CAP 5510: Introduction to Bioinformatics

Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS07.html
How to Represent Patterns

- Consensus sequence
- Alignments
- LOGO format
- Frequency Matrices
- Weight Matrices (Profiles, PSSMs, PWMs)
Pattern Representations

- Consensus sequences

Based on Pribnow, 1975:

- TACGAT
- TATAAT
- TATAAT
- GATACT
- TATGAT
- TATGTT

Consensus needs alignment.

- TATRNT Consensus w/ IUPAC
- TATAAT Multi-level Consensus
- G CGC T
Pattern Representations

- Consensus sequences
- **Weight Matrices (Profiles, PSSMs)**
  - Frequency Counts
  - Relative Frequency Measures
  - Normalized Measures
  - Log-transformed Measures
  - Information content
  - “Logo” technique
  - HMMs
Pattern Representation: Weight Matrix

Alignment

Consensus

Frequencies

Profile/ PSSM/PWM

Scoring a sequence against a profile

Visualizing a profile

[Wasserman, Sandelin, Nat Genet, 2004]
Formulae

- Prob of char $b$ in position $i$:
  \[ p(b, i) = \frac{f_{b,i}}{N} \]

- Corrected prob:
  \[ P(b, i) = \frac{f_{b,i} + s(b)}{N + \sum_{a \in A} s(a)} \]

- Weight matrix entry:

- Information content of position of $i$:
  \[ W_{b,i} = \log_2 \frac{P(b, i)}{BP(b)} \]

- 

\[ D_i = 2 + \sum_b P(b, i) \log_2 P(b, i) \]

[Wasserman, Sandelin, Nat Genet, 2004]
Probability of finding a sequence $w$ in some position of a DNA/protein sequence (assuming independence at each position)

$Pr(w_i) = BP(b) \ [\text{Background Frequency}]$

$$Pr(w) = \prod_{i=1}^{m} Pr(w_i)$$
 Statistical Evaluation

- Z-score of a motif with a certain frequency:

\[ z(w) = \frac{Obs(w) - Exp(w)}{\sqrt{Var(w)}} \]

- Information Content or Relative Entropy of an alignment or profile:

\[ IC(M) = \sum_{i=1}^{4} \sum_{j=1}^{m} m_{i,j} \log \frac{m_{i,j}}{b_i} \]

- Maximum a Posteriori (MAP) Score:

\[ MAP(M) = - \sum_{i=1}^{4} \sum_{j=1}^{m} n_{i,j} \log \frac{m_{i,j}}{b_i} \]

- Model Vs Background Score:

\[ L(w) = \frac{Pr(w | M)}{Pr(w | Bg)} = \prod_{j=1}^{m} \frac{m_{i,j}}{b_i} \]
Motifs are combinations of secondary structures in proteins with a specific structure and a specific function. They are also called super-secondary structures.

Examples: Helix-Turn-Helix, Zinc-finger, Homeobox domain, Hairpin-beta motif, Calcium-binding motif, Beta-alpha-beta motif, Coiled-coil motifs.

Several motifs may combine to form domains.
- Serine proteinase domain, Kringle domain, calcium-binding domain, homeobox domain.
Motif Detection

- **Profile Method**
  - If many examples of the motif are known, then
    - **Training**: build a Profile and compute a threshold
    - **Testing**: score against profile

- **Combinatorial Pattern Discovery Methods**
  - **Gibbs Sampling**
  - **Expectation Method**
  - **HMM**
How to evaluate these methods?

- Calculate TP, FP, TN, FN
- Compute sensitivity fraction of known sites predicted, specificity, and more.
  - Sensitivity = $\frac{TP}{TP+FN}$
  - Specificity = $\frac{TN}{TN+FN}$
  - Positive Predictive Value = $\frac{TP}{TP+FP}$
  - Performance Coefficient = $\frac{TP}{TP+FN+FP}$
  - Correlation Coefficient =
Motif Detection Problem

**Input:** Set, $S$, of known (aligned) examples of a motif $M$, A new protein sequence, $P$.

**Output:** Does $P$ have a copy of the motif $M$?

**Example:** Zinc Finger Motif

```
...YKCGLCERSFVEKSAFLSRHORVHKN...
```

**Input:** Database, $D$, of known protein sequences, A new protein sequence, $P$.

**Output:** What interesting patterns from $D$ are present in $P$?
Supervised Pattern Discovery

- **Input:** Alignment of known motifs, and Query sequence

- **Output:** Is the query sequence a motif?

  - Profile Method [Gribskov et al., 1996]
    - Build a profile from the alignment and score query sequence against the profile to decide if it “fits the profile”.
    - Need to pick a threshold score.

  - Enumerative/Combinatorial Methods
Profile HMMs

PROFILE METHOD, [M. Gribskov et al., ’90]

<table>
<thead>
<tr>
<th>Location in Seq</th>
<th>Sequence</th>
<th>Protein Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G V S A S A</td>
<td>Ka RbtR</td>
</tr>
<tr>
<td></td>
<td>G V S E M T</td>
<td>Ec DeoR</td>
</tr>
<tr>
<td></td>
<td>G V S P G T</td>
<td>Ec RpoD</td>
</tr>
<tr>
<td></td>
<td>G A G I A T</td>
<td>Ec TrpR</td>
</tr>
<tr>
<td></td>
<td>G C S R E T</td>
<td>Ec CAP</td>
</tr>
<tr>
<td></td>
<td>C L S P S R</td>
<td>Ec AraC</td>
</tr>
<tr>
<td></td>
<td>C L S P S R</td>
<td>St AraC</td>
</tr>
<tr>
<td></td>
<td>G V N K E T</td>
<td>Br MerR</td>
</tr>
</tbody>
</table>

Graphical representation of states 1 to 6 with transitions between them.
Combinatorial Method: GYM

Pattern Generation:

Aligned Motif Examples

Pattern Generator

Motif Detection:

Pattern Dictionary

New Protein Sequence

Motif Detector

Detection Results

[Narasimhan, Bu, Wang, Xu, Yang, Mathee, J Comput Biol, 2002]
Helix-Turn-Helix Motifs

- Structure
  - 3-helix complex
  - Length: 22 amino acids
  - Turn angle

- Function
  - Gene regulation by binding to DNA
Figure 7.10 The helix-turn-helix motif in lambda Cro bound to DNA (orange) with the two recognition helices (red) of the Cro dimer sitting in the major groove of DNA. The binding model, suggested by Brian Matthews, is shown schematically in (a) with connected circles for the Cg positions as they were model built into regular B-DNA. A schematic diagram of the Cro dimer is shown in (b) with different colors for the two subunits. A schematic space-filling model of the dimer of Cro bound to a bent B-DNA molecule is shown in (c). The sugar-phosphate backbone of DNA is red, and the bases are yellow. Protein atoms are colored red, blue, green, and white. [(a) Adapted from D. Ohlendorf et al., J. Mol. Evol. 19: 113, 1983. (c) Courtesy of Brian Matthews.]
## HTH Motifs: Examples

<table>
<thead>
<tr>
<th>Loc</th>
<th>Protein Name</th>
<th>Helix 2</th>
<th>Turn</th>
<th>Helix 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Cro</td>
<td>F G Q E K T A K D L G V Y Q S A I N K A I H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>434 Cro</td>
<td>M T Q T E L A T K A G V K Q Q S I Q L I E A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>P22 Cro</td>
<td>G T Q R A V A K A L G I S D A A V S Q W K E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Rep</td>
<td>L S Q E S V A D K M G M G Q S G V G A L F N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>434 Rep</td>
<td>L N Q A E L A Q K V G T T Q Q S I E Q L E N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>P22 Rep</td>
<td>I R Q A A L G K M V G V S N V A I S Q W E R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>CII</td>
<td>L G T E K T A E A V G V D K S Q I S R W K R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>LacR</td>
<td>V T L Y D V A E Y A G V S Y Q T V S R V V N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>167</td>
<td>CAP</td>
<td>I T R Q E I G Q I V G C S R E T V G R I L K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>TrpR</td>
<td>M S Q R E L K N E L G A G I A T I T R G S N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>BlaA Pv</td>
<td>L N F T K A A L E L Y V T Q G A V S Q Q V R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>TrpI Ps</td>
<td>N S V S Q A A E Q L H V T H G A V S R Q L K</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Combinatorial Method: GYM

- Combinations of residues in specific locations (may not be contiguous) contribute towards stabilizing a structure.
- Some reinforcing combinations are relatively rare.
- GYM algorithm is inspired by the A Priori algorithm [Agrawal et al., 1996]

[Narasimhan, Bu, Wang, Xu, Yang, Mathee, J Comput Biol, 2002]
### Patterns

<table>
<thead>
<tr>
<th>Loc</th>
<th>Protein Name</th>
<th>Helix 2</th>
<th>Turn</th>
<th>Helix 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Cro</td>
<td>F G Q E K T A K D L G V Y Q S A I N K A I H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>434 Cro</td>
<td>M T Q T E L A T K A G V K Q Q S I Q L I E A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>P22 Cro</td>
<td>G T Q R A V A K A L G I S D A A V S Q W K E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Rep</td>
<td>L S Q E S V A D K M G M G Q S G V G A L F N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>434 Rep</td>
<td>L N Q A E L A Q K V G T T Q Q S I E Q L E N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>P22 Rep</td>
<td>I R Q A A L G K M V G V S N V A I S Q W E R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>CII</td>
<td>L G T E K T A E A V G V D K S Q I S R W K R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>LacR</td>
<td>V T L Y D V A E Y A G V S Y Q T V S R V V N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>167</td>
<td>CAP</td>
<td>I T R Q E I G Q I V G C S R E T V G R I L K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>TrpR</td>
<td>M S Q R E L K N E L G A G I A T I T R G S N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>BlaA Pv</td>
<td>L N F T K A A L E L Y V T Q G A V S Q Q V R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>TrpI Ps</td>
<td>N S V S Q A A E Q L H V T H G A V S R Q L K</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Q1 G9 N20**
- **A5 G9 V10 I15**
Pattern Mining Algorithm

Algorithm Pattern-Mining
Input: Motif length \( m \), support threshold \( T \), list of aligned motifs \( M \).
Output: Dictionary \( L \) of frequent patterns.

1. \( L_1 := \) All frequent patterns of length 1
2. \textbf{for} \( i = 2 \) \textbf{to} \( m \) \textbf{do}
3. \( C_i := \text{Candidates}(L_{i-1}) \)
4. \( L_i := \) Frequent candidates from \( C_i \)
5. \textbf{if} \( |L_i| \leq 1 \) \textbf{then}
6. \textbf{return} \( L \) as the union of all \( L_j, j \leq i \).
### Candidates Function

<table>
<thead>
<tr>
<th>L₃</th>
<th>C₄</th>
<th>L₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>G₁, V₂, S₃</td>
<td>G₁, V₂, S₃, T₆</td>
<td>G₁, V₂, S₃, T₆</td>
</tr>
<tr>
<td>G₁, V₂, T₆</td>
<td>G₁, V₂, S₃, I₇</td>
<td>G₁, V₂, S₃, I₇</td>
</tr>
<tr>
<td>G₁, V₂, I₇</td>
<td>G₁, V₂, S₃, E₈</td>
<td>G₁, V₂, S₃, E₈</td>
</tr>
<tr>
<td>G₁, V₂, E₈</td>
<td>G₁, V₂, T₆, I₇</td>
<td>G₁, V₂, T₆, E₈</td>
</tr>
<tr>
<td>G₁, S₃, T₆</td>
<td>G₁, V₂, T₆, E₈</td>
<td>G₁, V₂, T₆, E₈</td>
</tr>
<tr>
<td>G₁, T₆, I₇</td>
<td>G₁, V₂, I₇, E₈</td>
<td>G₁, V₂, T₆, E₈</td>
</tr>
<tr>
<td>V₂, T₆, I₇</td>
<td>V₂, T₆, I₇, E₈</td>
<td>V₂, T₆, I₇, E₈</td>
</tr>
<tr>
<td>V₂, T₆, E₈</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Motif Detection Algorithm

Algorithm Motif-Detection

**Input**: Motif length m, threshold score T, pattern dictionary L, and input protein sequence P[1..n].

**Output**: Detected motif(s).

1. **for** each location i **do**
2. \[ S := \text{MatchScore}(P[i..i+m-1], L). \]
3. **if** \( (S > T) \) **then**
4. Report it as a possible motif
# Experimental Results: GYM 2.0

## Motif

<table>
<thead>
<tr>
<th>Protein Family</th>
<th>Number Tested</th>
<th>GYM = DE Agree</th>
<th>Number Annotated</th>
<th>GYM = Annot.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HTH Motif (22)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Master</td>
<td>88</td>
<td>88 (100 %)</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Sigma</td>
<td>314</td>
<td>284 + 23 (98 %)</td>
<td>96</td>
<td>82</td>
</tr>
<tr>
<td>Negates</td>
<td>93</td>
<td>86 (92 %)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LysR</td>
<td>130</td>
<td>127 (98 %)</td>
<td>95</td>
<td>93</td>
</tr>
<tr>
<td>AraC</td>
<td>68</td>
<td>57 (84 %)</td>
<td>41</td>
<td>34</td>
</tr>
<tr>
<td>Rreg</td>
<td>116</td>
<td>99 (85 %)</td>
<td>57</td>
<td>46</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>675</strong></td>
<td><strong>653 + 23 (94 %)</strong></td>
<td><strong>289</strong></td>
<td><strong>255 (88 %)</strong></td>
</tr>
</tbody>
</table>
Transcription Regulation

Basal TF Binding Sites

CAT Box

Gene-Specific TF Binding Sites

TATA Box

Gene

coding region

upstream region
Single Gene Activation
Multiple Gene Activation
Transcription Regulation

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
Predicting Motifs in Whole Genome

- **MEME**: EM algorithm [Bailey et al., 1994]
- **AlignACE**: Gibbs Sampling Approach [Hughes et al., 2000]
- **Consensus**: Greedy Algorithm Based [Hertz et al., 1990]
- **ANN-Spec**: Artificial Neural Network and a Gibbs sampling method [Workman et al., 2000]
- **YMF**: Enumerative search [Sinha et al., 2003]
- ...

2/26/08  CAP5510  32
Input: upstream sequences

\[ X = \{X_1, X_2, \ldots, X_n\} \]

Motif profile: \(4^k\) matrix \(\Pi (\Pi_{rp})\),

- \( r \in \{A, C, G, T\} \)
- \( 1 \leq p \leq k \)
- \( \Pi_{rp} = \Pr(\text{residue } r \text{ in position } p \text{ of motif}) \)

Background distribution:

- \( \Pi_{r0} = \Pr(\text{residue } r \text{ in background}) \)
EM Method: Hidden Information

- $Z = \{Z_{ij}\}$, where

  $$Z_{ij} = \begin{cases} 
  1, & \text{if motif instance starts at position } i \text{ of } X_j \\
  0, & \text{otherwise}
  \end{cases}$$

- Iterate over probabilistic models that could generate $X$ and $Z$, trying to converge on this solution.
Z-score of a motif with a certain frequency:

\[ z(w) = \frac{\text{Obs}(w) - \text{Exp}(w)}{\sqrt{\text{Var}(w)}} \]

Information Content or Relative Entropy of an alignment or profile:

\[ IC(M) = \sum_{i=1}^{4} \sum_{j=1}^{m} m_{i,j} \log \frac{m_{i,j}}{b_i} \]

Maximum a Posteriori (MAP) Score:

\[ MAP(M) = -\sum_{i=1}^{4} \sum_{j=1}^{m} n_{i,j} \log \frac{m_{i,j}}{b_i} \]

Model Vs Background Score:

\[ L(w) = \frac{\Pr(w | M)}{\Pr(w | Bg)} = \prod_{j=1}^{m} \frac{m_{i,j}}{b_i} \]
EM Algorithm

**Goal:** Find $\pi, Z$ that maximize $\Pr(X, Z | \pi)$

**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
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>1 GAATTCGTAATACGCACTGCTATTTATG GTTTG</td>
<td>CTGNNNNNNNNNNNNGAG</td>
</tr>
<tr>
<td>3 TT</td>
<td>TGGACA</td>
</tr>
<tr>
<td>42 CCCTAATGCGCTTTTAGGTATTACGTCAACATG</td>
<td>CTGNNNNNNNNNNNNGAG</td>
</tr>
<tr>
<td>81 TATAATACCATGGGGAAGCGGTTCATCAACCCGA</td>
<td>TATAT, &gt; mRNA start</td>
</tr>
<tr>
<td>−35</td>
<td>CTGNNNNNNNNNNNNGAG</td>
</tr>
<tr>
<td>+10 GGGGG Ribosomal binding site</td>
<td>GGAGG</td>
</tr>
<tr>
<td>121 GGCACACGCCCATGTAATACGTCAATCCCTAGTTGATGATC</td>
<td>AG</td>
</tr>
<tr>
<td>161 CCAGAAGCGAGATTCGT CTGCAG</td>
<td>OPEN READING FRAME</td>
</tr>
<tr>
<td>201 CGGGTGGTGCCGCATGGGAAGCGGTAAGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
<tr>
<td>241 TGCGGGCGACTGAGGAGTGACTGCAATACGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
<tr>
<td>281 CGGGGGAAGGCTGAGGACATGCAATACGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
<tr>
<td>321 TACGAGGCTGAGGCGGTAAGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
<tr>
<td>361 CATTGAGGCTGAGGCGGTAAGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
<tr>
<td>401 CGGGGTAAGGCATGCAATACGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
<tr>
<td>441 CGGGGTAAGGCATGCAATACGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
<tr>
<td>481 ACGGCGGTAAGGCATGCAATACGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
<tr>
<td>521 TACGAGGCTGAGGCGGTAAGTCGACGTAACCCGA</td>
<td>GAT</td>
</tr>
</tbody>
</table>

Shown are matches to approximate consensus binding sites for LexA repressor (CTGNNNNNNNNNNNNGAG), the −10 and −35 promoter regions relative to the start of the mRNA (TTGACA and TATAT), the ribosomal binding site on the mRNA (GGAGG), and the open reading frame (AGG...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.
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

<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
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?

RNA fragments with fluorescent tags from sample to be tested

Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow

RNA fragment hybridizes with DNA on GeneChip® array

Non-hybridized DNA

Hybridized DNA
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