CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS15.html

Gene Expression

- Process of transcription and/or translation of a gene is called gene expression.
- Every cell of an organism has the same genetic material, but different genes are expressed at different times.
- Patterns of gene expression in a cell is indicative of its state.

Hybridization

- If two complementary strands of DNA or mRNA are brought together under the right experimental conditions they will hybridize.
- A hybridizes to B ⇒
  - A is reverse complementary to B, or
  - A is reverse complementary to a subsequence of B.
- It is possible to experimentally verify whether A hybridizes to B, by labeling A or B with a radioactive or fluorescent tag, followed by excitation by laser.

Measuring gene expression

- Gene expression for a single gene can be measured by extracting mRNA from the cell and doing a simple hybridization experiment.
- Given a sample of cells, gene expression for every gene can be measured using a single microarray experiment.

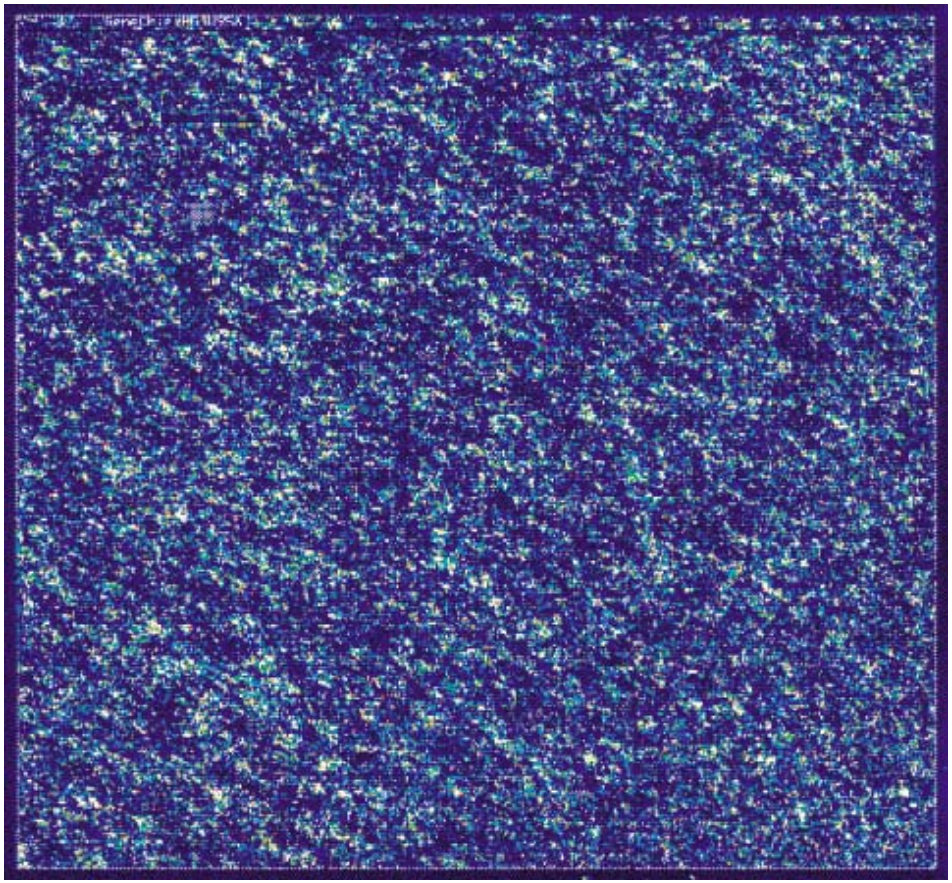
Microarray/DNA chip technology

- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between tissues, between drugs, across disease states

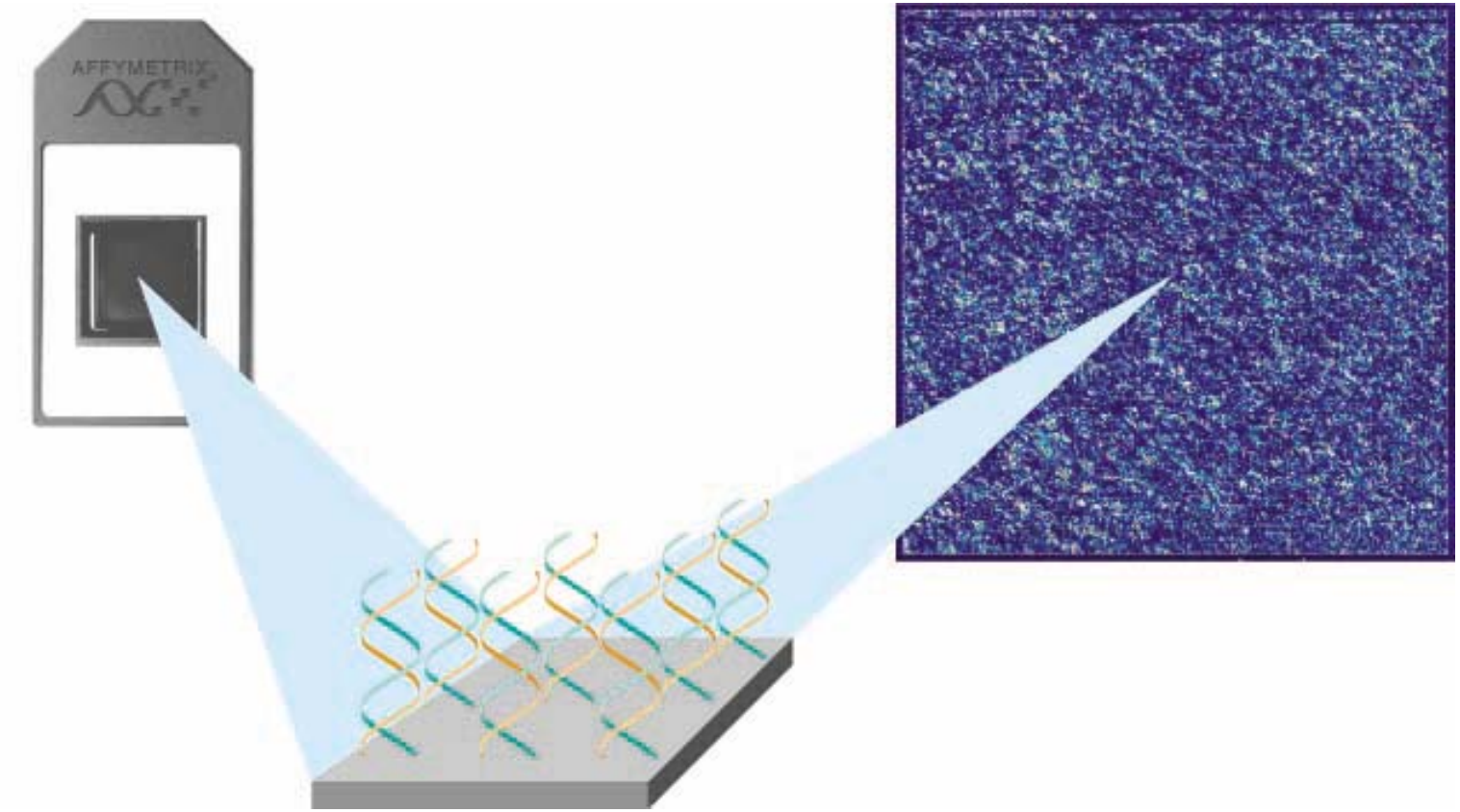
Microarray Data

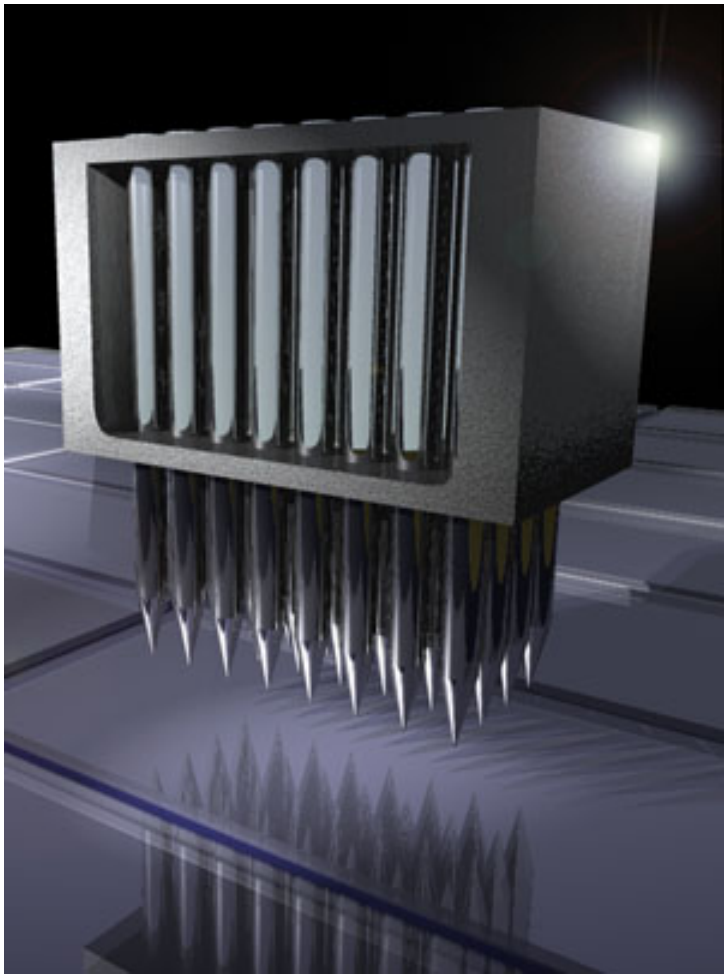
Gene	Expression Level
Gene1	
Gene2	
Gene3	
…	

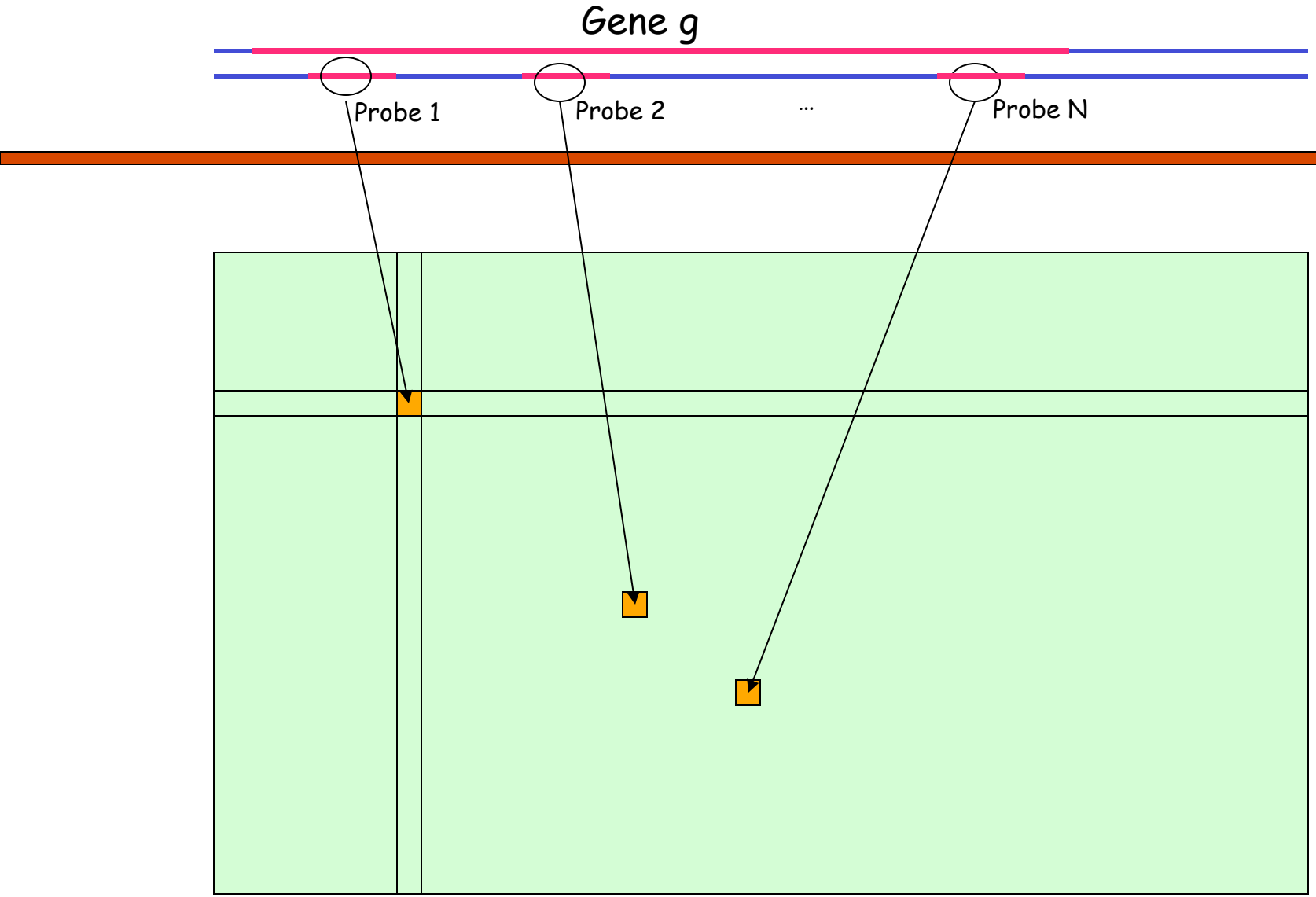
Gene Chips



DNA Chips & Images



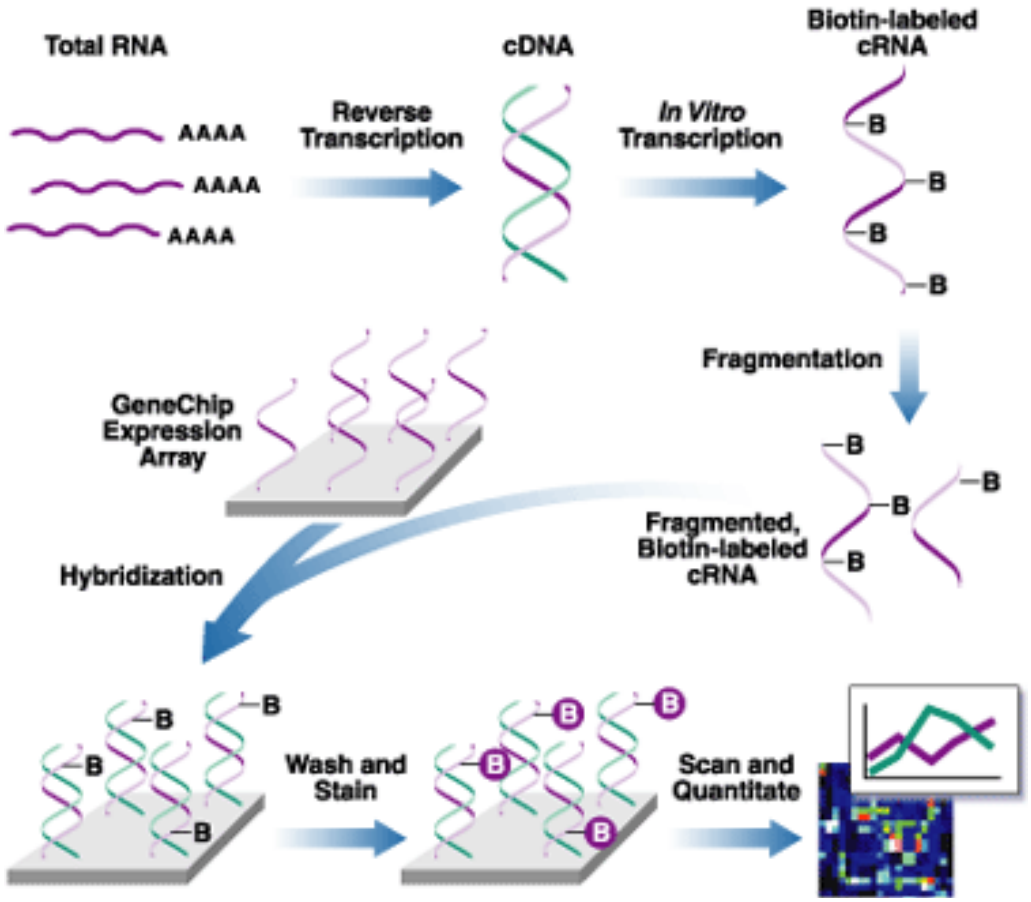



Gene g
Probe 1
Probe 2
...
Probe N

Microarray/DNA chips (Simplified)

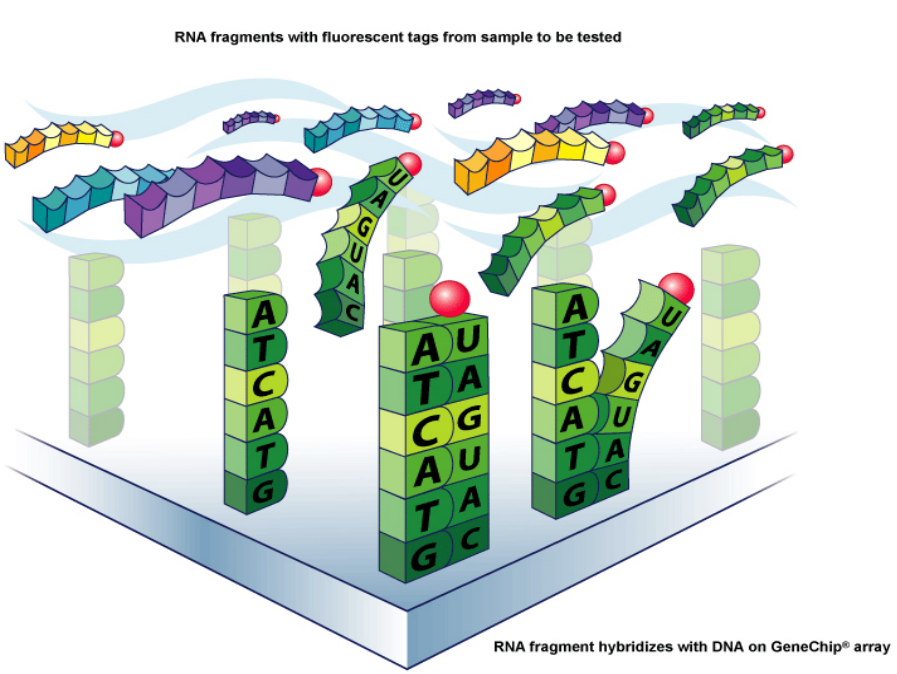
- Construct probes corresponding to reverse complements of genes of interest.
- Microscopic quantities of probes placed on solid surfaces at defined spots on the chip.
- Extract mRNA from sample cells and label them.
- Apply labeled sample (mRNA extracted from cells) to every spot, and allow hybridization.
- Wash off unhybridized material.
- Use optical detector to measure amount of fluorescence from each spot.

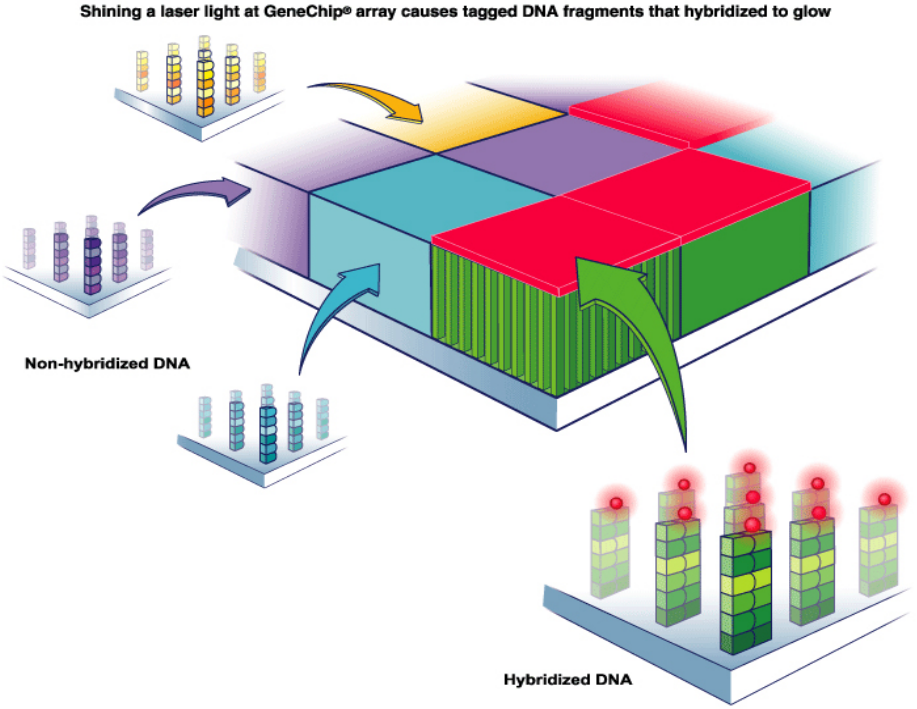
Affymetrix DNA chip schematic



www.affymetrix.com

What's on the slide?

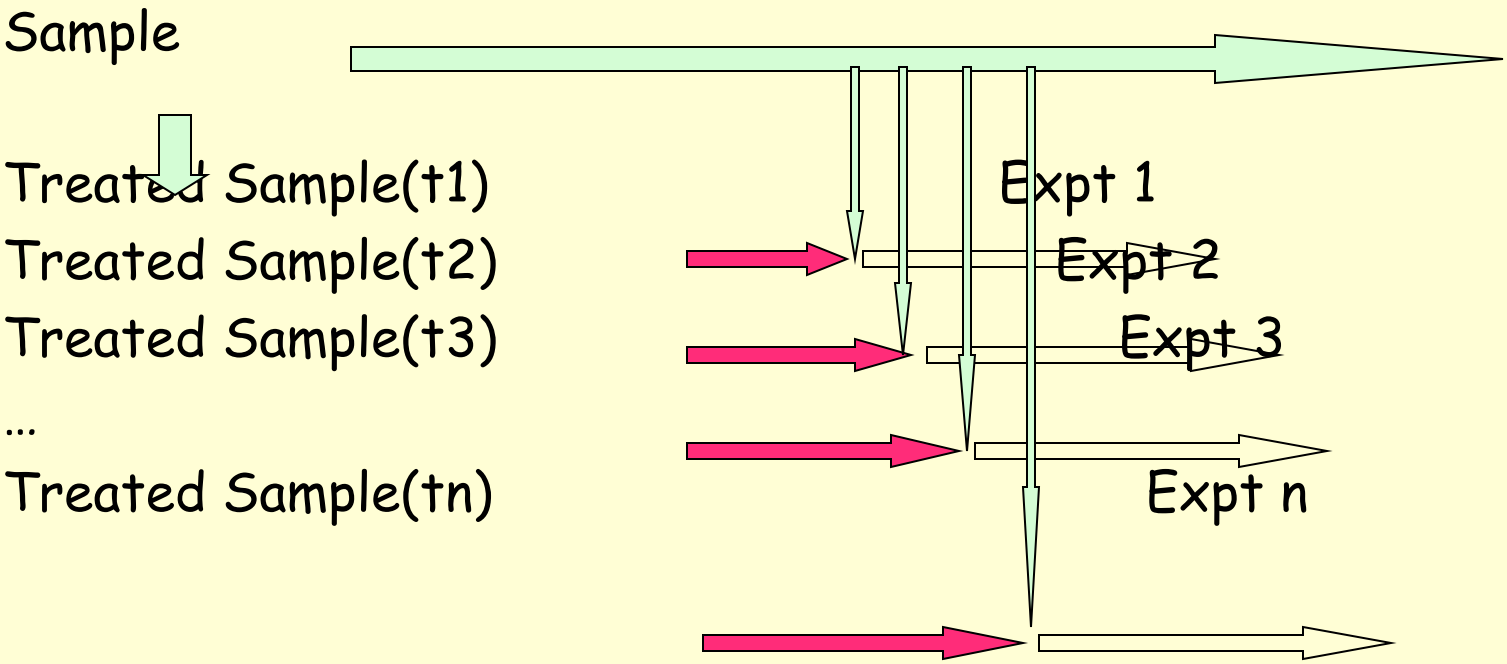


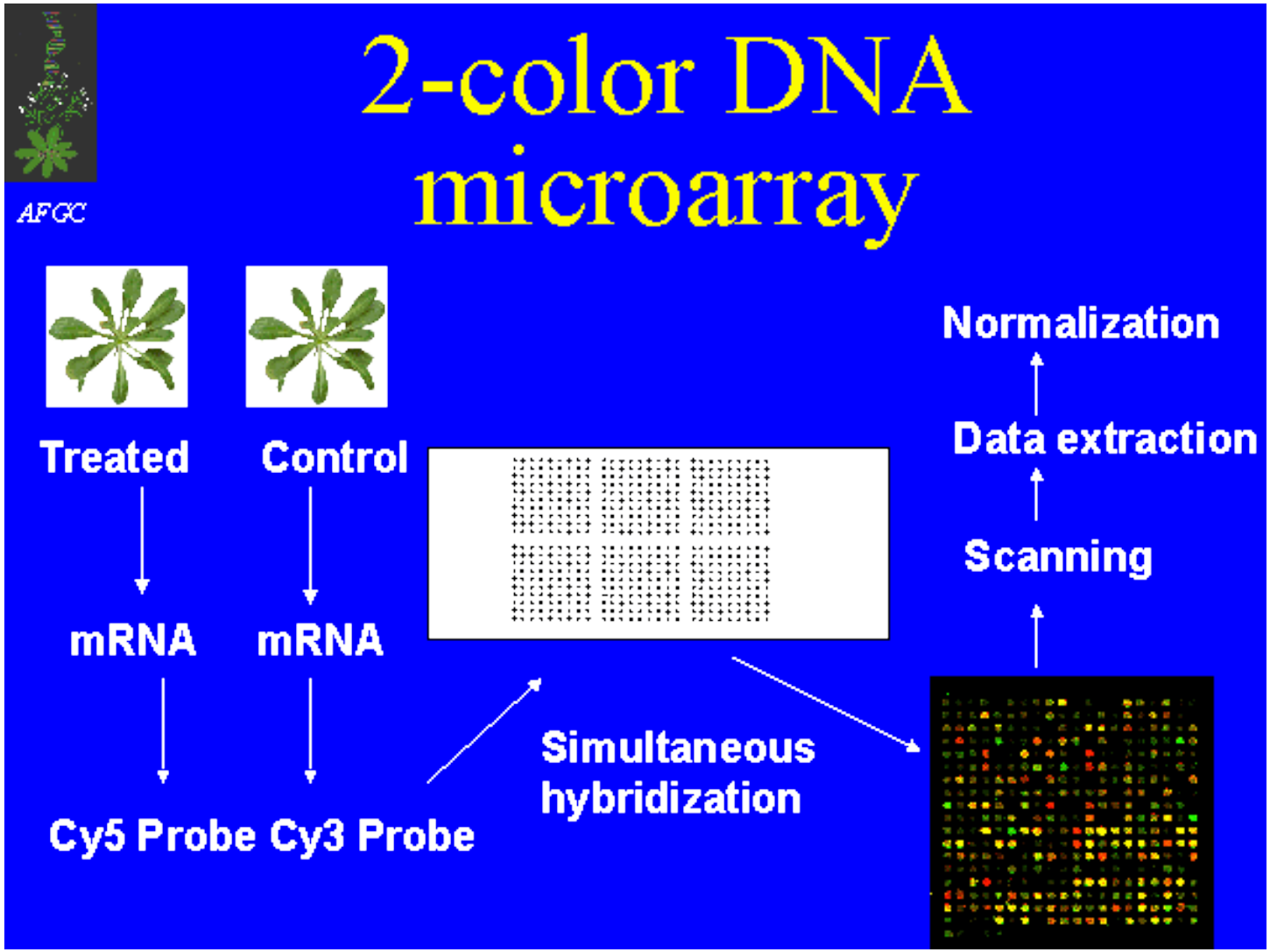


Microarrays: competing technologies

- Affymetrix & Agilent
- Differ in:
  - method to place DNA: Spotting vs. photolithography
  - Length of probe
  - Complete sequence vs. series of fragments

Study effect of treatment over time





http://www.arabidopsis.org/info/2010_projects/comp_proj/AFGC/RevisedAFGC/Friday/

How to compare 2 cell samples with Two-Color Microarrays?

- mRNA from sample 1 is extracted and labeled with a red fluorescent dye.
- mRNA from sample 2 is extracted and labeled with a green fluorescent dye.
- Mix the samples and apply it to every spot on the microarray. Hybridize sample mixture to probes.
- Use optical detector to measure the amount of green and red fluorescence at each spot.

Sources of Variations & Experimental Errors

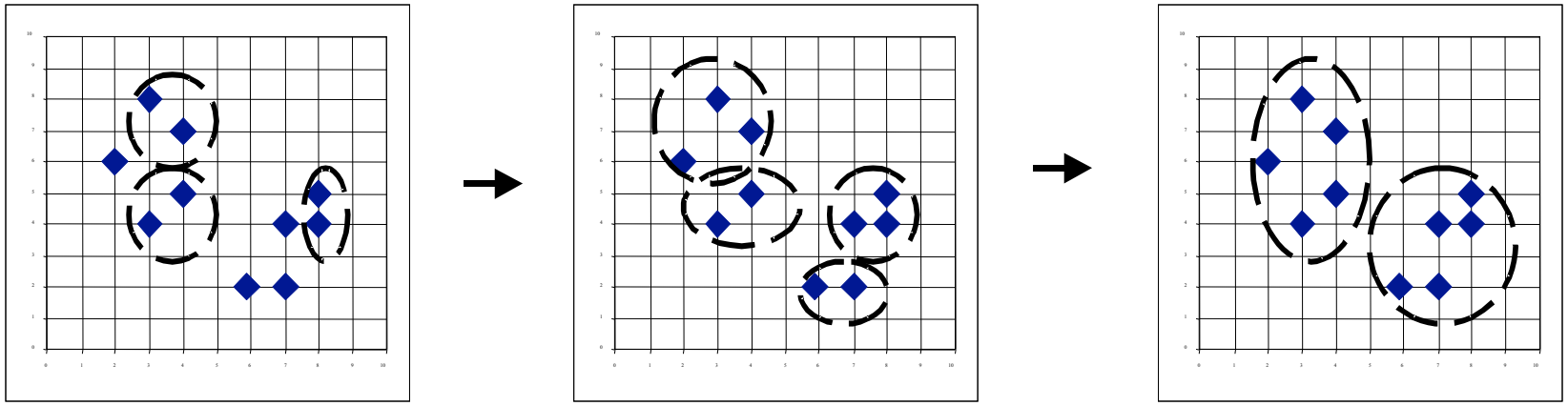
- Variations in cells/individuals
- Variations in mRNA extraction, isolation, introduction of dye, variation in dye incorporation, dye interference
- Variations in probe concentration, probe amounts, substrate surface characteristics
- Variations in hybridization conditions and kinetics
- Variations in optical measurements, spot misalignments, discretization effects, noise due to scanner lens and laser irregularities
- Cross-hybridization of sequences with high sequence identity
- Limit of factor 2 in precision of results
- Variation changes with intensity: larger variation at low or high expression levels

Need to Normalize data

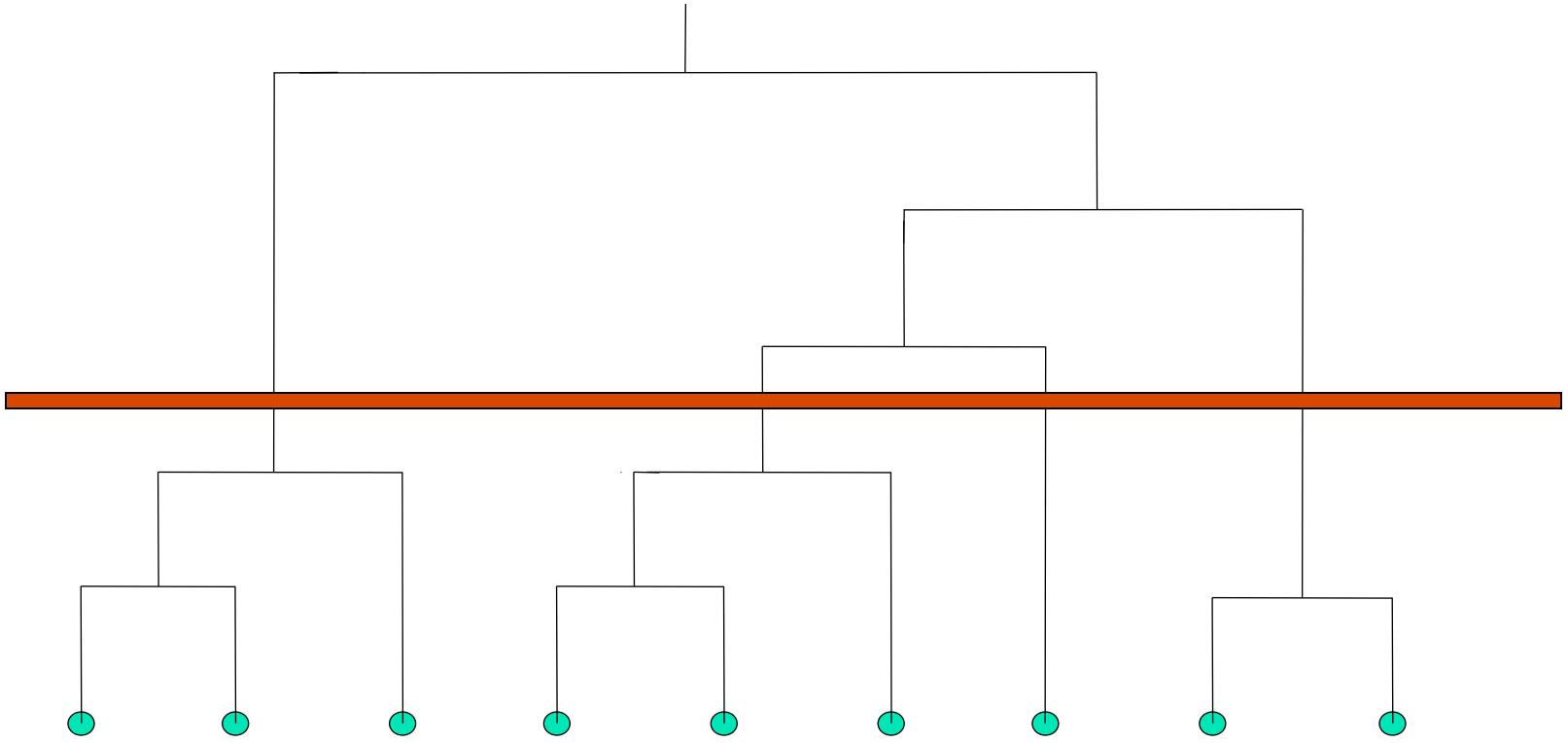
Clustering

- Clustering is a general method to study patterns in gene expressions.
- Several known methods:
  - Hierarchical Clustering (Bottom-Up Approach)
  - K-means Clustering (Top-Down Approach)
  - Self-Organizing Maps (SOM)

Hierarchical Clustering: Example



A Dendrogram



Hierarchical Clustering [Johnson, SC, 1967]

- Given n points in R^d, compute the distance between every pair of points
- While (not done)
  - Pick closest pair of points s_i and s_j and make them part of the same cluster.
  - Replace the pair by an average of the two s_{ij}

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Distance Metrics

- For clustering, define a distance function:
  - Euclidean distance metrics

$$D_k(X,Y) = \left[\sum_{i=1}^{d} (X_i - Y_i)^k\right]^{1/k}$$

k=2: Euclidean Distance

  - Pearson correlation coefficient

$$\rho_{xy} = \frac{1}{d}\sum_{i=1}^{d}\left(\frac{X_i - \overline{X}}{\sigma_x}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_y}\right)$$

$-1 \leq \rho_{xy} \geq 1$

EXHIBIT 3.4 *Joint Probability Model for the Ratings of Two People*

(a) $\rho_{XY} = 0$

	y			
x	1	2	3	Total
3	1/9	1/9	1/9	1/3
2	1/9	1/9	1/9	1/3
1	1/9	1/9	1/9	1/3
Total	1/3	1/3	1/3	1

(b) $\rho_{XY} = \frac{1}{2}$

	y			
x	1	2	3	Total
3	1/18	1/18	4/18	1/3
2	1/18	4/18	1/18	1/3
1	4/18	1/18	1/18	1/3
Total	1/3	1/3	1/3	1

(c) $\rho_{XY} = -\frac{1}{2}$

	y			
x	1	2	3	Total
3	4/18	1/18	1/18	1/3
2	1/18	4/18	1/18	1/3
1	1/18	1/18	4/18	1/3
Total	1/3	1/3	1/3	1

(d) $\rho_{XY} = \frac{4}{9}$

	y			
x	1	2	3	Total
3	1/27	2/27	6/27	1/3
2	2/27	5/27	2/27	1/3
1	6/27	2/27	1/27	1/3
Total	1/3	1/3	1/3	1

(e) $\rho_{XY} = -\frac{5}{9}$

	y			
x	1	2	3	Total
3	6/27	2/27	1/27	1/3
2	2/27	5/27	2/27	1/3
1	1/27	2/27	6/27	1/3
Total	1/3	1/3	1/3	1

(f) $\rho_{XY} = \frac{2}{3}$

	y			
x	1	2	3	Total
3	1/36	2/36	9/36	1/3
2	2/36	8/36	2/36	1/3
1	9/36	2/36	1/36	1/3
Total	1/3	1/3	1/3	1

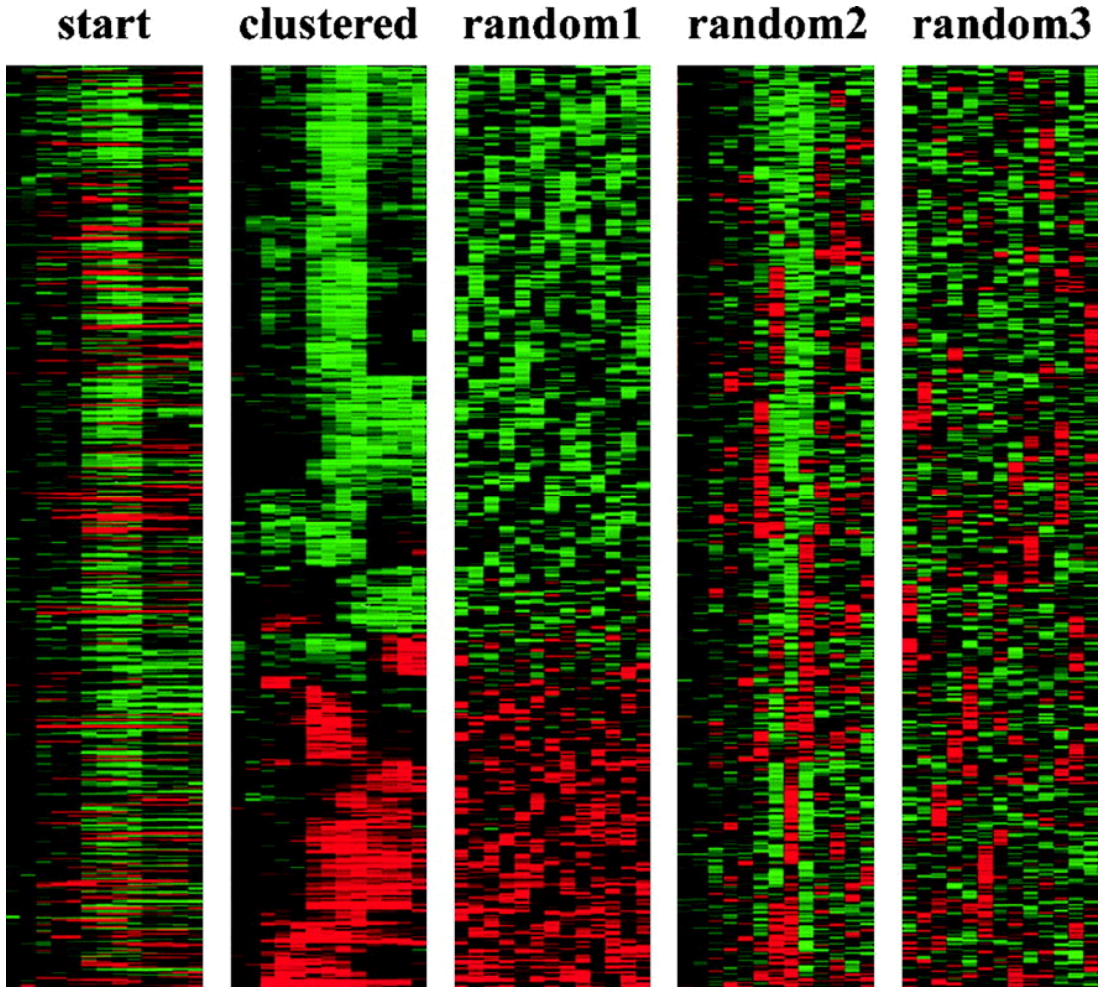
(g) $\rho_{XY} = -\frac{1}{3}$

	y			
x	1	2	3	Total
3	9/36	2/36	1/36	1/3
2	2/36	8/18	2/18	1/3
1	1/36	2/36	9/36	1/3
Total	1/3	1/3	1/3	1

Clustering of gene expressions

- Represent each gene as a vector or a point in d-space where d is the number of arrays or experiments being analyzed.

Clustering Random vs. Biological Data



From *Eisen MB, et al, PNAS 1998 95(25):14863-8*

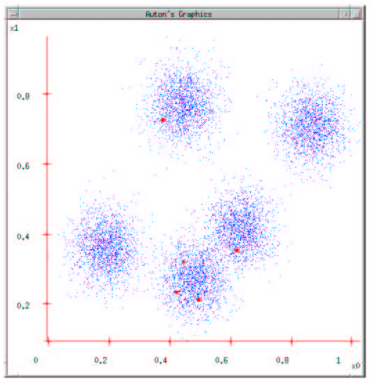
K-Means Clustering: Example

Example from Andrew Moore's tutorial on Clustering.

Start

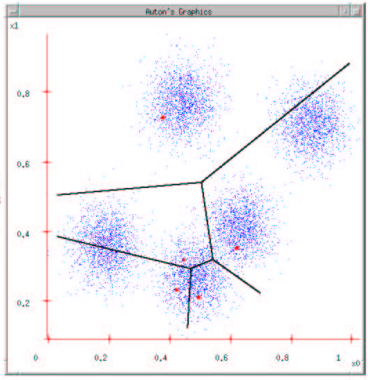
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations



K-means

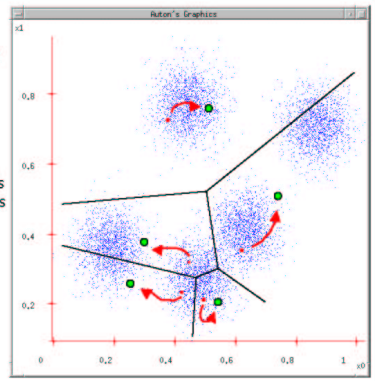
1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



4

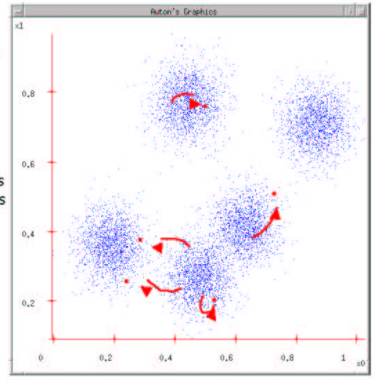
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!



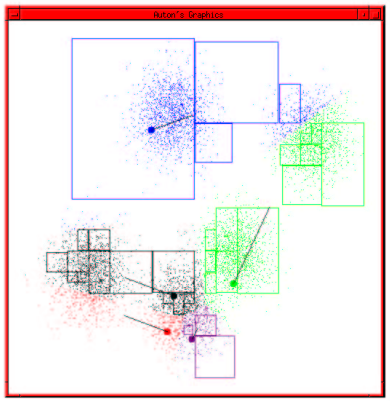
5

K-means Start

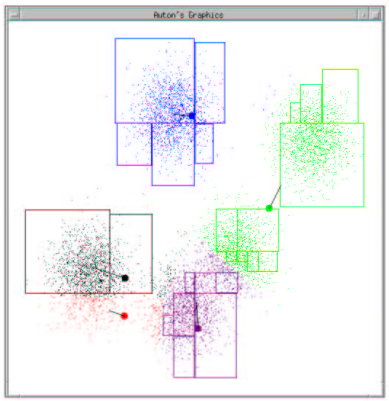
Advance apologies: in Black and White this example will deteriorate

Example generated by Dan Pelleg's super-duper fast K-means system:

Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999, (KDD99) (available on www.autonlab.org/pap.html)

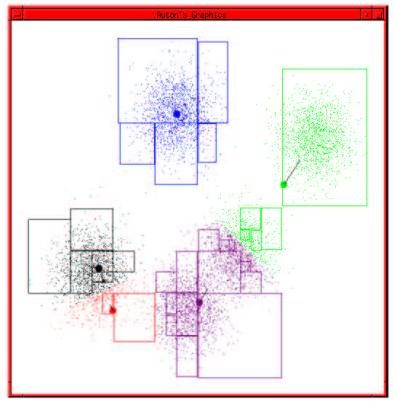


K-means continues ...

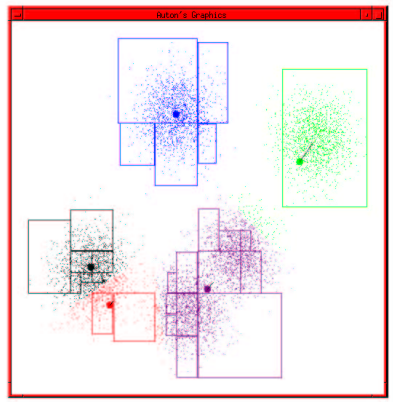


6

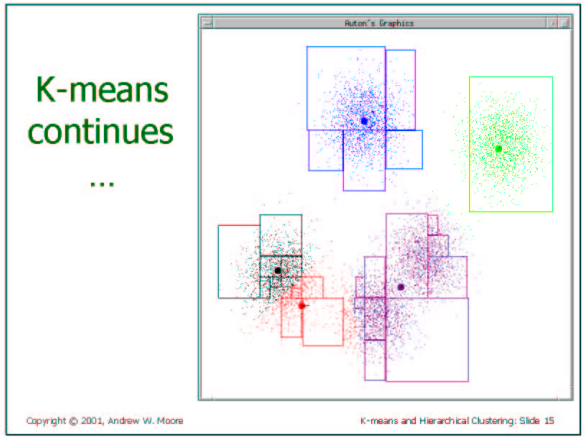
K-means continues ...

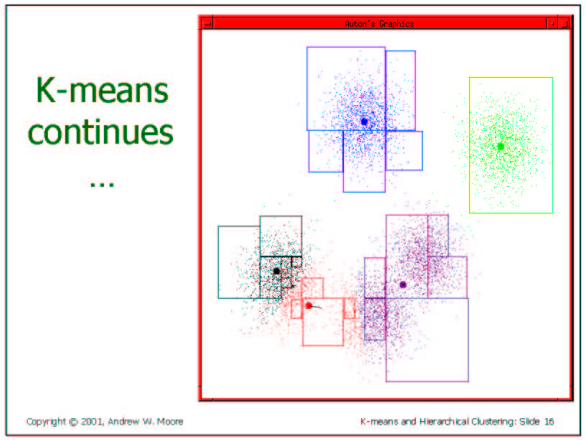


K-means continues ...

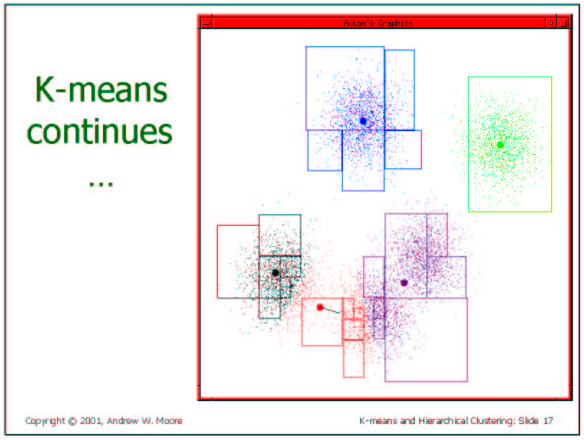


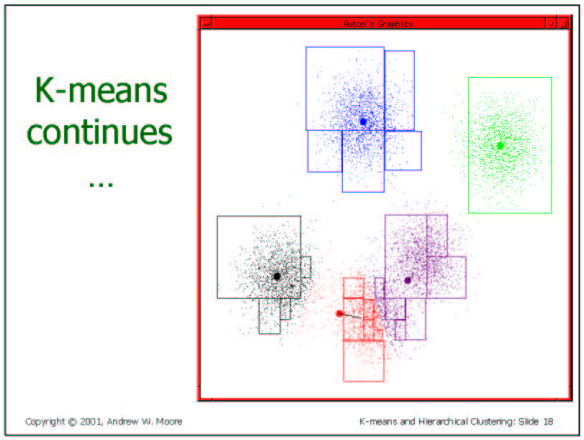
7


K-means
continues
...
Auton's Graphics
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 15


K-means
continues
...
Auton's Graphics
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 16

8


K-means
continues
...
Auton's Graphics
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 17

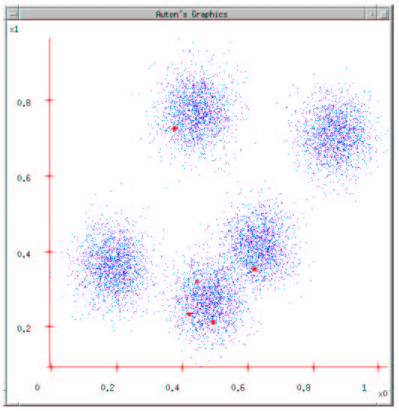

K-means
continues
...
Auton's Graphics
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 18

9

Start

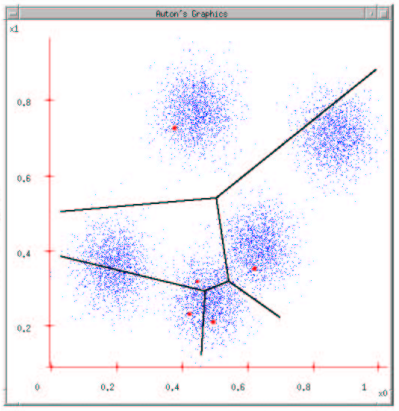
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations



K-means

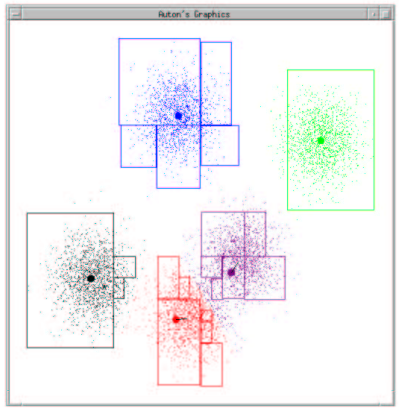
1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



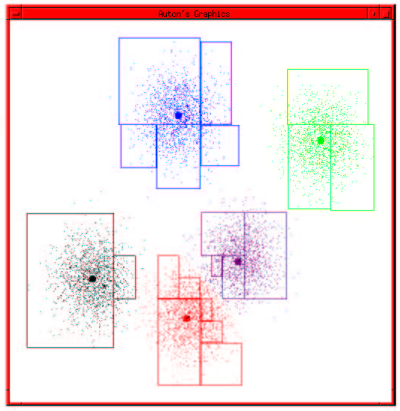
4

K-means continues

...



K-means terminates



10

End

K-Means Clustering [McQueen '67]

Repeat

- Start with randomly chosen cluster centers
- Assign points to give greatest increase in score
- Recompute cluster centers
- Reassign points

until (no changes)

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Comparisons

- Hierarchical clustering
  - Number of clusters not preset.
  - Complete hierarchy of clusters
  - Not very robust, not very efficient.
- K-Means
  - Need definition of a mean. Categorical data?
  - More efficient and often finds optimum clustering.

Functionally related genes behave similarly across experiments

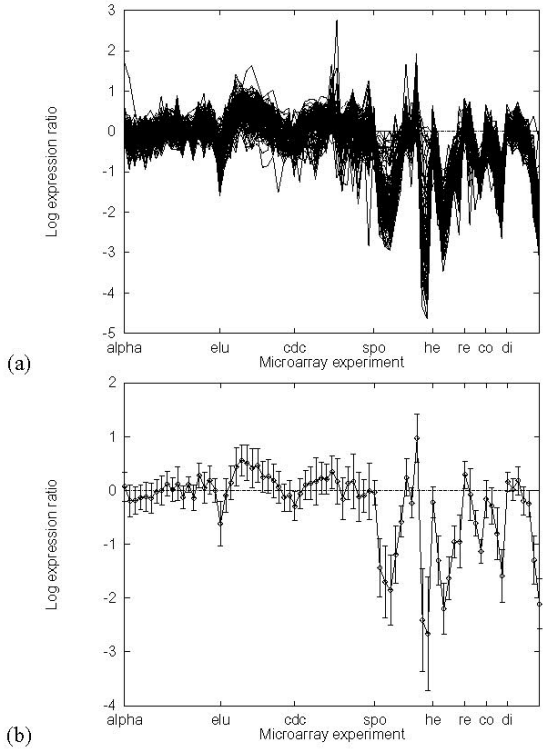


Figure 1: **Expression profiles of the cytoplasmic ribosomal proteins.** Figure (a) shows the expression profiles from the data in [Eisen et al., 1998] of 121 cytoplasmic ribosomal proteins, as classified by MYGD [MYGD, 1999]. The logarithm of the expression ratio is plotted as a function of DNA microarray experiment. Ticks along the X-axis represent the beginnings of experimental series. They are, from left to right, cell division cycle after synchronization with α factor arrest (alpha), cell division cycle after synchronization by centrifugal elutriation (elu), cell division cycle measured using a temperature sensitive *cdc15* mutant (cdc), sporulation (spo), heat shock (he), reducing shock (re), cold shock (co), and diauxic shift (di). Sporulation is the generation of a yeast spore by meiosis. Diauxic shift is the shift from anaerobic (fermentation) to aerobic (respiration) metabolism. The medium starts rich in glucose, and yeast cells ferment, producing ethanol. When the glucose is used up, they switch to ethanol as a source for carbon. Heat, cold, and reducing shock are various ways to stress the yeast cell. Figure (b) shows the average, plus or minus one standard deviation, of the data in Figure (a).

Self-Organizing Maps [Kohonen]

- Kind of neural network.
- Clusters data and find complex relationships between clusters.
- Helps reduce the dimensionality of the data.
- Map of 1 or 2 dimensions produced.
- Unsupervised Clustering
- Like K-Means, except for visualization

SOM Architectures

- 2-D Grid
- 3-D Grid
- Hexagonal Grid

SOM Algorithm

- Select SOM architecture, and initialize weight vectors and other parameters.
- **While** (stopping condition not satisfied) **do** for each input point x
  - winning node q has weight vector closest to x.
  - Update weight vector of q and its neighbors.
  - Reduce neighborhood size and learning rate.

SOM Algorithm Details

- Distance between x and weight vector: $\|x - w_i\|$
- Winning node: $q(x) = \min_i \|x - w_i\|$
- Weight update function (for neighbors):

$$w_i(k+1) = w_i(k) + \mu(k, x, i)[x(k) - w_i(k)]$$

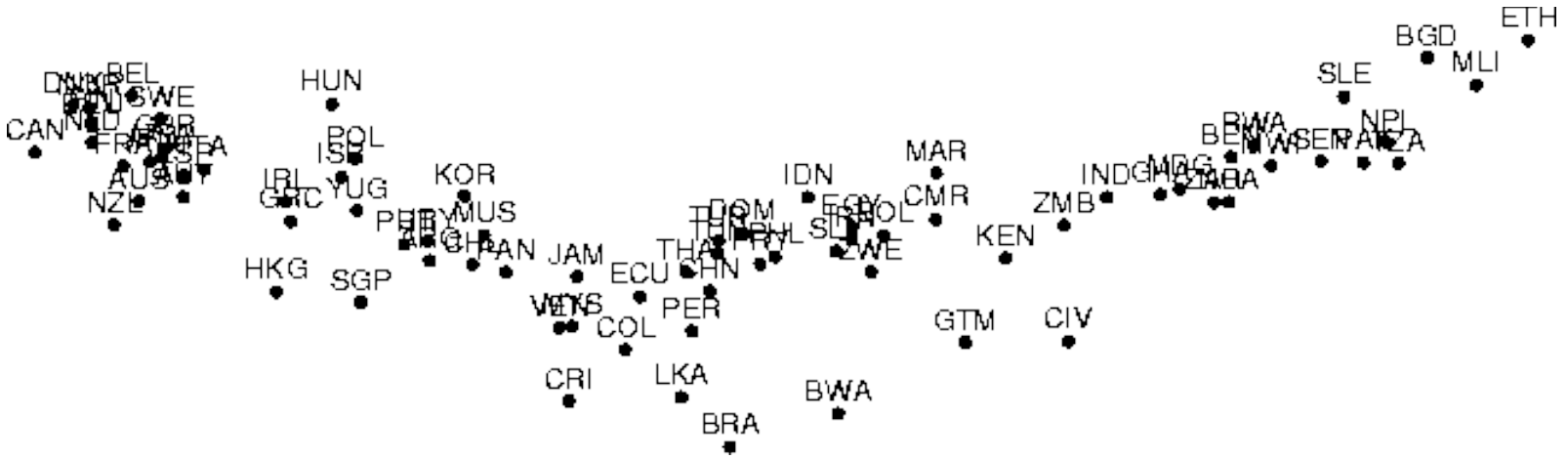
- Learning rate:

$$\mu(k, x, i) = \eta_0(k) \exp\left(\frac{-\|r_i - r_{q(x)}\|^2}{\sigma^2} \right)$$

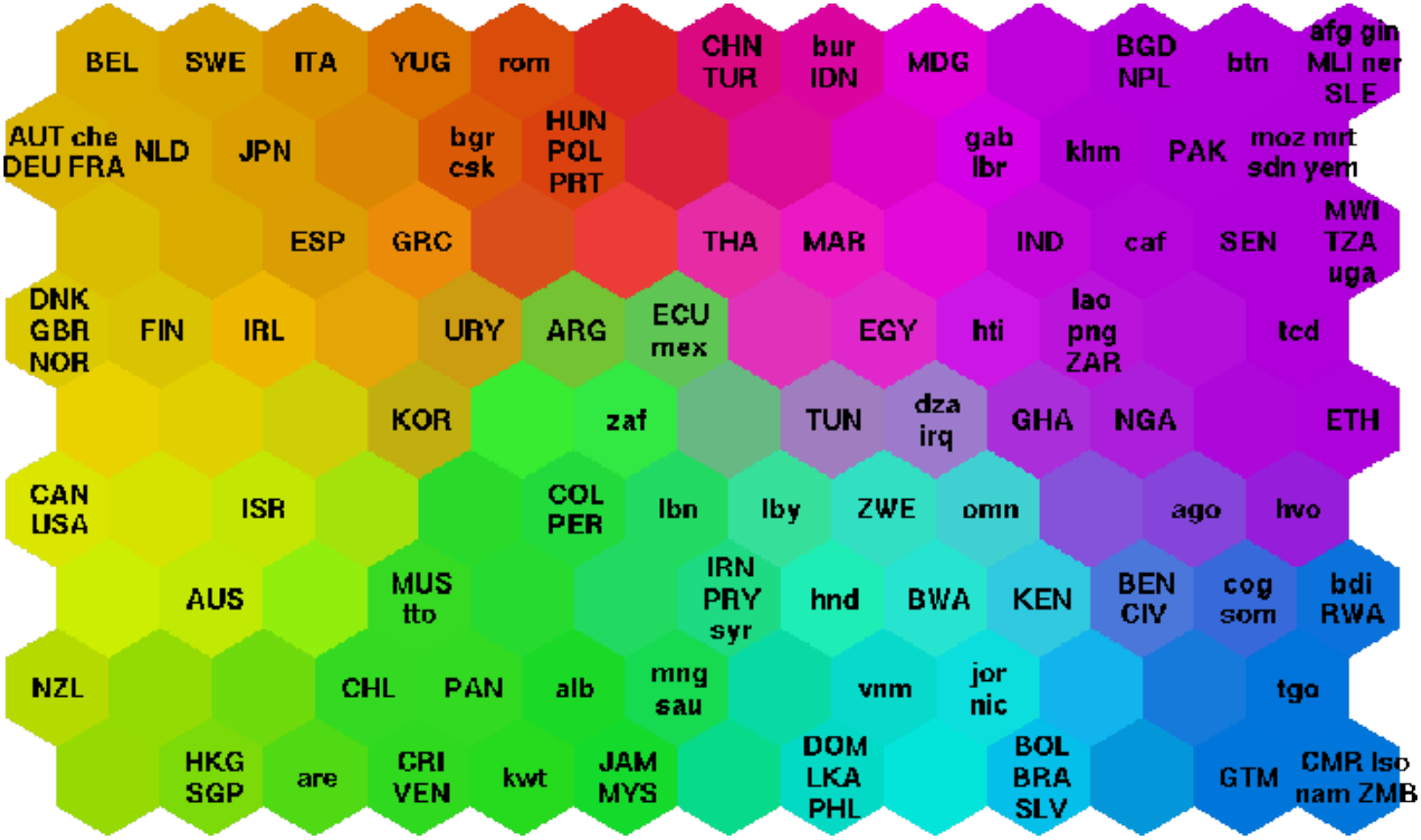
World Bank Statistics

- Data: World Bank statistics of countries in 1992.
- 39 indicators considered e.g., health, nutrition, educational services, etc.
- The complex joint effect of these factors can can be visualized by organizing the countries using the self-organizing map.

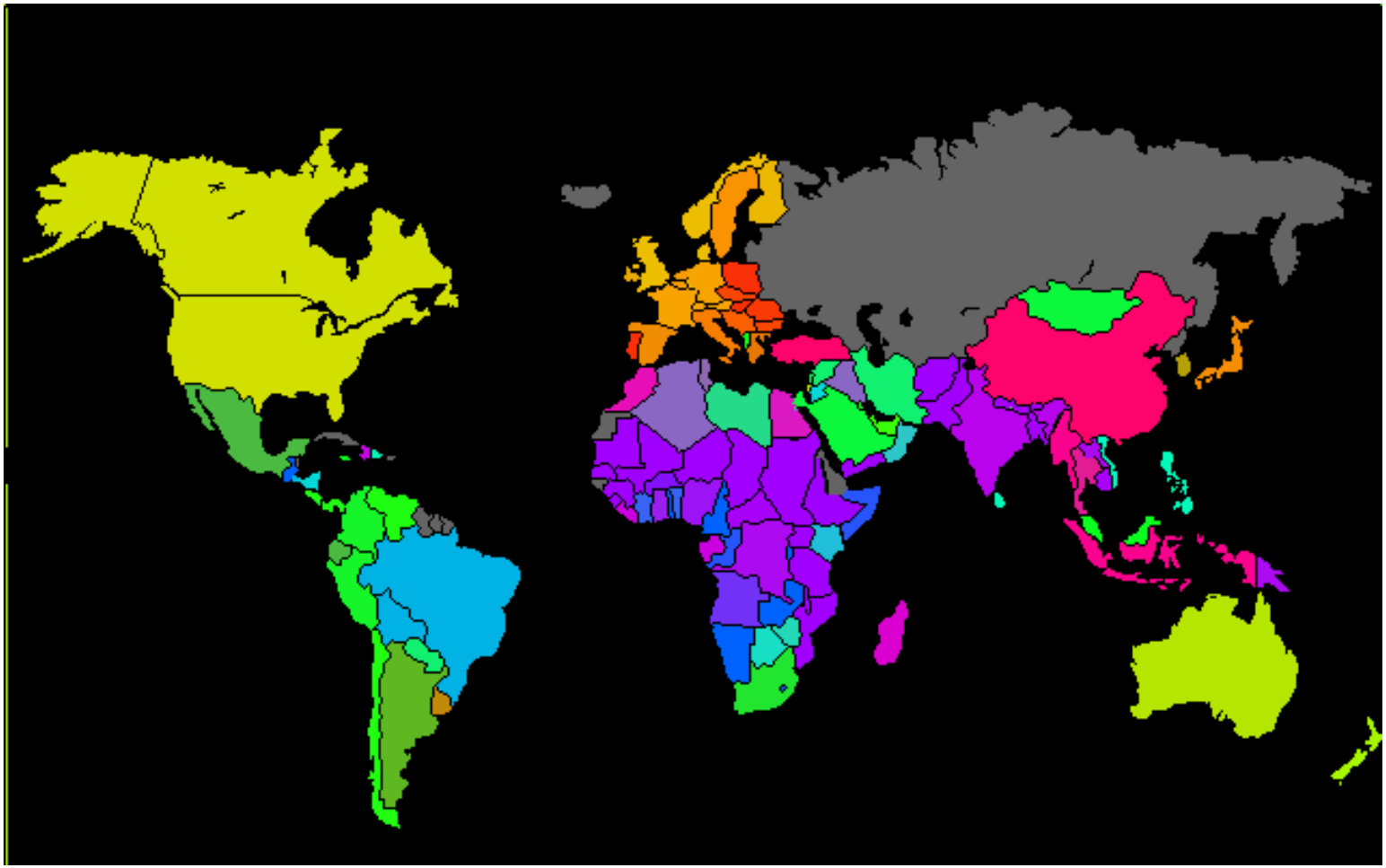
World Poverty PCA

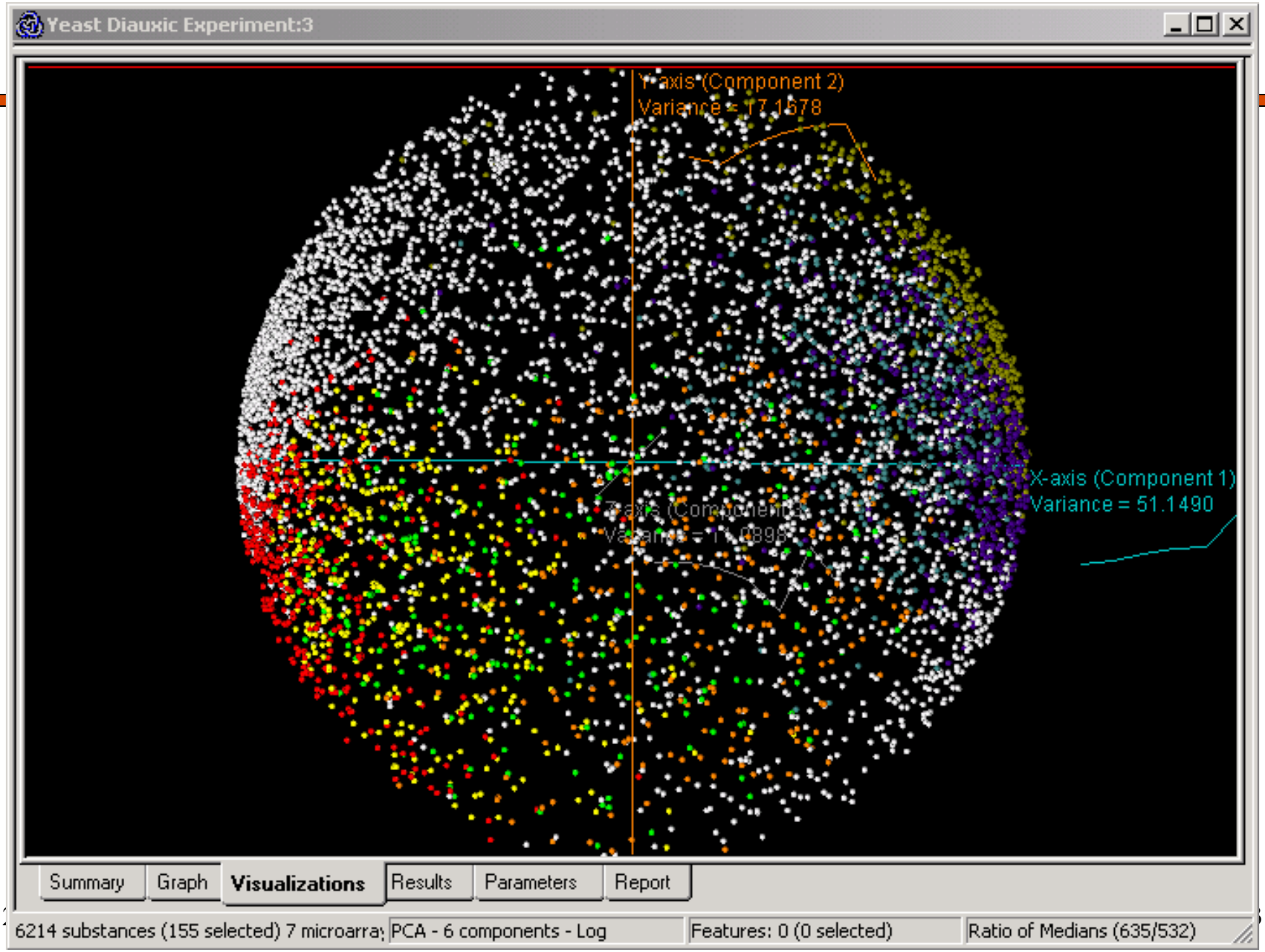


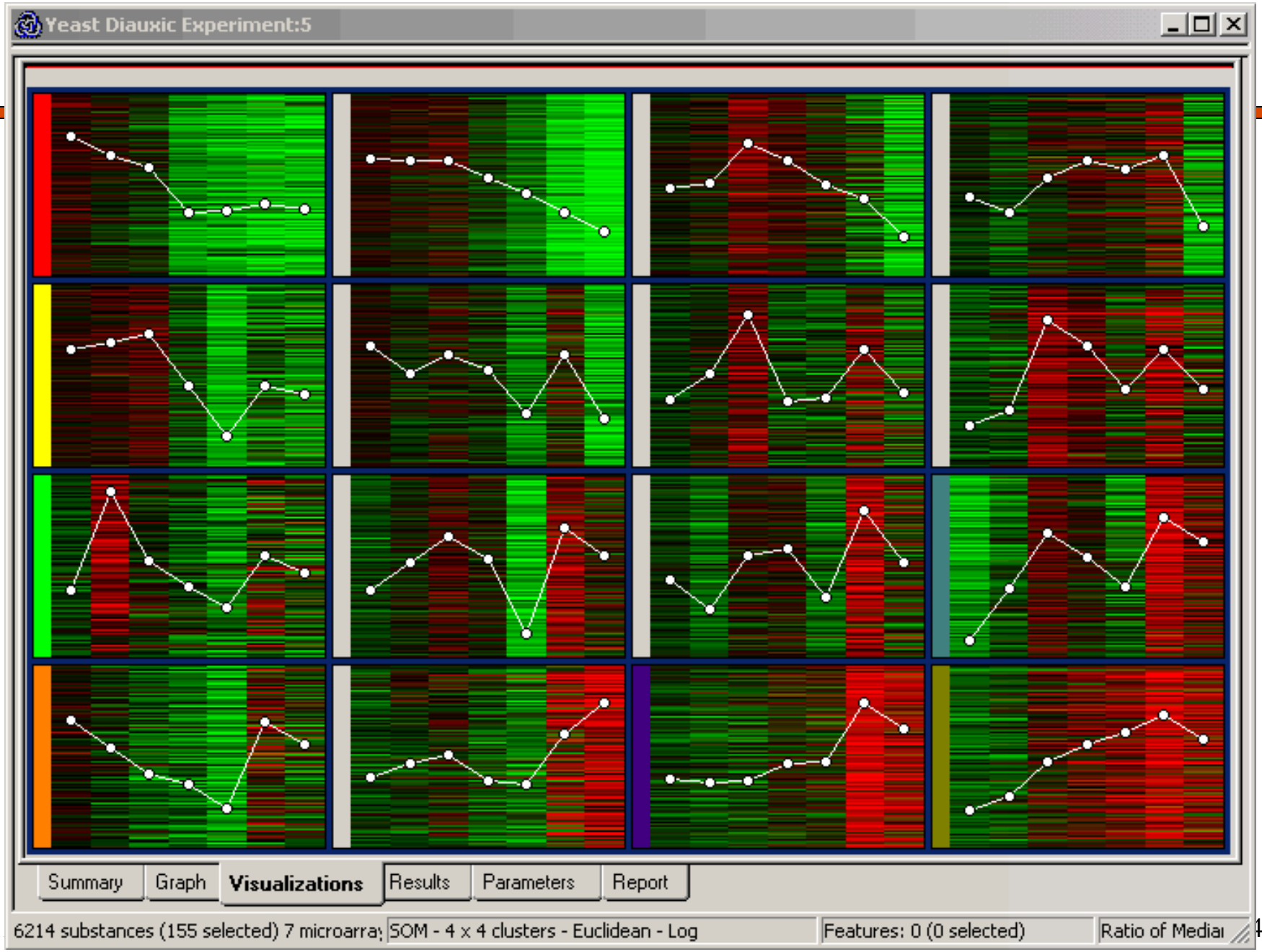
World Poverty SOM



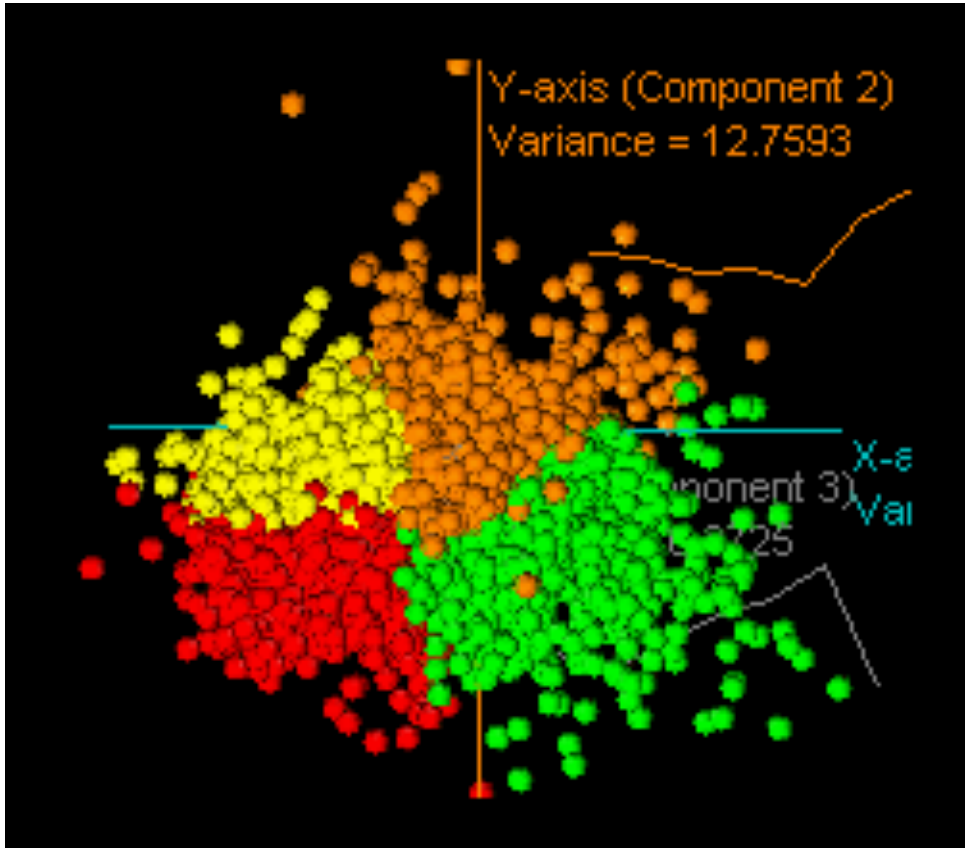
World Poverty Map



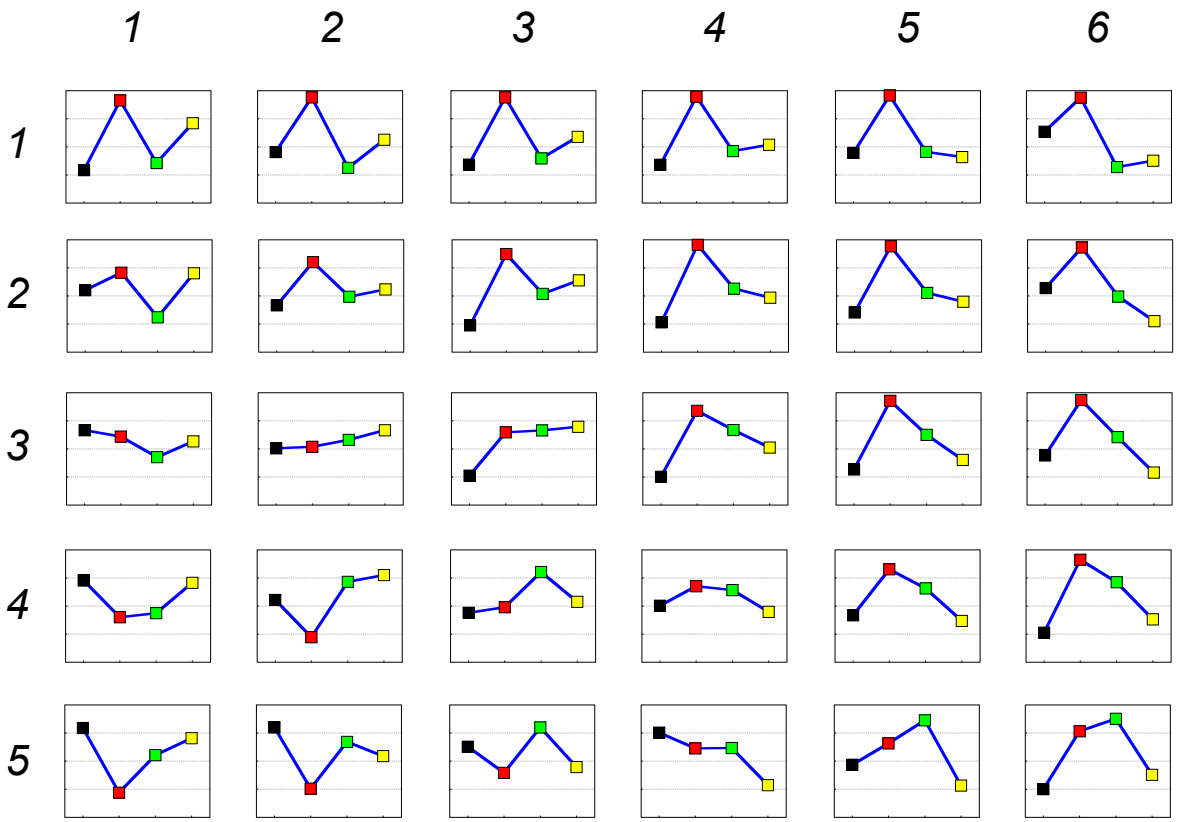

Yeast Diauxic Experiment:3
Y-axis (Component 2)
Variance = 17.1578
X-axis (Component 1)
Variance = 51.1490
Summary
Graph
Visualizations
Results
Parameters
Report
6214 substances (155 selected) 7 microarra
PCA - 6 components - Log
Features: 0 (0 selected)
Ratio of Medians (635/532)


Yeast Diauxic Experiment:5
Summary
Graph
Visualizations
Results
Parameters
Report
6214 substances (155 selected) 7 microarra
SOM - 4 x 4 clusters - Euclidean - Log
Features: 0 (0 selected)
Ratio of Media

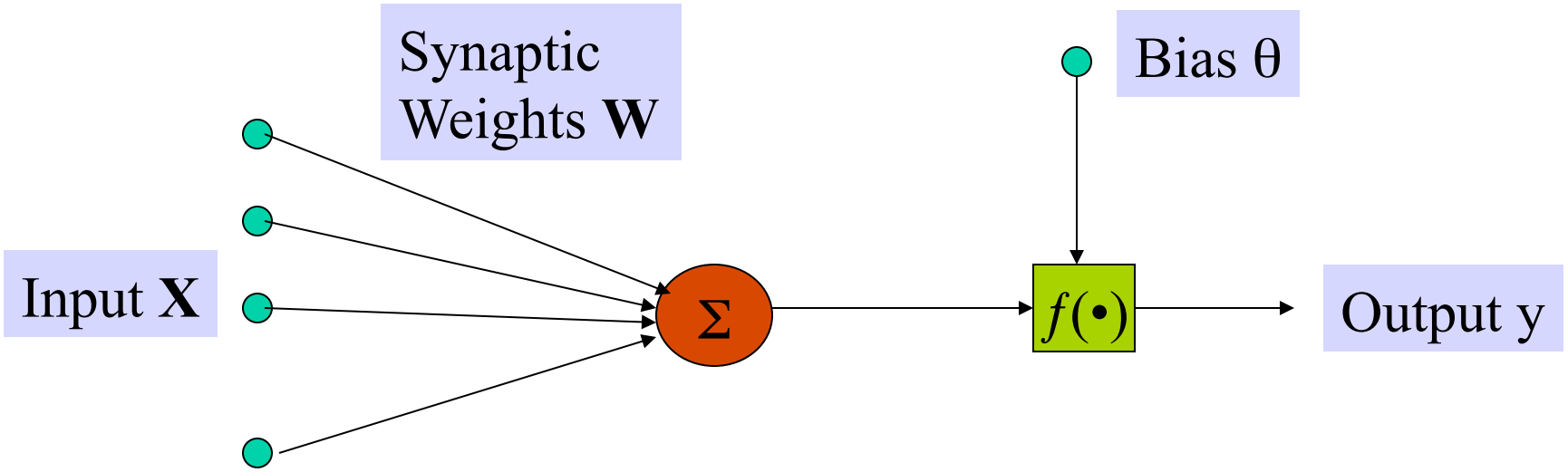
Viewing SOM Clusters on PCA axes



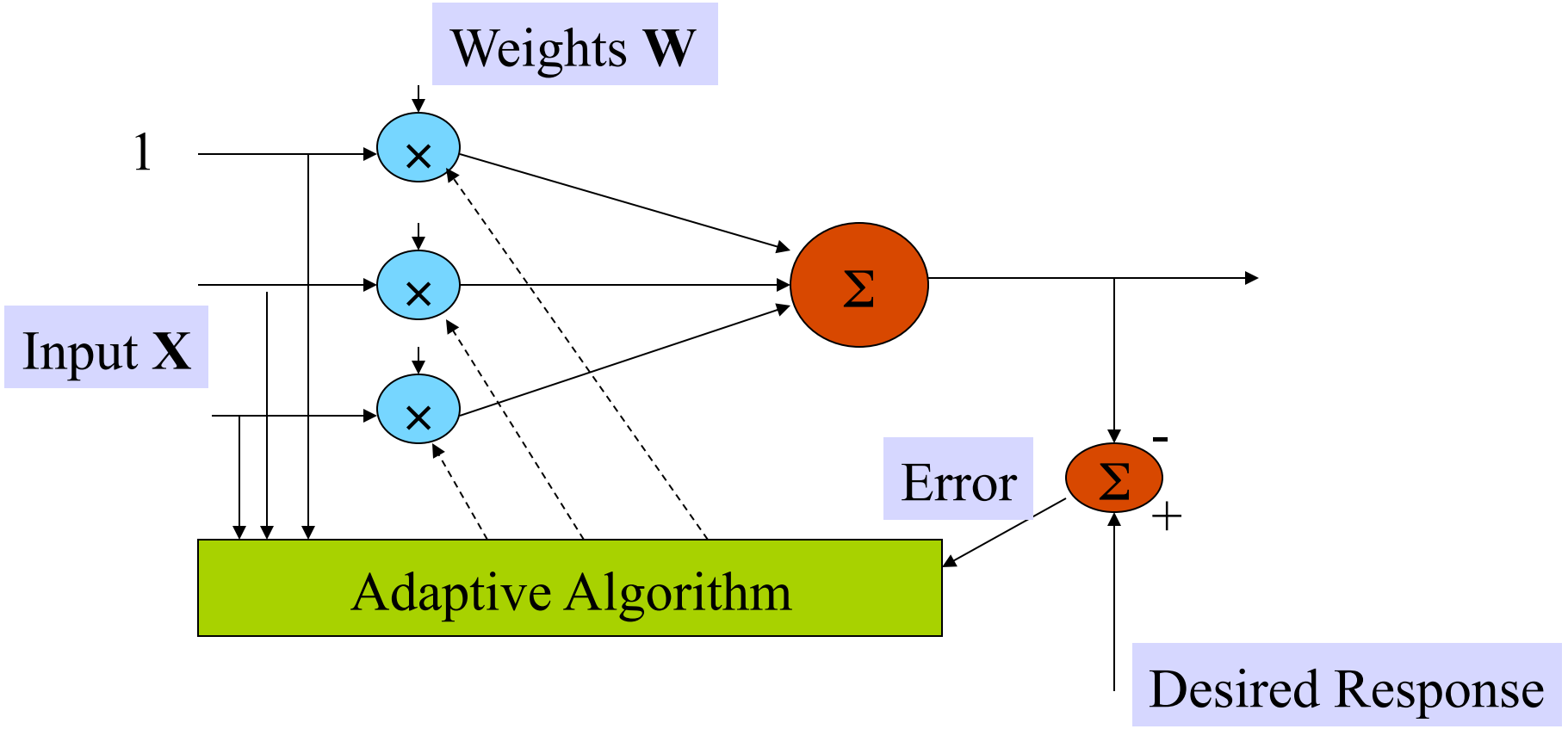
SOM Example [Xiao-rui He]



Neural Networks



Learning NN


Weights W
1
Input X
×
×
×
Σ
Error
Σ
-
+
Adaptive Algorithm
Desired Response

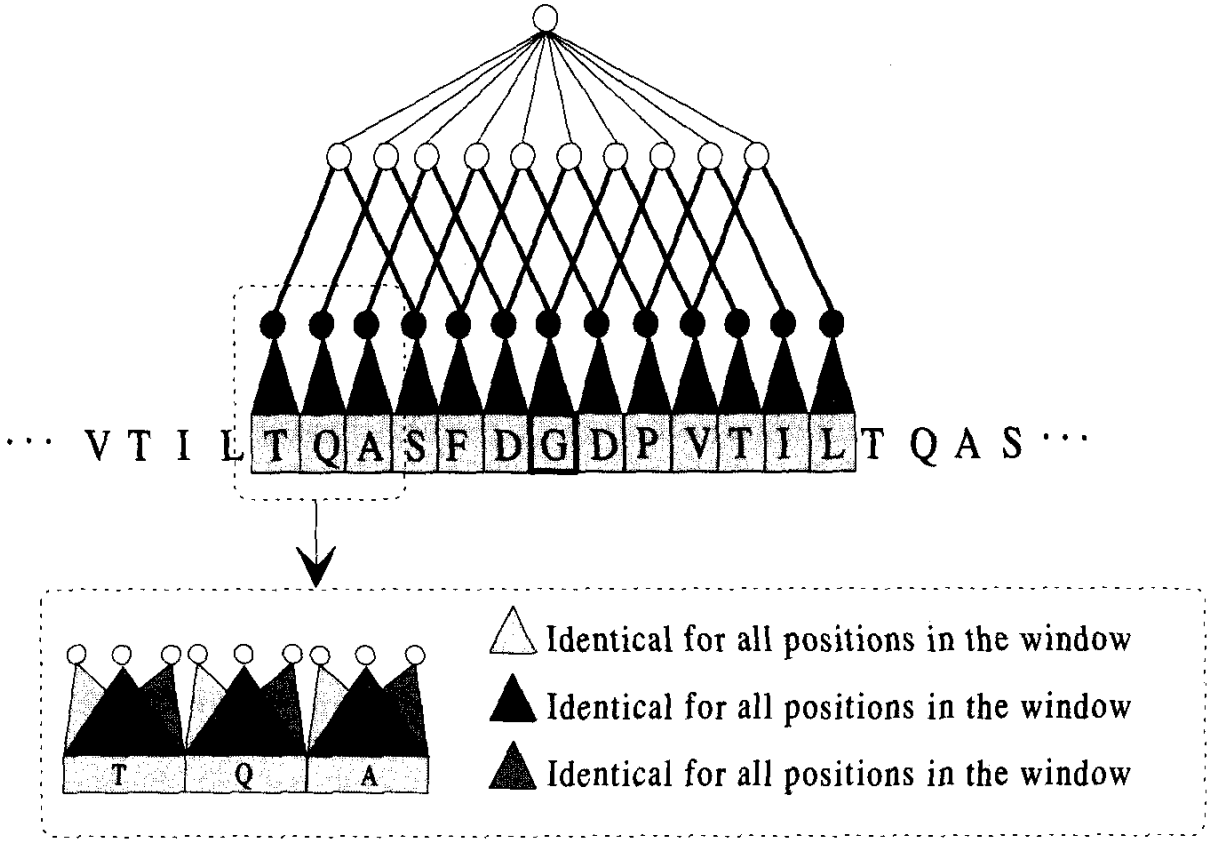
Types of NNs

- Recurrent NN
- Feed-forward NN
- Layered

Other issues

- Hidden layers possible
- Different activation functions possible

Application: Secondary Structure Prediction

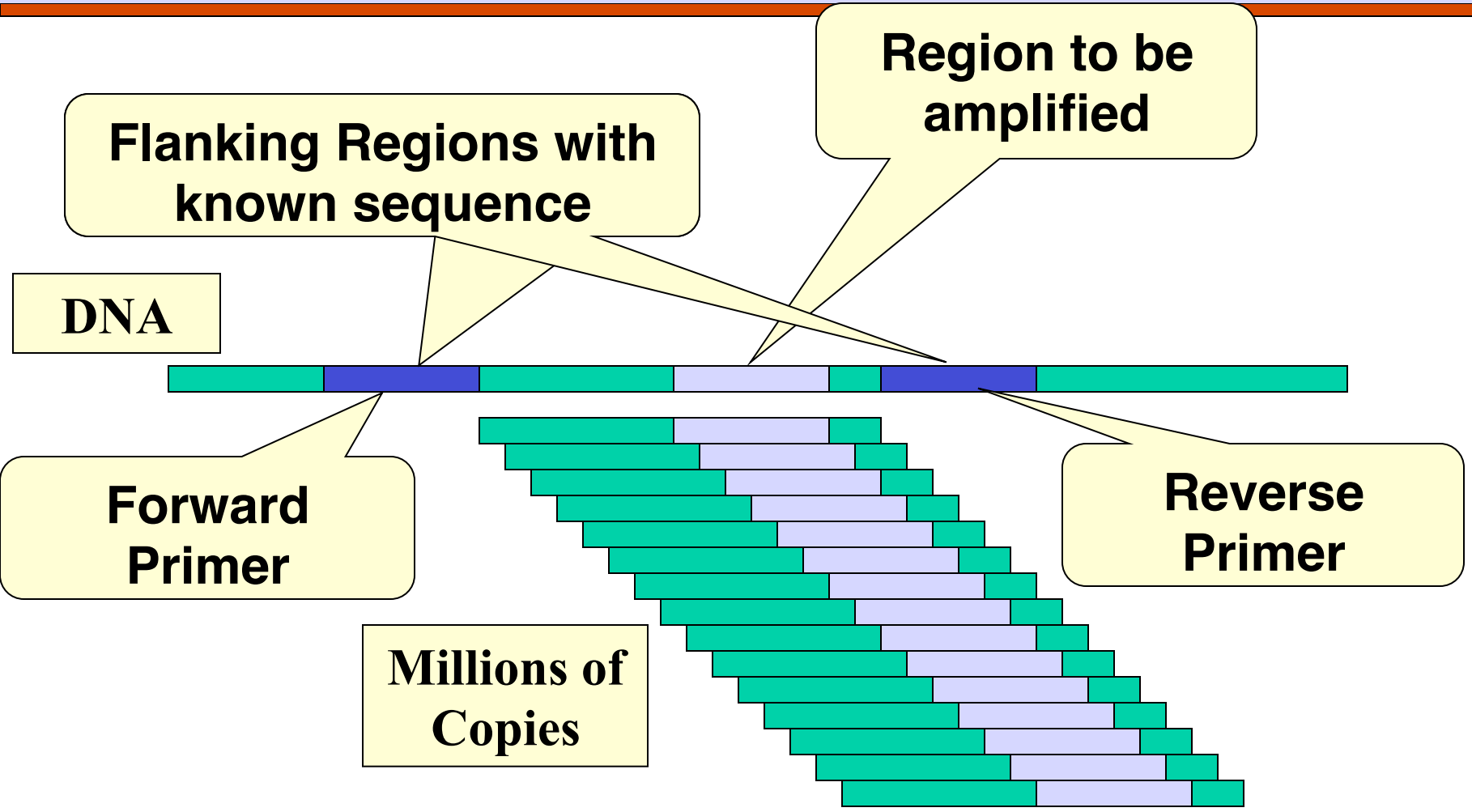


PCR and Sequencing

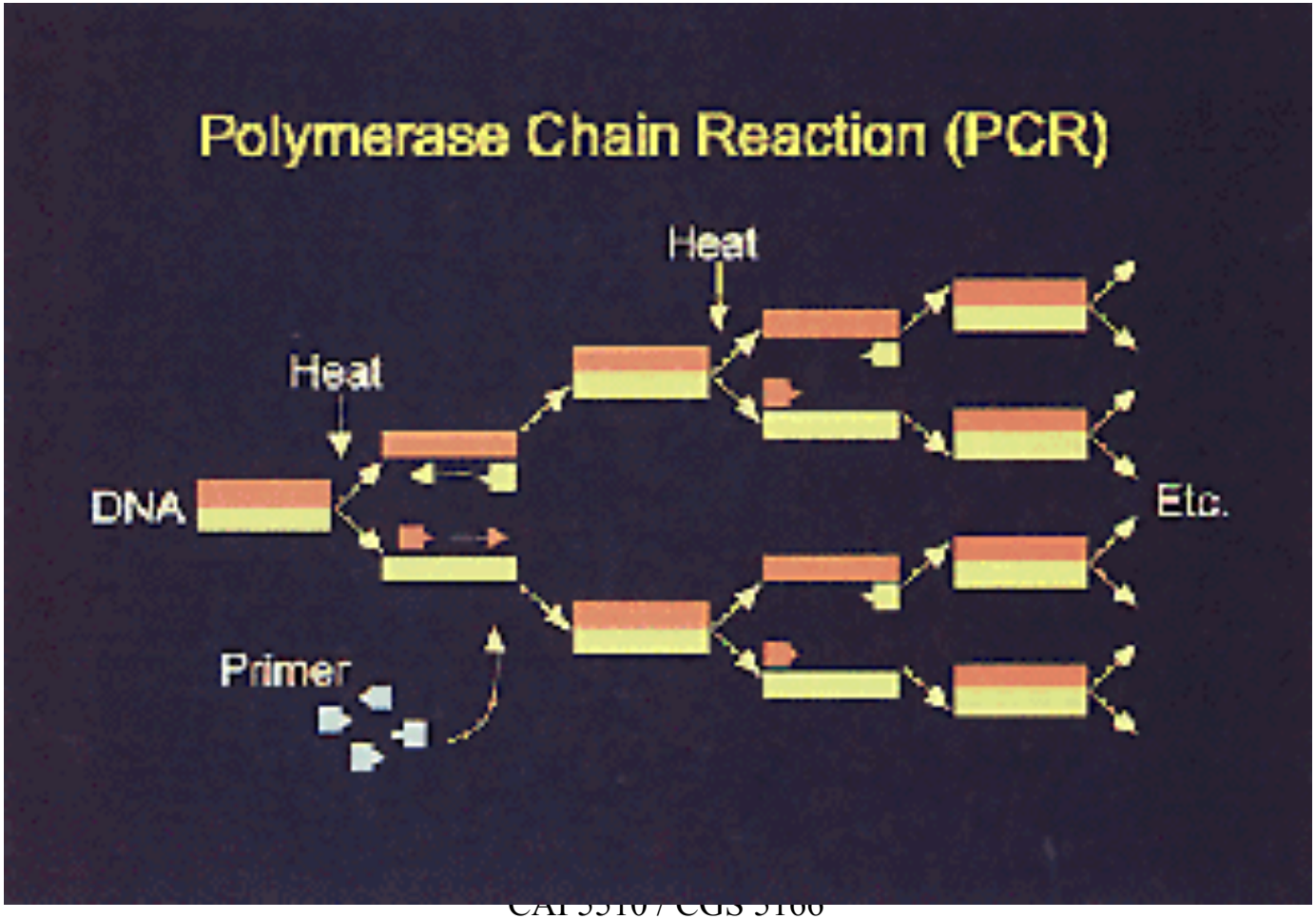
Polymerase Chain Reaction (PCR)

- For testing, large amount of DNA is needed
  - Identifying individuals for forensic purposes
    - (0.1 microliter of saliva contains enough epithelial cells)
  - Identifying pathogens (viruses and/or bacteria)
- PCR is a technique to amplify the number of copies of a specific region of DNA.
- Useful when exact DNA sequence is unknown
- Need to know "flanking" sequences
- Primers designed from "flanking" sequences

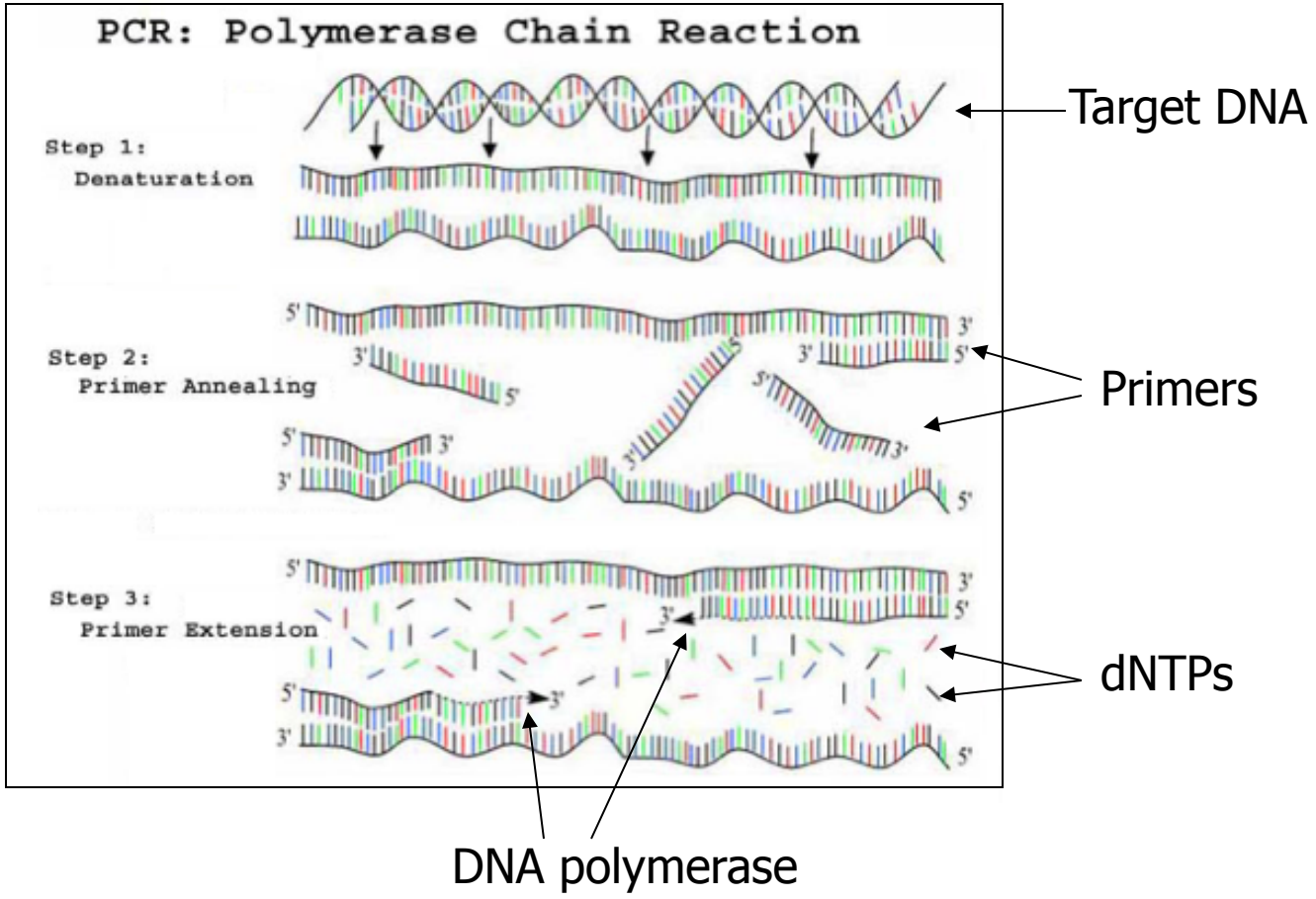
PCR


Region to be amplified
Flanking Regions with known sequence
DNA
Forward Primer
Reverse Primer
Millions of Copies

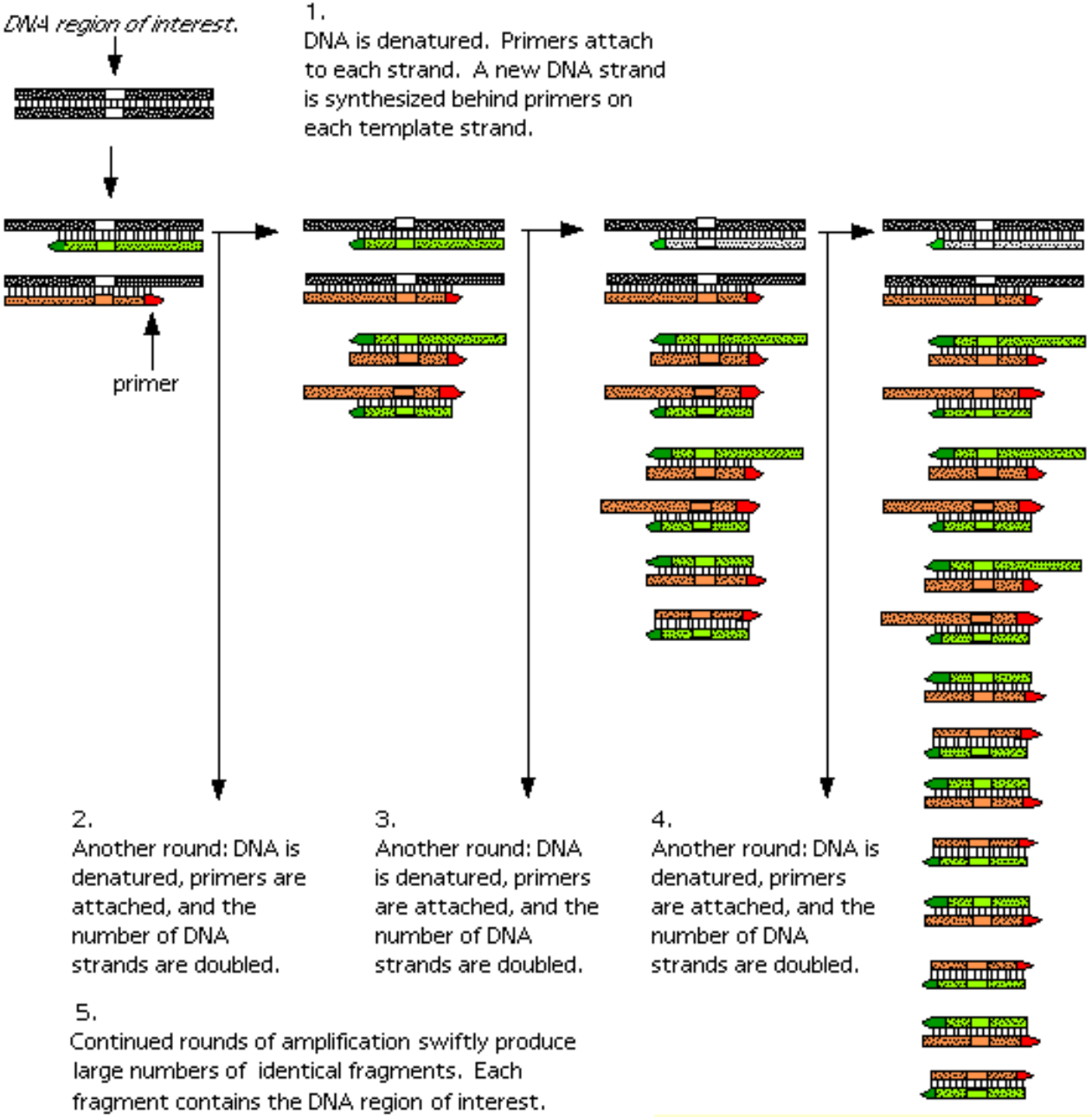
PCR



Schematic outline of a typical PCR cycle



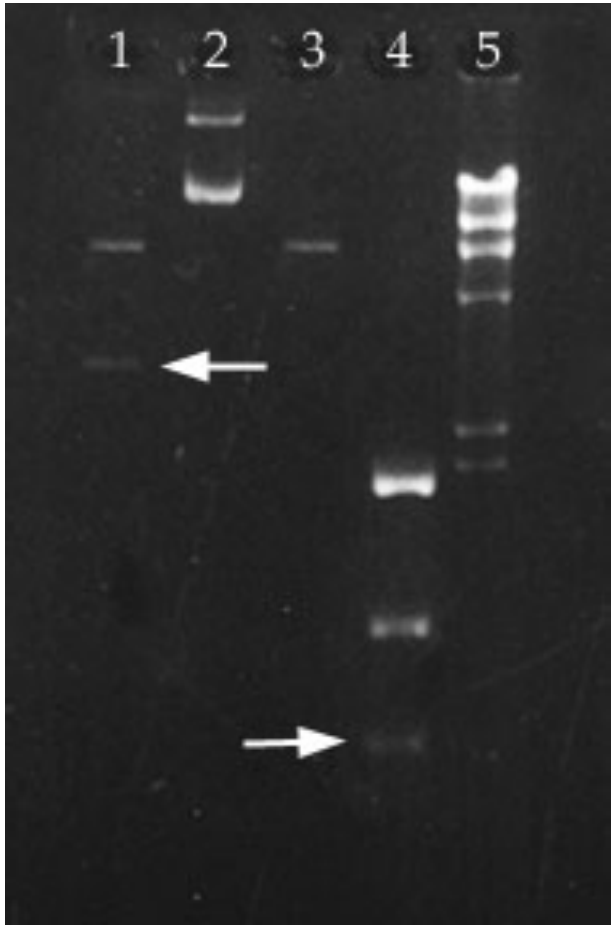
POLYMERASE CHAIN REACTION



Gel Electrophoresis

- Used to measure the lengths of DNA fragments.
- When voltage is applied to DNA, different size fragments migrate to different distances (smaller ones travel farther).

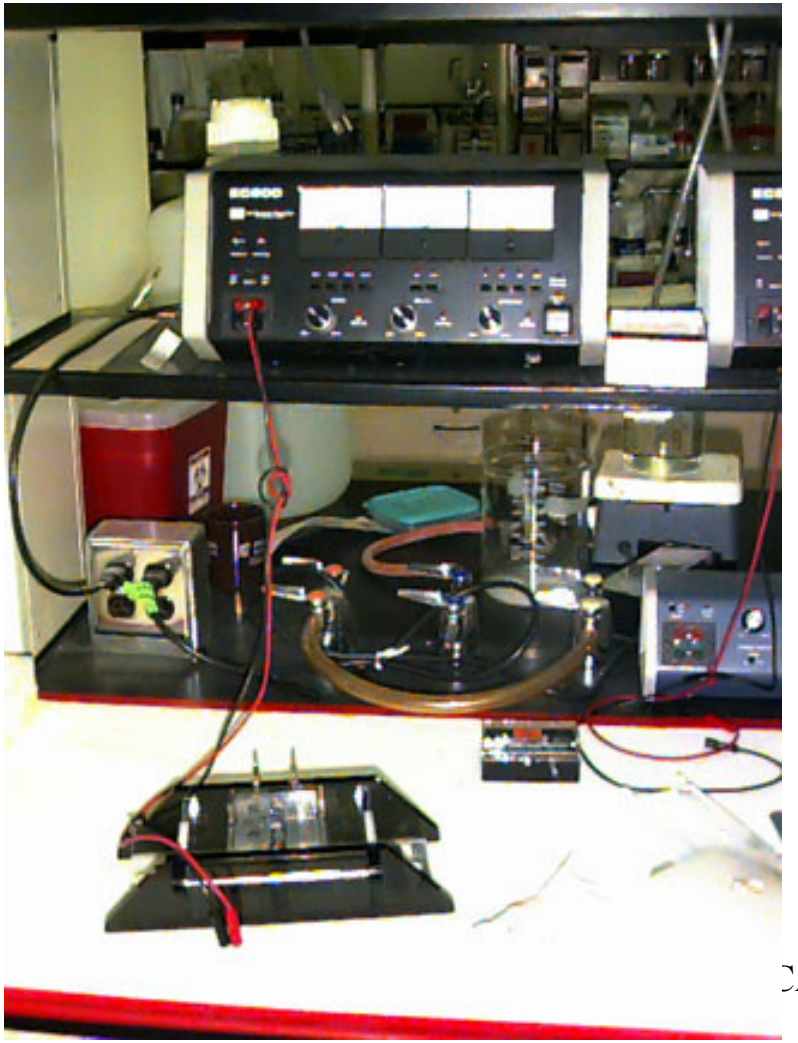
Gel Pictures

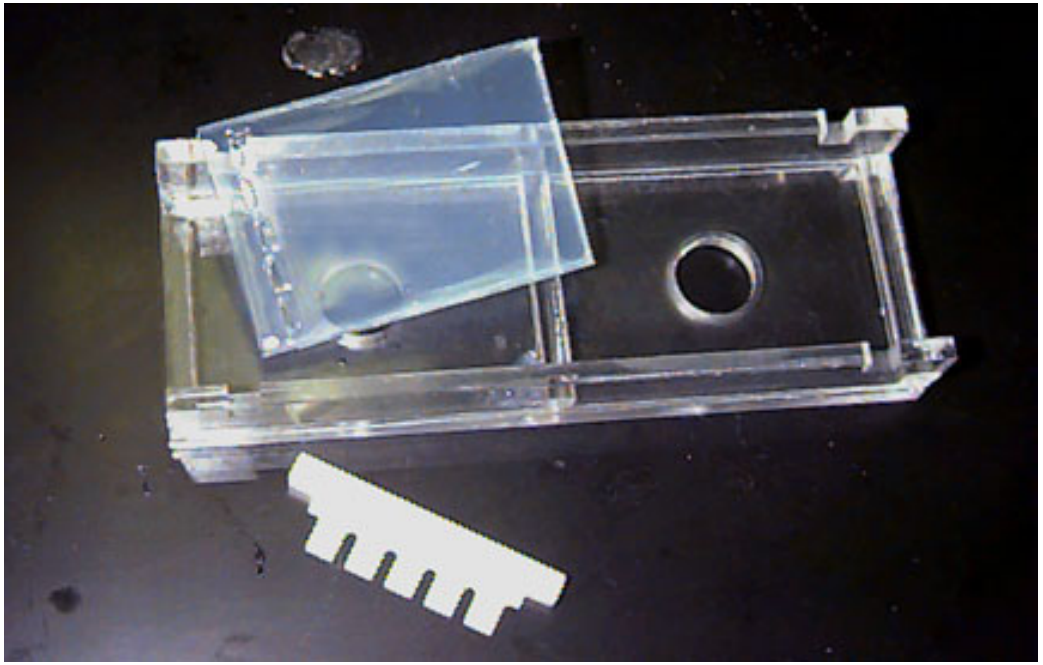


Gel Electrophoresis: Measure sizes of fragments

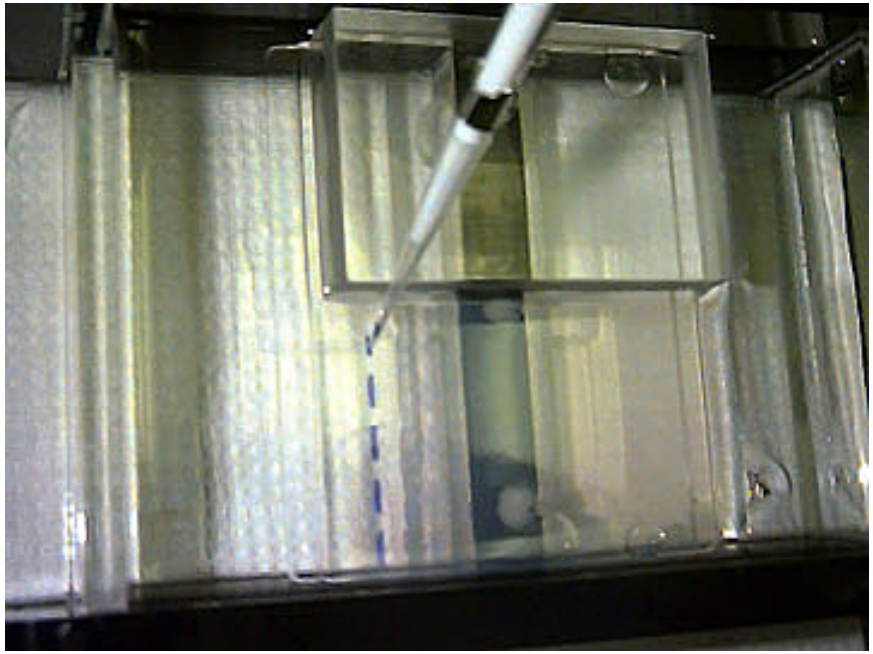
- The phosphate backbone makes DNA a highly negatively charged molecule.
- DNA can be separated according to its size.
- Gel: allow hot 1% solution of purifed agarose to cool and solidify/polymerize.
- DNA sample added to wells at the top of a gel and voltage is applied. Larger fragments migrate through the pores slower.
- Varying concentration of agarose makes different pore sizes & results.
- Proteins can be separated in much the same way, only acrylamide is used as the crosslinking agent.

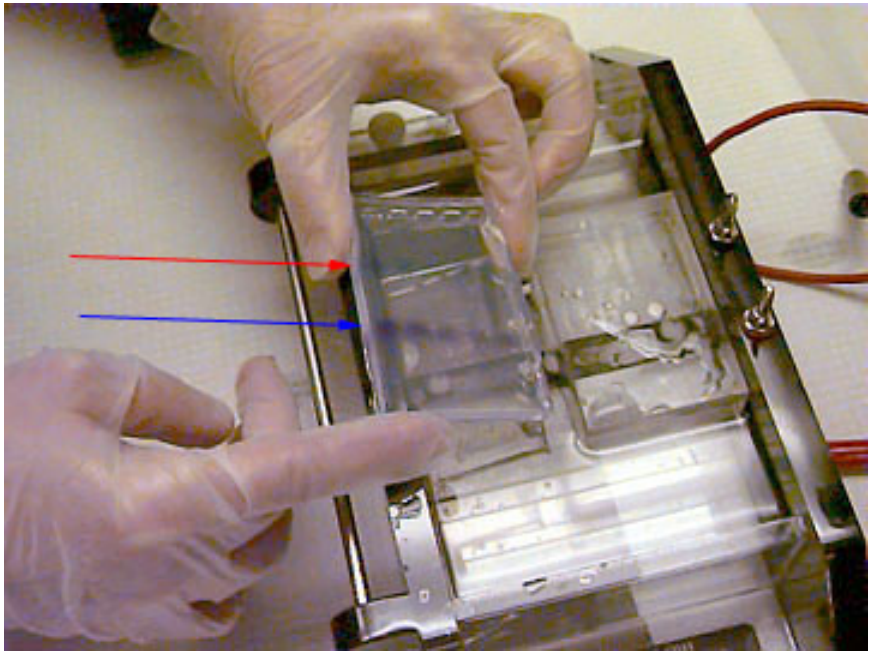
Gel Electrophoresis

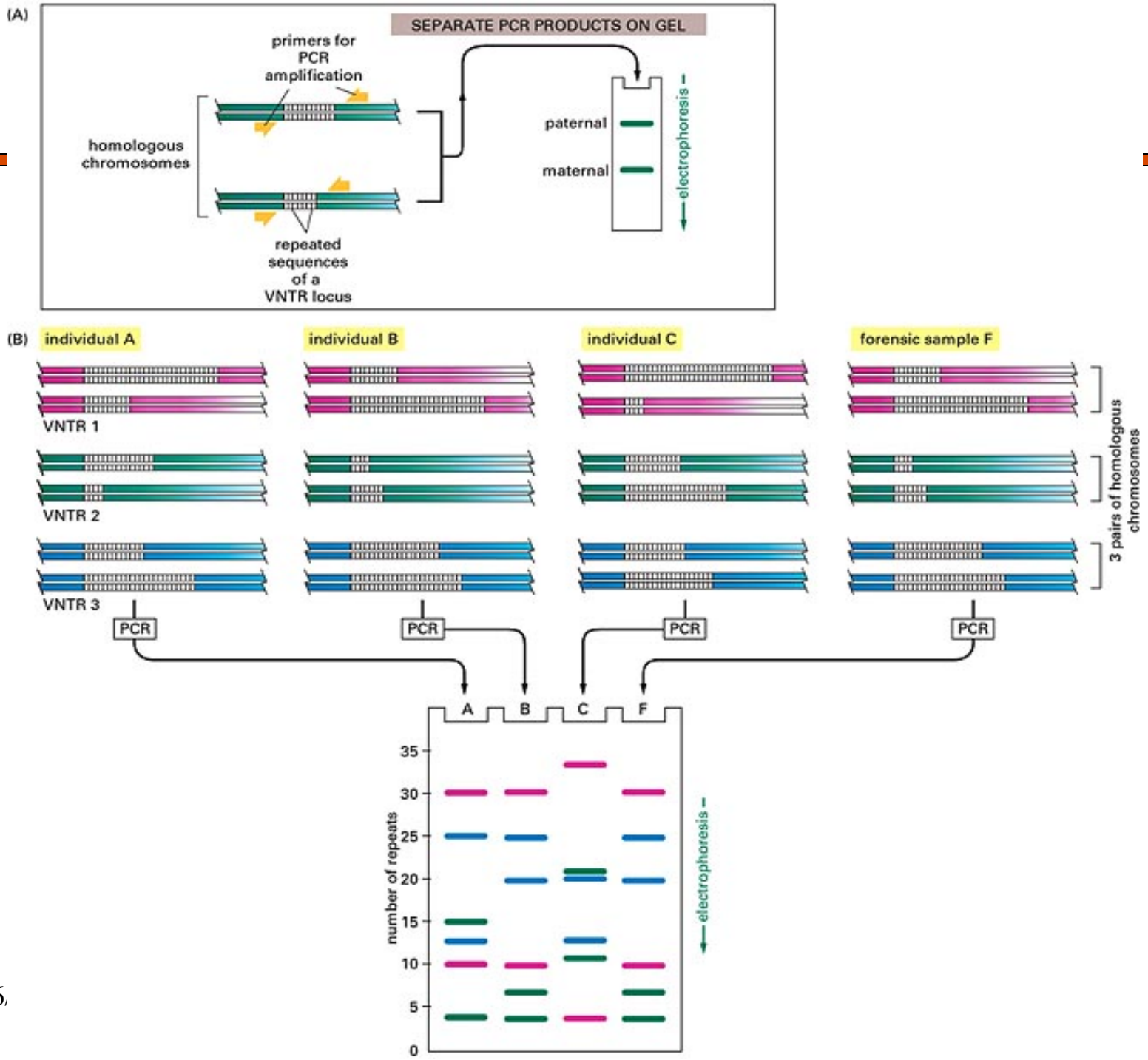




Gel Electrophoresis






(A)
SEPARATE PCR PRODUCTS ON GEL
primers for PCR amplification
homologous chromosomes
repeated sequences of a VNTR locus
paternal
maternal
electrophoresis
(B)
individual A
individual B
individual C
forensic sample F
VNTR 1
VNTR 2
VNTR 3
3 pairs of homologous chromosomes
PCR
PCR
PCR
PCR
A
B
C
F
number of repeats
35
30
25
20
15
10
5
0
electrophoresis

Sequencing

Why sequencing?

- Useful for further study:
  - Locate gene sequences, regulatory elements
  - Compare sequences to find similarities
  - Identify mutations
  - Use it as a basis for further experiments

Next 4 slides contains material prepared by Dr. Stan Metzenberg. Also see:
http://stat-www.berkeley.edu/users/terry/Classes/s260.1998/Week8b/week8b/node9.html

History

- Two methods independently developed in 1974
  - Maxam & Gilbert method
  - Sanger method: became the standard
- Nobel Prize in 1980

Original Sanger Method

- (Labeled) Primer is annealed to template strand of denatured DNA. This primer is specifically constructed so that its 3' end is located next to the DNA sequence of interest. Once the primer is attached to the DNA, the solution is divided into four tubes labeled "G", "A", "T" and "C". Then reagents are added to these samples as follows:
  - "G" tube: ddGTP, DNA polymerase, and all 4 dNTPs
  - "A" tube: ddATP, DNA polymerase, and all 4 dNTPs
  - "T" tube: ddTTP, DNA polymerase, and all 4 dNTPs
  - "C" tube: ddCTP, DNA polymerase, and all 4 dNTPs
- DNA is synthesized, & nucleotides are added to growing chain by the DNA polymerase. Occasionally, a ddNTP is incorporated in place of a dNTP, and the chain is terminated. Then run a gel.
- All sequences in a tube have same prefix and same last nucleotide.
- http://www.wellcome.ac.uk/Education-resources/Teaching-and-education/Animations/DNA/WTDV026689.htm

Sanger Method

- Example of sequences seen in gel from "G" tube:

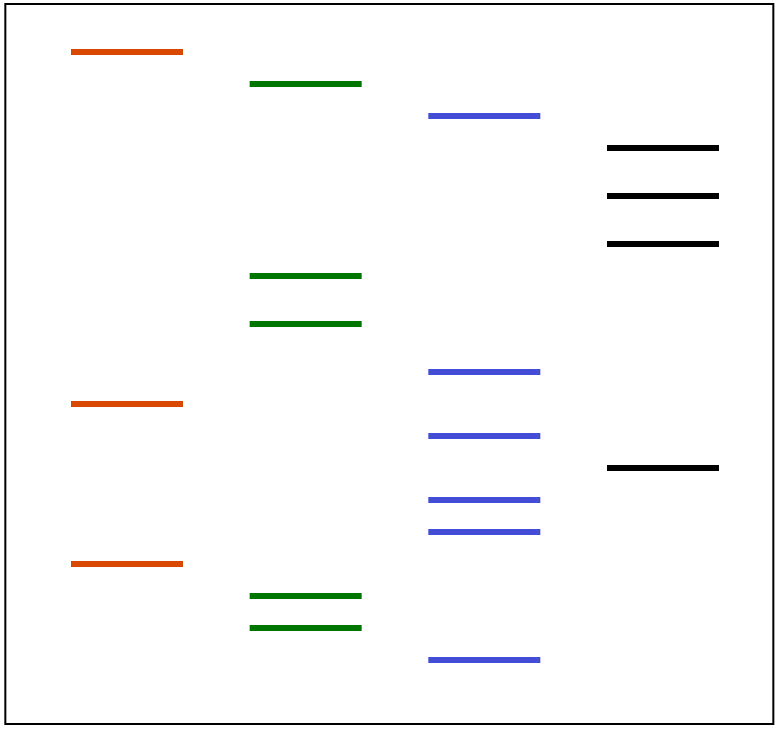
```
     5'-GAATGTCCTTTCTCTAAGTCCTAAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

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

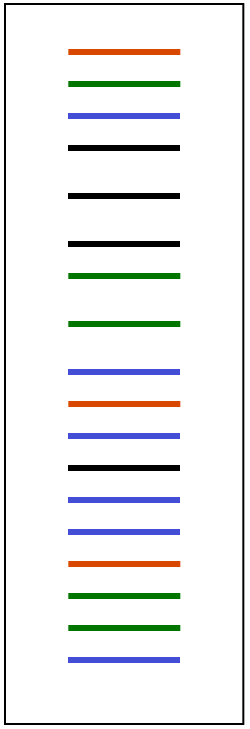
Modified Sanger

- Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color dye

Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTGCA





A C G T

Sequencing

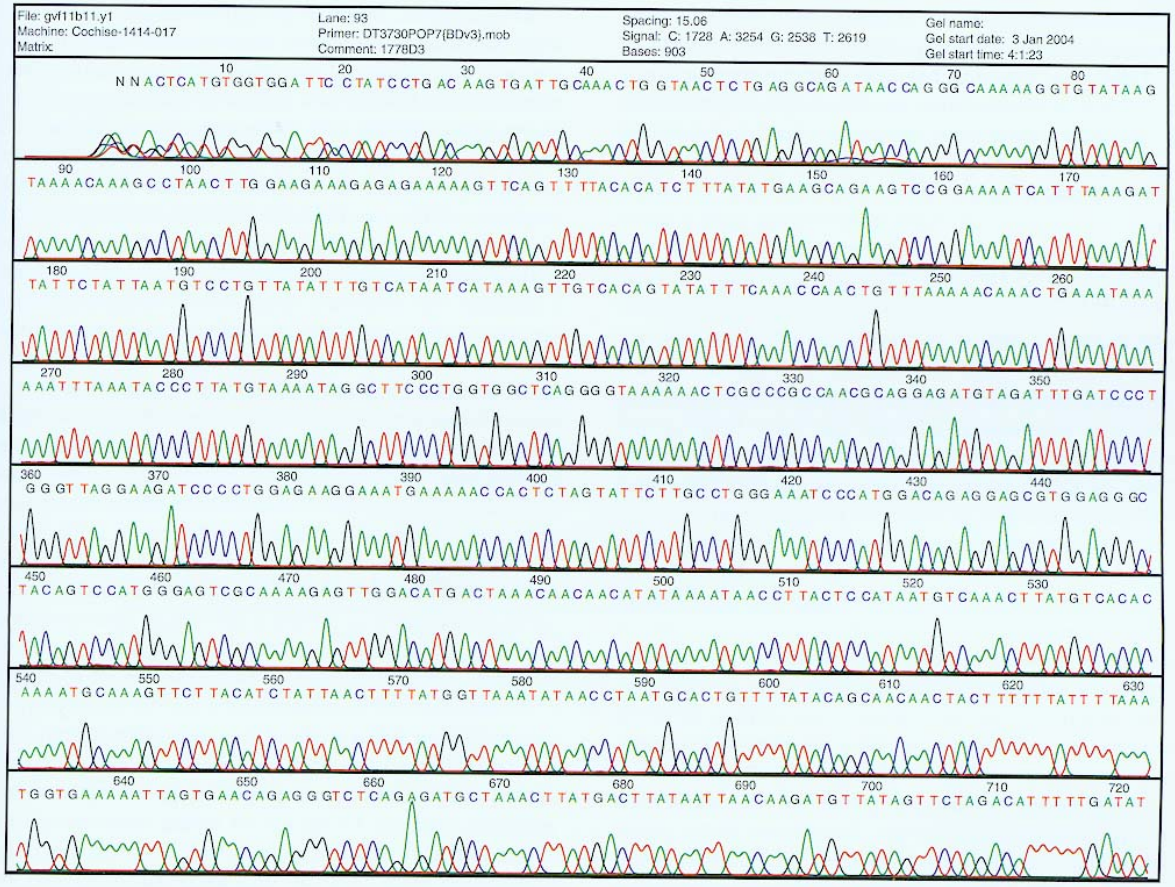
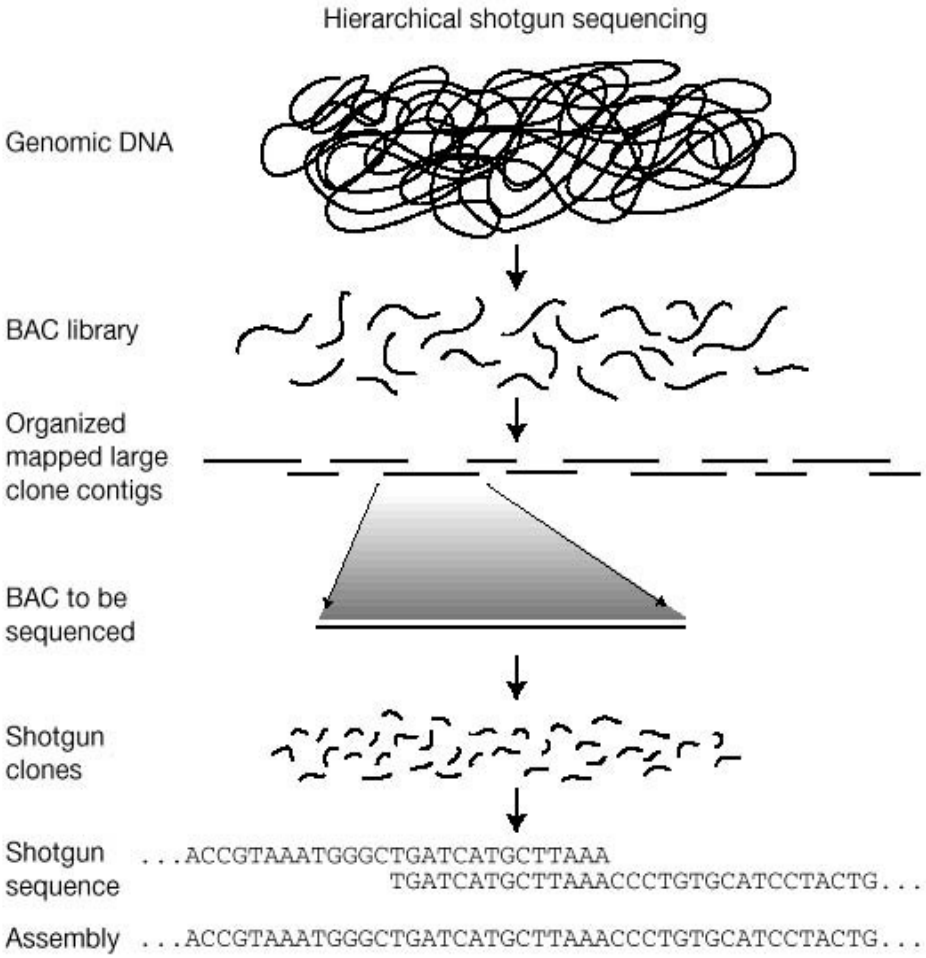


FIGURE 13.3 **A sample chromatogram, as viewed with the vtrace program (Ewing, 2002). Signal intensities corresponding to fragments ending with A (green), C (blue), G (black), and T (red) are shown out to approximately 722 bases.**

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Sequencing

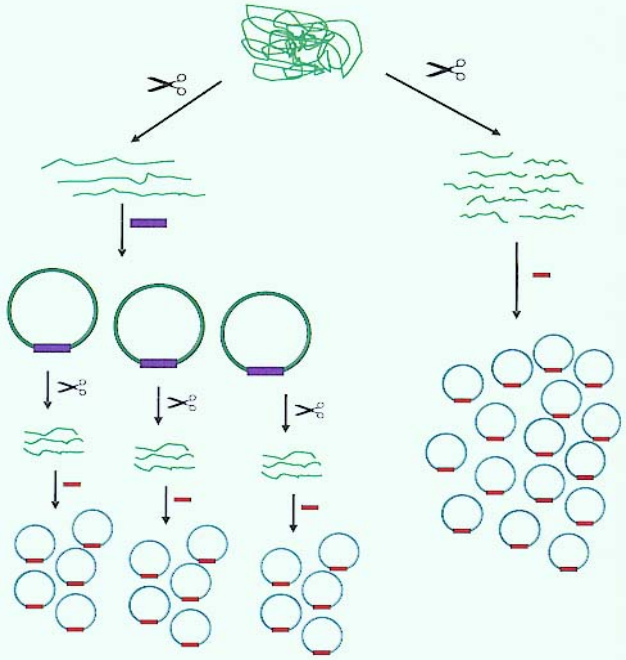
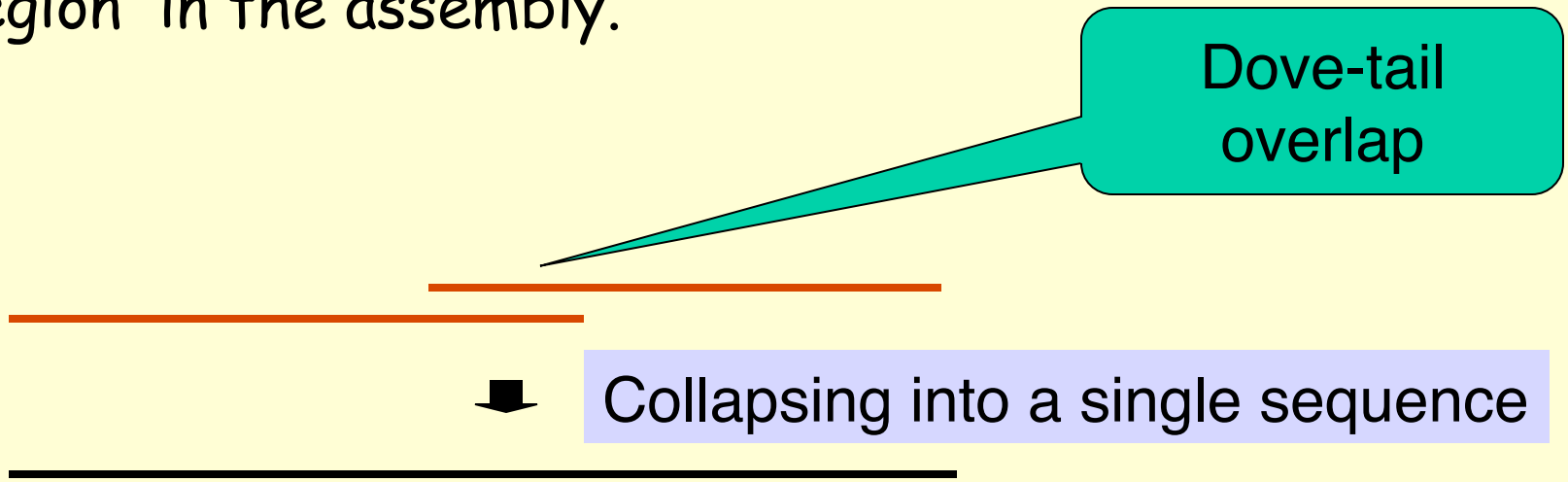


FIGURE 13.1 Shotgun cloning. Genomic DNA sequencing begins with isolated genomic DNA in green at the top of the figure. In the hierarchical clone-based shotgun approach on the left, DNA is sheared and the size is selected for large fragments on the order of 200 Kb, then ligated to a suitable vector, such as a BAC vector shown in blue. Individually isolated clones in turn are sheared independently, generating fragments of approximately 4 Kb, which are then ligated to a small-scale vector, typically a plasmid (red bar) suitable for sequencing reactions. The whole genome shotgun approach bypasses the intermediate large-insert clone and generates large numbers of small fragments, typically 4 Kb and 10 Kb.

Sequencing: Generate Contigs

- Short for "contiguous sequence". A continuously covered region in the assembly.

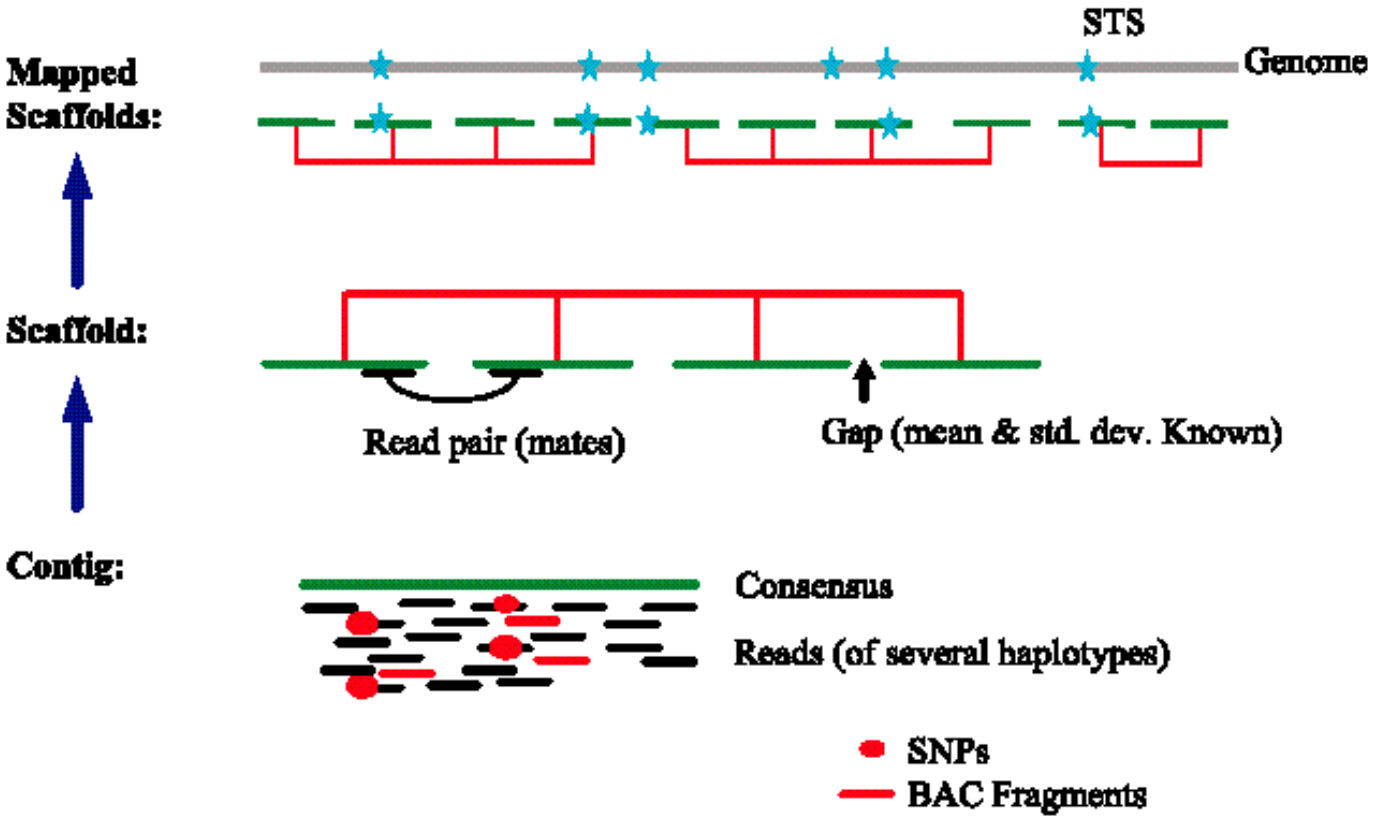


- Jang W et al (1999) Making effective use of human genomic sequence data. Trends Genet. 15(7): 284-6.
  Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11(9): 1541-8.

Paired Reads

- Scaffold (supercontig): formed when two contigs with no sequence overlap can be linked
  - Data from paired end reads help create scaffolds with known gaps
    - If two reads end up in two different contigs, then we can link contigs to form scaffold.

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Human Genome Project

- Many videos available on youtube.com, dnatube.com, and elsewhere.
- Find some and watch them.

Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10
-

  - --ACCGT--
  - ----CGTGC
  - TTAC-----
  - -TACCGT-
  - TTACCGTGC

Assembly: Complications

- Errors in input sequence fragments (~3%)
  - Indels or substitutions
- Contamination by host DNA
- Chimeric fragments (joining of non-contiguous fragments)
- Unknown orientation
- Repeats (long repeats)
  - Fragment contained in a repeat
  - Repeat copies not exact copies
  - Inherently ambiguous assemblies possible
  - Inverted repeats
- Inadequate Coverage

Assembly: Complications

w = AGTATTGGCAATC

z = AATCGATG

u = ATGCAAACCT

x = CCTTTTGG

y = TTGGCAATCACT

```
AGTATTGGCAATC---AATCGATG------------
--------------------ATGCAAACCT-----
----TTGGCAATCACT------------CCTTTTGG
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG
```

FIGURE 4.20

A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.

```
AGTATTGGCAATC--------CCTTTTGG--------
---------AATCGATG--------TTGGCAATCACT
--------------ATGCAAACCT-------------
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT
```

FIGURE 4.21

Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.

Assembly: Complications

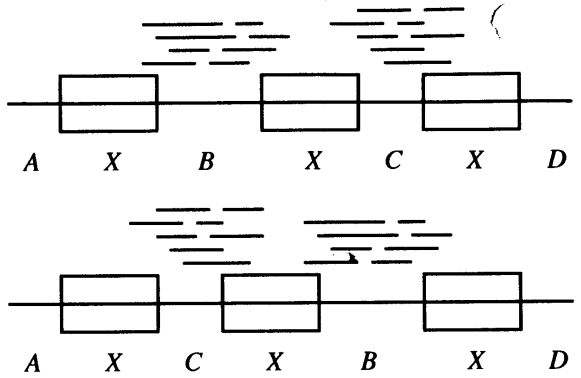


FIGURE 4.8

Target sequence leading to ambiguous assembly because of repeats of the form XXX.

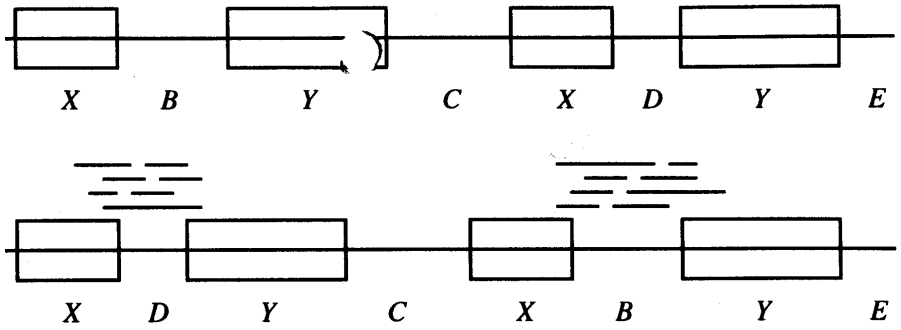


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

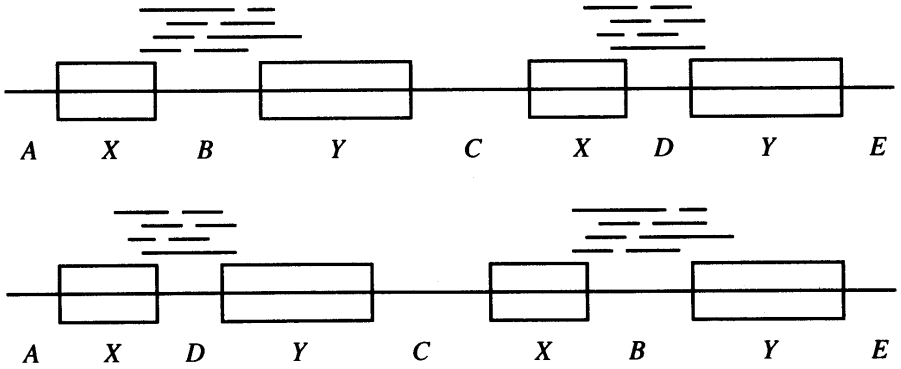


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

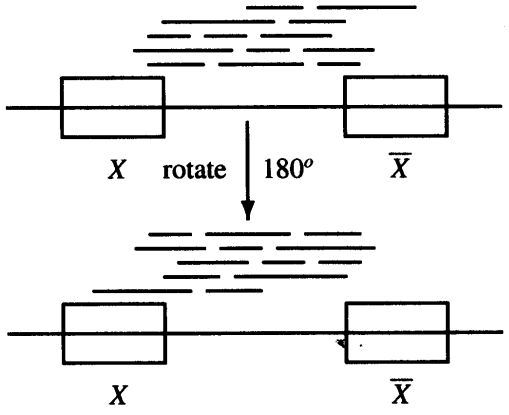


FIGURE 4.10

Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.