Gene Regulation and Microarrays

Overview

- A. Gene Expression and Regulation

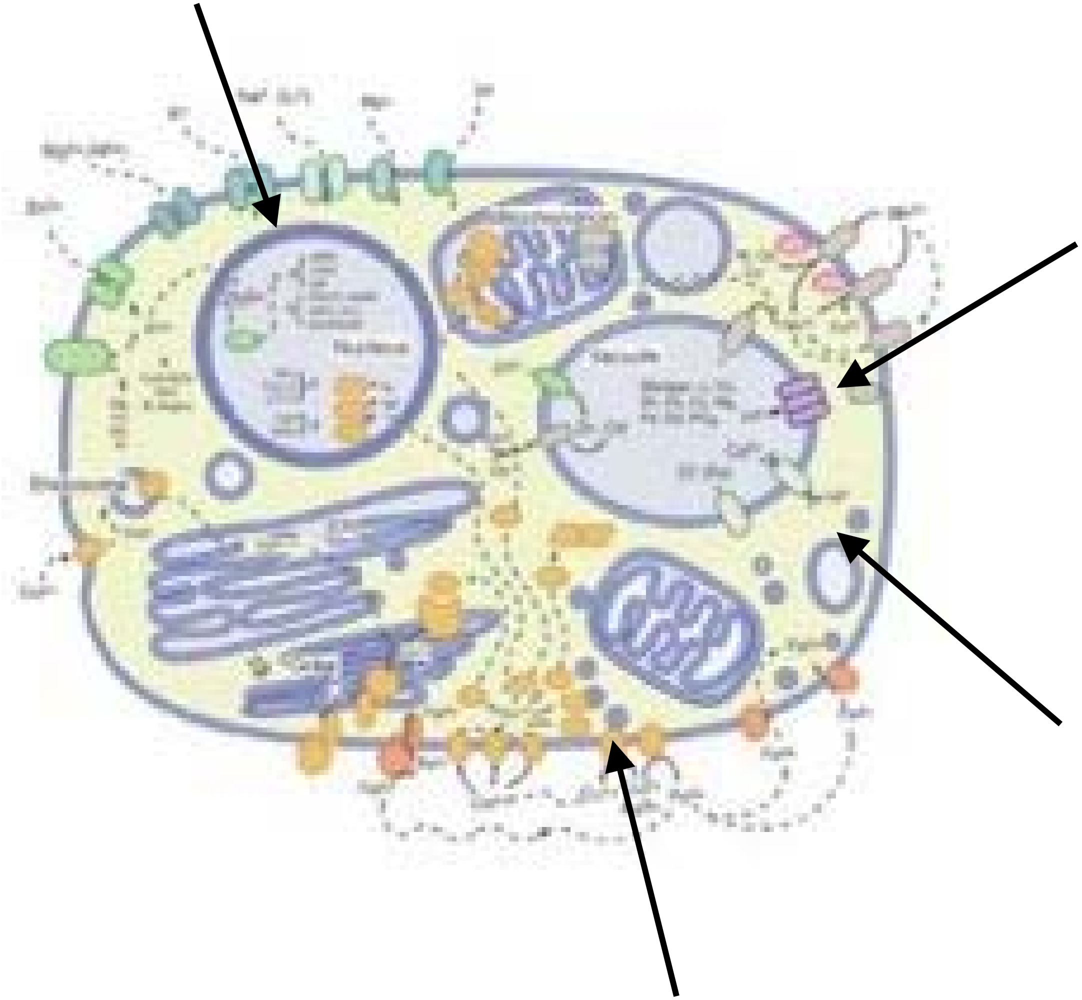
- B. Measuring Gene Expression: Microarrays

- C. Finding Regulatory Motifs

A. Regulation of Gene Expression

Cells respond to environment

Various external messages



Heat

Food Supply

Responds to environmental conditions

Genome is fixed – Cells are dynamic

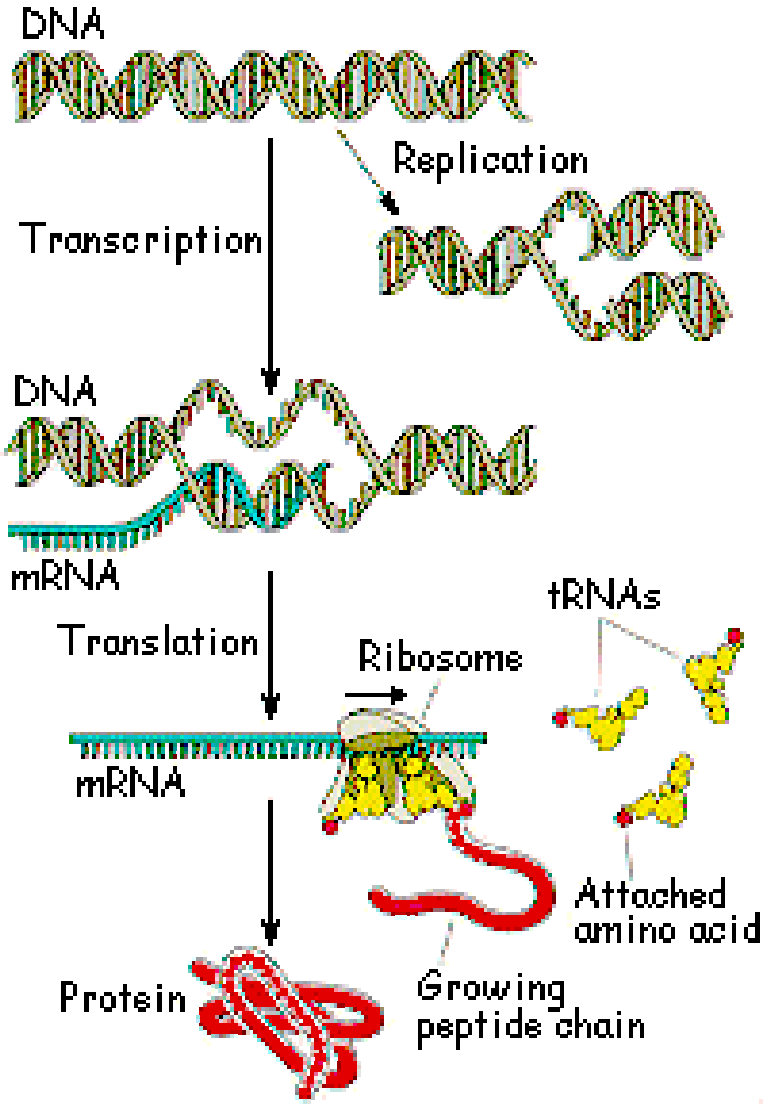
- A genome is static
  - Every cell in our body has a copy of same genome
- A cell is dynamic
  - Responds to external conditions
  - Most cells follow a cell cycle of division
- Cells differentiate during development

Gene regulation

- Gene regulation is responsible for dynamic cell

- Gene expression varies according to:
  - Cell type
  - Cell cycle
  - External conditions
  - Location

Where gene regulation takes place



- Opening of chromatin
- Transcription
- Translation
- Protein stability
- Protein modifications

Transcriptional Regulation

- Strongest regulation happens during transcription

- Best place to regulate:

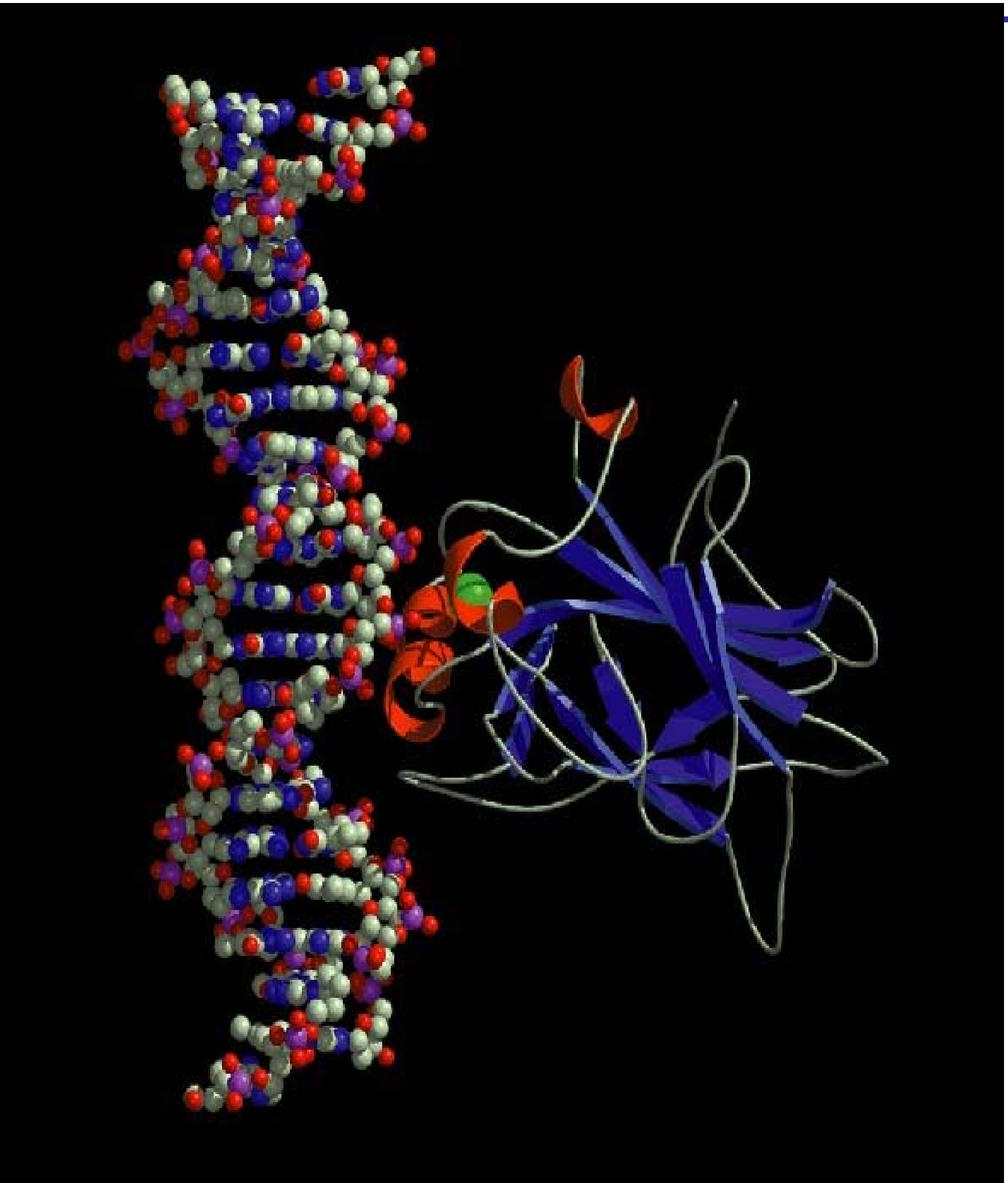
  No energy wasted making intermediate products

- However, slowest response time

  After a receptor notices a change:

  1. Cascade message to nucleus
  2. Open chromatin & bind transcription factors
  3. Recruit RNA polymerase and transcribe
  4. Splice mRNA and send to cytoplasm
  5. Translate into protein

Transcription Factors Binding to DNA



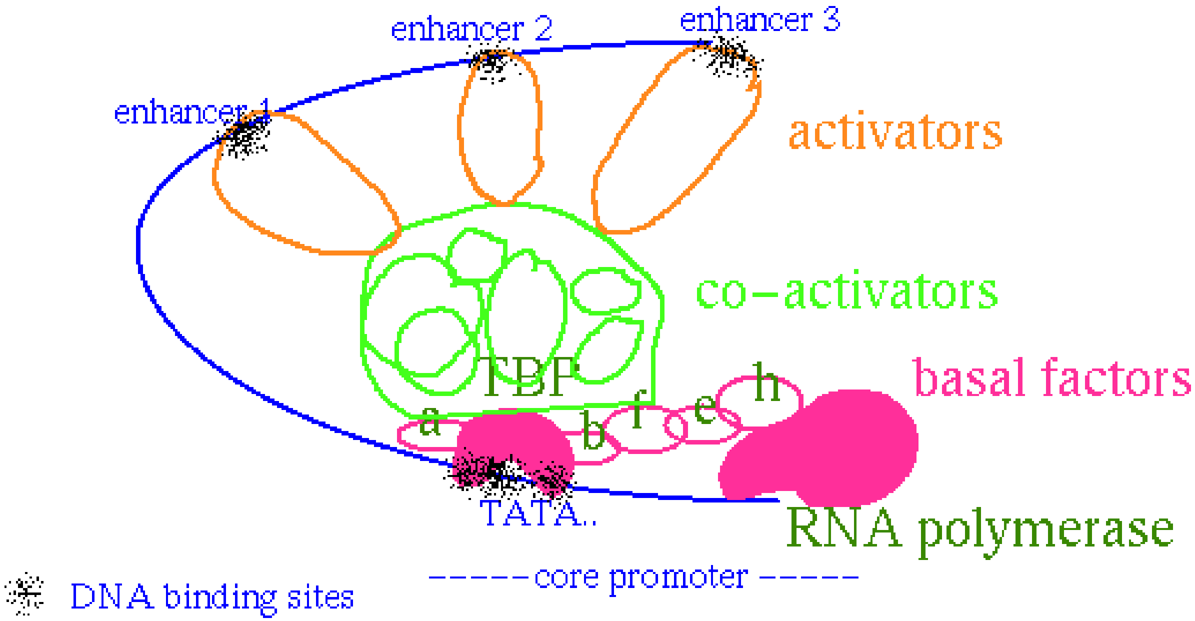
Transcription regulation:

Certain transcription factors bind DNA

Binding recognizes DNA substrings:

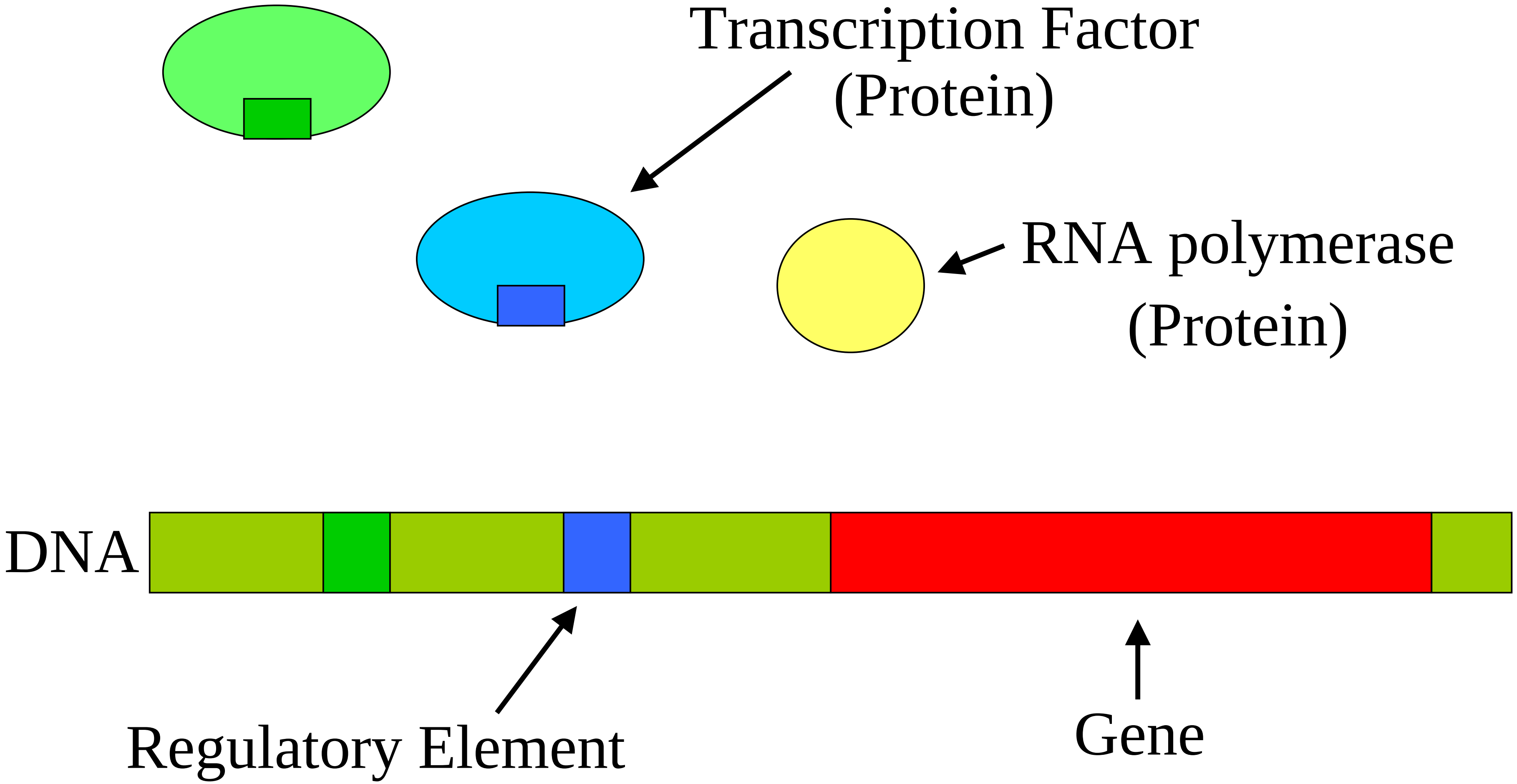
Regulatory motifs

Promoter and Enhancers

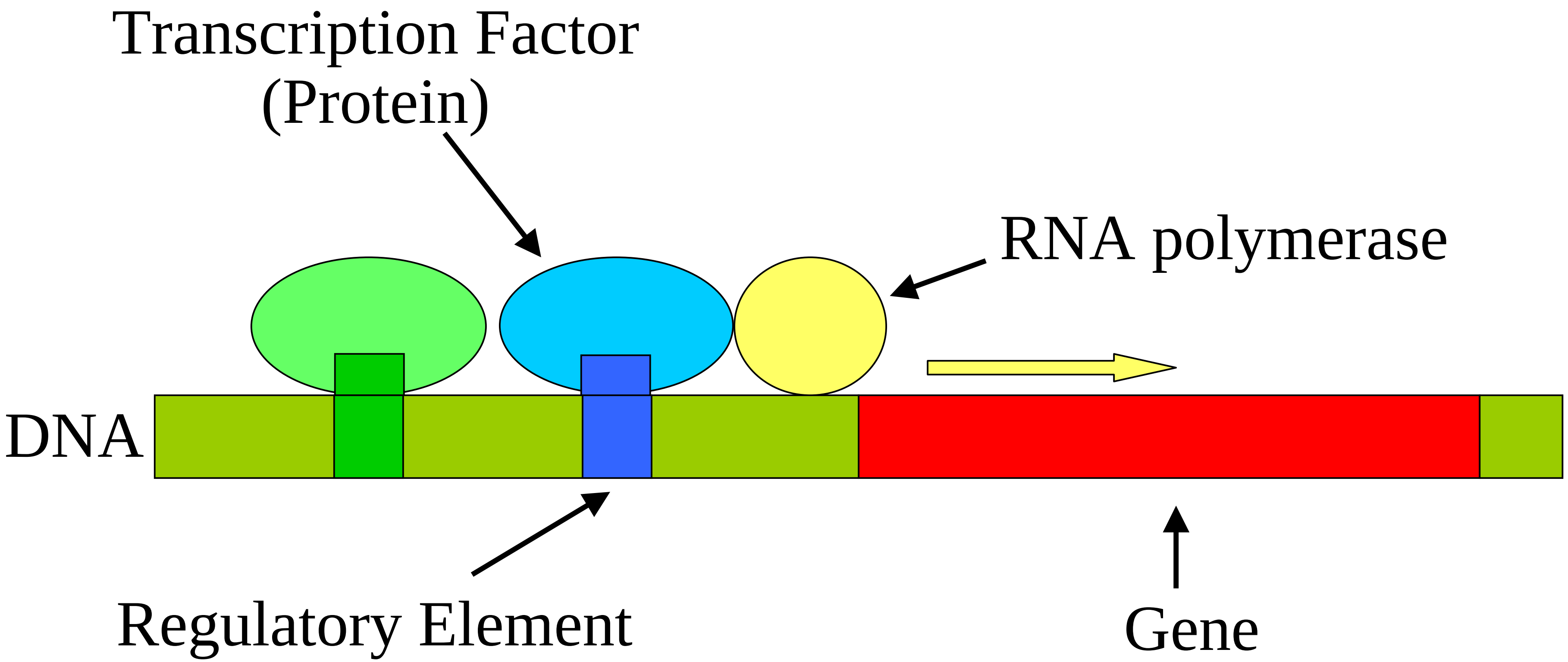


- **Promoter** necessary to start transcription
- **Enhancers** can affect transcription from afar

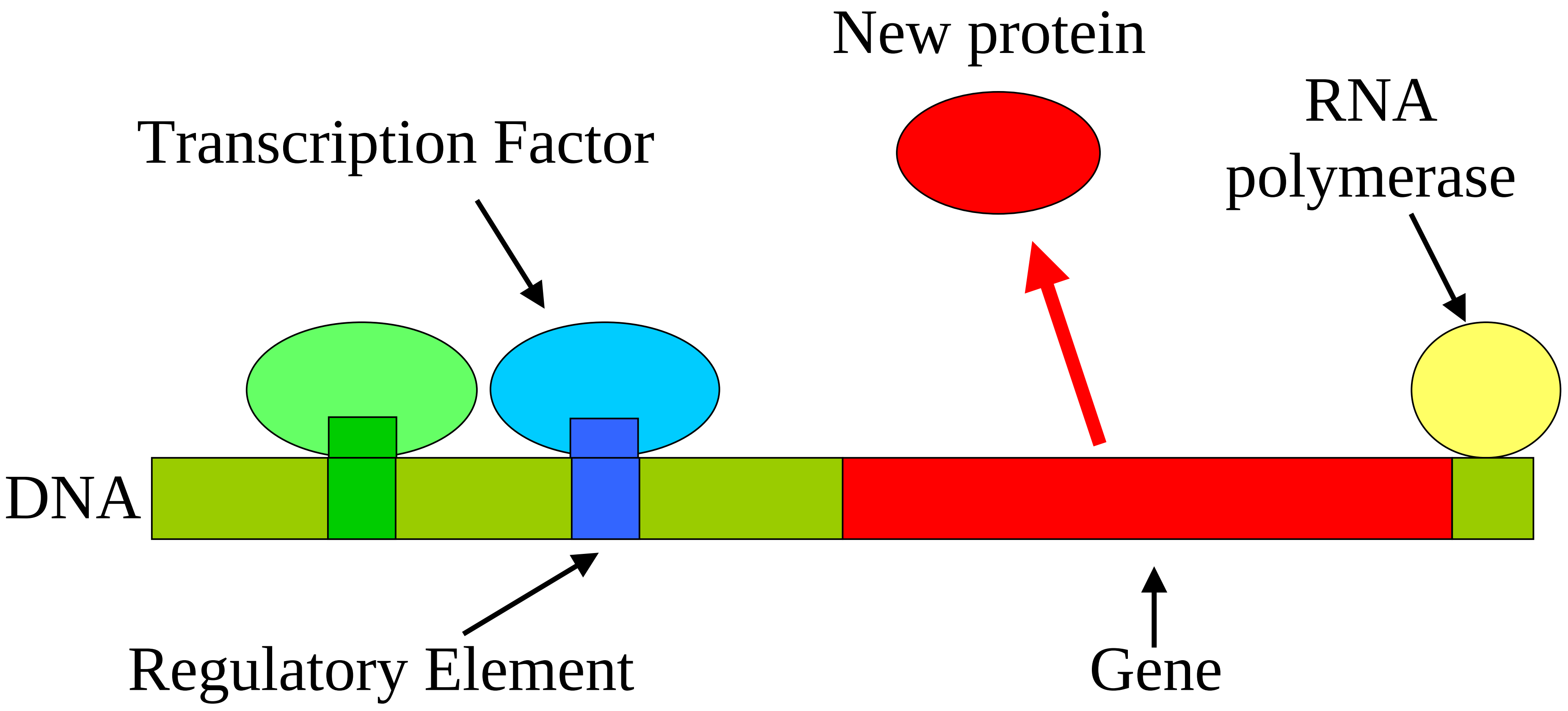
Regulation of Genes



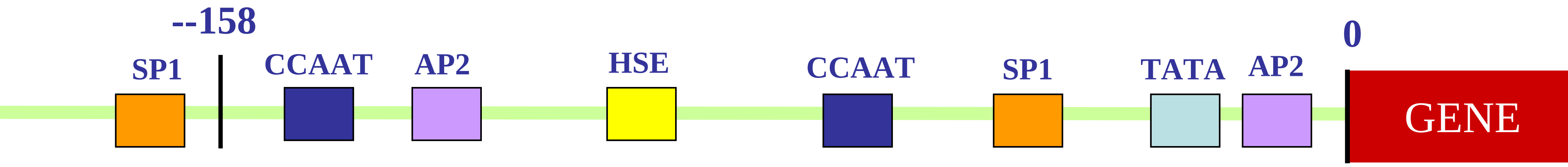
Regulation of Genes



Regulation of Genes



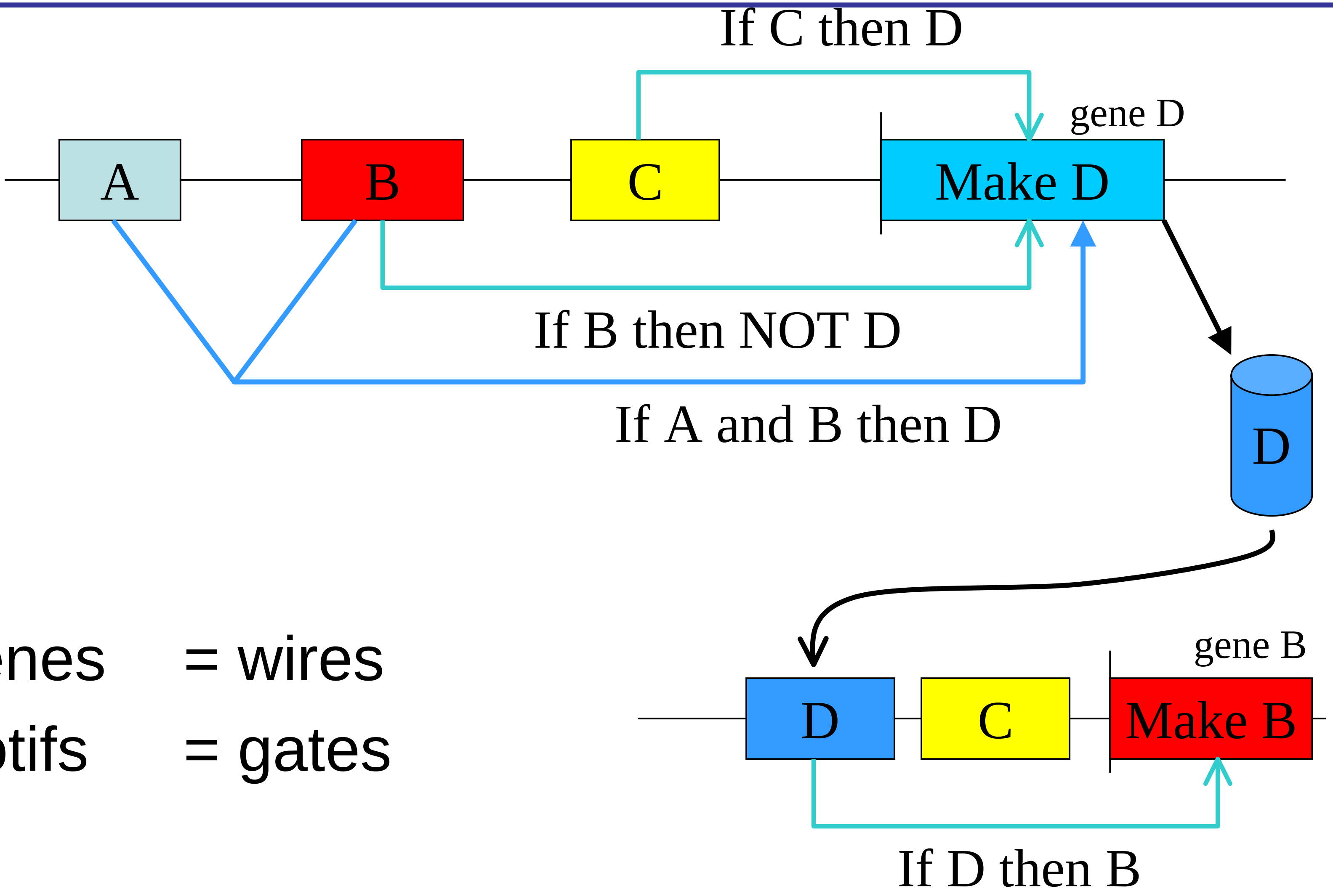
Example: A Human heat shock protein



promoter of heat shock hsp70

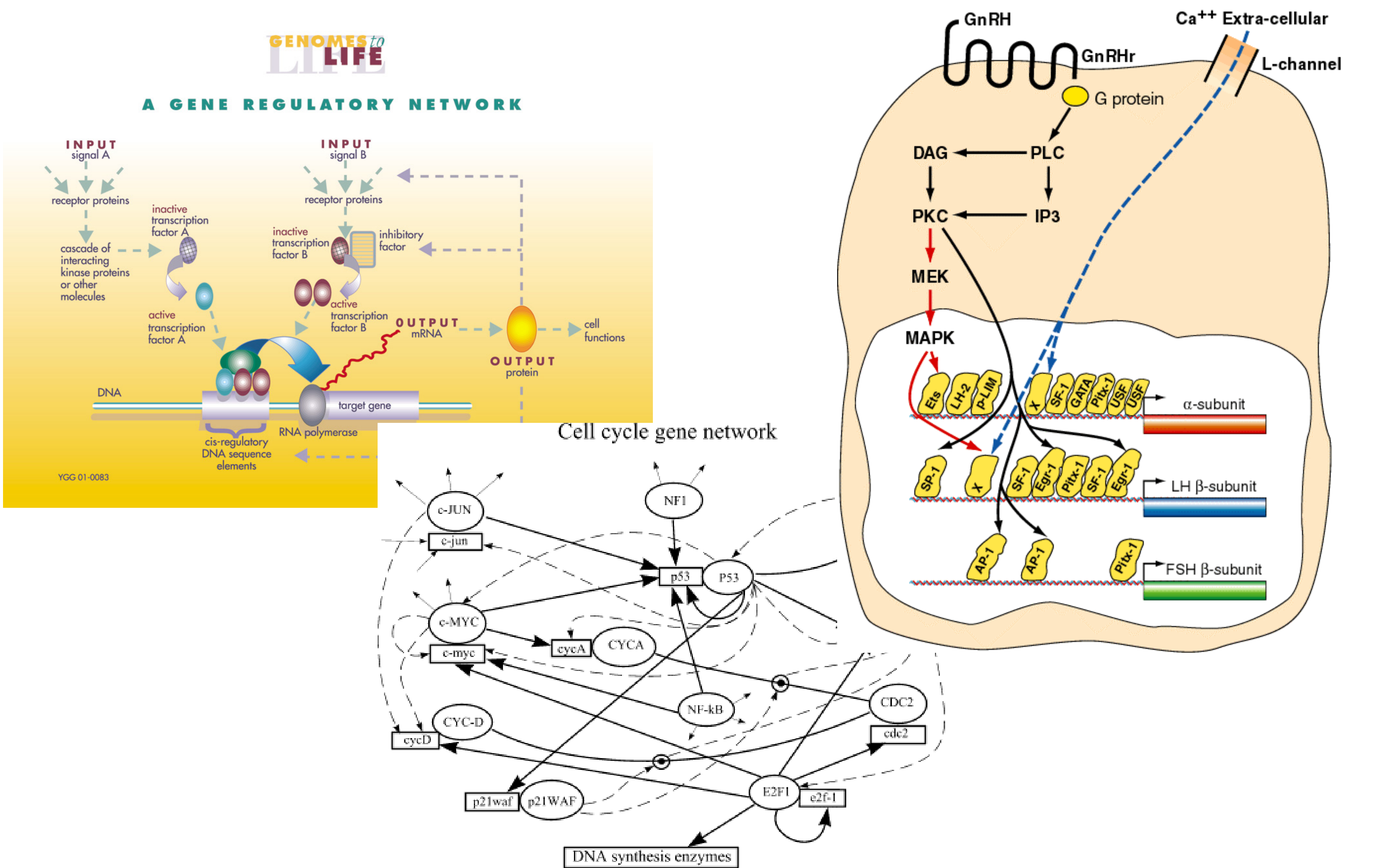
• TATA box:	positioning transcription start
• TATA, CCAAT:	constitutive transcription
• GRE:	glucocorticoid response
• MRE:	metal response
• HSE:	heat shock element

The Cell as a Regulatory Network



- Genes = wires
- Motifs = gates

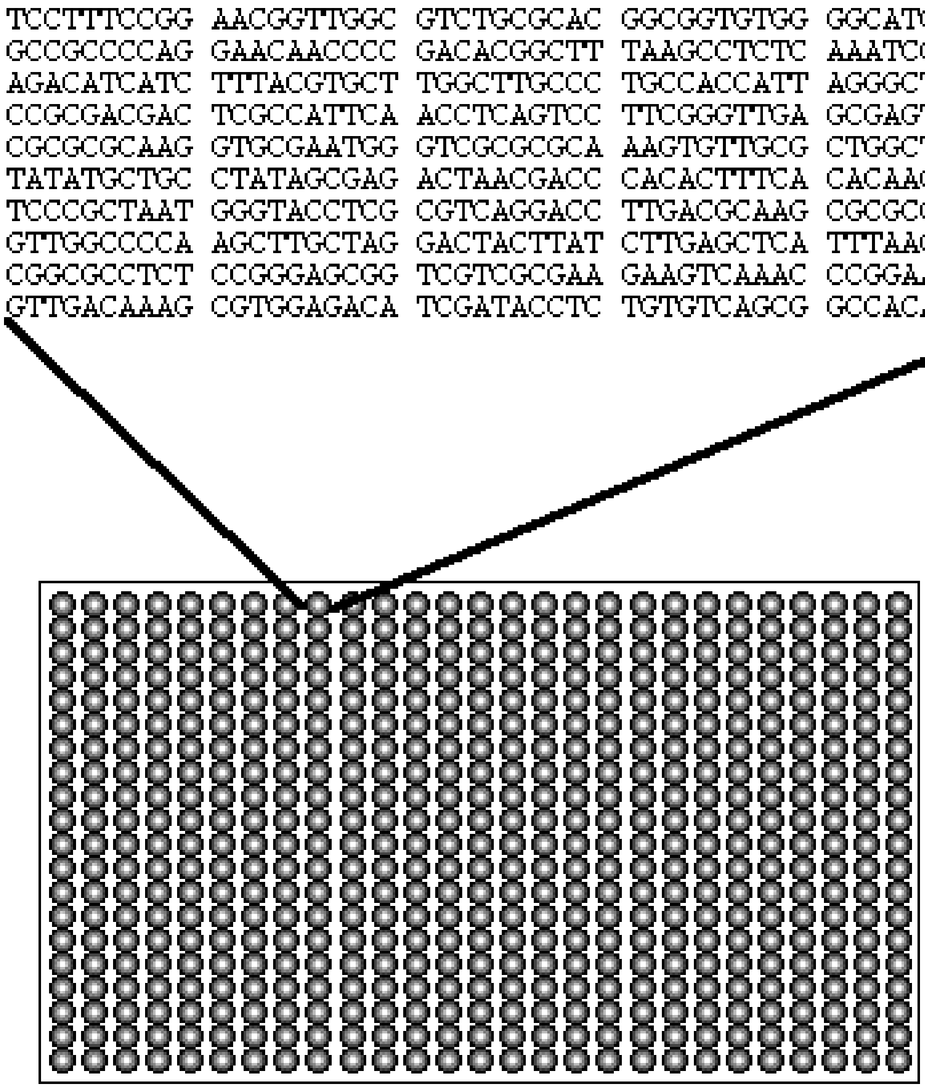
The Cell as a Regulatory Network (2)



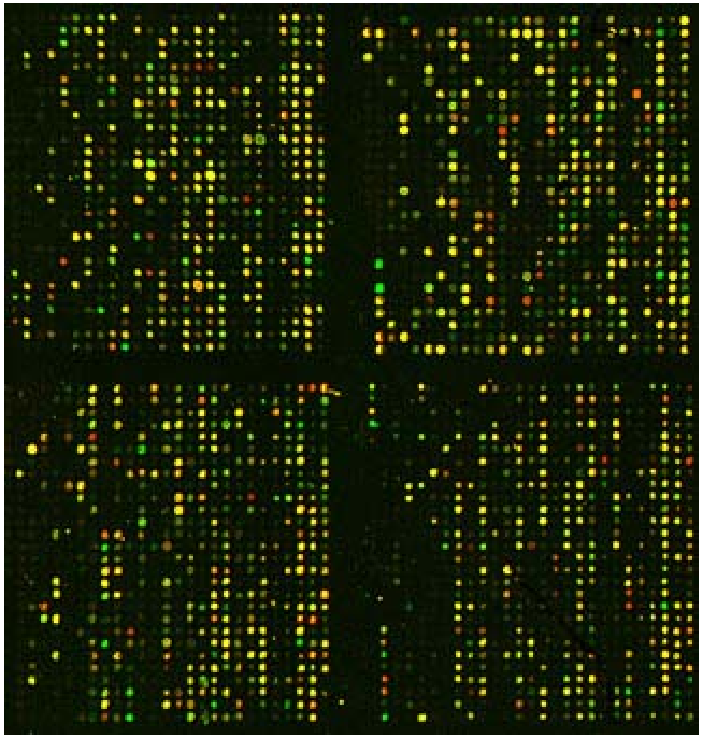
B. DNA Microarrays

Measuring gene transcription in a high-throughput fashion

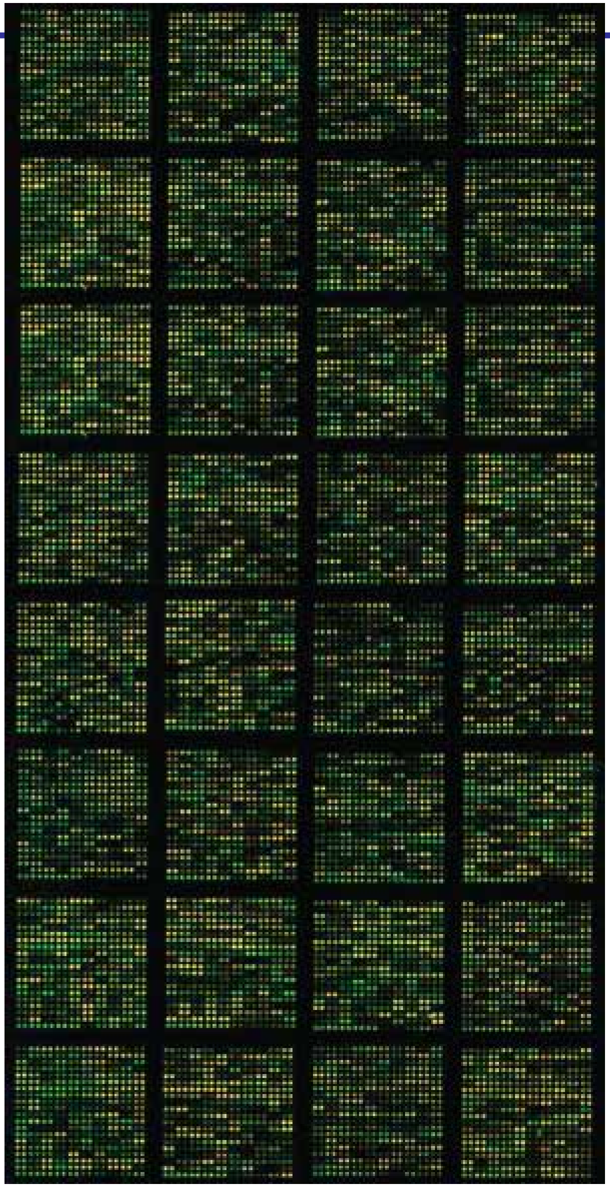
What is a microarray



```
TCCTTTCCGG AACGGTTGGC GTCTGCGCAC GGCGGTGTGG GGCATGACAT
GCCGCCCCAG GAACAACCCC GACACGGCTT TAAGCCTCTC AAATCGCTGT
AGACATCATC TTTACGTGCT TGGCTTGCCC TGCCACCATT AGGGCTGTTC
CCGCGACGAC TCGCCATTCA ACCTCAGTCC TTCGGGTTGA GCGAGTGGGT
CGCGCGCAAG GTGCGAATGG GTCGCGCGCA AAGTGTTGCG CTGGCTGTAT
TATATGCTGC CTATAGCGAG ACTAACGACC CACACTTTCA CACAAGGATT
TCCCGCTAAT GGGTACCTCG CGTCAGGACC TTGACGCAAG CGCGCCTTCG
GTTGGCCCCA AGCTTGCTAG GACTACTTAT CTTGAGCTCA TTTAACATCC
CGGCGCCTCT CCGGGAGCGG TCGTCGCGAA GAAGTCAAAC CCGGAACGGC
GTTGACAAAG CGTGGAGACA TCGATACCTC TGTGTCAGCG GCCACAAATC
```



What is a microarray (2)



- A 2D array of DNA sequences from thousands of genes
- Each spot has many copies of same gene
- Allow mRNAs from a sample to hybridize
- Measure number of hybridizations per spot

How to make a microarray

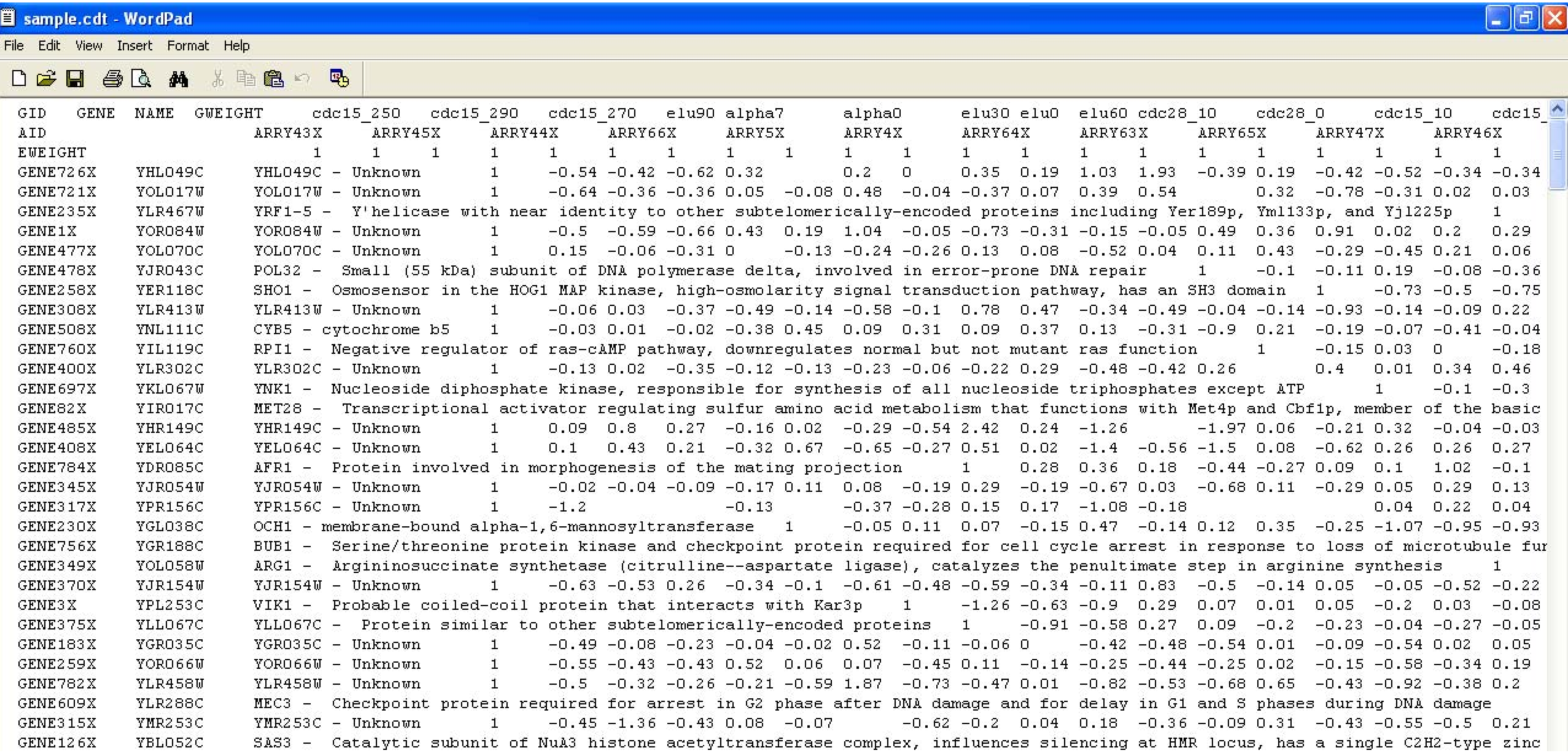
- Method 1: *DNA microarray* (Stanford)
  - Use PCR to amplify a 1Kb portion of each gene
  - Apply each sample on glass slide

- Method 2: *DNA Chip* (Affymetrix)
  - Grow oligonucleotides (25bp) on glass
  - Several words per gene (choose unique words)

If we know the gene sequences,

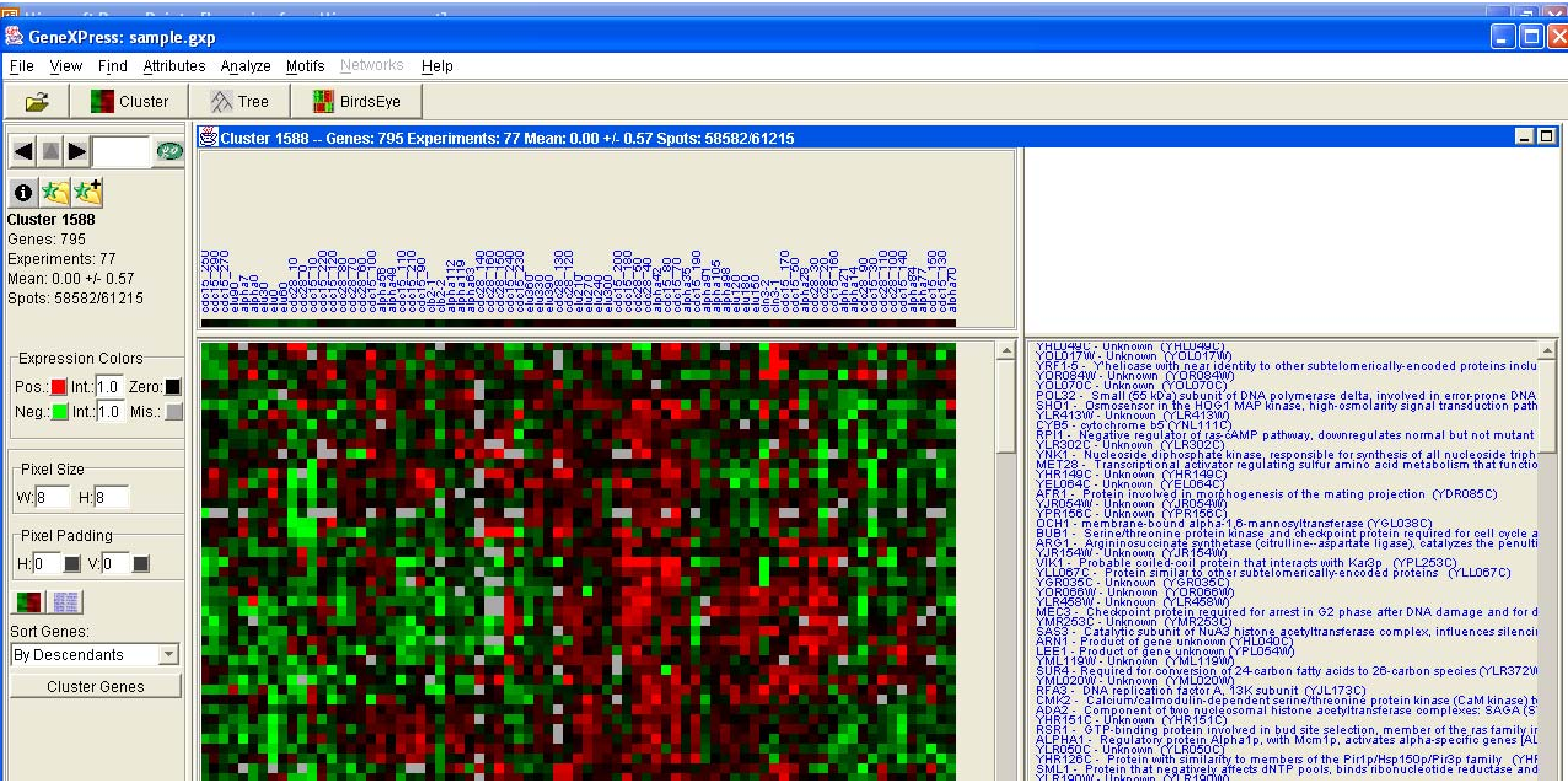
Can sample all genes in one experiment!

Sample Data



```
GID   GENE  NAME  GWEIGHT     cdc15_250   cdc15_290   cdc15_270   elu90 alpha7      alpha0      elu30 elu0  elu60 cdc28_10    cdc28_0     cdc15_10    cdc15_
AID                     ARRY43X     ARRY45X     ARRY44X     ARRY66X     ARRY5X      ARRY4X      ARRY64X     ARRY63X     ARRY65X     ARRY47X     ARRY46X
EWEIGHT                       1     1     1     1     1     1     1     1     1     1     1     1     1     1     1     1     1     1     1     1     1
GENE726X    YHL049C     YHL049C - Unknown       1     -0.54 -0.42 -0.62 0.32        0.2   0     0.35  0.19  1.03  1.93  -0.39 0.19  -0.42 -0.52 -0.34 -0.34
GENE721X    YOL017W     YOL017W - Unknown       1     -0.64 -0.36 -0.36 0.05  -0.08 0.48  -0.04 -0.37 0.07  0.39  0.54        0.32  -0.78 -0.31 0.02  0.03
GENE235X    YLR467W     YRF1-5 -  Y'helicase with near identity to other subtelomerically-encoded proteins including Yer189p, Yml133p, and Yjl225p    1
GENE1X      YOR084W     YOR084W - Unknown       1     -0.5  -0.59 -0.66 0.43  0.19  1.04  -0.05 -0.73 -0.31 -0.15 -0.05 0.49  0.36  0.91  0.02  0.2   0.29
GENE477X    YOL070C     YOL070C - Unknown       1     0.15  -0.06 -0.31 0     -0.13 -0.24 -0.26 0.13  0.08  -0.52 0.04  0.11  0.43  -0.29 -0.45 0.21  0.06
GENE478X    YJR043C     POL32 -  Small (55 kDa) subunit of DNA polymerase delta, involved in error-prone DNA repair     1     -0.1  -0.11 0.19  -0.08 -0.36
GENE258X    YER118C     SHO1 -  Osmosensor in the HOG1 MAP kinase, high-osmolarity signal transduction pathway, has an SH3 domain   1     -0.73 -0.5  -0.75
GENE308X    YLR413W     YLR413W - Unknown       1     -0.06 0.03  -0.37 -0.49 -0.14 -0.58 -0.1  0.78  0.47  -0.34 -0.49 -0.04 -0.14 -0.93 -0.14 -0.09 0.22
GENE508X    YNL111C     CYB5 - cytochrome b5    1     -0.03 0.01  -0.02 -0.38 0.45  0.09  0.31  0.09  0.37  0.13  -0.31 -0.9  0.21  -0.19 -0.07 -0.41 -0.04
GENE760X    YIL119C     RPI1 -  Negative regulator of ras-cAMP pathway, downregulates normal but not mutant ras function      1     -0.15 0.03  0     -0.18
GENE400X    YLR302C     YLR302C - Unknown       1     -0.13 0.02  -0.35 -0.12 -0.13 -0.23 -0.06 -0.22 0.29  -0.48 -0.42 0.26        0.4   0.01  0.34  0.46
GENE697X    YKL067W     YNK1 -  Nucleoside diphosphate kinase, responsible for synthesis of all nucleoside triphosphates except ATP       1     -0.1  -0.3
GENE82X     YIR017C     MET28 -  Transcriptional activator regulating sulfur amino acid metabolism that functions with Met4p and Cbf1p, member of the basic
GENE485X    YHR149C     YHR149C - Unknown       1     0.09  0.8   0.27  -0.16 0.02  -0.29 -0.54 2.42  0.24  -1.26       -1.97 0.06  -0.21 0.32  -0.04 -0.03
GENE408X    YEL064C     YEL064C - Unknown       1     0.1   0.43  0.21  -0.32 0.67  -0.65 -0.27 0.51  0.02  -1.4  -0.56 -1.5  0.08  -0.62 0.26  0.26  0.27
GENE784X    YDR085C     AFR1 -  Protein involved in morphogenesis of the mating projection      1     0.28  0.36  0.18  -0.44 -0.27 0.09  0.1   1.02  -0.1
GENE345X    YJR054W     YJR054W - Unknown       1     -0.02 -0.04 -0.09 -0.17 0.11  0.08  -0.19 0.29  -0.19 -0.67 0.03  -0.68 0.11  -0.29 0.05  0.29  0.13
GENE317X    YPR156C     YPR156C - Unknown       1     -1.2              -0.13       -0.37 -0.28 0.15  0.17  -1.08 -0.18                   0.04  0.22  0.04
GENE230X    YGL038C     OCH1 - membrane-bound alpha-1,6-mannosyltransferase   1     -0.05 0.11  0.07  -0.15 0.47  -0.14 0.12  0.35  -0.25 -1.07 -0.95 -0.93
GENE756X    YGR188C     BUB1 -  Serine/threonine protein kinase and checkpoint protein required for cell cycle arrest in response to loss of microtubule fur
GENE349X    YOL058W     ARG1 -  Argininosuccinate synthetase (citrulline--aspartate ligase), catalyzes the penultimate step in arginine synthesis     1
GENE370X    YJR154W     YJR154W - Unknown       1     -0.63 -0.53 0.26  -0.34 -0.1  -0.61 -0.48 -0.59 -0.34 -0.11 0.83  -0.5  -0.14 0.05  -0.05 -0.52 -0.22
GENE3X      YPL253C     VIK1 -  Probable coiled-coil protein that interacts with Kar3p    1     -1.26 -0.63 -0.9  0.29  0.07  0.01  0.05  -0.2  0.03  -0.08
GENE375X    YLL067C     YLL067C -  Protein similar to other subtelomerically-encoded proteins   1     -0.91 -0.58 0.27  0.09  -0.2  -0.23 -0.04 -0.27 -0.05
GENE183X    YGR035C     YGR035C - Unknown       1     -0.49 -0.08 -0.23 -0.04 -0.02 0.52  -0.11 -0.06 0     -0.42 -0.48 -0.54 0.01  -0.09 -0.54 0.02  0.05
GENE259X    YOR066W     YOR066W - Unknown       1     -0.55 -0.43 -0.43 0.52  0.06  0.07  -0.45 0.11  -0.14 -0.25 -0.44 -0.25 0.02  -0.15 -0.58 -0.34 0.19
GENE782X    YLR458W     YLR458W - Unknown       1     -0.5  -0.32 -0.26 -0.21 -0.59 1.87  -0.73 -0.47 0.01  -0.82 -0.53 -0.68 0.65  -0.43 -0.92 -0.38 0.2
GENE609X    YLR288C     MEC3 -  Checkpoint protein required for arrest in G2 phase after DNA damage and for delay in G1 and S phases during DNA damage
GENE315X    YMR253C     YMR253C - Unknown       1     -0.45 -1.36 -0.43 0.08  -0.07       -0.62 -0.2  0.04  0.18  -0.36 -0.09 0.31  -0.43 -0.55 -0.5  0.21
GENE126X    YBL052C     SAS3 -  Catalytic subunit of NuA3 histone acetyltransferase complex, influences silencing at HMR locus, has a single C2H2-type zinc
```

Visualization Tools



Cluster 1588 -- Genes: 795 Experiments: 77 Mean: 0.00 +/- 0.57 Spots: 58582/61215

Cluster 1588
Genes: 795
Experiments: 77
Mean: 0.00 +/- 0.57
Spots: 58582/61215

Pos.: 🟥 Int.: 1.0 Zero: ■
Neg.: 🟩 Int.: 1.0 Mis.: ▨

W: 8 H: 8

H: 0 V: 0

Sort Genes: By Descendants

```
YHL049C - Unknown  (YHL049C)
YOL017W - Unknown  (YOL017W)
YRF1-5 -  Y'helicase with near identity to other subtelomerically-encoded proteins inclu
YOR084W - Unknown  (YOR084W)
YOL070C - Unknown  (YOL070C)
POL32 -  Small (55 kDa) subunit of DNA polymerase delta, involved in error-prone DNA
SHO1 -  Osmosensor in the HOG1 MAP kinase, high-osmolarity signal transduction path
YLR413W - Unknown  (YLR413W)
CYB5 - cytochrome b5 (YNL111C)
RPI1 -  Negative regulator of ras-cAMP pathway, downregulates normal but not mutant
YLR302C - Unknown  (YLR302C)
YNK1 -  Nucleoside diphosphate kinase, responsible for synthesis of all nucleoside triph
MET28 -  Transcriptional activator regulating sulfur amino acid metabolism that functio
YHR149C - Unknown  (YHR149C)
YEL064C - Unknown  (YEL064C)
AFR1 -  Protein involved in morphogenesis of the mating projection  (YDR085C)
YJR054W - Unknown  (YJR054W)
YPR156C - Unknown  (YPR156C)
OCH1 - membrane-bound alpha-1,6-mannosyltransferase (YGL038C)
BUB1 -  Serine/threonine protein kinase and checkpoint protein required for cell cycle a
ARG1 -  Argininosuccinate synthetase (citrulline--aspartate ligase), catalyzes the penulti
YJR154W - Unknown  (YJR154W)
VIK1 -  Probable coiled-coil protein that interacts with Kar3p  (YPL253C)
YLL067C -  Protein similar to other subtelomerically-encoded proteins  (YLL067C)
YGR035C - Unknown  (YGR035C)
YOR066W - Unknown  (YOR066W)
YLR458W - Unknown  (YLR458W)
MEC3 -  Checkpoint protein required for arrest in G2 phase after DNA damage and for d
YMR253C - Unknown  (YMR253C)
SAS3 -  Catalytic subunit of NuA3 histone acetyltransferase complex, influences silenci
ARN1 - Product of gene unknown (YHL040C)
LEE1 - Product of gene unknown (YPL054W)
YML119W - Unknown  (YML119W)
SUR4 - Required for conversion of 24-carbon fatty acids to 26-carbon species (YLR372W
YML020W - Unknown  (YML020W)
RFA3 -  DNA replication factor A, 13K subunit  (YJL173C)
CMK2 -  Calcium/calmodulin-dependent serine/threonine protein kinase (CaM kinase) t
ADA2 -  Component of two nucleosomal histone acetyltransferase complexes: SAGA (S
YHR151C - Unknown  (YHR151C)
RSR1 -  GTP-binding protein involved in bud site selection, member of the ras family ir
ALPHA1 -  Regulatory protein Alpha1p, with Mcm1p, activates alpha-specific genes [AL
YLR050C - Unknown  (YLR050C)
YHR126C -  Protein with similarity to members of the Pir1p/Hsp150p/Pir3p family  (YHF
SML1 -  Protein that negatively affects dNTP pools, binds ribonucleotide reductase and
YLR190W - Unknown  (YLR190W)
```

Goal of Microarray Experiments

- Measure level of gene expression across many different conditions:
  - Expression Matrix M: {genes}×{conditions}:

    M_{ij} = |gene_i| in condition_j

- Deduce gene function
  - Genes with similar function are expressed under similar conditions

- Deduce gene regulatory networks – parts and connections-level description of biology

Analysis of Microarray Data

- Clustering
  - **Idea:** Groups of genes that share similar function have similar expression patterns
    - Hierarchical clustering
    - k-means
    - Bayesian approaches
    - Projection techniques
      - Principal Component Analysis
      - Independent Component Analysis
- Classification
  - **Idea:** A cell can be in one of several states
    - (Diseased vs. Healthy, Cancer X vs. Cancer Y vs. Normal)
  - Can we train an algorithm to use the gene expression patterns to determine which state a cell is in?
    - Support Vector Machines
    - Decision Trees
    - Neural Networks
    - K-Nearest Neighbors

Hierarchical Agglomerative Clustering

Michael Eisen, 1998

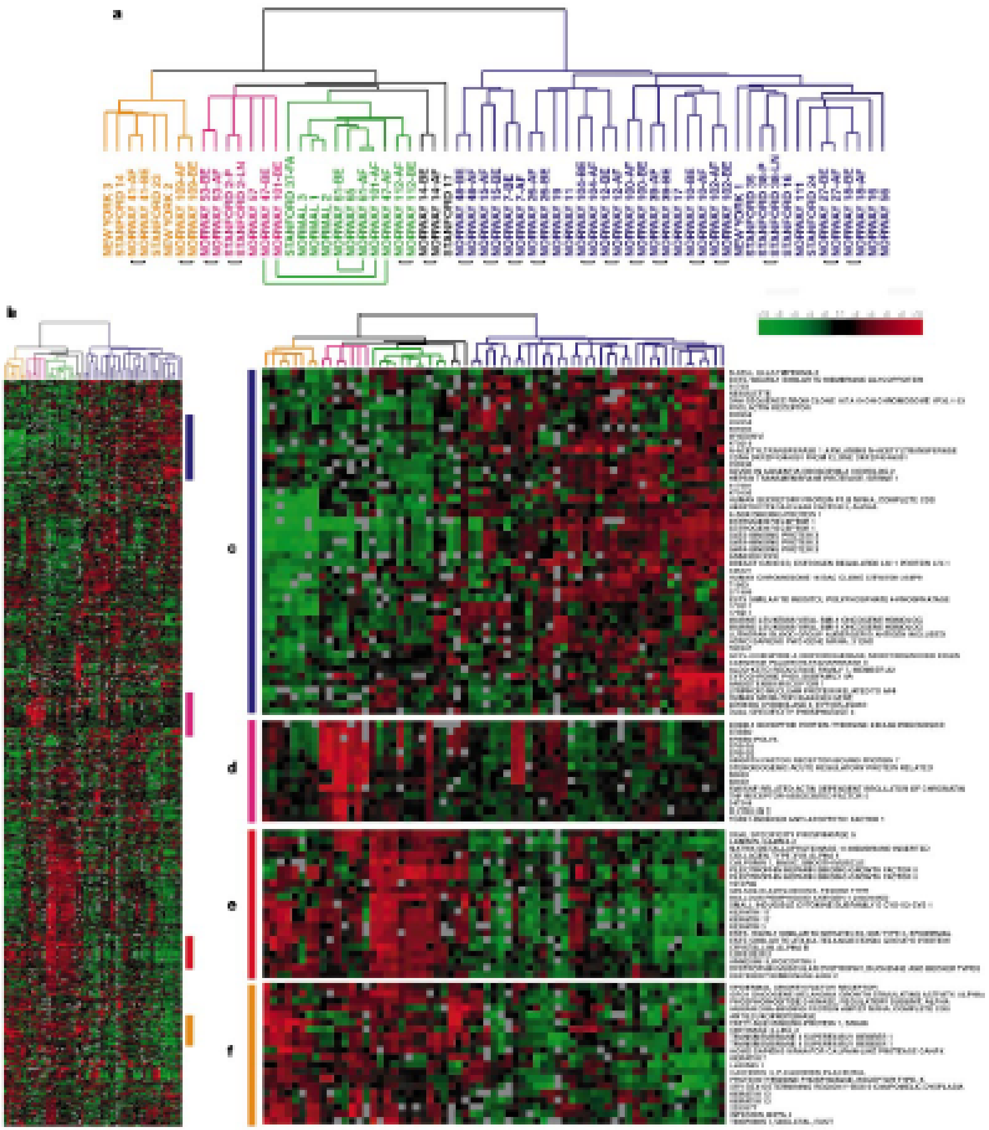
- Hierarchical Agglomerative Clustering
  - Step 1: Similarity score between all pairs of genes
    - Pearson Correlation

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\left(n-1\right)S_X S_Y}$$

  - Step 2: Find the two most similar genes, replace with a node that contains the average
    - Builds a tree of genes
  - Step 3: Repeat.

  - Can do the same with experiments

Results of Clustering Gene Expression



- CLUSTER is simple and easy to use
- De facto standard for microarray analysis

Time: $O(N^2M)$

N: #genes
M: #conditions

K-Means Clustering Algorithm

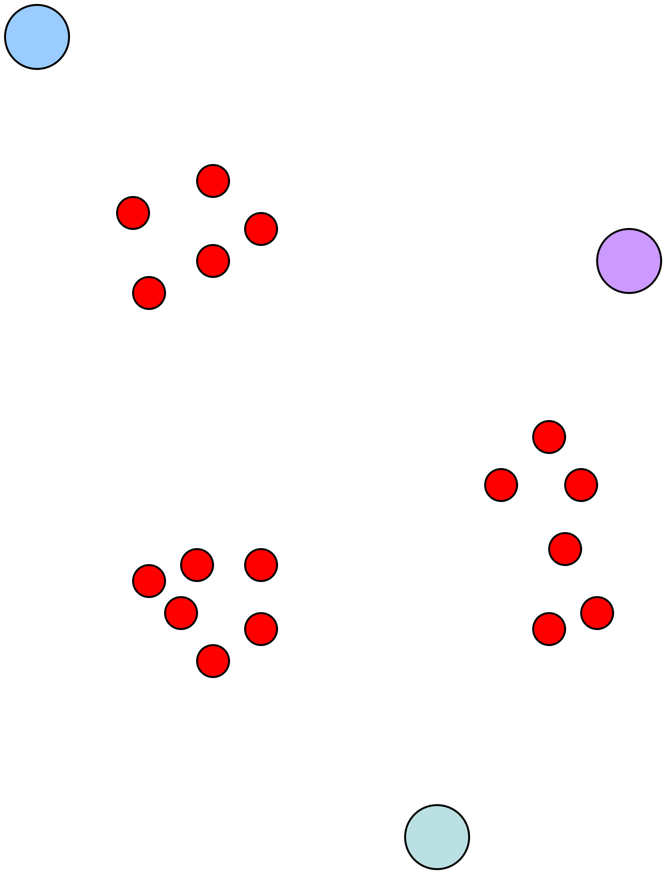
- Randomly initialize k cluster means
- Iterate:
  - Assign each genes to the nearest cluster mean
  - Recompute cluster means
- Stop when clustering converges

Notes:

- Really fast
- Genes are partitioned into clusters
- How do we select k?

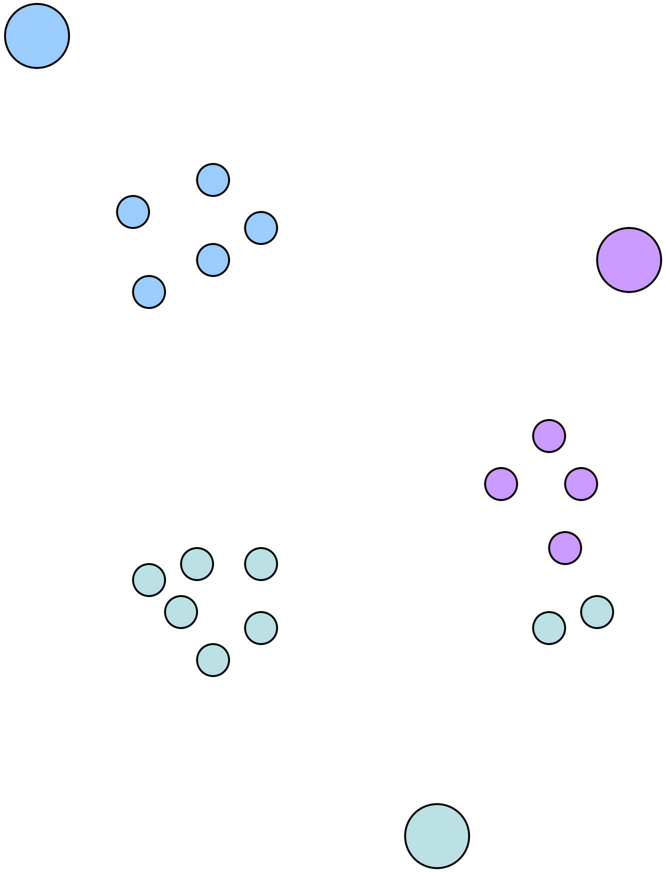
K-Means Algorithm

- Randomly Initialize Clusters



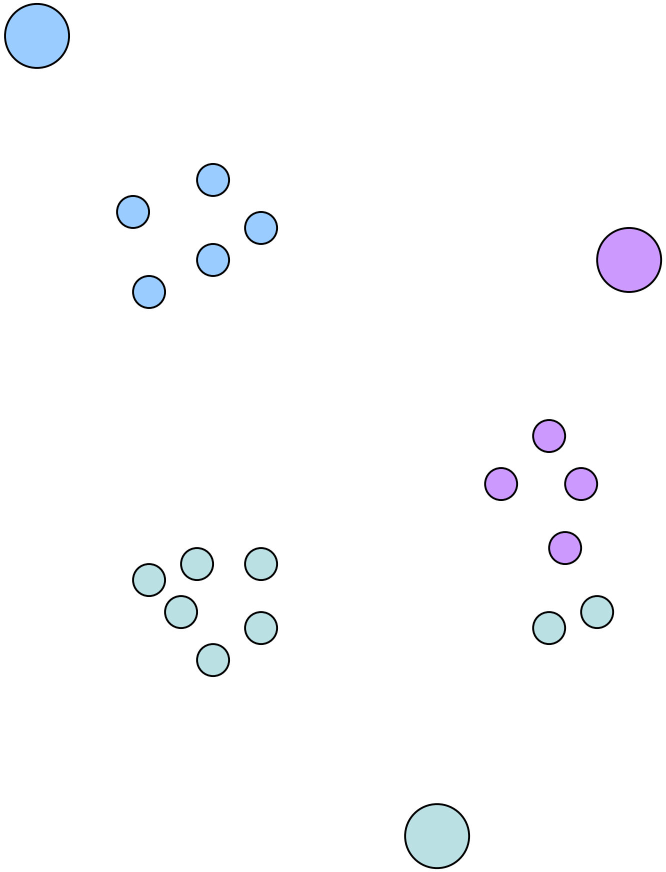
K-Means Algorithm

- Assign data points to nearest clusters



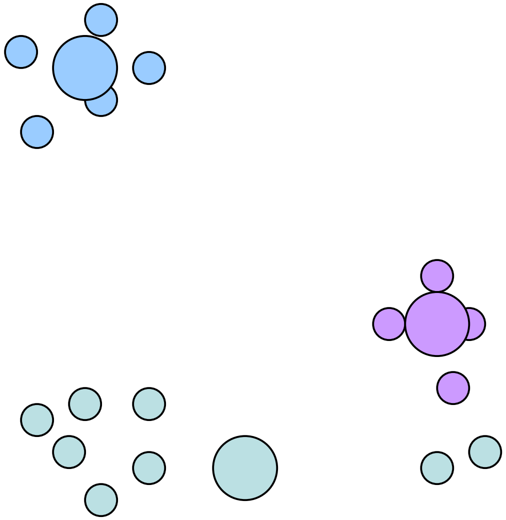
K-Means Algorithm

- Recalculate Clusters



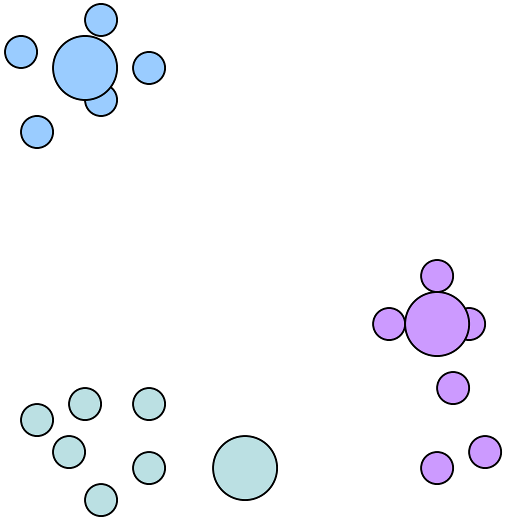
K-Means Algorithm

- Recalculate Clusters



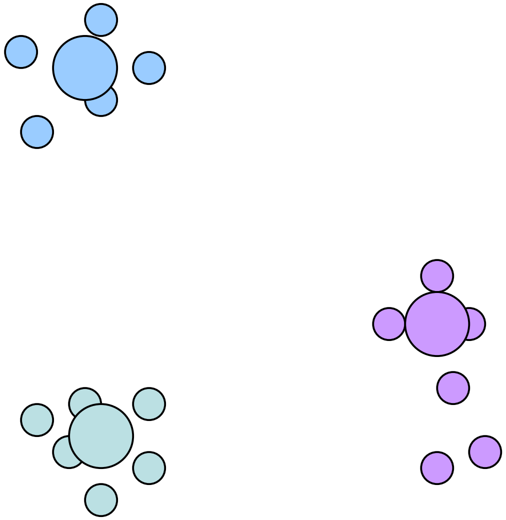
K-Means Algorithm

- Repeat



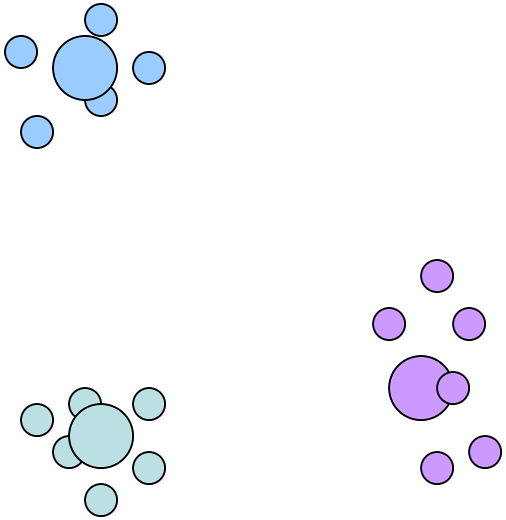
K-Means Algorithm

- Repeat



K-Means Algorithm

- Repeat … until convergence



Time: O(KNM) per iteration

N: #genes
M: #conditions

Multiple-pass K-Means clustering

(A Gasch, MB Eisen 2002)

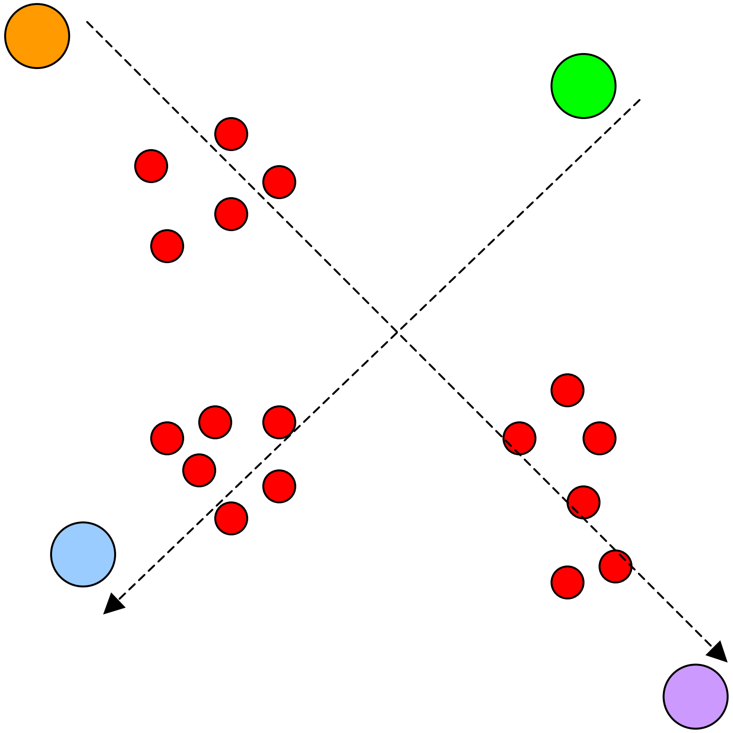
- Each gene can belong to many clusters
- Soft (fuzzy) assignment of genes to clusters
  - Each gene has 1.0 membership units, allocated amongst clusters based on correlation with means
- Cluster means are calculated by taking the weighted average of all the genes in the cluster

Algorithm:

- Use PCA to initialize cluster means
- 3 applications of k-means clustering, find k/3 clusters per application
  - In each application, start with brand new clusters and initializations
- And a few more heuristic tricks

Initialization

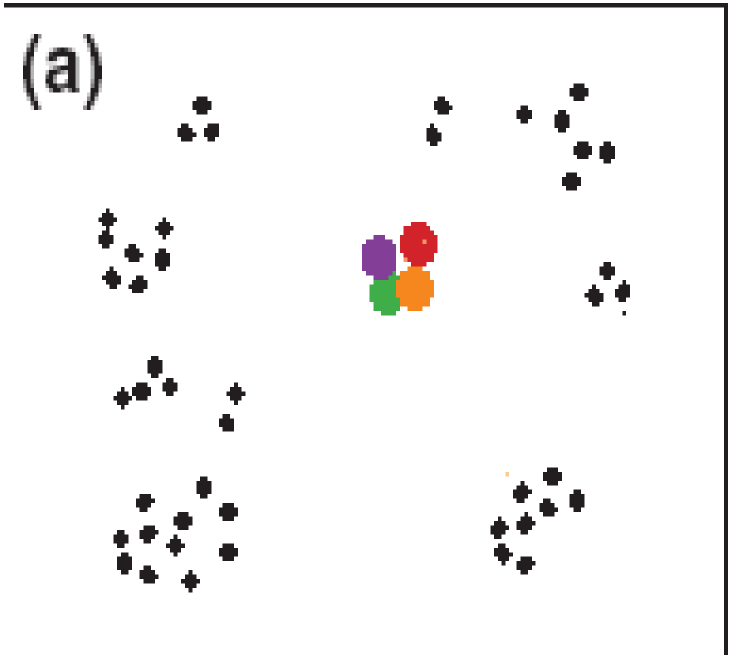
- Use PCA to find a few eigenvectors for initialization
- These features capture the directions of maximum variance
- Must be orthonormal



Example

Initialization

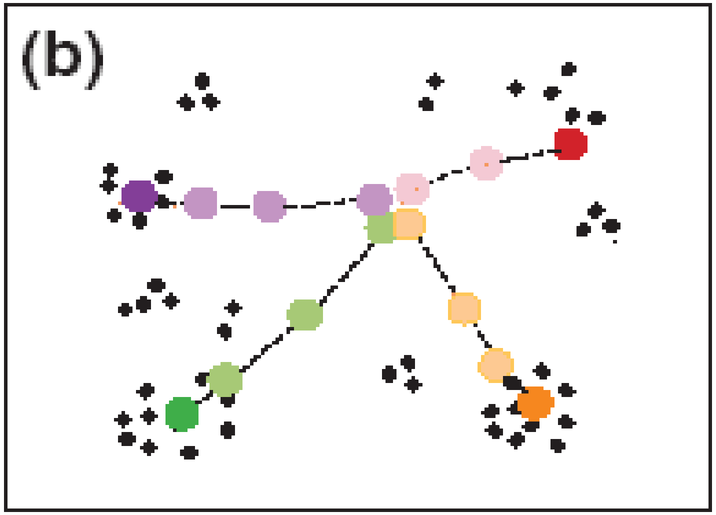
- k/3 centroids defined from k/3 first eigenvectors



Example

- First application of clustering

$$J(F,V) = \sum_{i=1}^{N} \sum_{j=1}^{K} m^2_{XiVj} \, d^2_{XiVj}$$



Objective function to minimize, **J(F, V)**

X genes

F assignment of genes to clusters

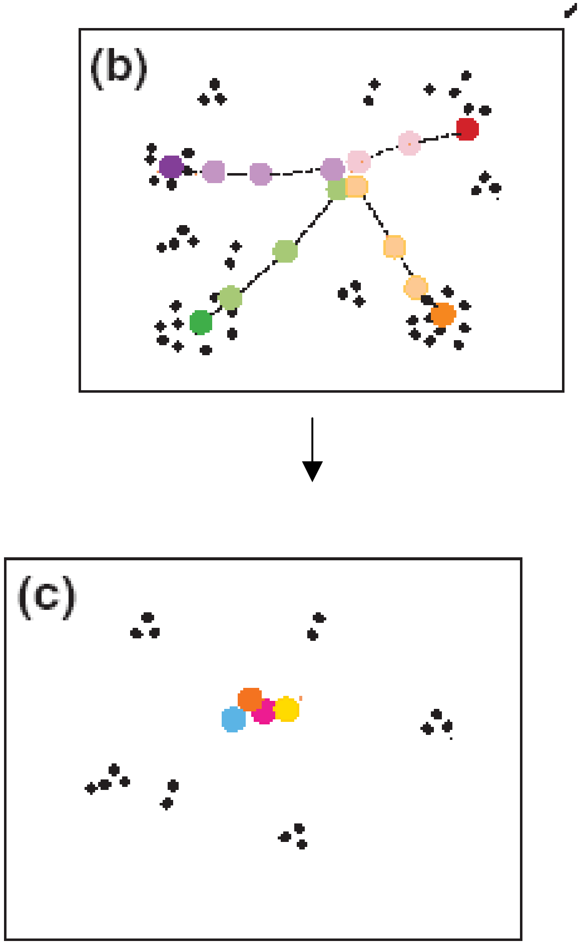
$\mathbf{m}_{XV}$ assign. coeff. of gene X_i to cluster V_j

$\mathbf{d}_{XV}$ distance of gene X_i with centroid V_j

Iteration of the approach

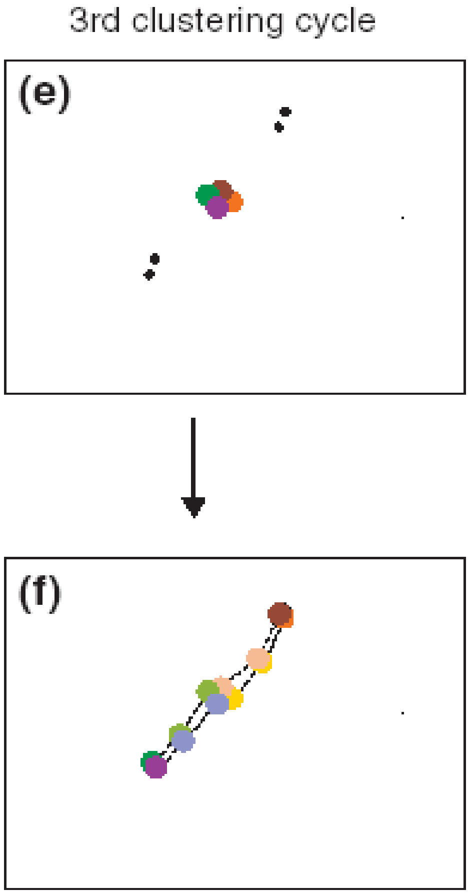
- Remove genes that have a Pearson Correlation with a particular cluster greater than .7
  - Intuition: These strong signal from these genes has been accounted for

- Repeat



Removing Duplicate Centroids

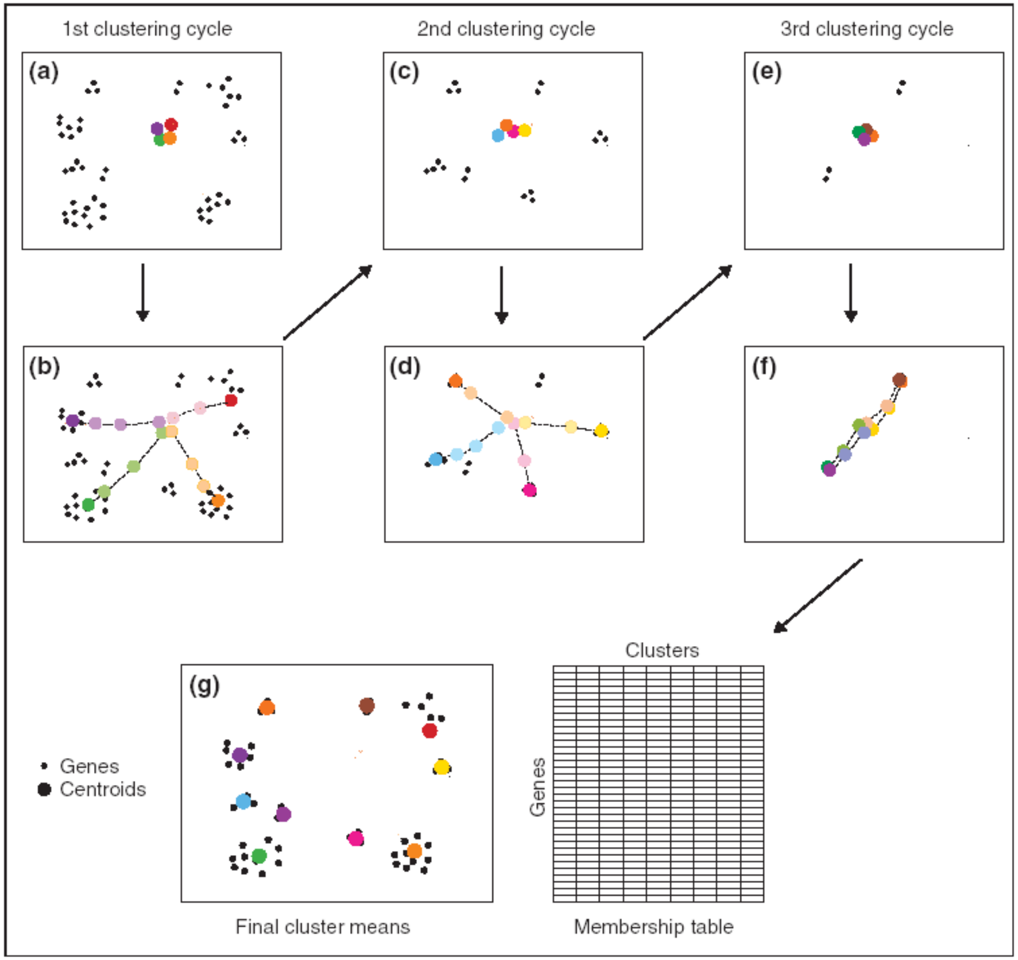
- Remove centroids with Pearson correlation > 0.9
- Allows selecting a large initial number of clusters, since duplicates will be removed



Repeat 3 times

Output

1) Cluster means
2) Gene assignments to clusters



4. Analysis of Clustered Data

- Statistical Significance of Clusters
  - Gene Ontology/ KEGG databases

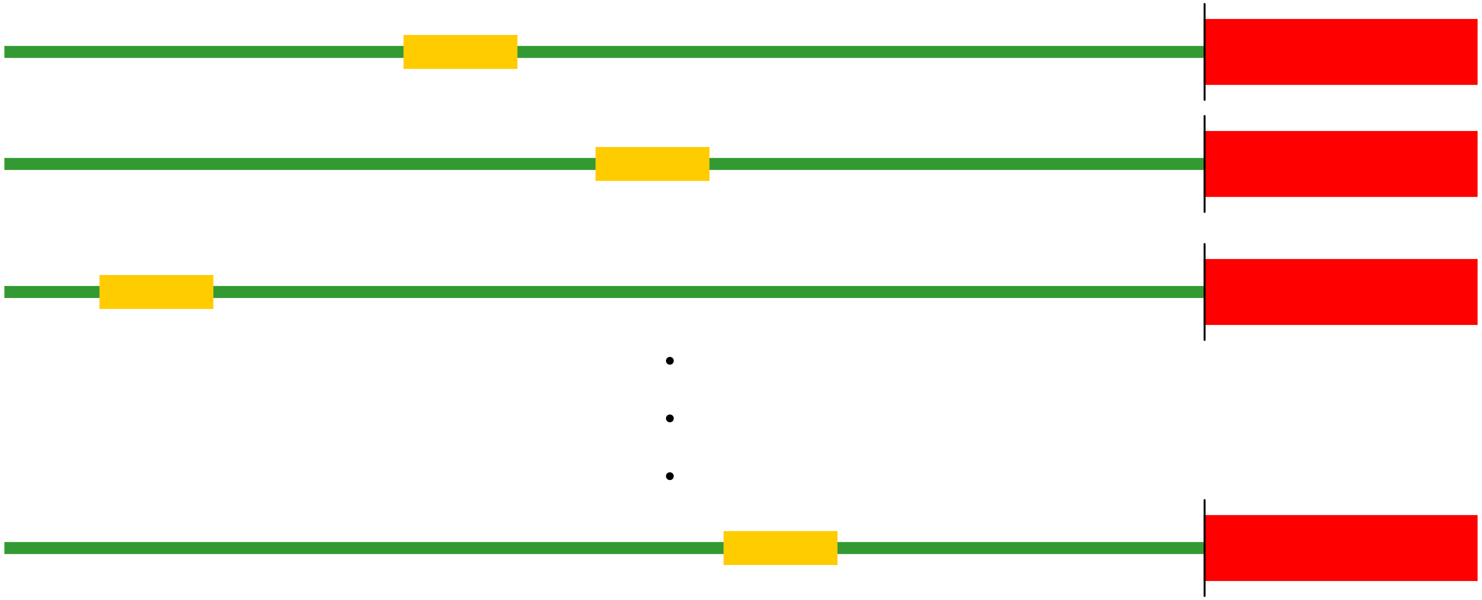
- Regulatory motifs responsible for common expression

- Regulatory Networks

- Experimental Verification

C. Finding Regulatory Motifs

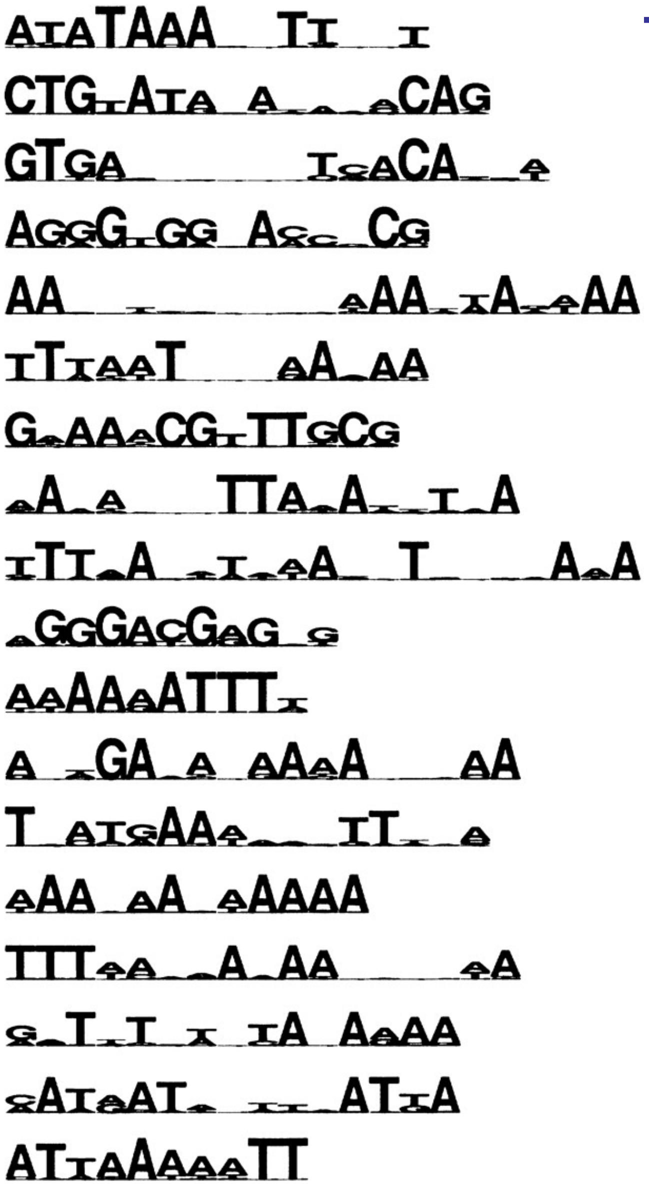
Finding Regulatory Motifs



Given a collection of genes with common expression,

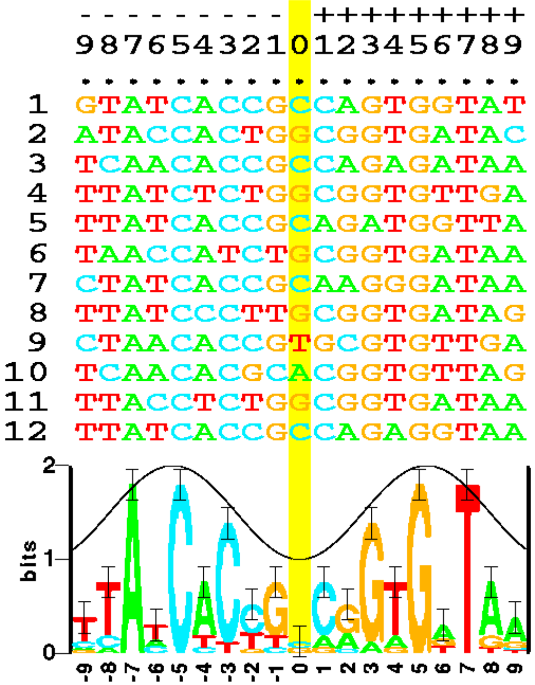
Find the TF-binding motif in common

Characteristics of Regulatory Motifs



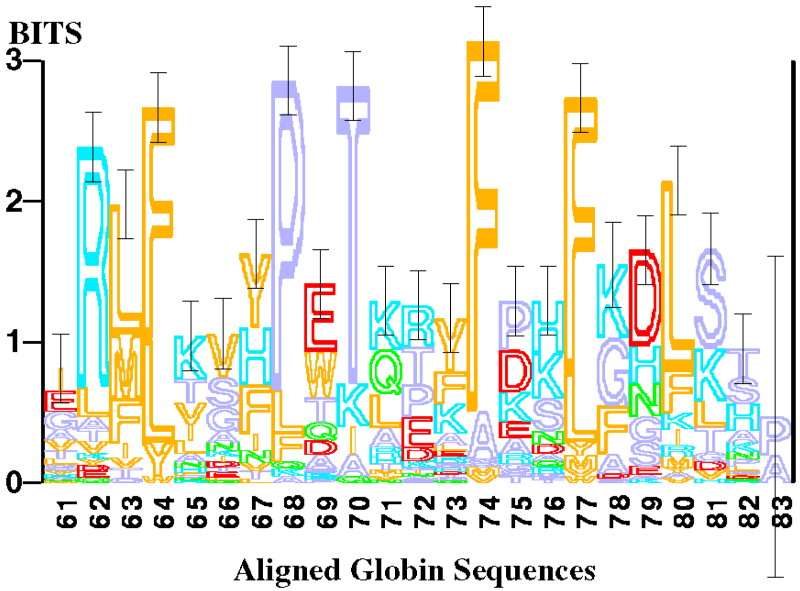
- Tiny
- Highly Variable
- ~Constant Size
  - Because a constant-size transcription factor binds
- Often repeated
- Low-complexity-ish

Sequence Logos



```
    - - - - - - - - - 0 + + + + + + + + +
    9876543210123456789
    . . . . . . . . . . . . . . . . . . .
 1  GTATCACCGCCAGTGGTAT
 2  ATACCACTGGCGGTGATAC
 3  TCAACACCGCCAGAGATAA
 4  TTATCTCTGGCGGTGTTGA
 5  TTATCACCGCAGATGGTTA
 6  TAACCATCTGCGGTGATAA
 7  CTATCACCGCAAGGGATAA
 8  TTATCCCTTGCGGTGATAG
 9  CTAACACCGTGCGTGTTGA
10  TCAACACGCACGGTGTTAG
11  TTACCTCTGGCGGTGATAA
12  TTATCACCGCCAGAGGTAA
```

12 Lambda cI and cro binding sites



Aligned Globin Sequences

- Information at pos'n I, H(i) $= - \Sigma_{\{\text{letter a}\}}$ Prob(a, i) $\log_2$ Prob(a, i)
- Height of x at pos'n i, L(a, i) = Prob(a, i) (2 – H(i))
  - Examples:
    - Prob(A, i) = 1; H(i) = 0; L(A, i) = 2
    - A: ½; C: ¼; G: ¼; H(i) = 1.5; L(A, i) = ¼; L(not T, i) = ¼

Problem Definition

Given a collection of promoter sequences $s_1,\ldots, s_N$ of genes with common expression

Probabilistic

Motif: M_{ij}; $1 \le i \le W$

$1 \le j \le 4$

M_{ij} = Prob[letter j, pos i]

Find best M, and positions $p_1,\ldots, p_N$ in sequences

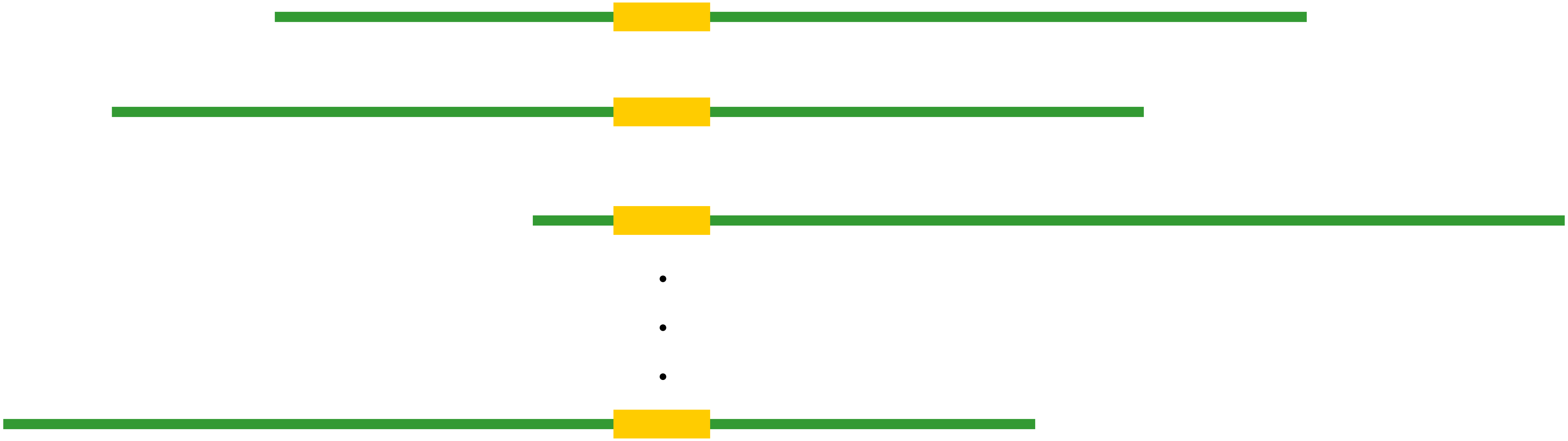
Combinatorial

Motif M: $m_1\ldots m_W$

Some of the m_i's blank

Find M that occurs in all s_i with $\le$ k differences

Essentially a Multiple Local Alignment



- Find “best” multiple local alignment

Alignment score defined differently in probabilistic/combinatorial cases

Algorithms

- Probabilistic

  1. Expectation Maximization:
     MEME
  2. Gibbs Sampling:
     AlignACE, BioProspector

- Exhaustive
  CONSENSUS, TEIRESIAS, SP-STAR, MDscan

Discrete Approaches to Motif Finding

Discrete Formulations

Given sequences $S = \{x^1, \ldots, x^n\}$

- A motif W is a consensus string $w_1 \ldots w_K$

- Find motif W^* with “best” match to $x^1, \ldots, x^n$

Definition of “best”:

$d(W, x^i)$ = min hamming dist. between W and a word in x^i

$d(W, S) = \Sigma_i \, d(W, x^i)$