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
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 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 P₁ and P₂ 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/sequence.html
Other Motifs in DNA Sequences: Human Splice Junctions

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

Gene-Specific TF Binding Sites

Basal TF Binding Sites

CAT Box

TATA Box

Gene

upstream region

coding region
Prokaryotic Gene Characteristics

### DNA PATTERNS IN THE *E. coli* lexA GENE

<table>
<thead>
<tr>
<th>GENE SEQUENCE</th>
<th>PATTERN</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATTGCTGATTGTTAATTCCTCGGTATTGTTGATGCTCTGTGCTCAGTTATAGTT</td>
<td>CTGHRNHNNRNRNRCAG</td>
</tr>
<tr>
<td>GCCAACTTATGCTTTGCGGCTTCGACATCAGCACTGTTCA</td>
<td>TTGACA</td>
</tr>
<tr>
<td>TATAATTCTATTCTGAC</td>
<td>TTGACA</td>
</tr>
<tr>
<td>GGGGG Ribosomal binding site</td>
<td>TATTTT</td>
</tr>
<tr>
<td>AUG</td>
<td>AUG</td>
</tr>
<tr>
<td>OPEN READING FRAME</td>
<td></td>
</tr>
</tbody>
</table>

Shown are matches to approximate consensus binding sites for LexA repressor (CTGHRNRNMNRRNCAG), the -10 and -35 promoter regions relative to the start of the mRNA (TTGACA and TATATT), the ribosomal binding site on the mRNA (GGAGG), and the open reading frame (AUG...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 TGTCTT is present. The CRE that binds factor C/EBP is also shown. (B) The 2300-bp promoter of the developmentally regulated gene enha16 of the sea urchin (Bloouri 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 Bloouri and Davidson 2002 ©2002 Elsevier).
Single Gene Activation

Transcription Factor

TF binding site

Gene
Multiple Gene Activation

Co-regulated genes

- TF
- TF binding site
- Gene

Transcription Factor
Transcription Regulation

**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
**Goal**: Find $\theta$, $Z$ that maximize $\Pr(X, Z \mid \theta)$

**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 ⇒
  - 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.
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

<table>
<thead>
<tr>
<th>Gene</th>
<th>Expression Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene1</td>
<td></td>
</tr>
<tr>
<td>Gene2</td>
<td></td>
</tr>
<tr>
<td>Gene3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
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

Sample

Treated Sample(t1)  Expt 1
Treated Sample(t2)  Expt 2
Treated Sample(t3)  Expt 3
...
Treated Sample(tn)  Expt n
2-color DNA microarray

Treated → mRNA → Cy5 Probe → Data extraction → Scanning → simultaneous hybridization
Control → mRNA → Cy3 Probe

Normalization
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 $\mathbb{R}^d$, 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
Example from Andrew Moore’s tutorial on Clustering.
Start

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!
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)

K-means continues ...

K-means terminates

Start

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.
The following slides come from a series of talks by Rafael Irizarry from Johns Hopkins.

Much of the material can be found in detail in the following papers from [http://www.biostat.jhsph.edu/~ririzarr/papers/]


Inference Process

From Talk by Irizzary
Affymetrix Genechip Design

Reference sequence

...TGTGATGGTGATGTGGTCAGAAGGCCTCCGATGCGCCGATTGAGAAT...

GTACTACCCAGTCTTCGGAGGCTA
GTACTACCCAGTGTTCGGAGGCTA

Perfectmatch Mismatch

NSB & SB
NSB

From Talk by Irizzary
Workflow: Analyzing Affy data

1. Raw data (.DAT files)
2. Probe intensities (.CEL files)
3. Expression measures (tables)
4. Rank (list)
5. Candidate genes (short list)

- Image analysis
- Pre-processing normalization
- Statistical test
- Choose filter
  - Significance level

From Talk by Irizzary
Affy Files

- **DAT** file: image file, about 10 million pixels, 30-50 MB
- **CEL** file: cell intensity file with probe level PM and MM values
- **CDF** file: chip description file describing which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs)

From Talk by Irizzary
Each probe cell: 10 X 10 pixels
Gridding estimates location of probe cell centers
Signal is computed by
  - Ignoring outer 36 pixels leaving a 8 X 8 pixel area
  - Taking the 75 percentile of the signal from the 8 X 8 pixel area
Background signal is computed as the average of the lowest 2% probe cell values, which is then subtracted from the individual signals

From Talk by Irizzary
Standard Normalization Procedure

- Log-transform the data
- Ensure that the average intensity and the standard deviation are the same across all arrays.
- This requires the choice of a baseline array, which may or may not be obvious.
Analyzing Affy data

- **MAS 4.0**
  - Works with PM-MM
  - Negative values result very often
  - Very noisy for low expressed genes
  - Averages without log-transformation

- **dChip [Li & Wong, PNAS 98(1):31-36]**
  - Accounts for probe effect
  - Uses non-linear normalization
  - Multi-chip analysis reveals outliers

- **MAS 5.0**
  - Improves on problems with MAS 4.0

From Talk by Irizzary
Why you use log-transforms?

From Talk by Irizzary
Problem with using (transformed) PM-MM

Sometimes MM is larger than PM!

From Talk by Irizzary
Bimodality for large expression values

From Talk by Irizzary
MAS 5.0 is Affymetrix software for microarray data analysis.
Ad hoc background procedure used
For summarization, they use:
- Signal = TukeyBiweight{\log(\text{PM}_j - \text{MM}_j^*)}
- Tukey Biweight: \( B(x) = (1 - (x/c)^2)^2 \), if \( x < c \)
  \[ = 0 \] otherwise
Ad hoc scale normalization used

From Talk by Irizzary & PhD thesis by Astrand
2 replicate arrays

Expression from corresponding probes are highly correlated

Expression not correlated when probes randomly partitioned

Correlation is higher than 0.99

Correlation drops to 0.55

From Talk by Irizzary
We have to deal with variations!
MvA Plots

A = \{ \log_2(\text{expression 2}) + \log_2(\text{expression 1}) \} / 2

M = \log_2(\text{expression 2} / \text{expression 1})
Spike-in Experiment

- Replicate RNA samples were hybridized to various arrays
- Some probe sets were spiked in at different concentrations across the different arrays
- Goal was to see if these spiked probe sets “stood out” as differentially expressed
Analyzing Spike-in data with MAS 5.0
Robust Multiarray normalization (RMA)

- **Background correction** separately for each array
  - Find \( E\{\text{Sig} \mid \text{Sig+Bgd} = \text{PM}\} \)
  - Bgd is normal and Sig is exponential

- Uses **quantile normalization** to achieve “identical empirical distributions of intensities” on all arrays

- **Summarization**: Performed separately for each probe set by fitting probe level additive model

- Uses **median polish** algorithm to robustly estimate expression on a specific chip

- Also see **GCRMA** [Wu, Irizzary et al., 2004]

From Talk by Irizzary & PhD thesis by Astrand
Analyzing Spike-in data with RMA
MvA and q-q plots

From Talk by Irizzary

MAS 4.0

MAS 5.0

07/01/09 Q'BIC Bioinformatics
MvA and q-q Plots

From Talk by Irizzary

07/01/09

Q'BIC Bioinformatics
Before and after quantile normalization

Fig. 2. 10 pairwise $M$ versus $A$ plots using liver (at concentration 10) dilution series data for unadjusted data.

Fig. 3. 10 pairwise $M$ versus $A$ plots using liver (at concentration 10) dilution series data after quantile normalization.
Bioconductor

- **Bioconductor** is an open source and open development software project for the analysis of biomedical and genomic data.
- World-wide project started in 2001
- **R** and the **R package system** are used to design and distribute software
- Commercial version of Bioconductor software called **ArrayAnalyzer**

From Talk by Irizzary
R: A Statistical Programming Language

- Try the tutorial at: [http://www.cyclismo.org/tutorial/R/]
- Also at: [http://www.math.ilstu.edu/dhkim/Rstuff/Rtutor.html]