeled cells. Many different colors can be plotted on this graph, but cells should not be labeled by more than one color.

In figure 3, we see a different way to display the same data. The X-axis plots the intensity of green fluorescence while the Y-axis plots the intensity of red fluorescence. The individual black dots represent individual cells and we are not supposed to count the dots but just look at the relative density of dots in each quadrant. From this graph, we can see there were no cells labeled both red and green (top right) and many cells that were unlabeled (bottom left). The number of green-labeled cells (bottom right) is about the same as the number of unlabeled cells, but the number of red-labeled cells (top left) is about twice that of the other two categories of cells. Again, we can see that the level of fluorescence was higher in the green cells than the red ones. This method of graphing the data is especially useful if cells are present that have been labeled both red and green.

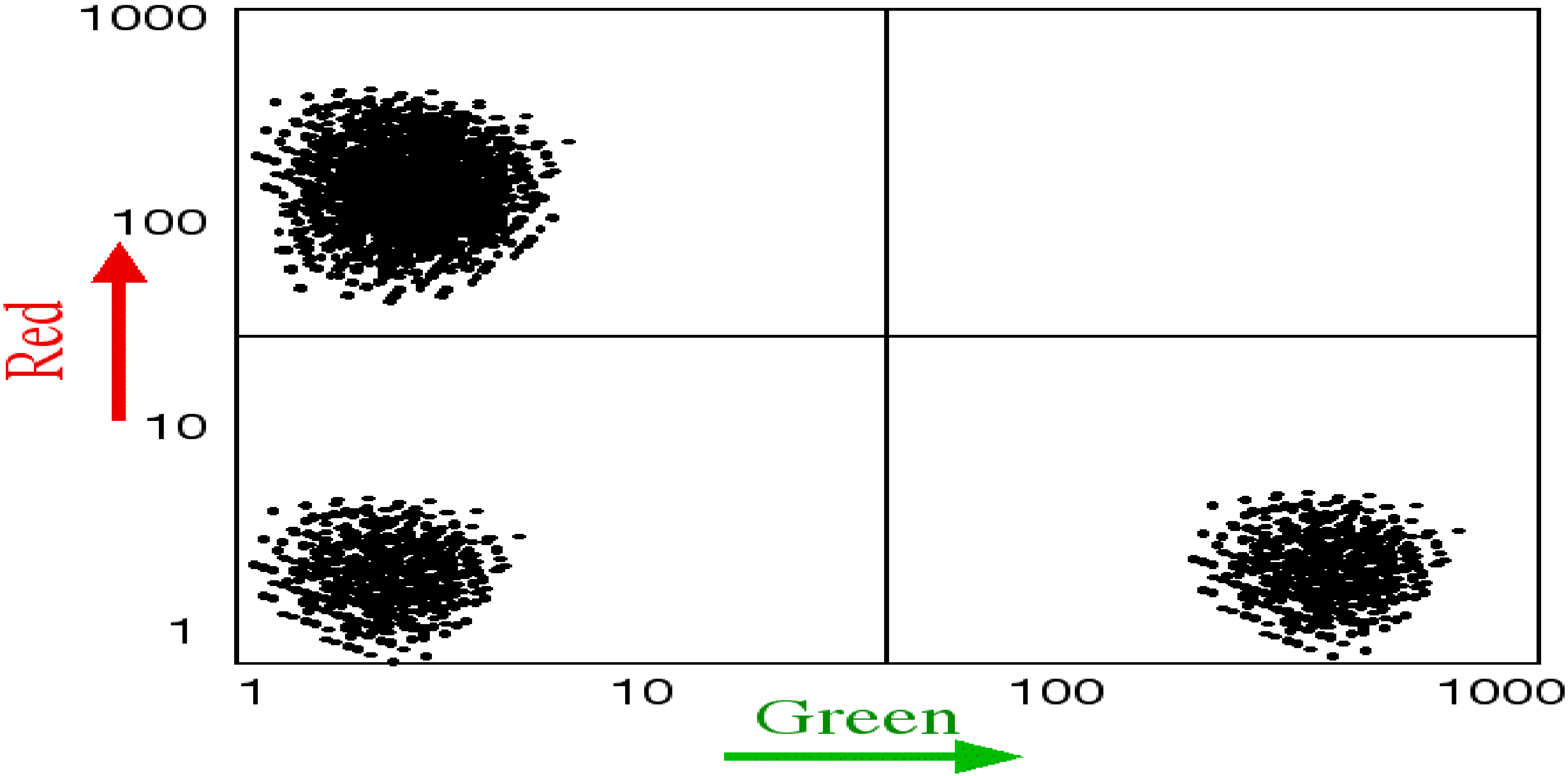


Figure 3. Quantification of FACS data. This graph compares the number of cells labeled by two colors - red (Y-axis) and green (X-axis). The intensity of the emitted light increases as indicated by the arrows. The number of cells at each intensity is shown by the number of dots where each dot represents a single cell. This graph does not work for more than two colors but it works well when individual cells can be labeled by both colors at the same time.

Microsatellite DNA Methodology

Microsatellites (sometimes referred to as a variable number of tandem repeats or VNTRs) are short segments of DNA that have a repeated sequence such as CACACACA, and they tend to occur in non-coding DNA.
In some microsatellites, the repeated unit (e.g. CA) may occur four times, in others it may be seven, or two, or thirty. In diploid organisms each individual animal will have two copies of any particular microsatellite segment. For example, a father might have a genotype of 12 repeats and 19 repeats, a mother might have 18 repeats and 15 repeats while their first born might have repeats of 12 and 15. On rare occasions, microsatellites can cause the DNA polymerase to make an extra copy of CA.

Over time, as animals in a population breed, they will recombine their microsatellites during sexual reproduction and the population will maintain a variety of microsatellites that is characteristic for that population and distinct from other populations which do not interbreed.

The most common way to detect microsatellites is to design PCR primers that are unique to one locus in the genome and that base pair on either side of the repeated portion (figure 1). Therefore, a single pair of PCR primers will work for every individual in the species and produce different sized products for each of the different length microsatellites.

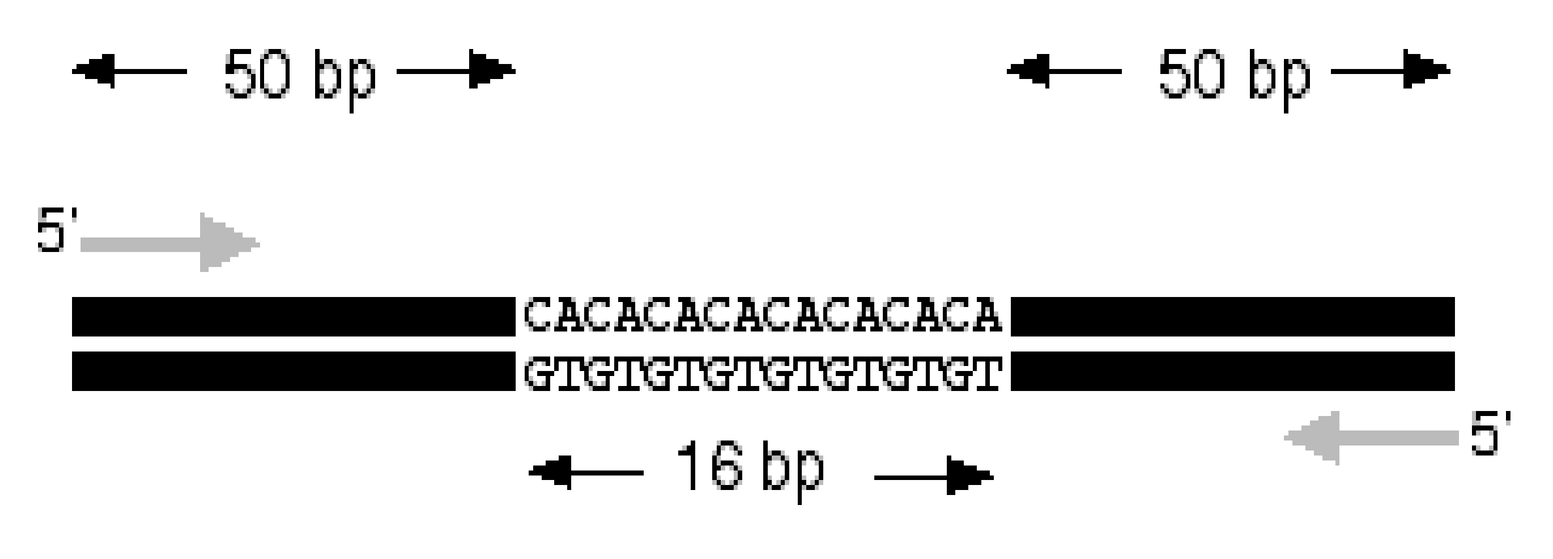


Figure 1. Detecting microsatellites from genomic DNA. Two PCR primers (forward and reverse gray arrows) are designed to flank the microsatellite region. If there were zero repeats, the PCR product would be 100 bp in length. Therefore, by determining the size of each PCR product (in this case 116 bp), you can calculate how many CA repeats are present in each microsatellite (8 CA repeats in this example).

The PCR products are then separated by either gel electrophoresis or capillary electrophoresis. Either way, the investigator can determine the size of the PCR product and thus how many times the dinucleotide "CA" was repeated for each allele (figure 2). It would be nice if microsatellite data produced only two bands but often there are minor bands in addition to the major bands; they are called stutter bands and they usually differ from the major bands by two nucleotides.

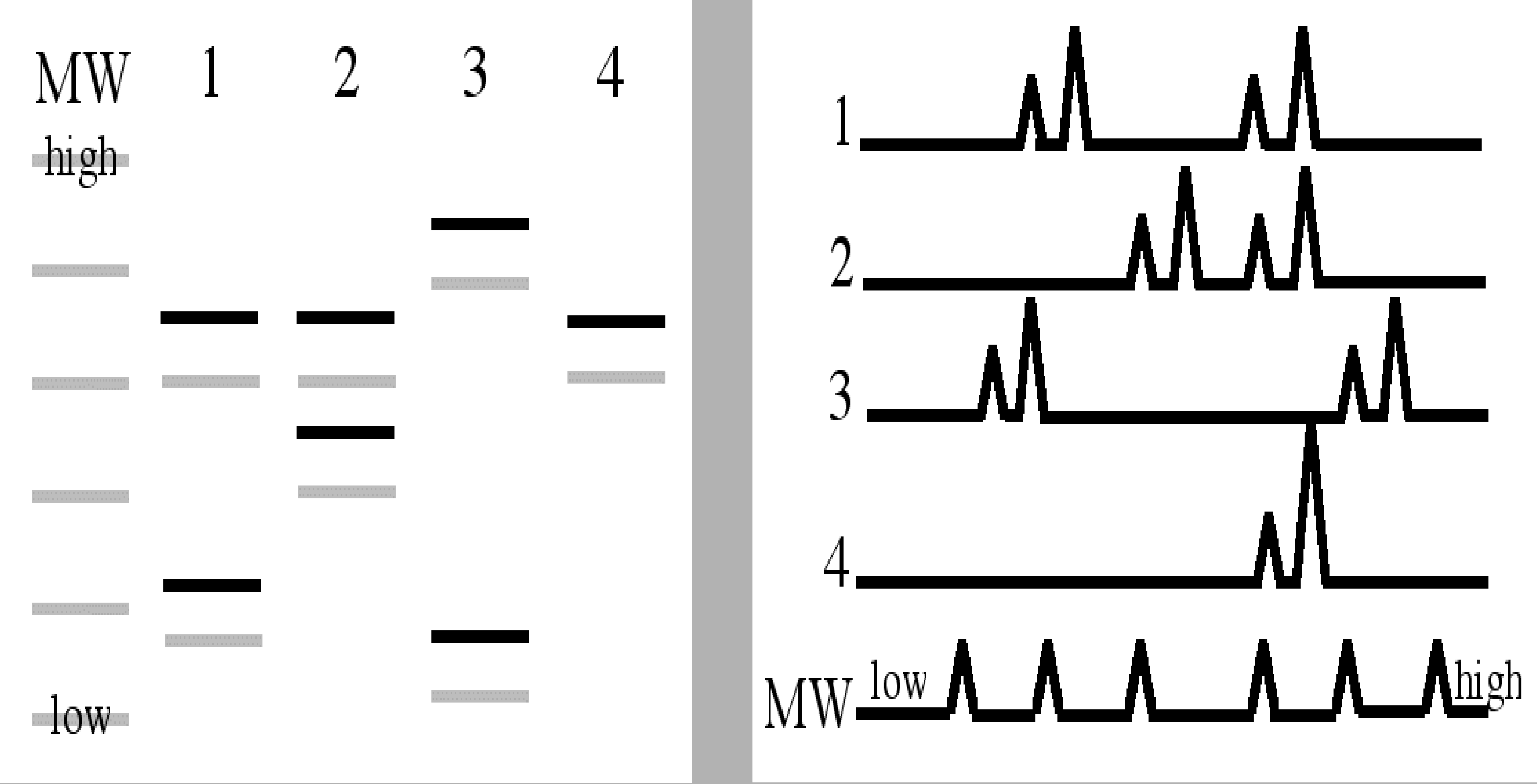


Figure 2. Stylized examples of microsatellite data. Left half: four sets of data were produced by gel electrophoresis and so you can see the major (black) and stutter (gray) bands. MW; molecular weight standards. Right half: These data were produced by analysis on an automated capillary electrophoresis-based DNA sequencer. The data are line graphs with the location of each peak on the X-axis representing a different sized PCR product and the height of each peak indicates the amount of PCR product. The major bands produce higher peaks than the stutter peaks.

Capillary Electrophoresis

Because the capillary tube has a high surface to volume ratio (25-100 µm diameter), it radiates heat readily and thus samples do not over heat. Detection of the migrating molecules is accomplished by shining a light source through a portion of the tubing and detecting the light emitted from the other side (figure 1).

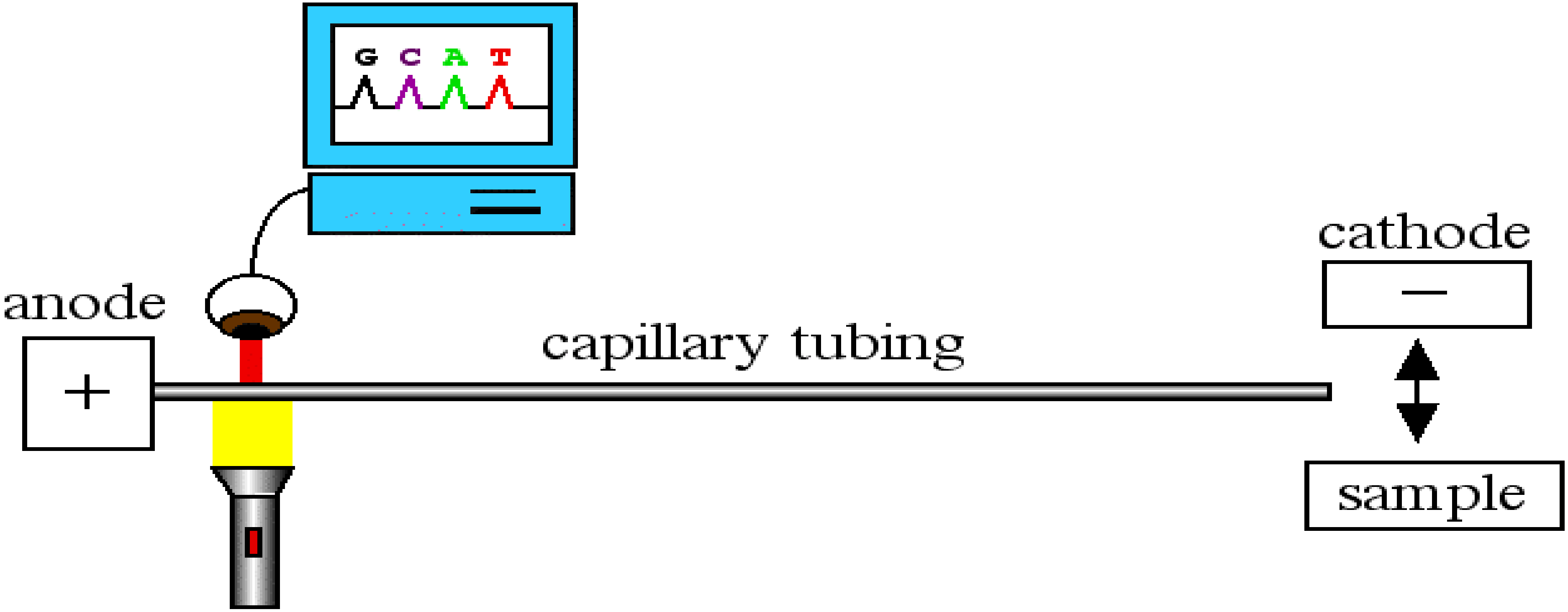
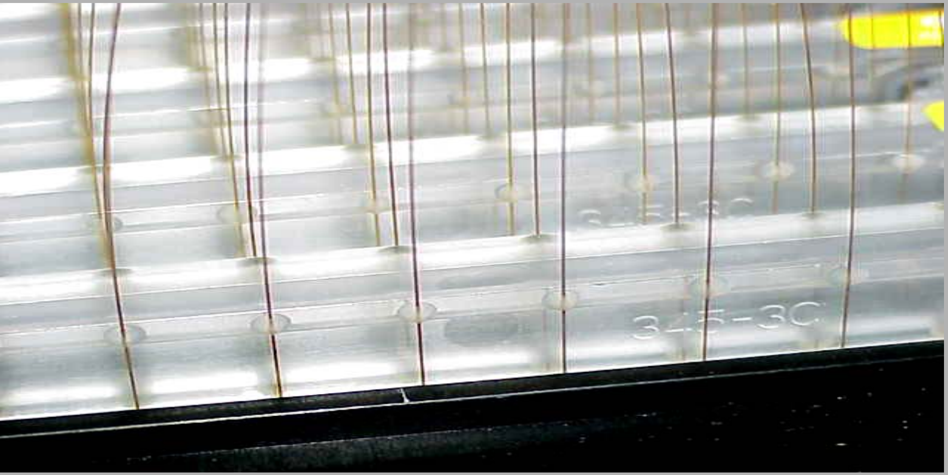
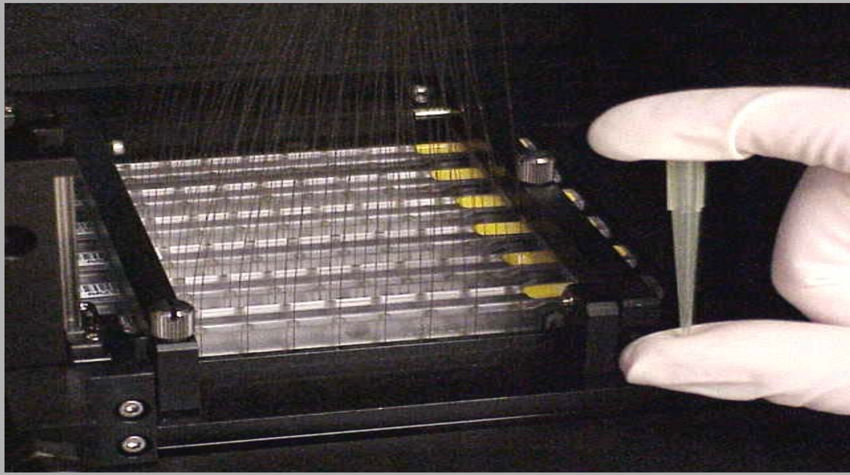


Figure 1. Schematic of capillary electrophoresis system. Samples enter the tube from the right and travel to the left to the detection system which records the chromatogram output on a computer.

Sample run times are very Short . Samples are applied to the capillary tubes when the cathode buffer is moved aside and sample chamber placed at the opening of the capillary tube. Either pressure is applied to the sample and 10 - 100 nL is injected or an electrical current is applied through the sample and only the charged molecules enter the capillary. Once the electrophoretic separation is completed, the contents of the capillary are flushed out and fresh matrix fills the tube. Replacing the matrix within the capillary minimizes the possiblity of contaminating samples between runs.





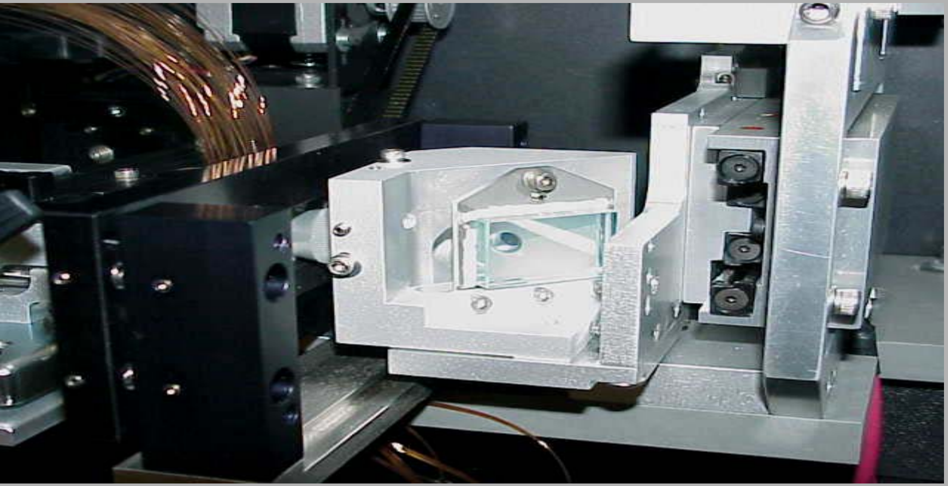


Figure 2. Megabase DNA sequencer images. Clockwise from top left: close up of capillary tubes descending into a 96-well plate containing samples; yellow tip shown with all 96 capillary tubes for size comparison.
From bottom left: optical detection system that shines focues light into the tubes and detects the amount and color of the emitted light.