BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan

ECS 254A; Phone: x3748

giri@cs.fiu.edu

http://www.cs.fiu.edu/~giri/teach/BSC4934_Su10.html

July 2010

HMM for Sequence Alignment

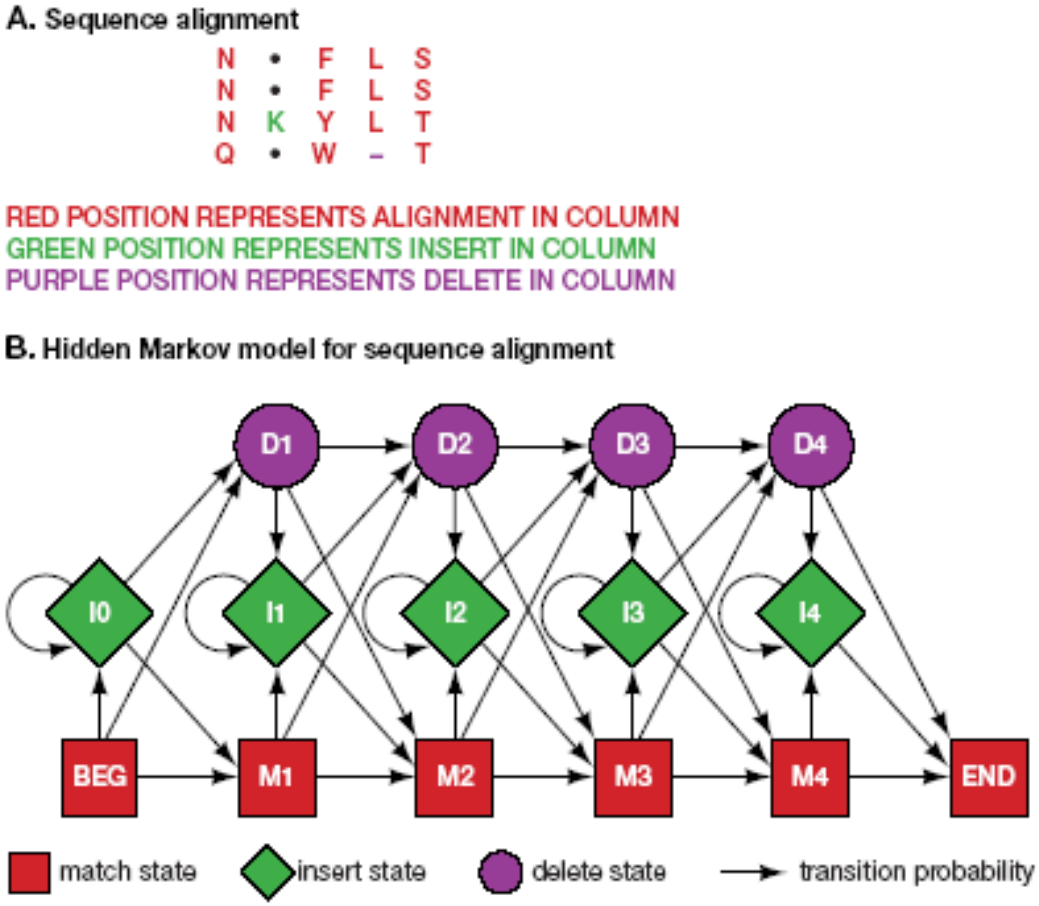


FIGURE 5.16. Relationship between the sequence alignment and the hidden Markov model of the alignment (Krogh et al. 1994). This particular form for the HMM was chosen to represent the sequence, structural, and functional variation expected in proteins. The model accommodates the identities, mismatches, insertions, and deletions expected in a group of related proteins. (*A*) A section of an msa. The illustration shows the columns generated in an msa. Each column may include matches and mismatches (*red* positions), insertions (*green* positions), and deletions (*purple* positions). (*B*) The HMM. Each column in the model represents the possibility of a match, insert, or delete in each column of the alignment in *A*. The HMM is a probabilistic representation of a section of the msa. Sequences can be generated from the HMM by starting at the beginning state labeled BEG and then by following any one of many pathways from one type of sequence variation to another (states) along the state transition arrows and terminating in the ending state labeled END. Any sequence can be generated by the model and each pathway has a probability associated with it. Each square match state stores an amino acid distribution such that the probability of finding an amino acid depends on

G-Protein Couple Receptors

- Transmembrane proteins with 7 α-helices and 6 loops; many subfamilies
- Highly variable: 200-1200 aa in length, some have only 20% identity.
- [Baldi & Chauvin, '94] HMM for GPCRs
- HMM constructed with 430 match states (avg length of sequences) ; Training: with 142 sequences, 12 iterations

GPCR - Analysis

- Compute main state entropy values

$$H_i = -\sum_a e_{ia} \log e_{ia}$$

- For every sequence from test set (142) & random set (1600) & all SWISS-PROT proteins
  - Compute the negative log of probability of the most probable path π

$$Score(S) = -\log\left(P(\pi \mid S, M)\right)$$

Entropy

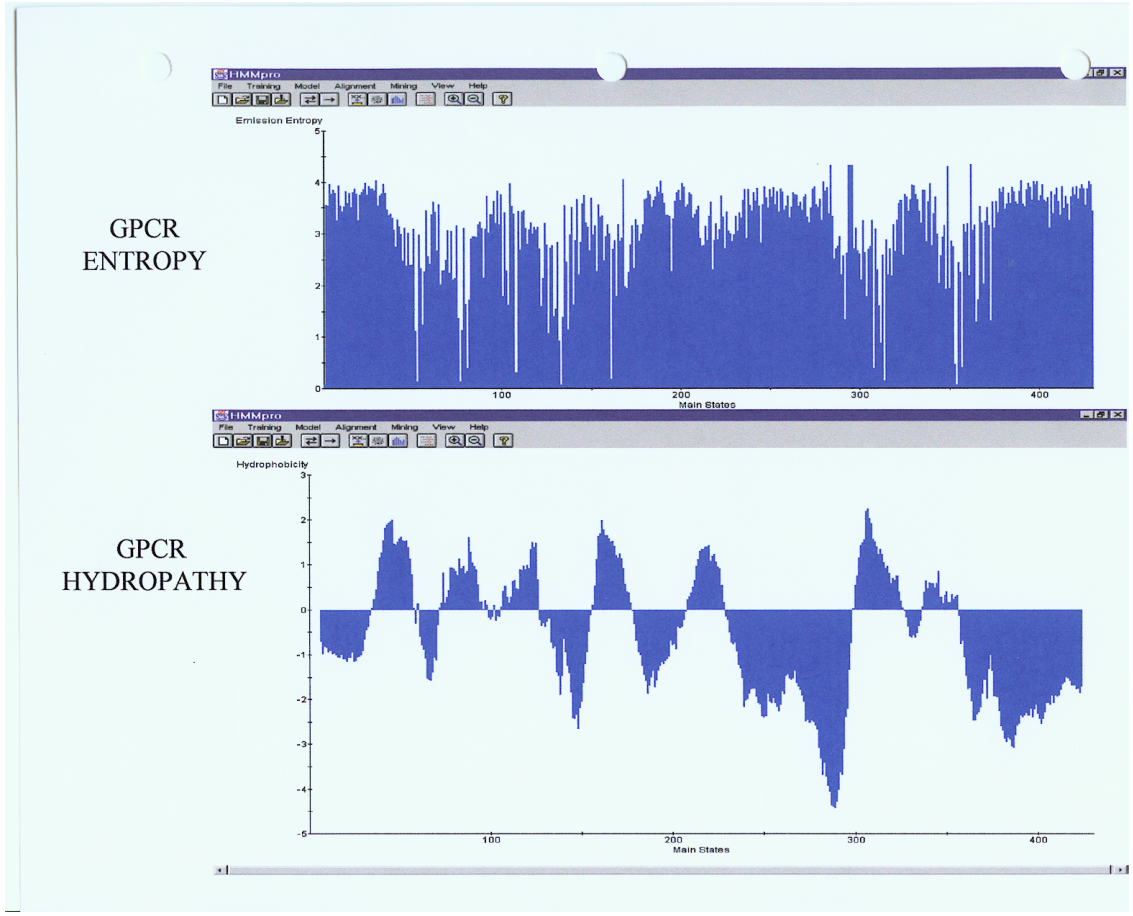
- Entropy measures the variability observed in given data.

$$E = -\sum_{c} p_c \log p_c$$

- Entropy is useful in multiple alignments & profiles.

- Entropy is max when uncertainty is max.

GPCR Analysis



Entropy

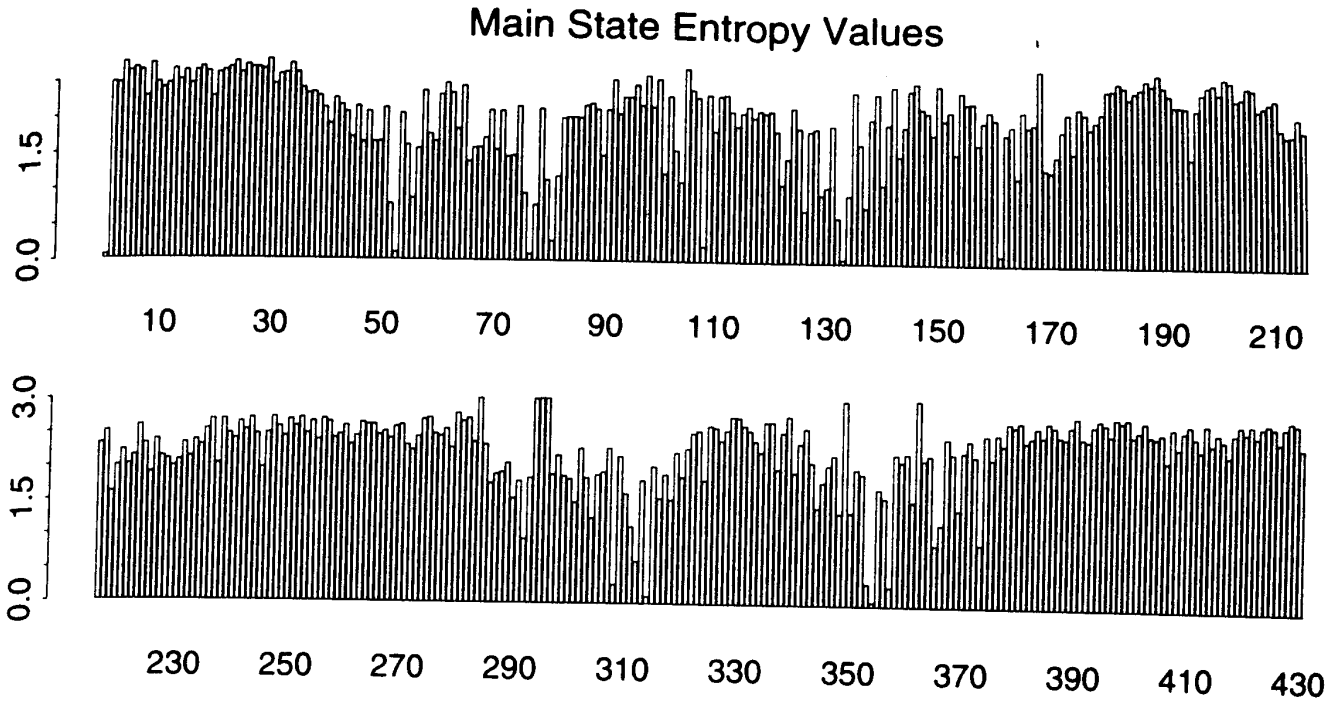


Figure 8.1: Entropy Profile of the Emission Probability Distributions Associated with the Main States of the HMM After 12 Cycles of Training.

GPCR Analysis (Cont'd)

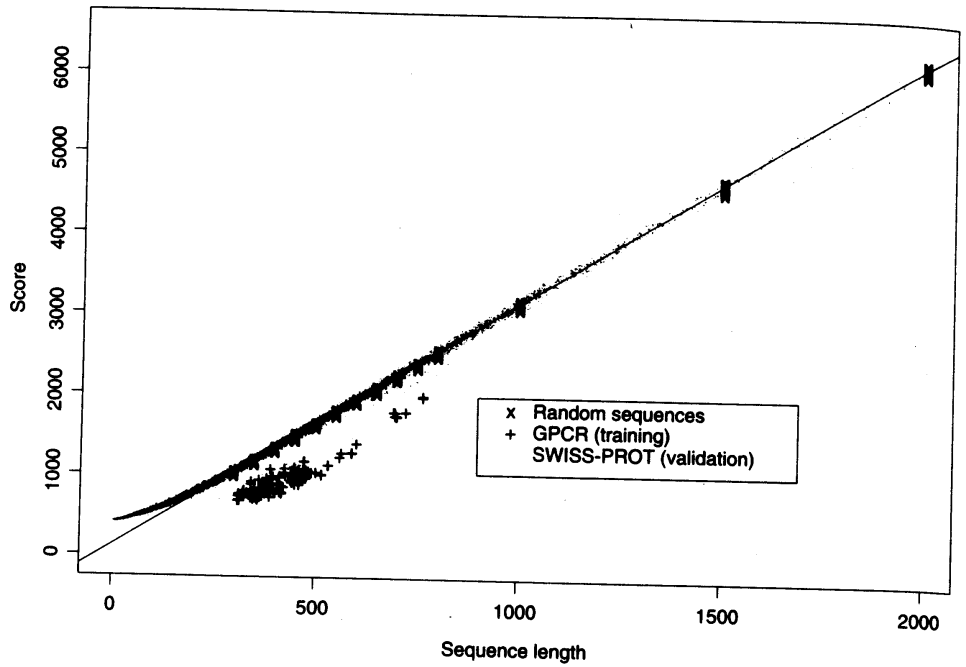


Figure 8.2: Scores (Negative Log-likelihoods of Optimal Viterbi Paths). Represented sequences consist of 142 GPCR training sequences, all sequences from the SWISS-PROT database of length less than or equal to 2000, and 220 randomly generated sequences with same average composition as the GPCRs of length 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800 (20 at each length). The regression line was obtained from the 220 random sequences. The horizontal distances in the histogram correspond to normalized scores (6).

Applications of HMM for GPCR

- Bacteriorhodopsin
  - Transmembrane protein with 7 domains
  - But it is not a GPCR
  - Compute score and discover that it is close to the regression line. Hence not a GPCR.
- Thyrotropin receptor precursors
  - All have long initial loop on INSERT STATE 20.
  - Also clustering possible based on distance to regression line.

HMMs – Advantages

- Sound statistical foundations
- Efficient learning algorithms
- Consistent treatment for insert/delete penalties for alignments in the form of locally learnable probabilities
- Capable of handling inputs of variable length
- Can be built in a modular & hierarchical fashion; can be combined into libraries.
- Wide variety of applications: Multiple Alignment, Data mining & classification, Structural Analysis, Pattern discovery, Gene prediction.

HMMs – Disadvantages

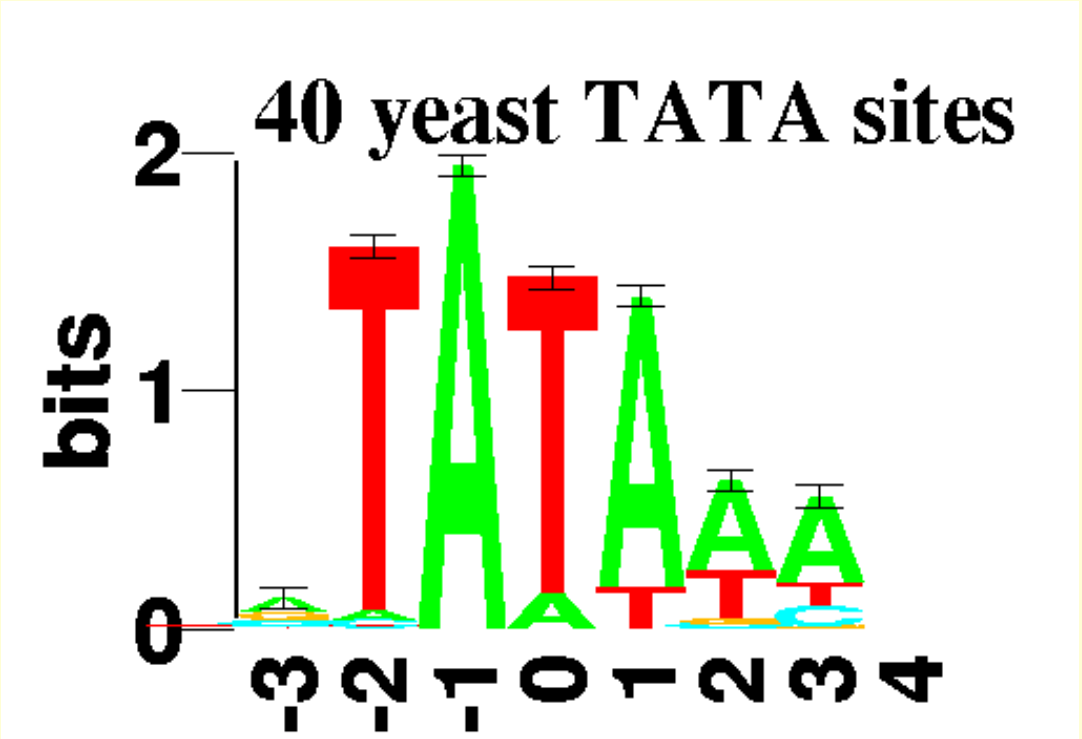
- Large # of parameters.
- Cannot express dependencies & correlations between hidden states.

Patterns in DNA Sequences

- Signals in DNA sequence control events
  - Start and end of genes
  - Start and end of introns
  - Transcription factor binding sites (regulatory elements)
  - Ribosome binding sites
- Detection of these patterns are useful for
  - Understanding gene structure
  - Understanding gene regulation

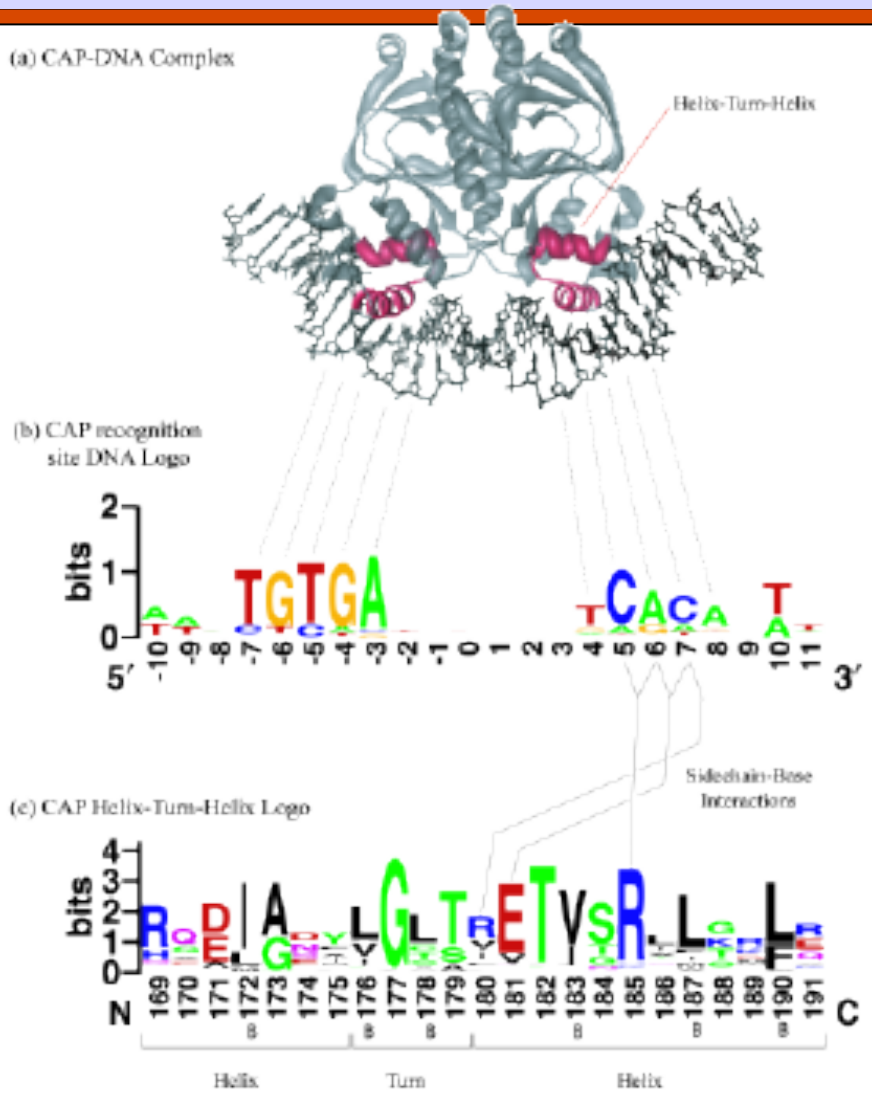
Motifs in DNA Sequences

- Given a collection of DNA sequences of promoter regions, locate the transcription factor binding sites (also called regulatory elements)
  - Example:



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

Motifs



http://weblogo.berkeley.edu/examples.html

Motifs in DNA Sequences

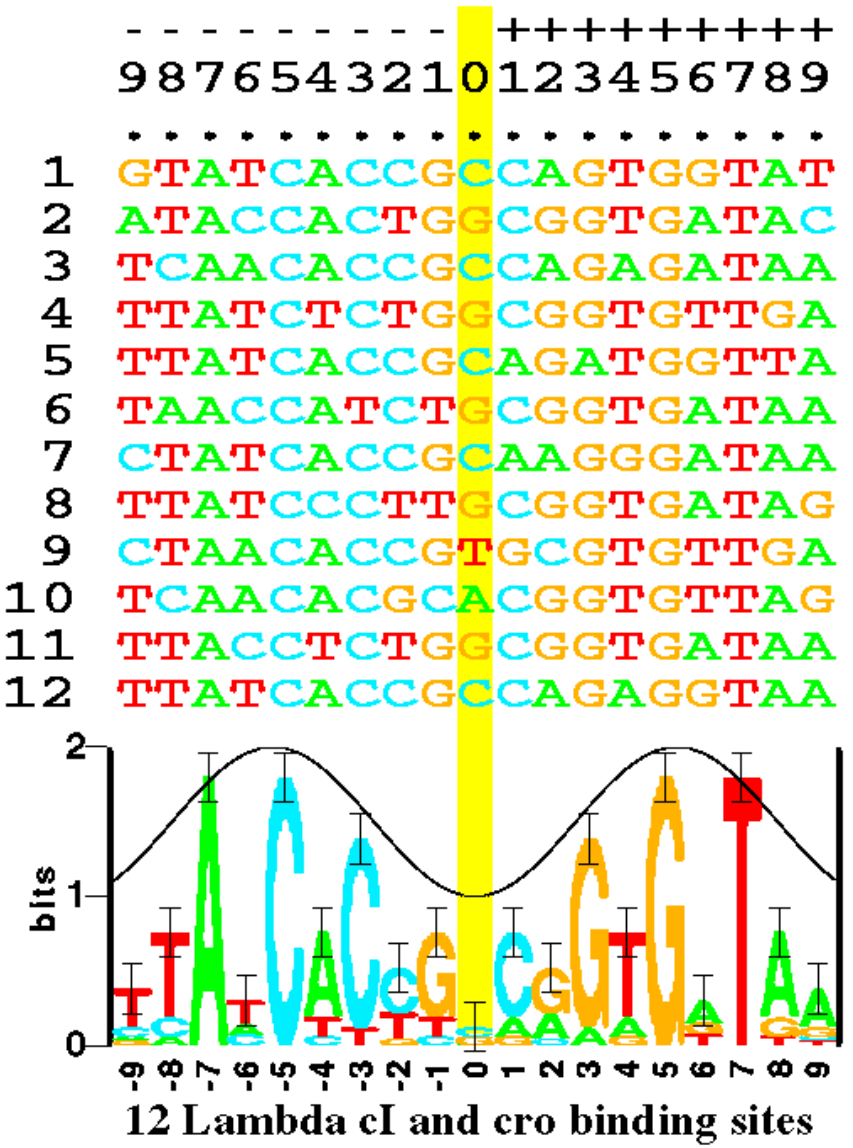
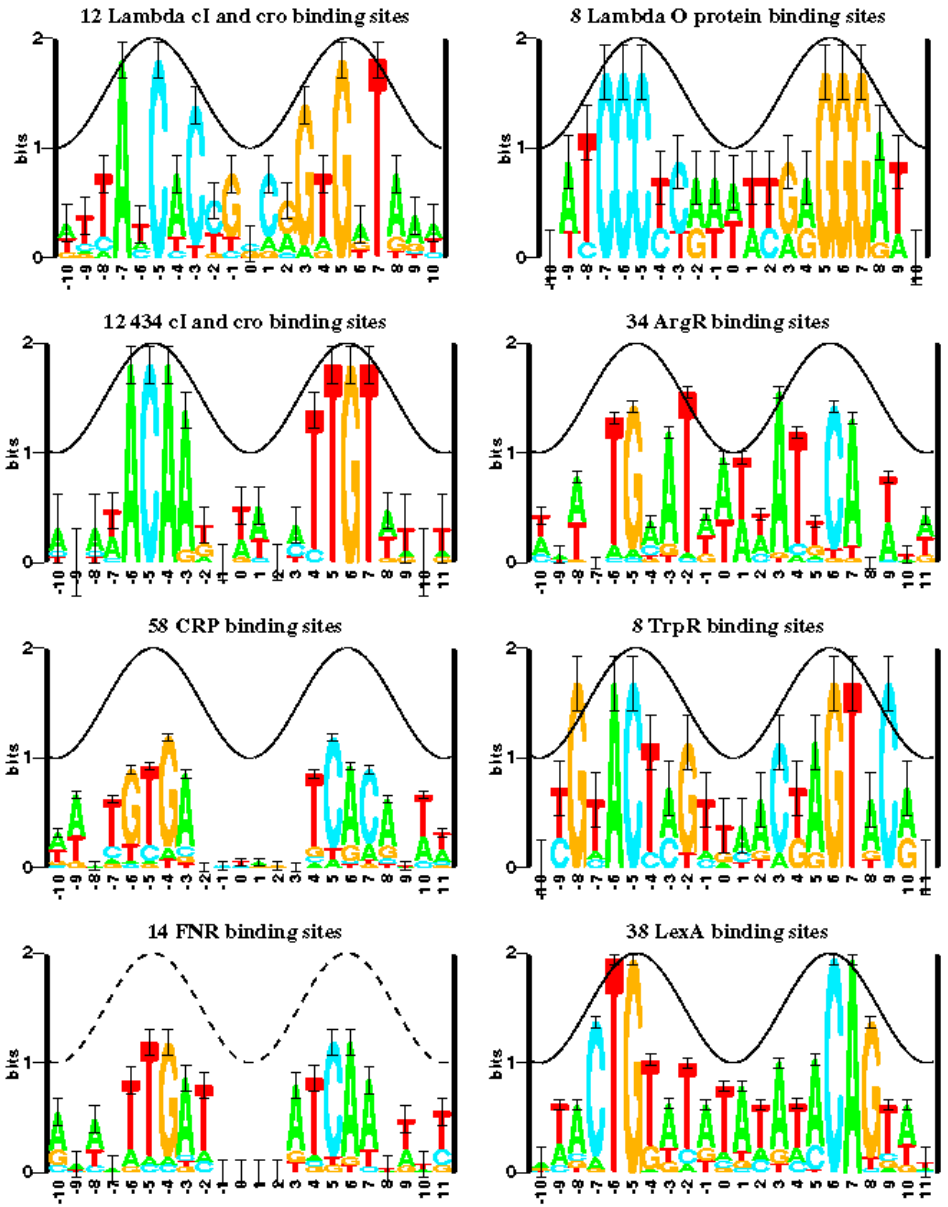


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [17].

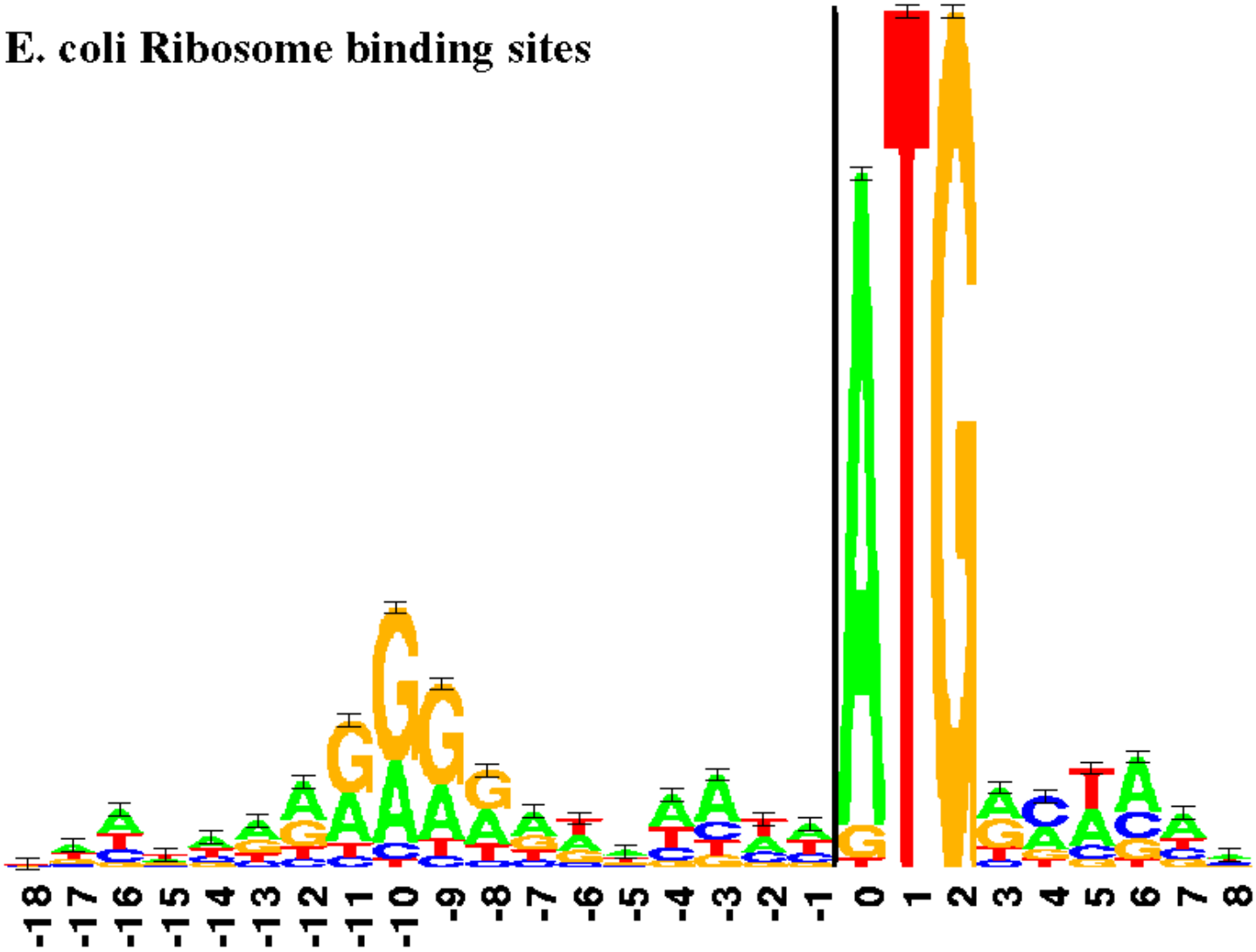
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

More Motifs in *E. Coli* DNA Sequences



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

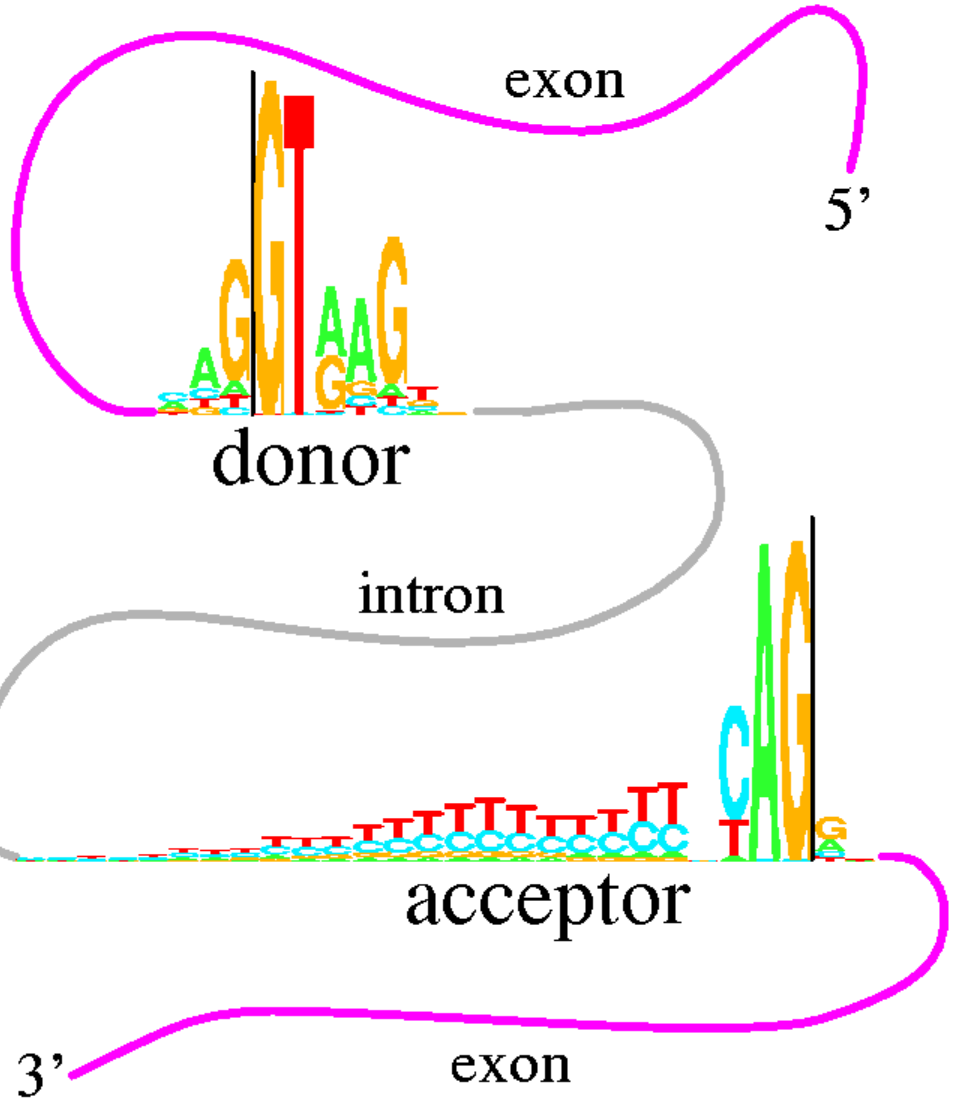
E. coli Ribosome binding sites



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

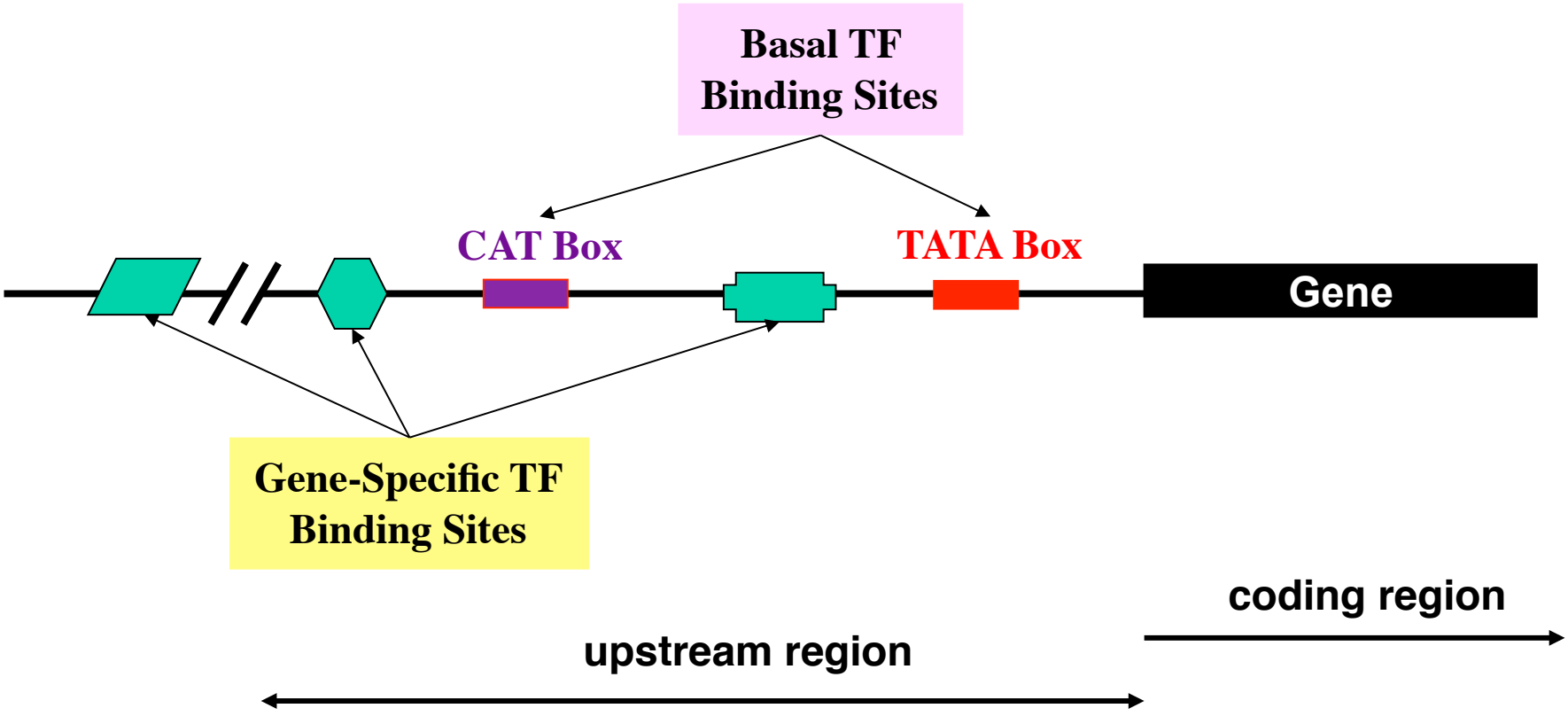
Other Motifs in DNA Sequences: Human Splice Junctions

This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

Transcription Regulation



Prokaryotic Gene Characteristics

DNA PATTERNS IN THE *E. coli lexA* GENE

	GENE SEQUENCE	PATTERN
1	GAATTCGATAAATCTCTGGTTTATTGTGCAGTTTATGGTT	CTGNNNNNNNNNNCAG
	TT	TTGACA
41	CCAAAATCGCCTTTTGCTGTATATACTCACAGCATAACTG	CTGNNNNNNNNNNCAG
	CCAA -35 -10 TATACT >	TATAAT, > mRNA start
81	TATATACACCCAGGGGGCGGAATGAAAGCGTTAACGGCCA	CTGNNNNNNNNNNCAG
	+10 GGGGG Ribosomal binding site	GGAGG
121	GGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATCAG	
161	CCAGACAGGTATGCCGCCGACGCGTGCGGAAATCGCGCAG	ATG
201	CGTTTGGGGTTCCGTTCCCCAAACGCGGCTGAAGAACATC	
241	TGAAGGCGCTGGCACGCAAAGGCGTTATTGAAATTGTTTC	
281	CGGCGCATCACGCGGGATTCGTCTGTTGCAGGAAGAGGAA	
321	GAAGGGTTGCCGCTGGTAGGTCGTGTGGCTGCCGGTGAAC	
361	CACTTCTGGCGCAACAGCATATTGAAGGTCATTATCAGGT	OPEN READING FRAME
401	CGATCCTTCCTTATTCAAGCCGAATGCTGATTTCCTGCTG	
441	CGCGTCAGCGGGATGTCGATGAAAGATATCGGCATTATGG	
481	ATGGTGACTTGCTGGCAGTGCATAAAACTCAGGATGTACG	
521	TAACGGTCAGGTCGTTGTCGCACGTATTGATGACGAAGTT	
561	ACCGTTTAGCGCCCTGAAAAACAGGGCAATAAAGTCGAAC	
601	TGTTGCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGA	
641	CCTTCGTCAGCAGAGCTTCACCATTGAAGGGCTGGCGGTT	
681	GGGGTTATTCGCAACGGCGACTGGCTGTAACATATCTCTG	TAA
721	AGACCGCGATGCCGCCTGGCGTCGCGGTTTGTTTTTCATC	
761	TCTCTTCATCAGGCTTGTCTGCATGGCATTCCTCACTTCA	
801	TCTGATAAAGCACTCTGGCATCTCGCCTTACCCATGATTT	
841	TCTCCAATATCACCGTTCCGTTGCTGGGACTGGTCGATAC	
881	GGCGGTAATTGGTCATCTTGATAGCCCGGTTTATTTGGGC	
921	GGCGTGGCGGTTGGCGCAACGGCGGACCAGCT	

Shown are matches to approximate consensus binding sites for LexA repressor (CTGNNNNNNNNNNCAG), the -10 amd -35 promoter regions relative to the start of the mRNA (TTGACA and TATAAT), the ribosomal binding site on the mRNA (GGAGG), and the open reading frame (ATG...TAA). Only the second two of the predicted LexA binding sites actually bind the repressor.

FIGURE 9.6. The promoter and open reading frame of the *E. coli lexA* gene.

Motifs in DNA Sequences

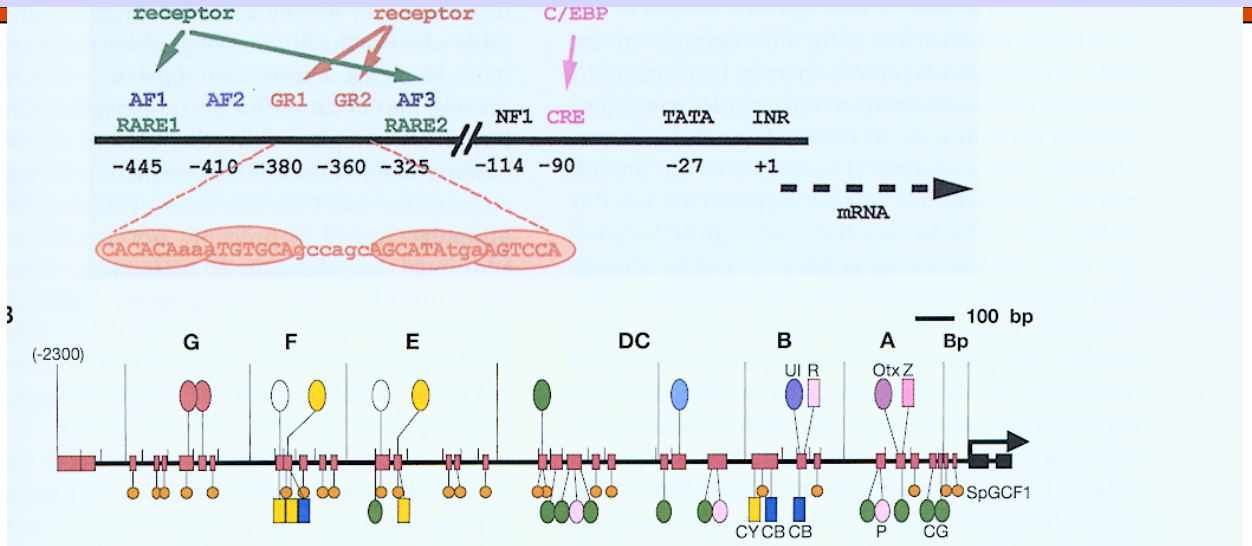
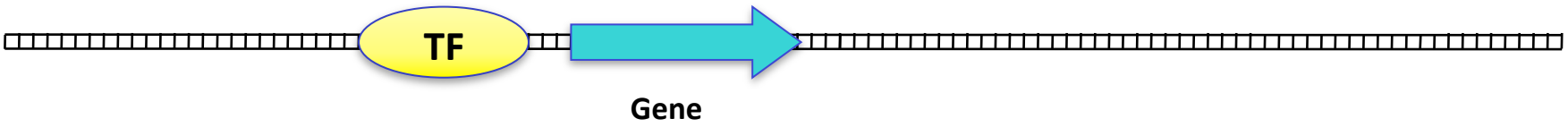


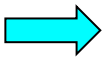
FIGURE 9.13. Regulatory elements of two promoters. (*A*) The rat *pepCK* gene. The relative positions of the TF-binding sites are illustrated (Yamada et al. 1999). The glucocorticoid response unit (GRU) includes three accessory factor–binding sites (AF1, AF2, and AF3), two glucocorticoid response elements (GR1 and GR2), and a cAMP response element (CRE). A dimer of glucocorticoid receptors bound to each GR element is depicted. The retinoic response unit (RAU) includes two retinoic acid response elements (RARE1 and RARE2) that coincide with the AF1 and AF3, respectively (Sugiyama et al. 1998). The sequences of the two GR sites and the binding of the receptor to these sites are shown. These sites deviate from the consensus sites and depend on their activity on accessory proteins bound to other sites in the GRU. This dependence on accessory proteins is reduced if a more consensus-like (canonical) GR element comprising the sequence TGTTCT is present. The CRE that binds factor C/EBP is also shown. (*B*) The 2300-bp promoter of the developmentally regulated gene *endo16* of the sea urchin (Bolouri and Davidson 2002). Different colors indicate different binding sites for distinct proteins and proteins shown above the line bind at unique locations, below the line at several locations. The regions A–G are functional modules that determine the expression of the gene in a particular tissue at a particular time of development and may either serve to induce transcription of the gene as a necessary developmental step (A, B, and G) or repress transcription (C–F) in tissues when it is not appropriate. (Reprinted, with permission, from Bolouri and Davidson 2002 [©2002 Elsevier].)

Single Gene Activation



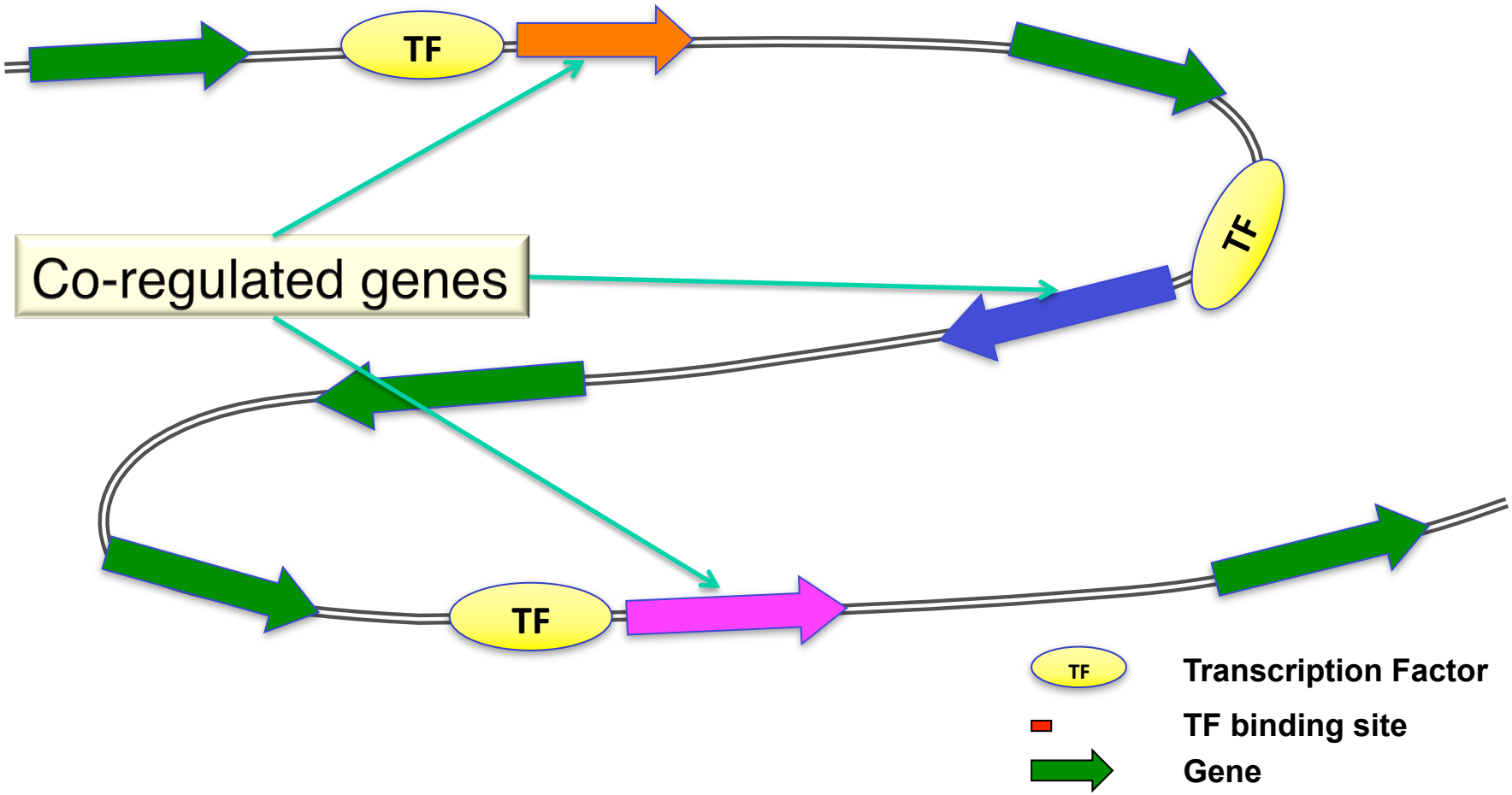
Transcription Factor

TF binding site

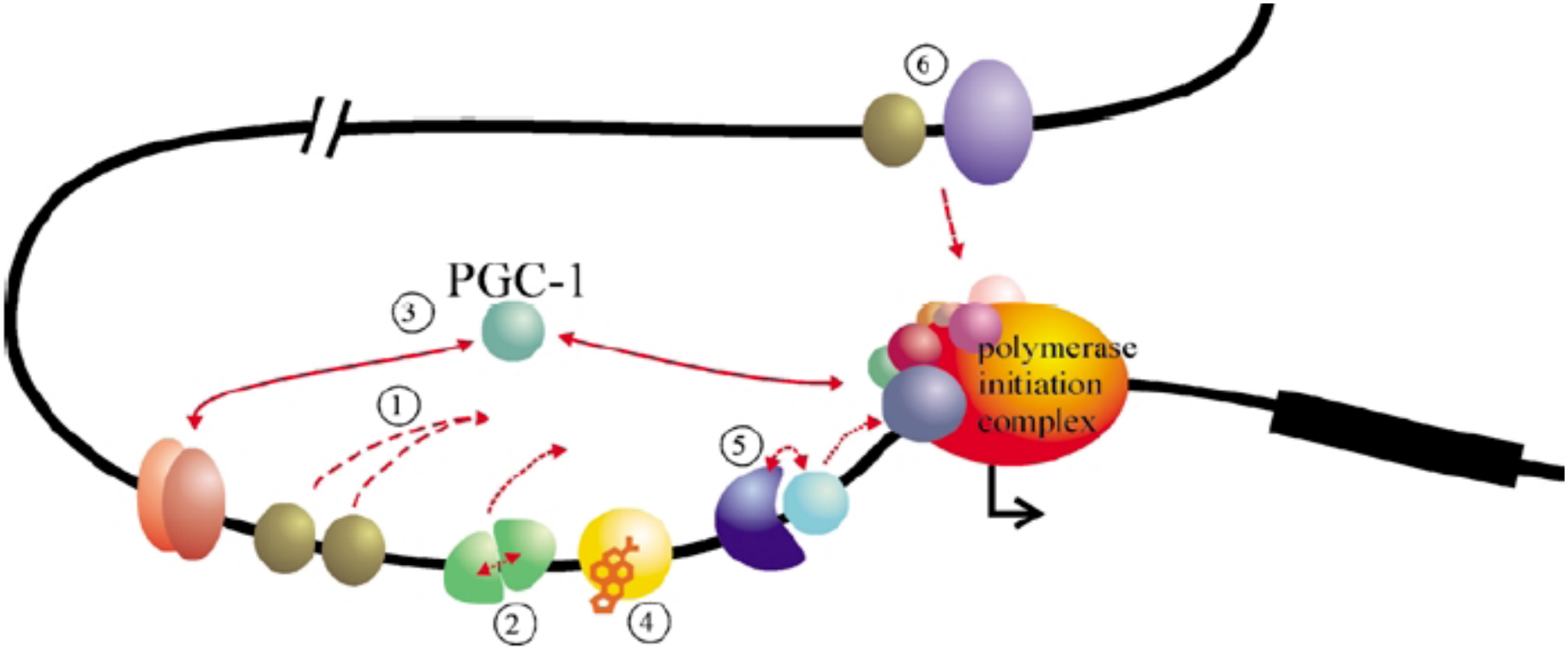


Gene

Multiple Gene Activation



Transcription Regulation

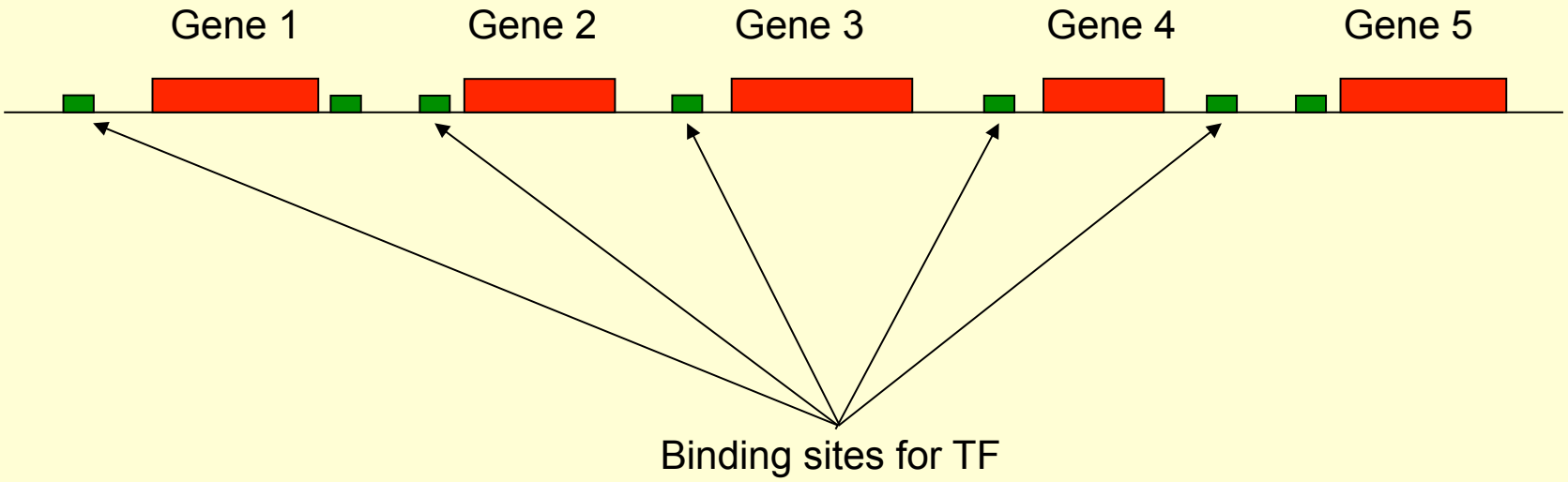


[Goffart *et al. Exp. Physiology* (2003)]

Motif-prediction: Whole genome

Problem: Given the upstream regions of all genes in the genome, find all over-represented sequence signatures.

Basic Principle: If a TF co-regulates many genes, then all these genes should have at least 1 binding site for it in their upstream region.



Motif Detection (TFBMs)

- See evaluation by Tompa et al.
  - [bio.cs.washington.edu/assessment]
- Gibbs Sampling Methods: AlignACE, GLAM, SeSiMCMC, MotifSampler
- Weight Matrix Methods: ANN-Spec, Consensus,
- EM: Improbizer, MEME
- Combinatorial & Misc.: MITRA, oligo/dyad, QuickScore, Weeder, YMF

EM Algorithm

Goal: Find θ, Z that maximize Pr (X, Z | θ)

Initialize: random profile

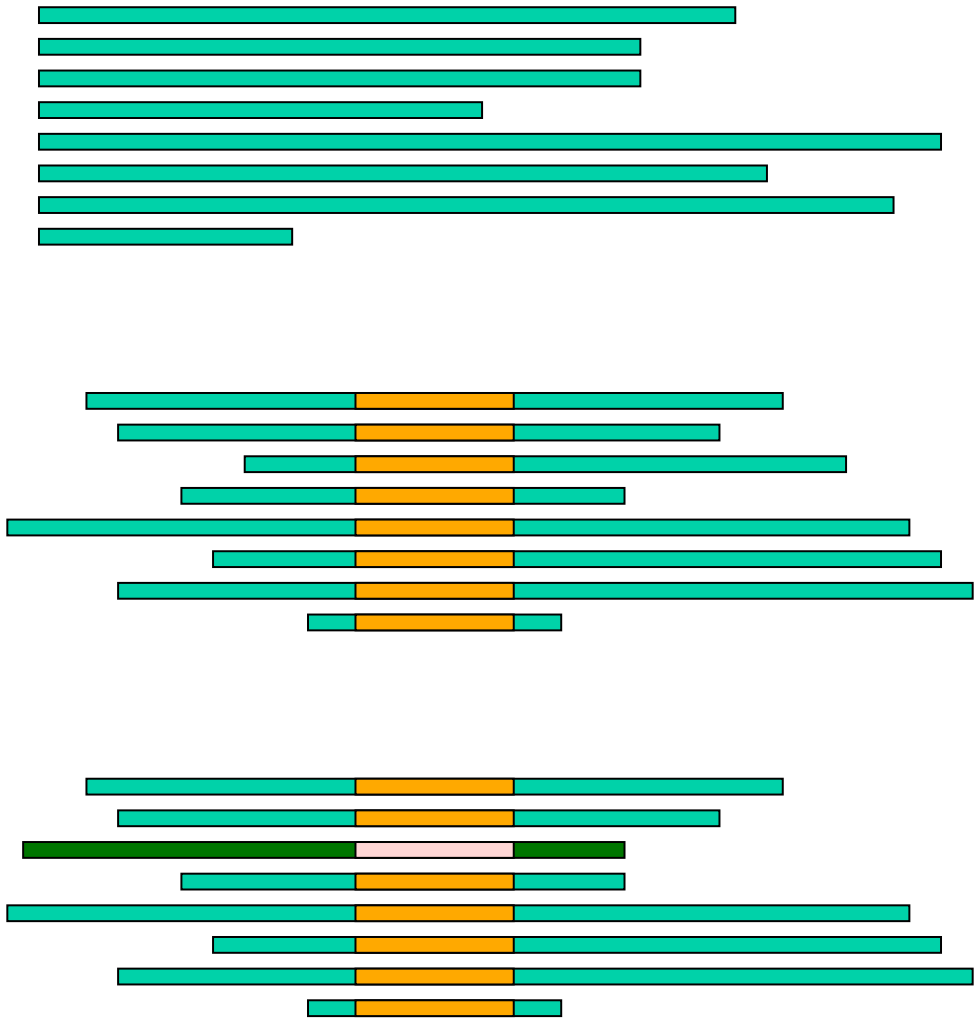
E-step: Using profile, compute a likelihood value z_{ij} for each *m*-window at position *i* in input sequence *j*.

M-step: Build a new profile by using every *m*-window, but weighting each one with value z_{ij}.

Stop if converged

MEME [Bailey, Elkan 1994]

Gibbs Sampling for Motif Detection



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* $\Rightarrow$
  - *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.

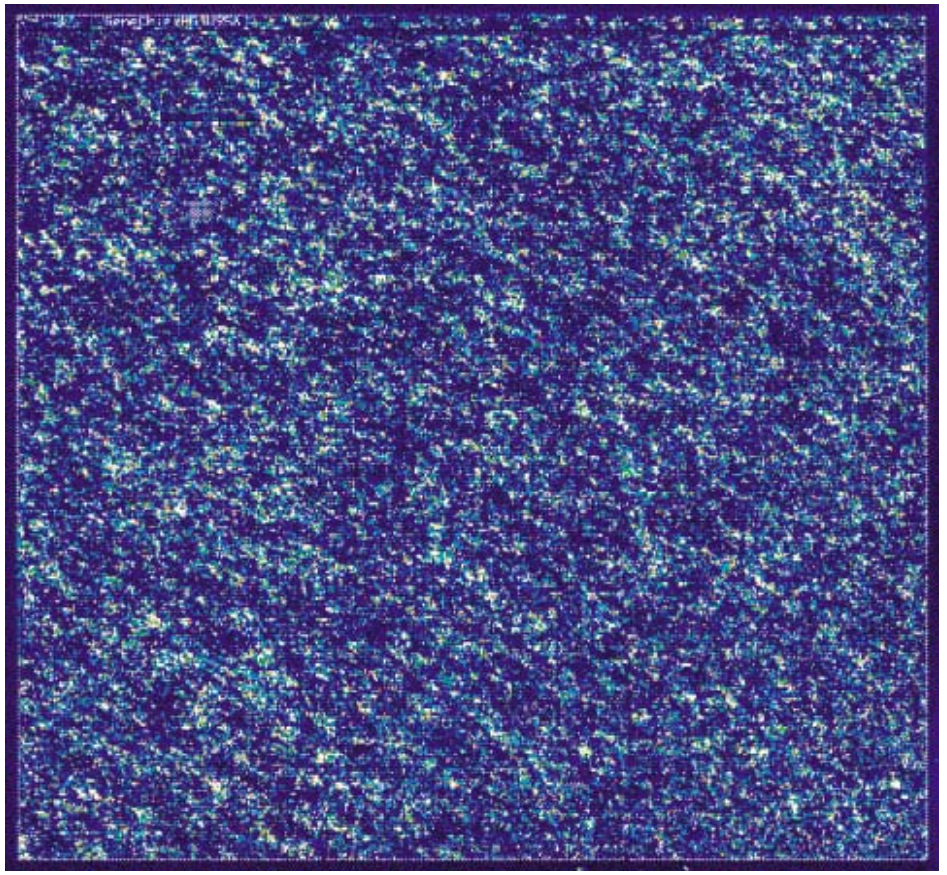
Microarray/DNA chip technology

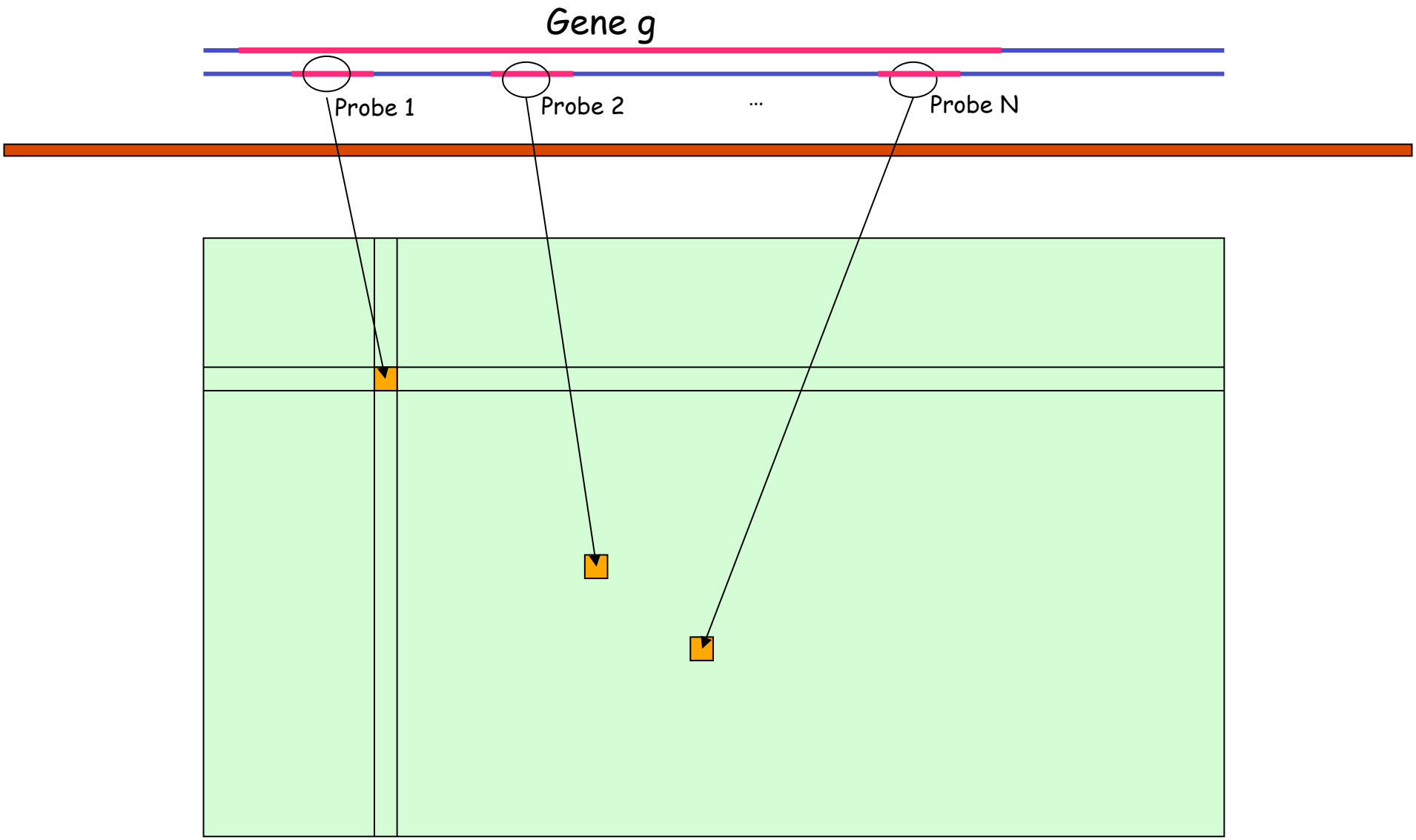
- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between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

Microarray Data

Gene	Expression Level
Gene1	
Gene2	
Gene3	
…	

Gene Chips

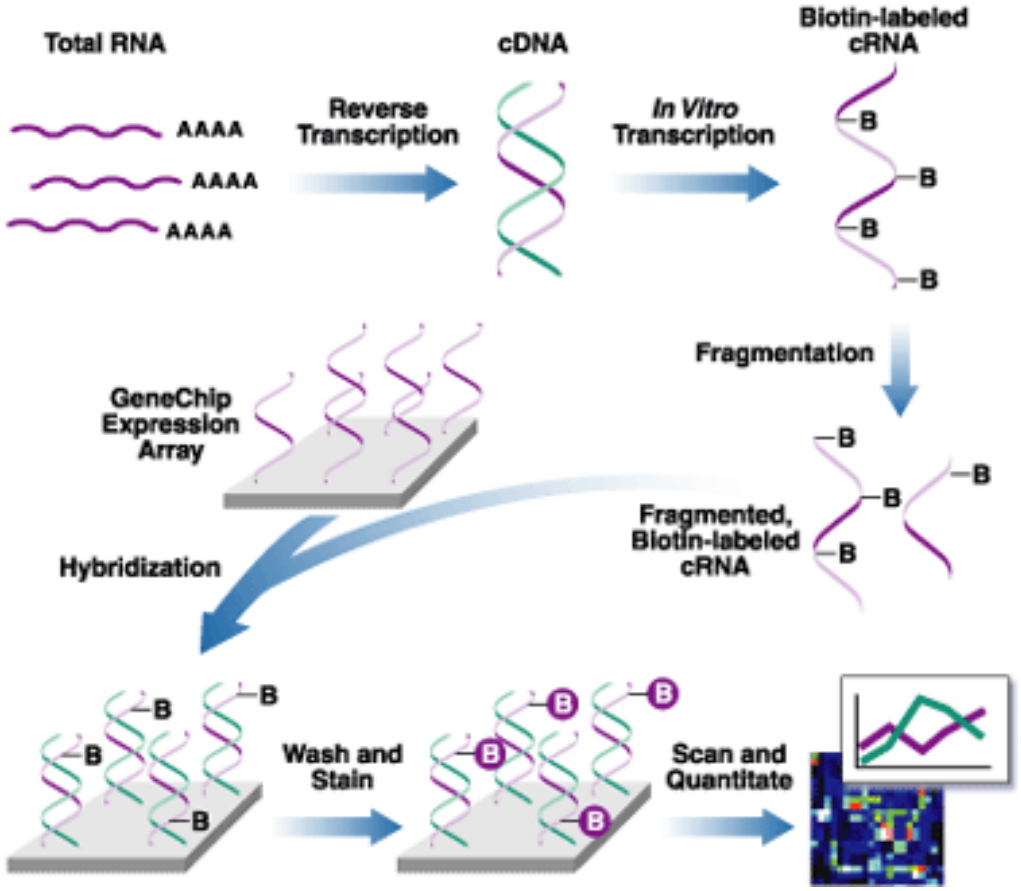



Gene g
Probe 1
Probe 2
...
Probe N

Microarray/DNA chips (Simplified)

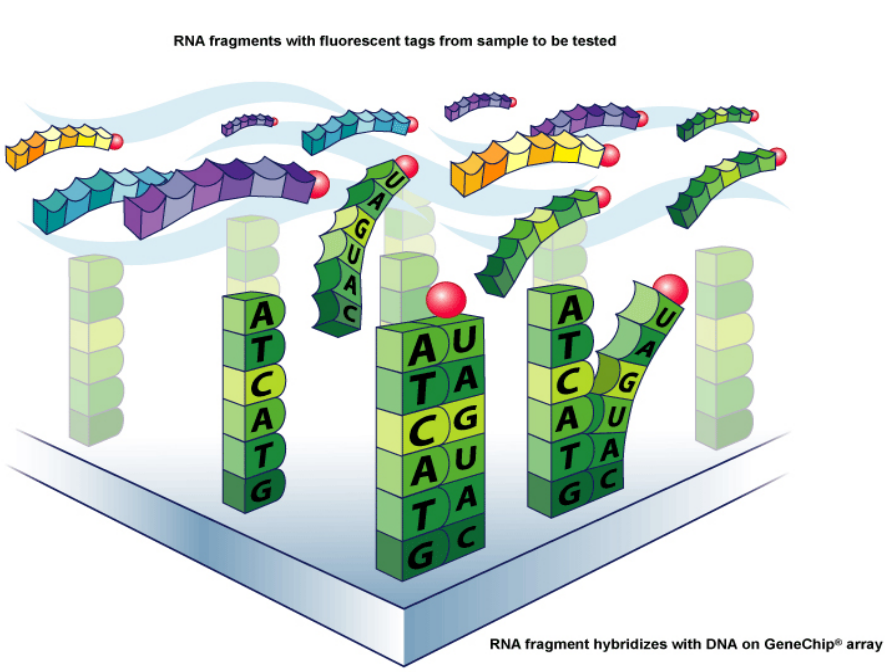
- Construct probes corresponding to reverse complements of genes of interest.
- Microscopic quantities of probes placed on solid surfaces at defined spots on the chip.
- Extract mRNA from sample cells and label them.
- Apply labeled sample (mRNA extracted from cells) to every spot, and allow hybridization.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

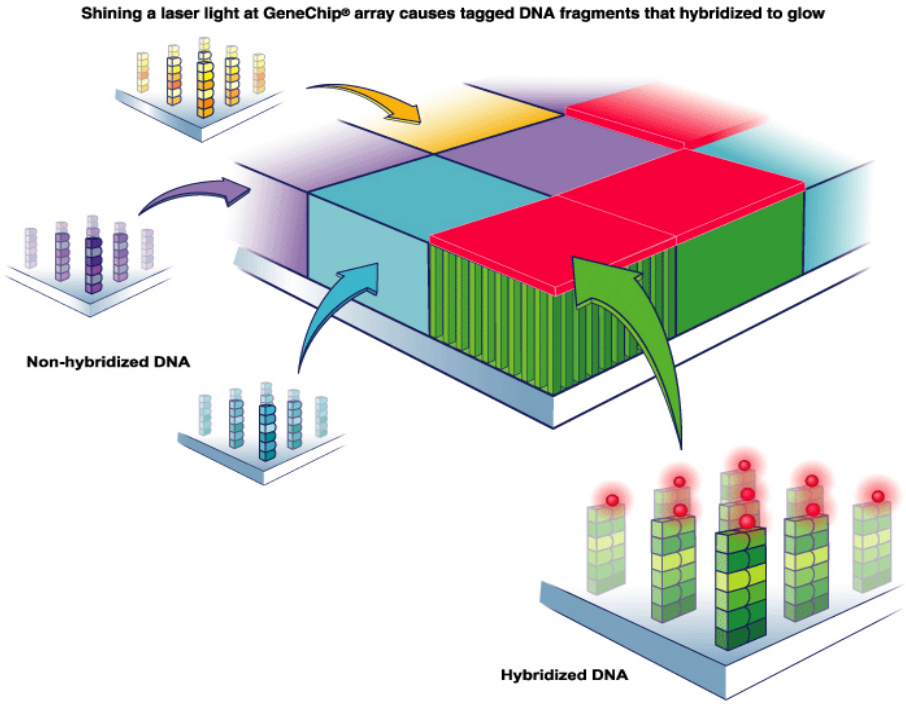
Affymetrix DNA chip schematic



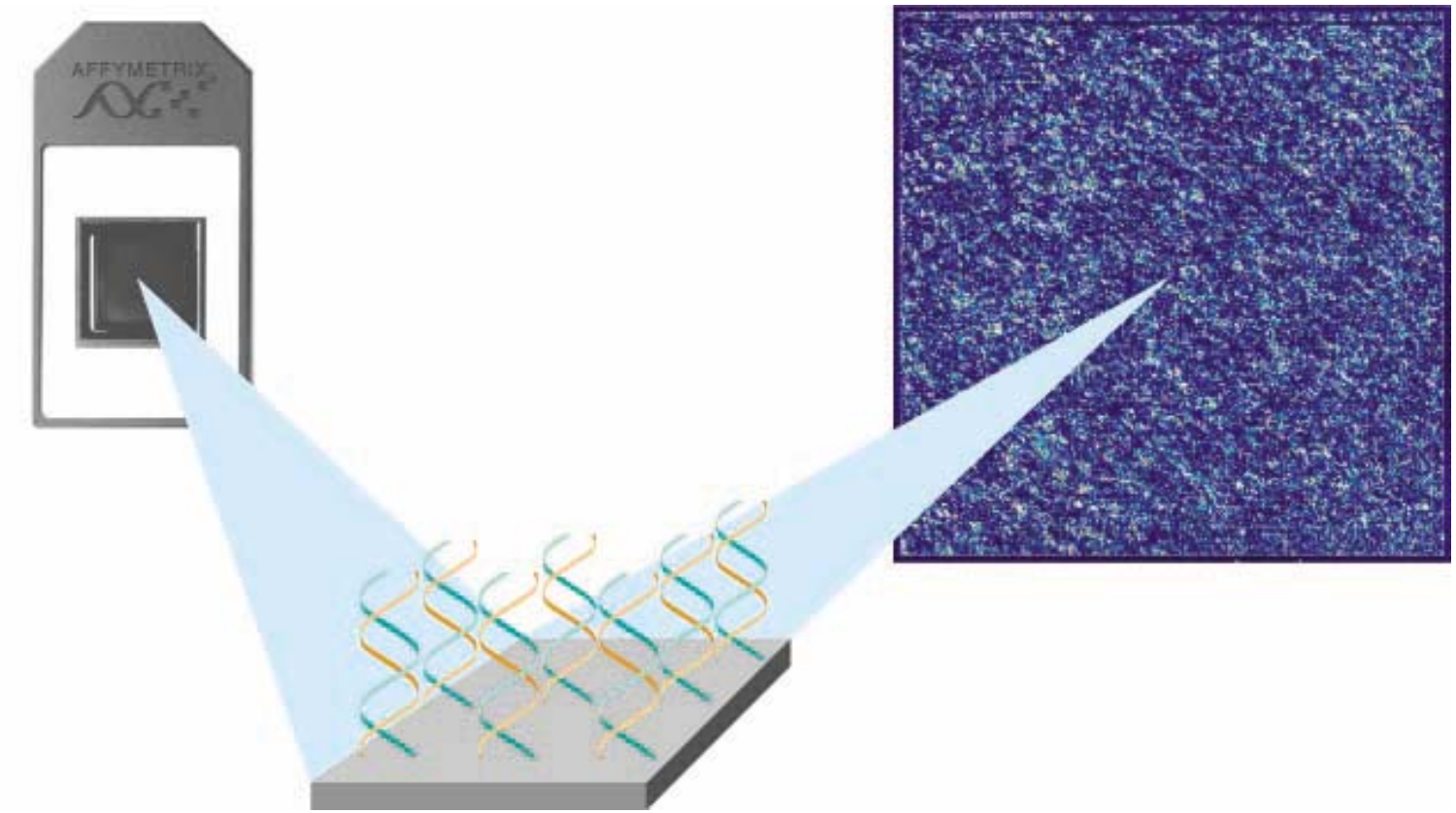
www.affymetrix.com

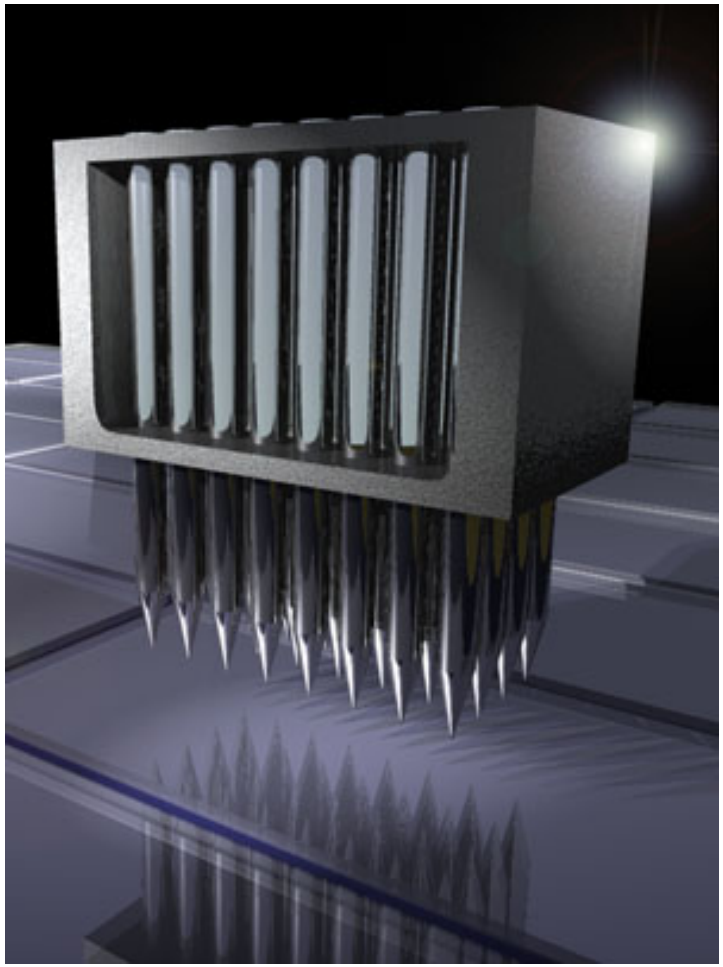
What's on the slide?





DNA Chips & Images

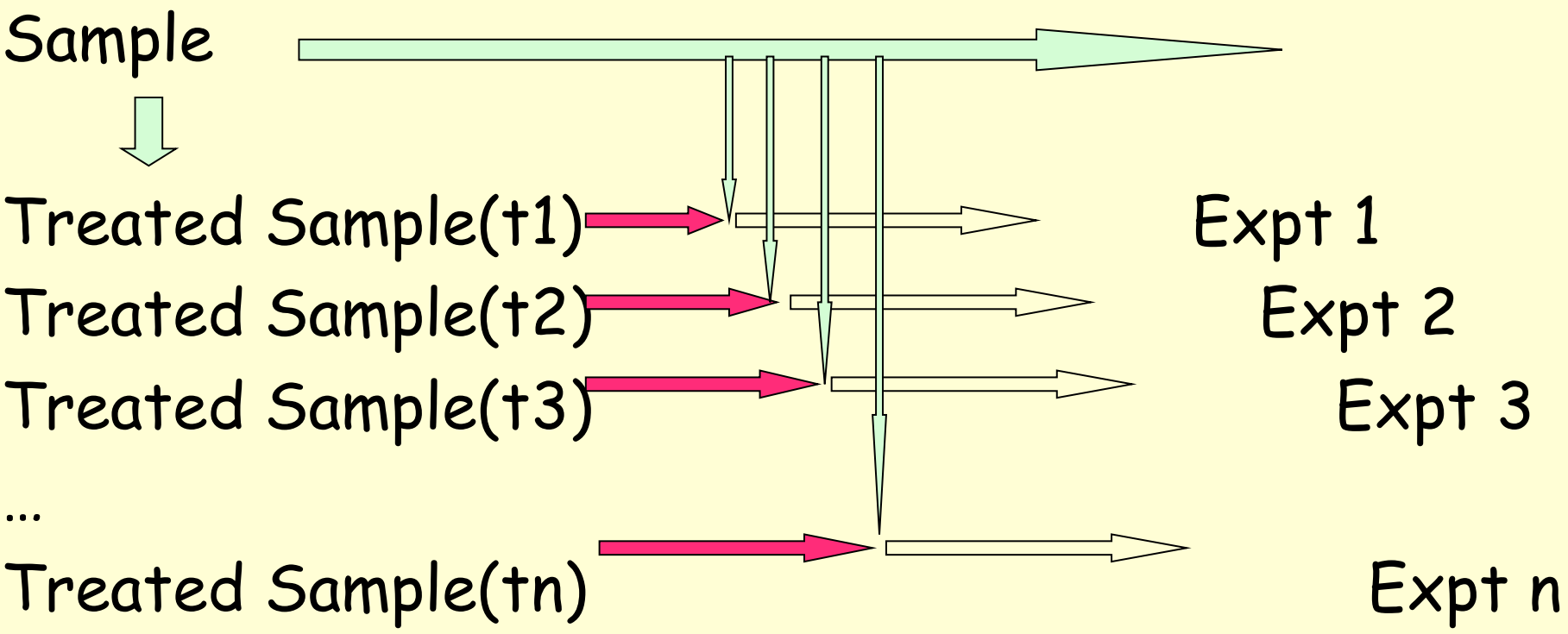


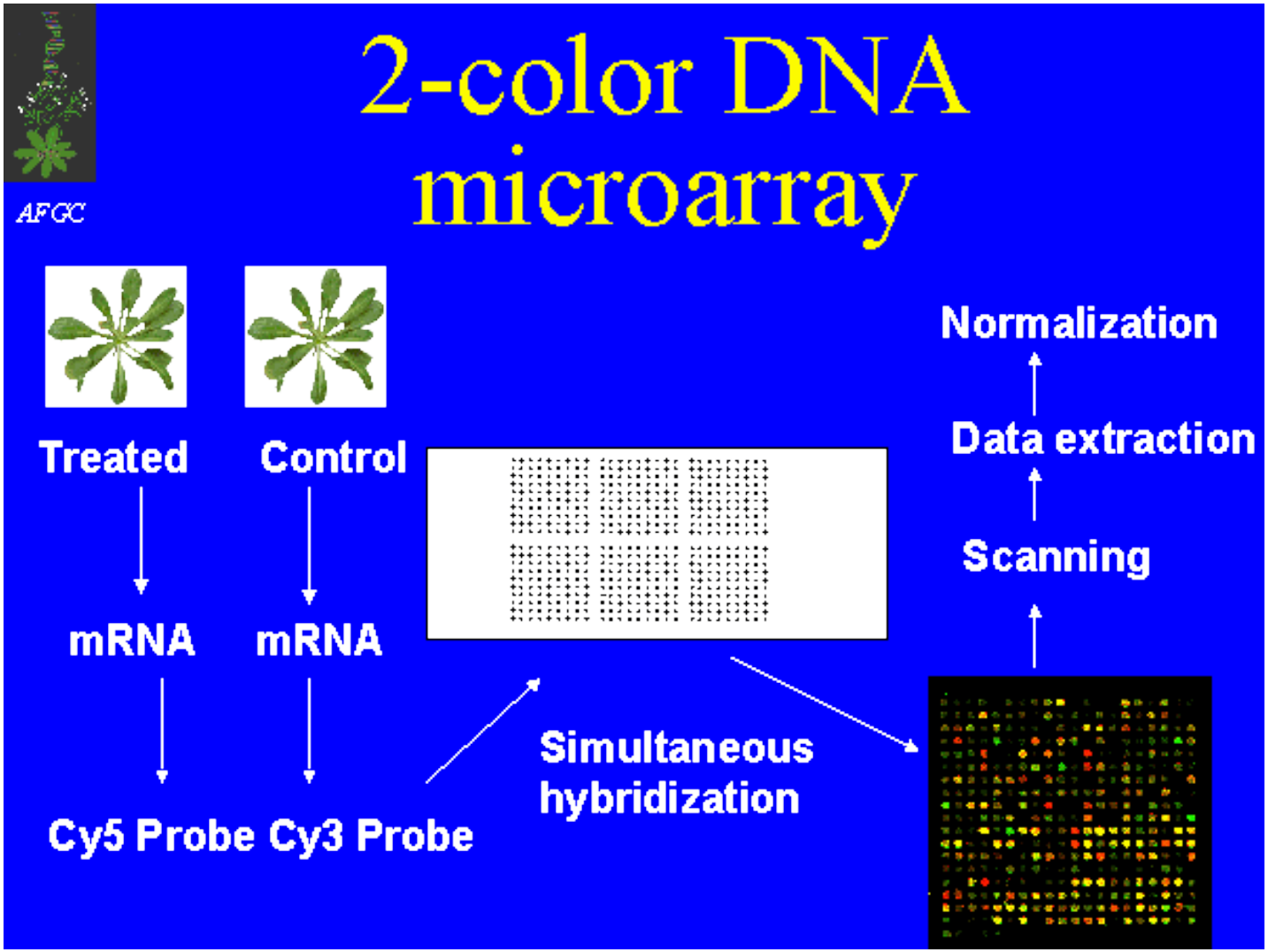


Microarrays: competing technologies

- 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