ECS 289A

Lecture 2: The Genome and BioTechnologies

Admin

- Projects: 1 per group, 1 project proposal at end of 3rd week of class, 1 progress report (if desired), 1 project report at the last day class.
- Presentations: 1 per student, 15-20min. Think about topics. Anything goes: algorithmic, statistics, biology, medicine, etc.
    - Microarray analysis
        - novel statistical analysis
        - gene classification
        - clustering
    - Experiment analysis and design
    - Promoter sequence analysis
    - Gene Regulation inference
    - Gene Networks and pathways
    - Data integration: sequence + expression + protein + annotation

- Good Source for papers:

  http://linkage.rockefeller.edu/wli/microarray/

- Todays proposed presentations (for next week):

- Genome Assembly software,

Genome Sequence Assembly:Algorithms and Issues, 2002, *Mihai Pop, Steven L. Salzberg, Martin Shumway, IEEE Computer, v35(7)*

- Microarray analysis software,

http://genome-www5.stanford.edu/MicroArray/SMD/restech.html

Genomes

Organization and complexity

- Genomes are the union of all DNA in an organism (there are different types of DNA: nuclear and mitochondrial)
- Only small % (2%) of the human genome is genes. The rest contains various promoter regions and “junk” (>50%)
- Genome sizes vary among organisms, shortest for Phages and Viruses, longest for mammals and some plants (figure from Baldi)

Evolution

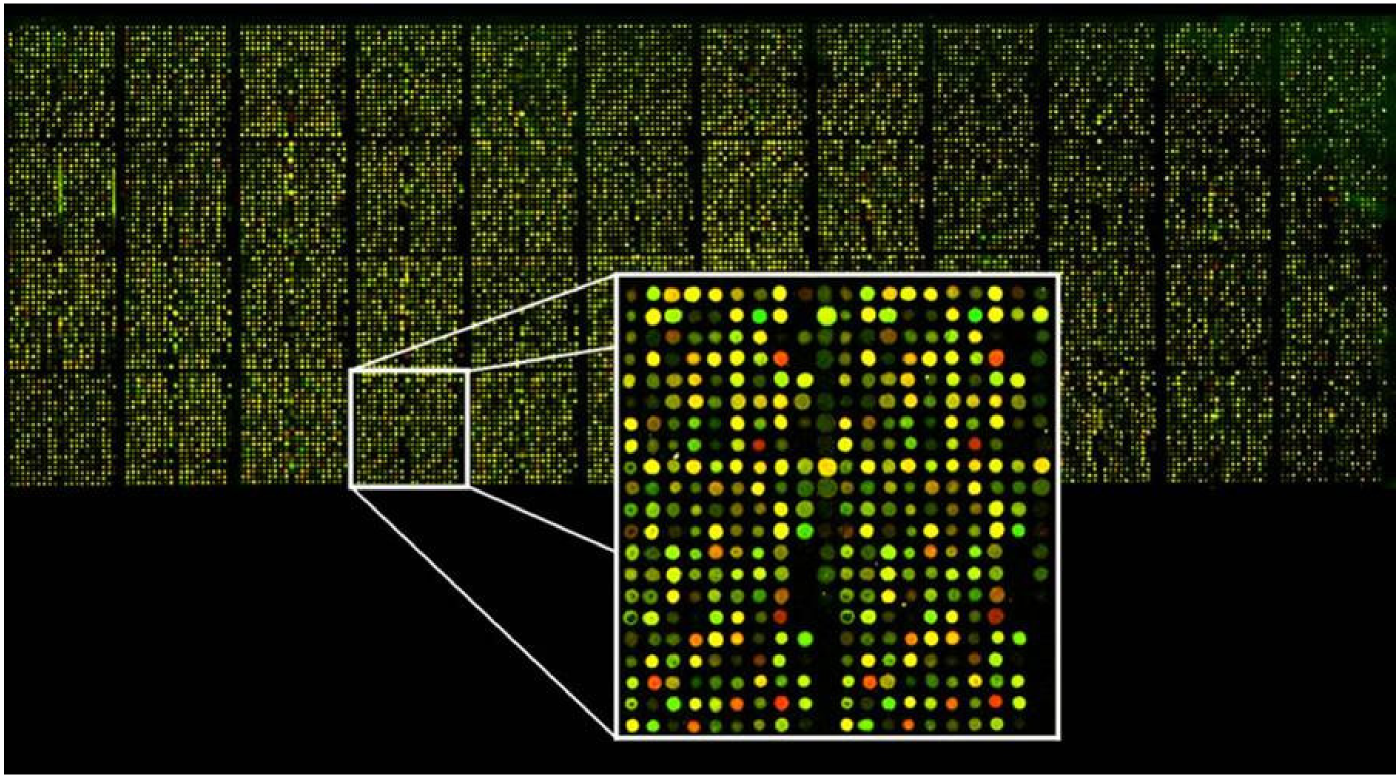
- Changes in the genomes
- Mutations: changes in genome driven by random or particular events. Can be single base change or larger events.
- Recombination: mixing of genomes to produce a new one
- Natural selection: beneficial changes are passed on

Similarity of genomes (i.e. organisms)

- Evolution implies that different organisms would have common ancestors
- Thus similarity comparisons (homology searches) provide clues to evolutionary ancestry (mention phylogeny)

Transcriptome

All possible gene expressions in the organism



Organization and Complexity

- Transcriptome is the measurable level of all different mRNA’s in an organism
- One DNA template multiple mRNAs: alternative splicing
- DNA to mRNA: one way street because of alternative splicing
- The “when and where” of mRNA concentration is coded in the promoter regions, and possibly elsewhere

Evolution

- Evolution of gene expression under emergent properties like network organization

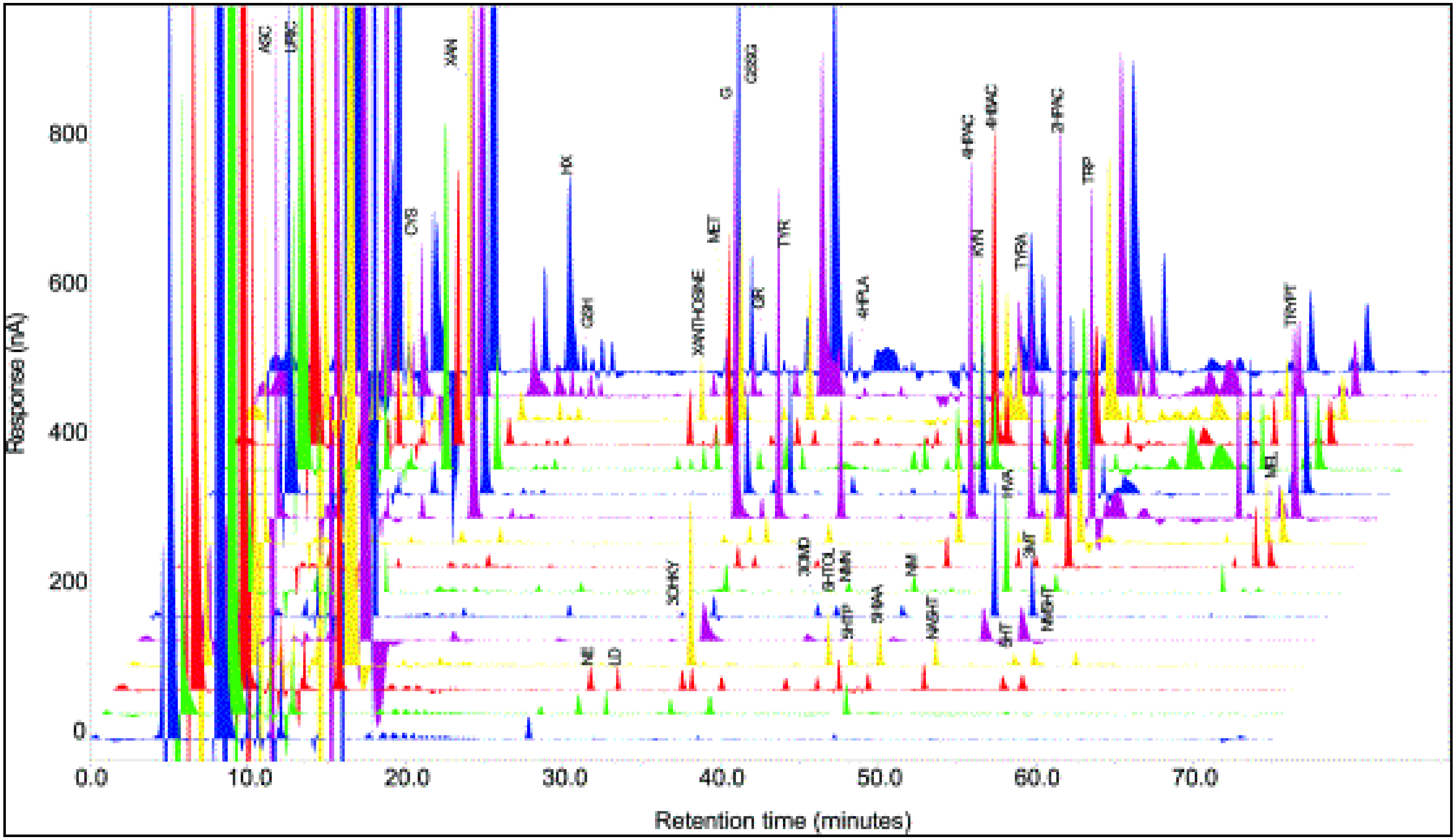
Similarity of Organisms

- Comparison of gene expression from a “system’s perspective”

Proteome

Localization, abundance, and interaction of all proteins in an organism

- Structure: Amino acid sequence, 3D crystal structure
- Structure => Function?
- Sequence homology not always good indicator of functional similarity
- Study of protein expression



Single Protein and Protein Complexes Profiling

BioTechnologies

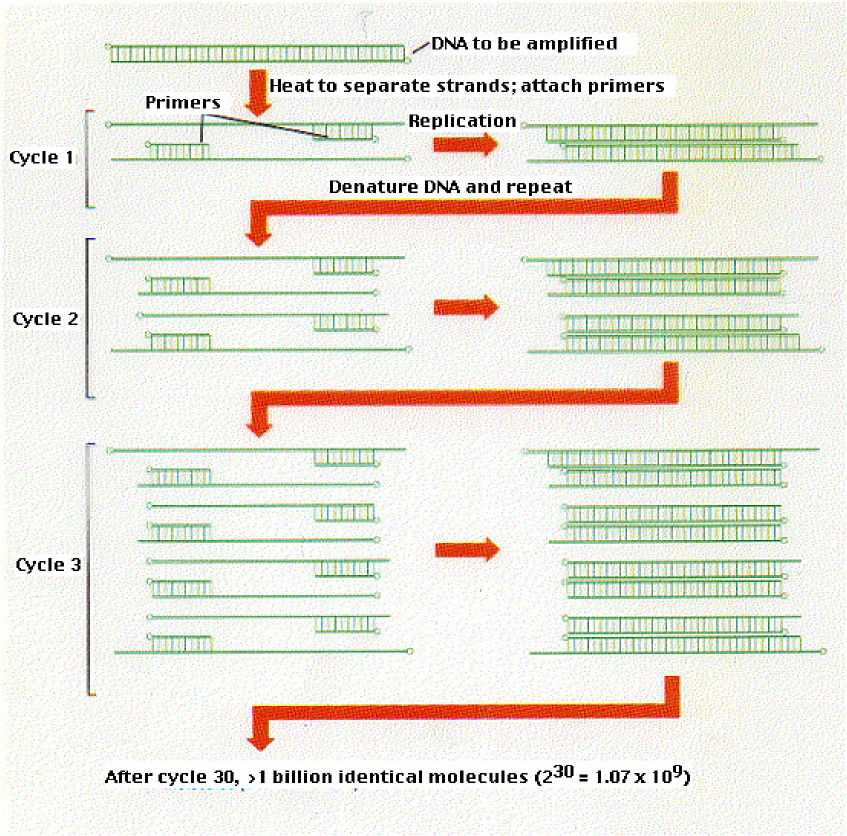
- Observing the Central Dogma: sequence, gene and protein expression, DNA-protein and protein-protein interactions
- PCR, DNA Sequencing, DNA Microarrays, Chromatin ImmunoPrecipitation
- Large-Scale Technologies:
  - Thousands of measured variables
  - Require computational processing

Seeing the minute: PCR

Producing multiple copies of given DNA fragment (amplification)

Start: double stranded DNA molecule

1. Separate strands into templates by heating the mixture
2. Cool to allow “primers” to attach to single strands
3. The primers identify the starting points for DNA synthesis
4. DNA synthesis of strands complementary to the templates
5. Repeat 1.


DNA to be amplified
Heat to separate strands; attach primers
Primers
Replication
Cycle 1
Denature DNA and repeat
Cycle 2
Cycle 3
After cycle 30, >1 billion identical molecules (2^{30} = 1.07 x 10^9)

PCR properties

- The primers can determine the amplified DNA fragment if chosen to flank that region
- n steps of the above produce 2^n copies of the intended DNA fragment

  2^30 ~ 10^9

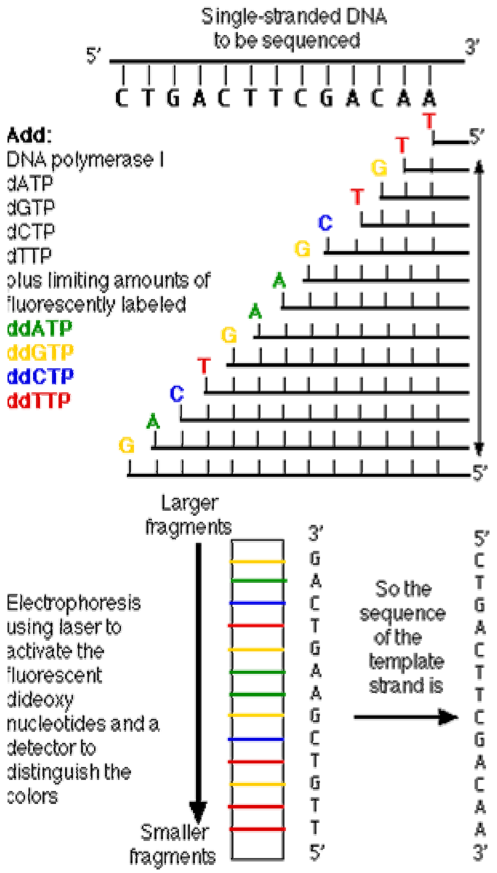
Growing DNA: Synthesizers

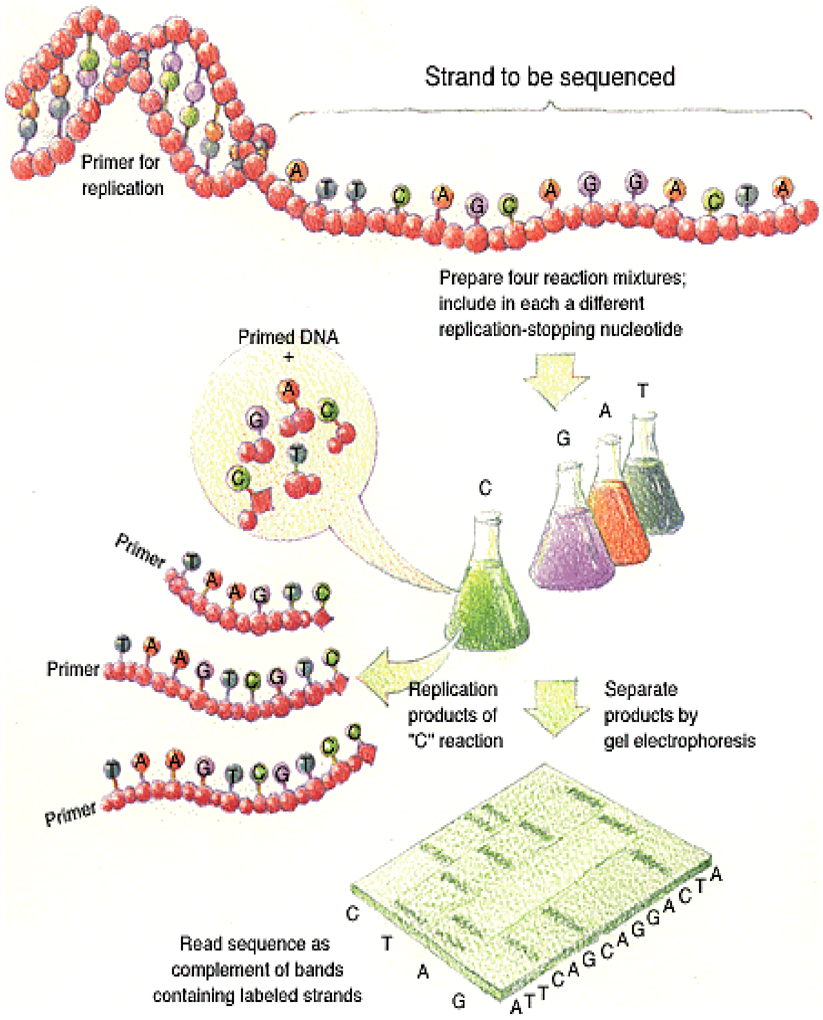
ABI 3900 High-Throughput DNA Synthesizer

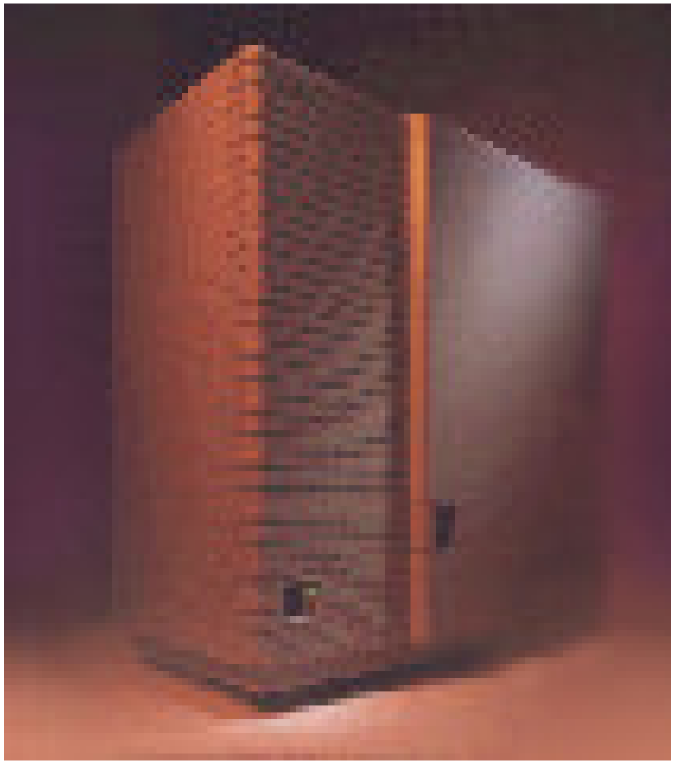
DNA Sequencing

Reading the string: exact positions of the base pairs A, C, G, T

- Digest the DNA to be sequenced into small, 500 - 700 bp fragments
- Replicate sample (fragment) into four bins
- Each bin has a sufficient amount of all four bases and Polymerase
- Bin associated with base x has in addition a special version of x, a stopping version, which stops replication
- The stopping bases are also fluorescently labeled
- DNA replication creates fragments of different lengths in the bins, but all fragments in a bin end in the same labeled base
- Using Gel electrophoresis the fragments are separated by length, thus identifying the base at any given length.




Strand to be sequenced
Primer for replication
A T T C A G C A G G A C T A
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA +
A G C T C
G A T
C
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
Read sequence as complement of bands containing labeled strands
C T A G
ATTCAGCAGGACTA



ABI Prism 377, modern capilary sequencer

Putting the pieces together

- Fragment assembly (figure)
- Gaps and overlaps

  - Lander-Waterman Equation

$$gaps = ne^{-n(l-t)/T}$$

  - Coverage: ratio of sequenced length vs. genome length (figure)

$$\mathrm{cov}\, erage = nl / T$$

Sequencing approaches

- Shotgun sequencing: “random” overlapping fragments (Celera)

- Mapped sequencing: shorter sequences are anchored (Human Genome Consortium)

Microarrays

- Testing for the presence of a sequence fragment

  - De novo sequencing

  - Gene expression

- Hypothesis generation vs. promisse of complete description on a large scale
- Possibility to do 100000 experiments at a time!

What are Microarrays good for?

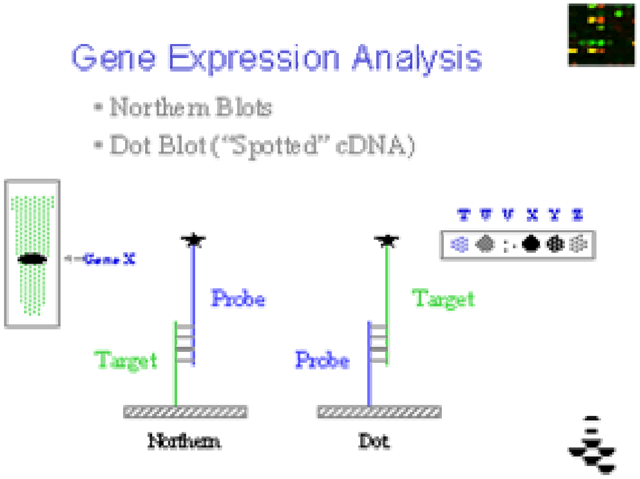
- Identifying differentially expressed genes

  - Genes that behave differently to treatments in same organisms

  - Different organisms
- Identifying naturally oscillating genes in the cell: example cell cycling genes in yeast
- Identifying SNPs
- Tumor vs normal cells

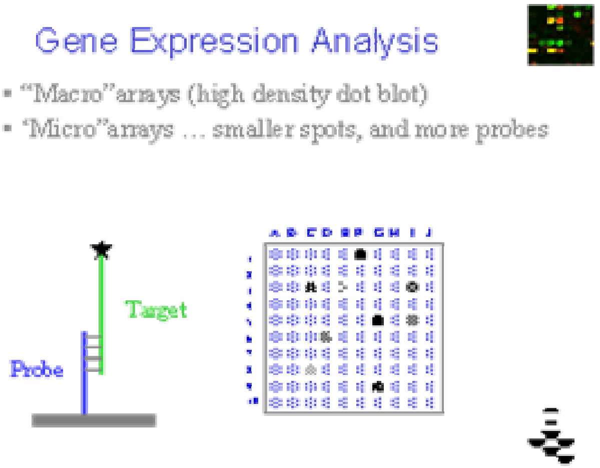
How do they work?

(Source: SUNYSB microarray facility tour)



Gene Expression Analysis

- Northern Blots
- Dot Blot (“Spotted” cDNA)



Gene Expression Analysis

- “Macro”arrays (high density dot blot)
- “Micro”arrays … smaller spots, and more probes

- Single stranded DNA/RNA molecules (probes) attached to a plate hybridize to their complement
- Probes attached in a square matrix typically
- Probes exposed to prepared solution (cellular extract) called a target
- They hybridize with their complements from the target
- Targets are labeled (usually by fluorescence)
- Reading the arrays: observing the color at each probe site
- Color indicates (relative) concentration of probe’s complement

Microarray Formats

Spotted Microarrays

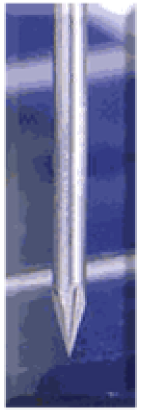
- Ed Southern 25 years ago, Patrick Brown recently

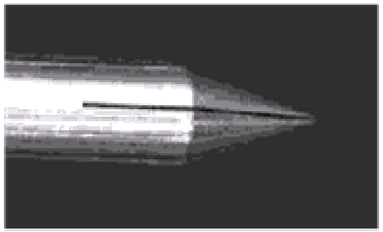
- Glass slide DNA arrays

- 100,000 sites per 1cm^2

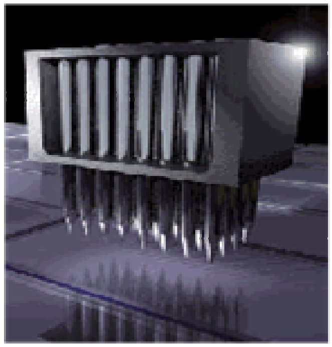
Printing a Microarray

“Spotting” Tips



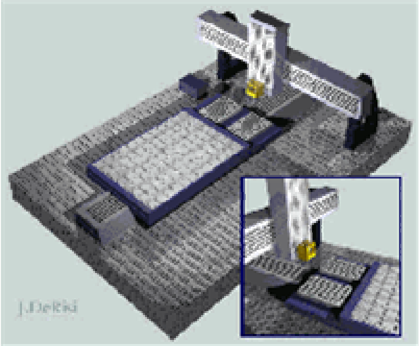


“Spotting” Head



... more arrays per batch

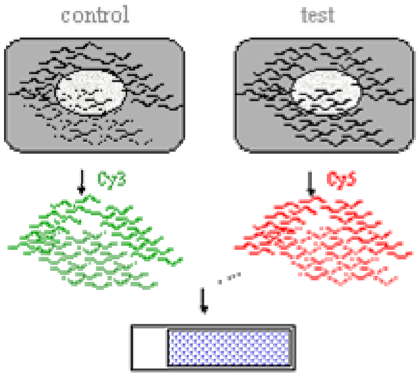
Arrayer ... DIY c/o Pat Brown (Stanford)



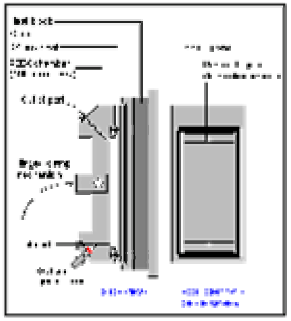
•http://cmgm.stanford.edu/pbrown/

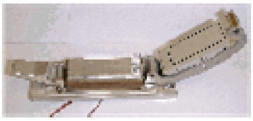
Experiment and Reading

mRNA Target Labeling



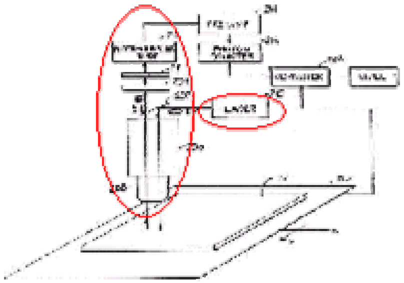
ASP - Automated Slide Processor





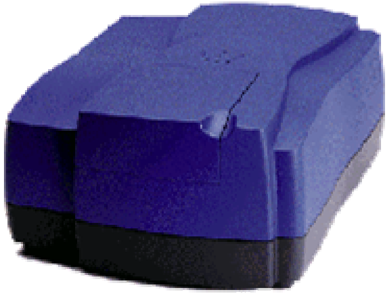
- Automated pre-hyb
- Auto hybridization
- Auto mix & wash

Scanner

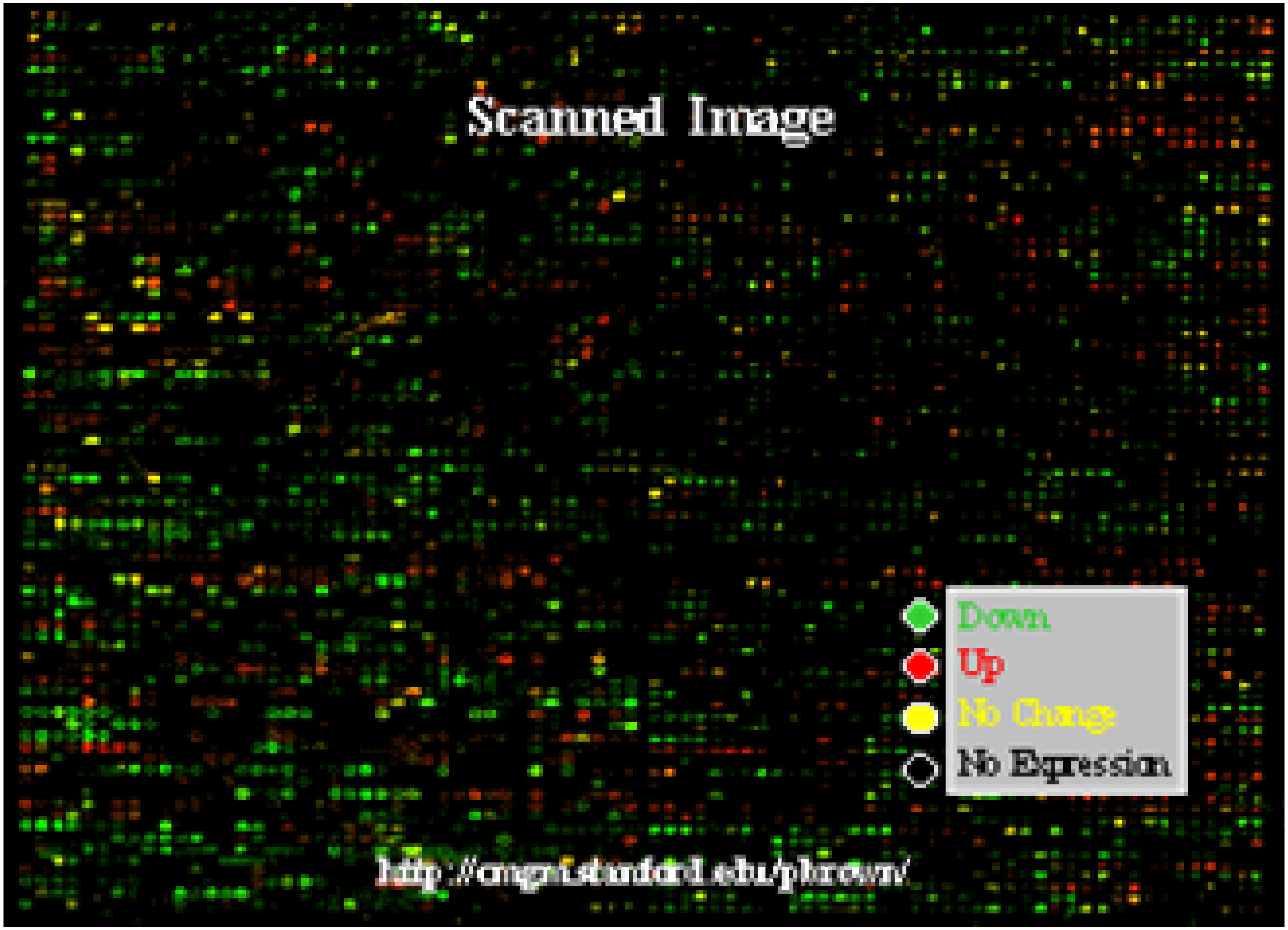


genpakARRAY 21

Scanner



Results

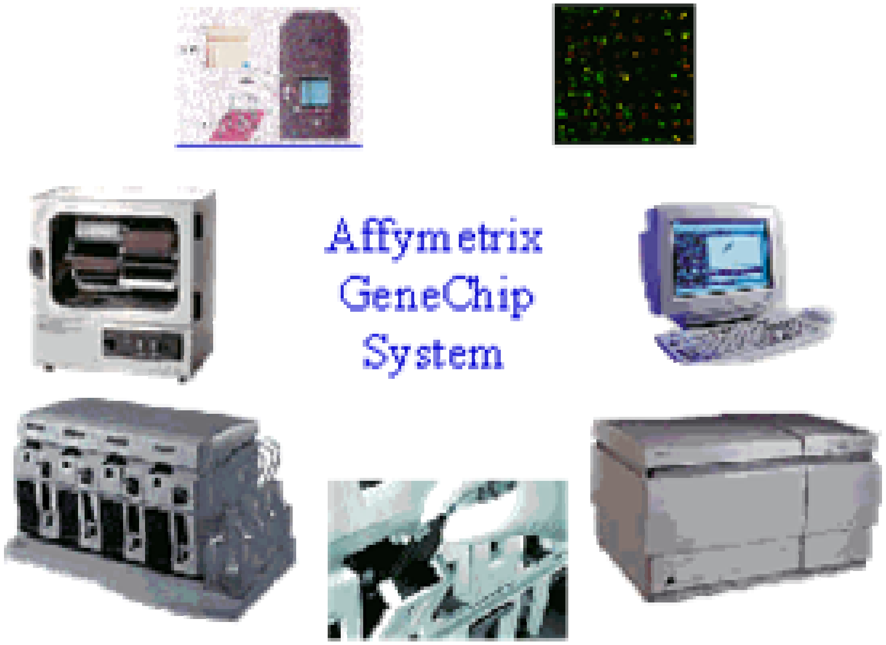

Scanned Image
Down
Up
No Change
No Expression
http://cmgm.stanford.edu/pbrown/

- How do they work?
- - single stranded DNA/RNA molecules (probes) attached to a plate hybridize to their complement
- - probes attached in a square matrix typically
- - Probes exposed to prepared solution (cellular extract) called a target
- - They hybridize with their complements from the target
- - Targets are labeled (usually by fluorescence)
- - Reading the arrays: observing the color at each probe site
- - Color indicates concentration of probe’s complement
- - Some techniques yield absolute concentrations others relative
- - Relative concentrations: two dye mixtures

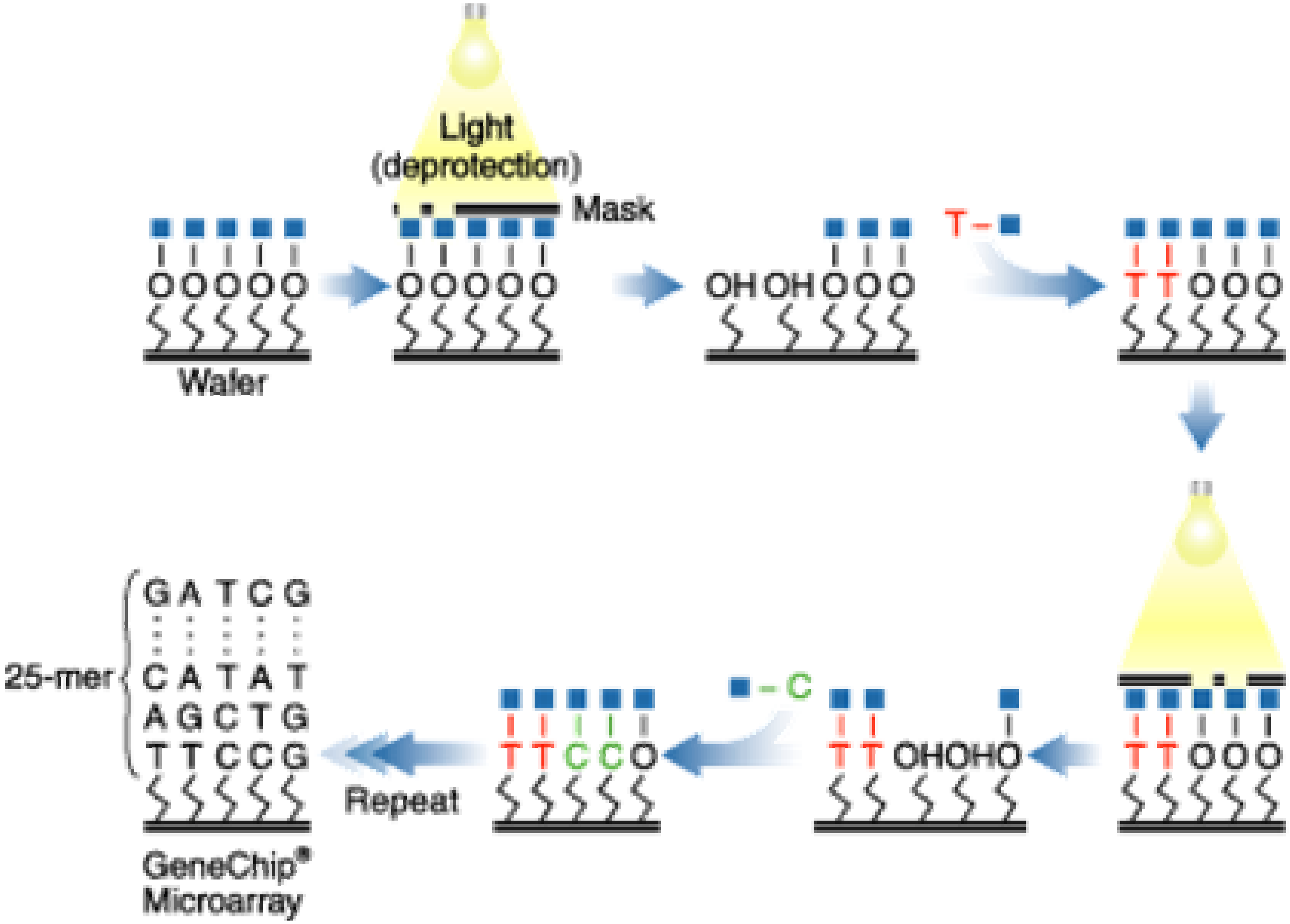
Gene Chips

Oligonucleotide arrays

- Photolithographic method (Affymetrix Inc.) just like computer chips
- Masks used to synthesize oligonucleotides to a chip
- 1,000,000 sites per 1cm^2.



Photolithography



Other Microarray Technologies

- Ink jet (Agilent),
- Addressable beads (Lynx),
- etc.

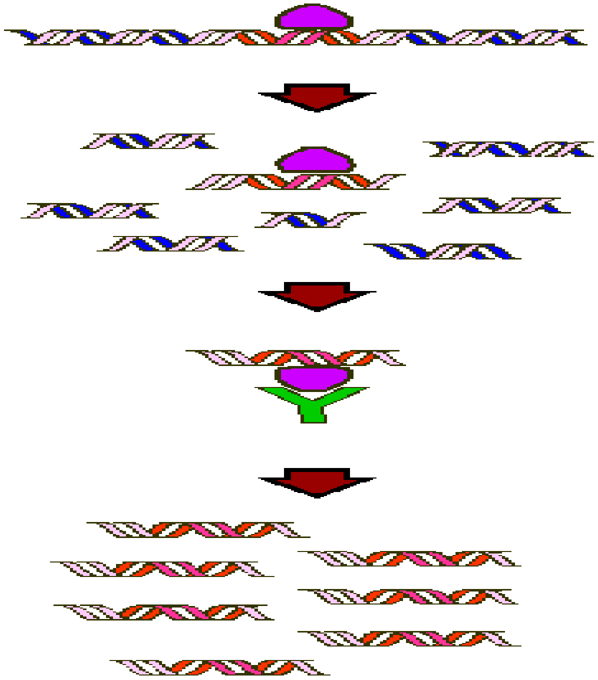
Sources of Error in Microarrays

- Length of probes
- Cross and self hybridization
- Environmental conditions

Algorithmic Problems

- Probe design
- Plate design
- Data Analysis:
  - classification,
  - clustering,
  - regulation inference,
  - gene networks

ChIP



Protein bound to DNA

Sheared DNA

Antibody binding and protein release

PCR the isolated DNA

ChIP: Chromatin Immuno-Precipitation
DNA-Protein Interactions

Protein Expression Arrays

- Abundance of peptides and polypeptides
- Much more difficult to work with, especially analyze