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

```
[Pribnow, 1975]
TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT
------
TATAAT Consensus
```

```
TATRNT Consensus
       w/ IUPAC
```

```
TATAAT Multi-level
G CGC  Consensus
    T
```

Needs Alignment

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

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

Consensus

b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

Frequencies

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Profile/ PSSM/PWM

d Position weight matrix (PWM)

A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

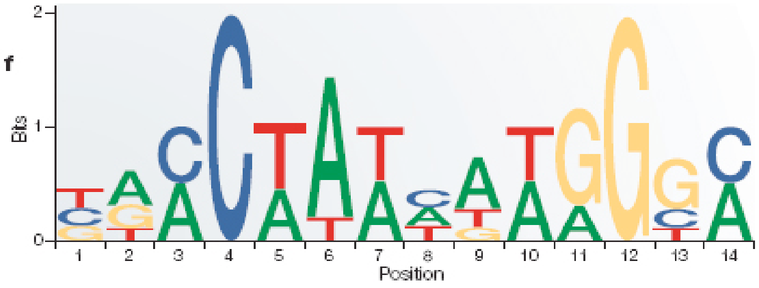
Scoring a sequence against a profile

e Site scoring

0.45	−0.66	0.79	1.68	0.45	−0.66	0.79	0.45	−0.66	0.79	0.00	1.68	−0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum

Visualizing a profile



[Wasserman, Sandelin, Nat Genet, 2004]

Formulae

- Prob of char b in position i:

$$p(b,i) = \frac{f_{b,i}}{N}$$

Frequency

# Sequences

- Corrected prob:

$$P(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{a \in \mathbf{A}} s(a)}$$

PseudoCount

- Weight matrix entry:

$$W_{b,i} = \log_2 \frac{P(b,i)}{BP(b)}$$

Background Frequency

- Information content of position of i:

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

[Wasserman, Sandelin, Nat Genet, 2004]

Statistical Evaluation Fundamentals

- Probability of finding a sequence w in some position of a DNA/protein sequence (assuming independence at each position)

- $\Pr(w_i)$ = BP(b) [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:

$$L(w) = \frac{\Pr(w \mid M)}{\Pr(w \mid Bg)} = \prod_{j=1}^{m} \frac{m_{i,j}}{b_i}$$

Pattern Discovery in Protein Sequences

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 = TP/(TP+FN)
  - Specificity = TN/(TN+FN)
  - Positive Predictive Value = TP/(TP+FP)
  - Performance Coefficient = 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

```
...YKCGLCERSFVEKSALSRHORVHKN...
     3  6           19  23
```

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

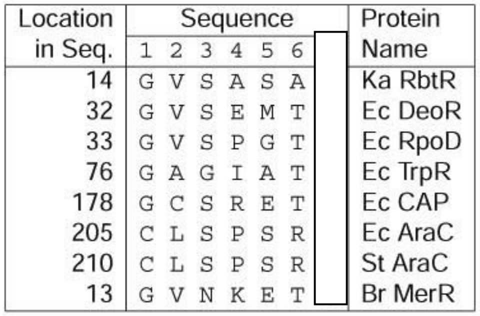
- Input: Alignment of known motifs, and
  Query sequence

  Output: Is the query sequence a motif?

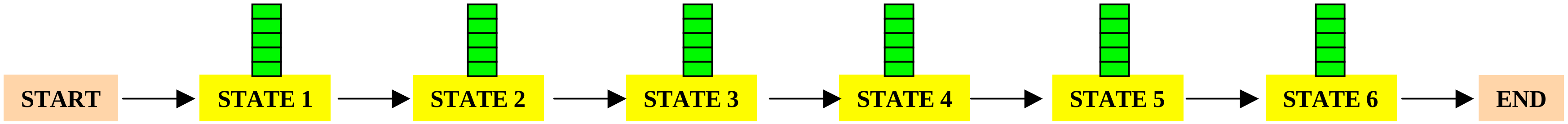
  - Profile Method [Gribskov et al., 1996]
    - Build a profile from the alignment and score query sequence against the profile to decide if it "fits the profile".
    - Need to pick a threshold score.
  - Enumerative/Combinatorial Methods

Profile HMMs

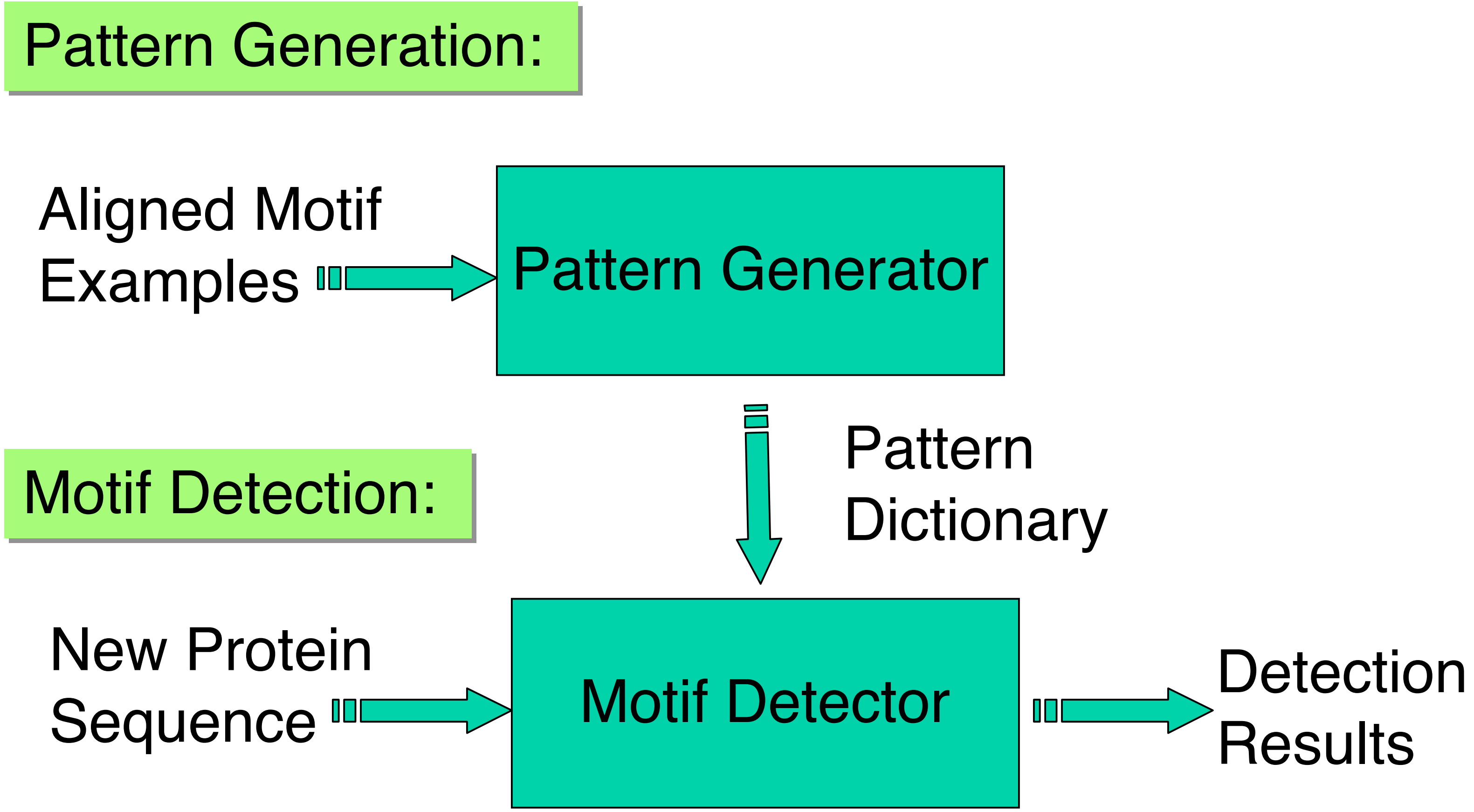
PROFILE METHOD, [M. Gribskov et al., '90]



Location in Seq.	Sequence 1	2	3	4	5	6	Protein Name
14	G	V	S	A	S	A	Ka RbtR
32	G	V	S	E	M	T	Ec DeoR
33	G	V	S	P	G	T	Ec RpoD
76	G	A	G	I	A	T	Ec TrpR
178	G	C	S	R	E	T	Ec CAP
205	C	L	S	P	S	R	Ec AraC
210	C	L	S	P	S	R	St AraC
13	G	V	N	K	E	T	Br MerR



Combinatorial Method: GYM



Pattern Generation:

Aligned Motif Examples

Pattern Generator

Pattern Dictionary

Motif Detection:

New Protein Sequence

Motif Detector

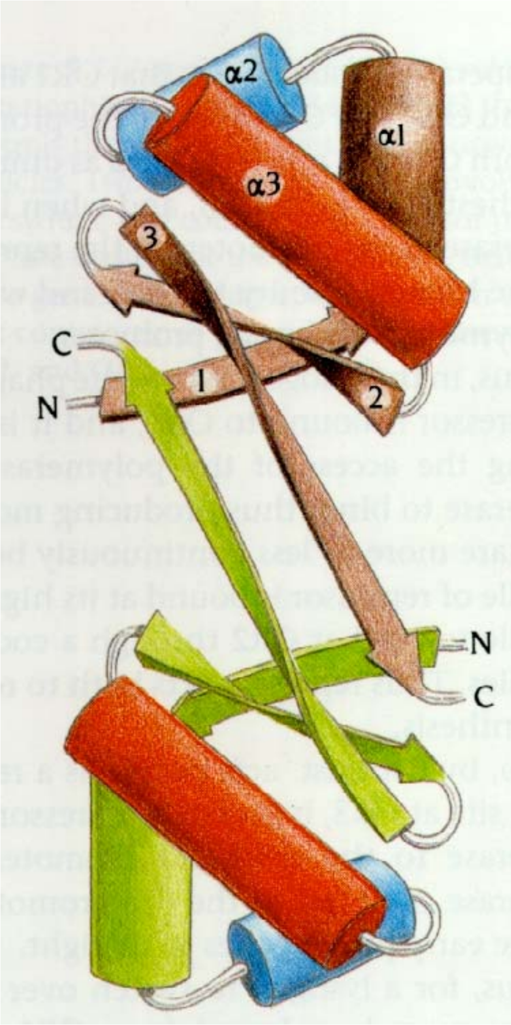
Detection Results

[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



Branden & Tooze

DNA Binding at HTH Motif

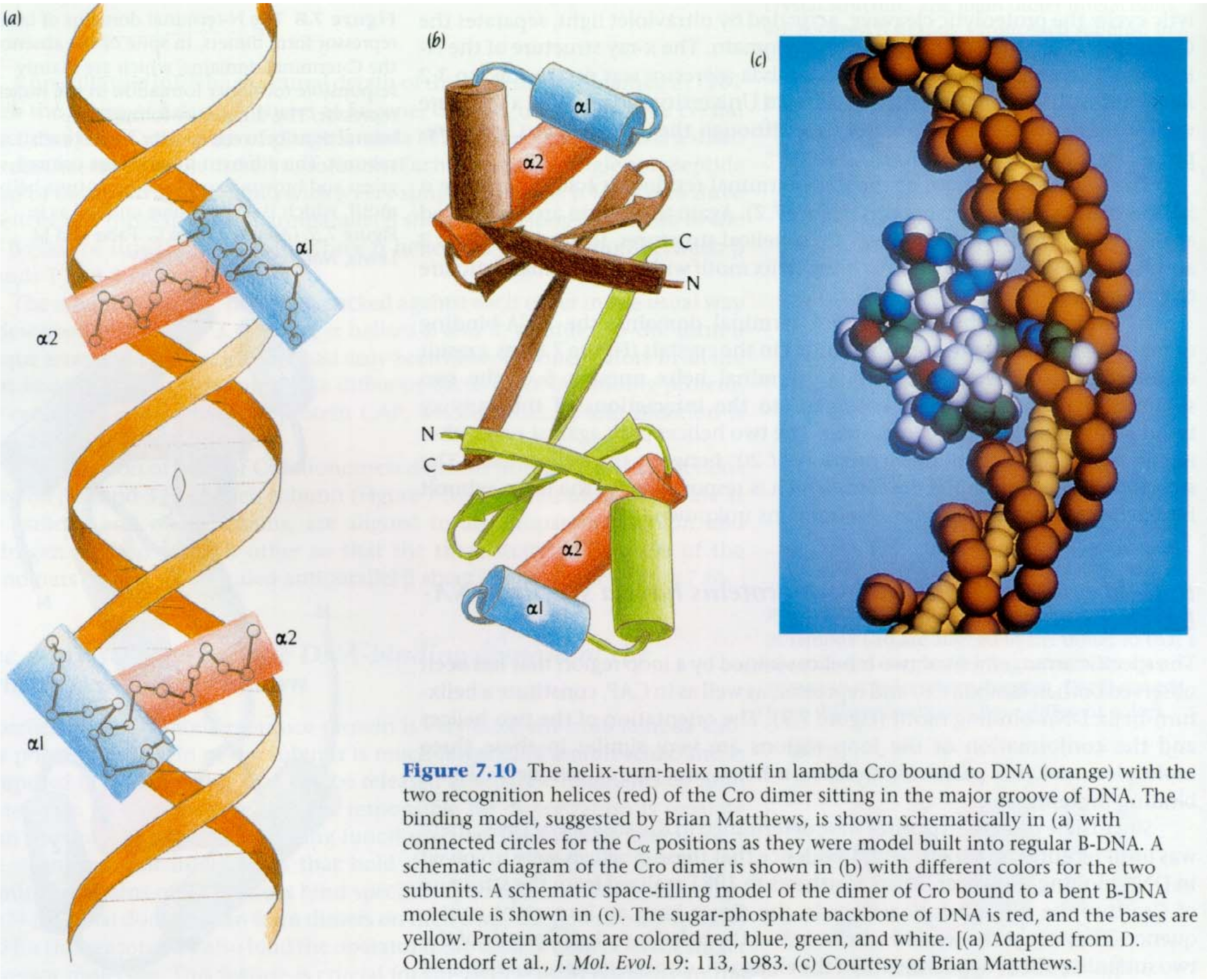


Figure 7.10 The helix-turn-helix motif in lambda Cro bound to DNA (orange) with the two recognition helices (red) of the Cro dimer sitting in the major groove of DNA. The binding model, suggested by Brian Matthews, is shown schematically in (a) with connected circles for the C_α positions as they were 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.]

Branden & Tooze

HTH Motifs: Examples

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

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 APriori algorithm [Agrawal et al., 1996]

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

Patterns

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

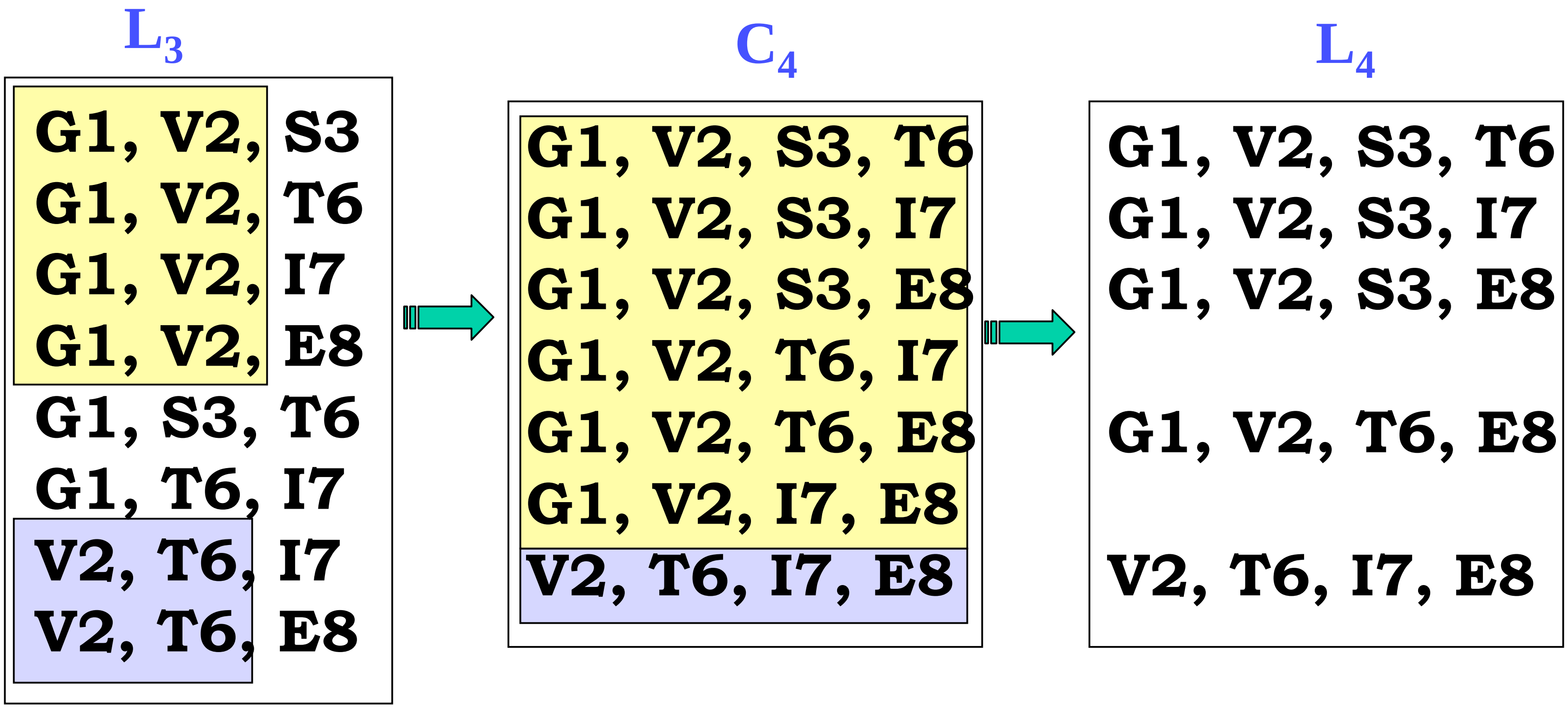
- ● **Q1 G9 N20**
- ● **A5 G9 V10 I15**

Pattern Mining Algorithm

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

```
$L_1$ := All frequent patterns of length 1
for i = 2 to m do
    $C_i$ := Candidates($L_{i-1}$)
    $L_i$ := Frequent candidates from $C_i$
    if ($|L_i| <= 1$) then
        return L as the union of all $L_j$ , $j <= i$.
```

Candidates Function



L_3

G1, V2, S3
G1, V2, T6
G1, V2, I7
G1, V2, E8
G1, S3, T6
G1, T6, I7
V2, T6, I7
V2, T6, E8

C_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8
G1, V2, T6, I7
G1, V2, T6, E8
G1, V2, I7, E8
V2, T6, I7, E8

L_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8
G1, V2, T6, E8
V2, T6, I7, E8

Motif Detection Algorithm

Algorithm Motif-Detection

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

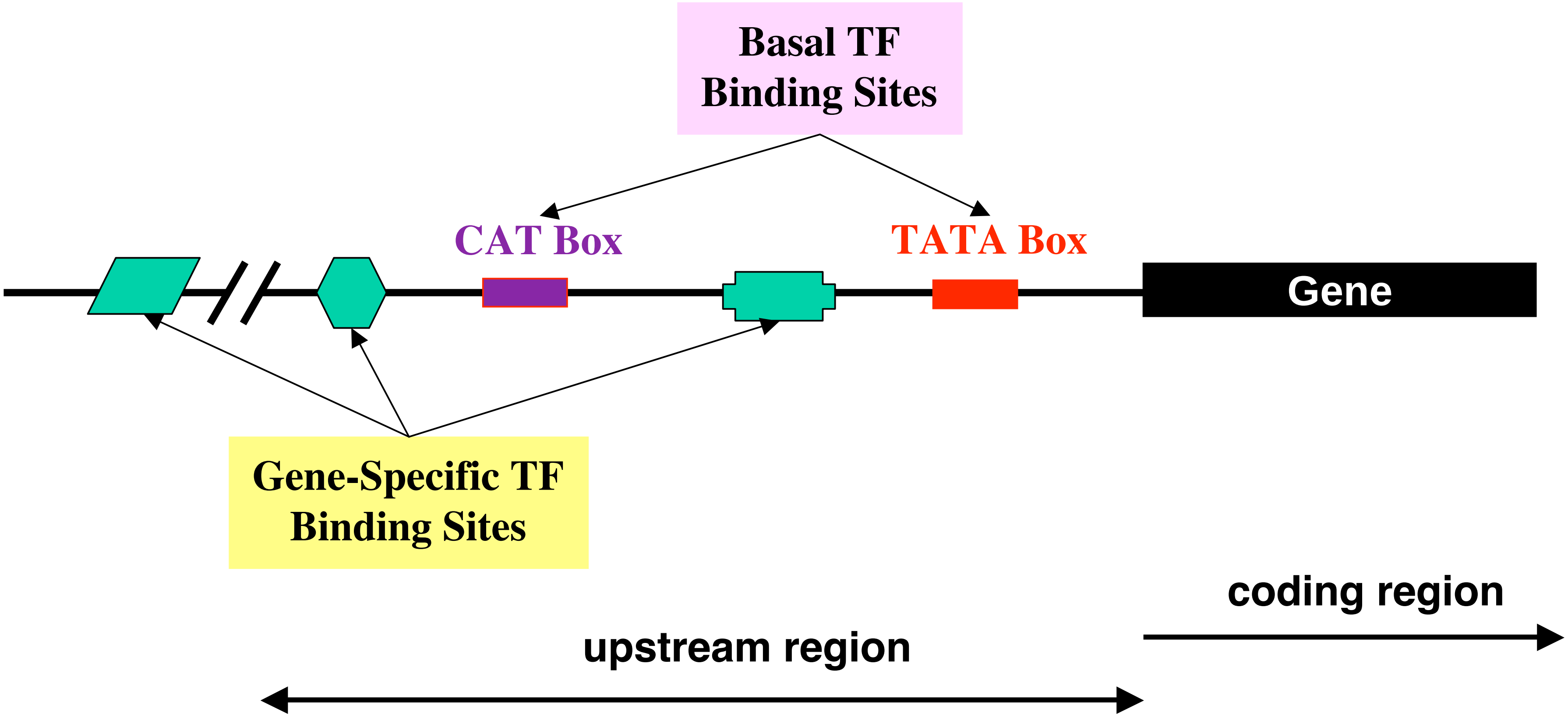
Output : Detected motif(s).

```
for each location i do
    S := MatchScore(P[i..i+m-1], L).
    if (S > T) then
        Report it as a possible motif
```

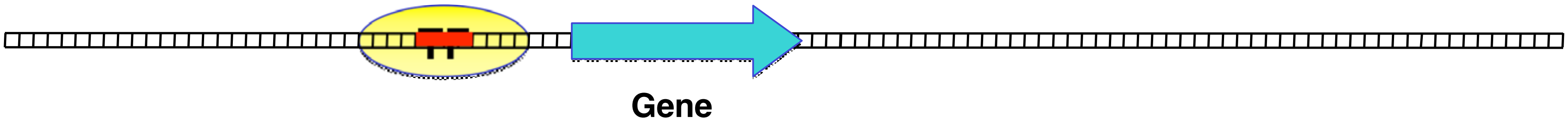
Experimental Results: **GYM 2.0**

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

Transcription Regulation



Single Gene Activation

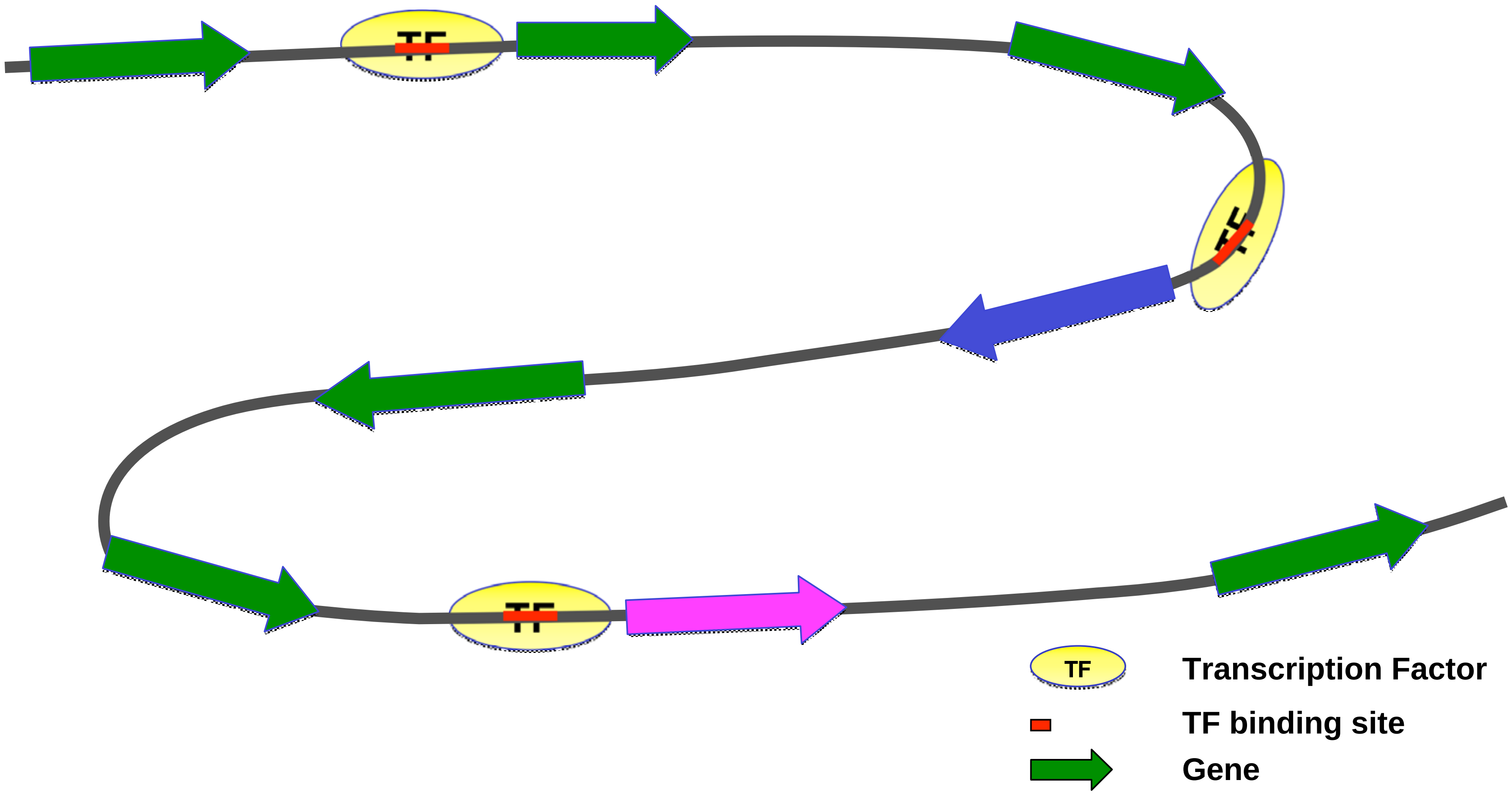


TF Transcription Factor

🟥 TF binding site

➡ Gene

Multiple Gene Activation

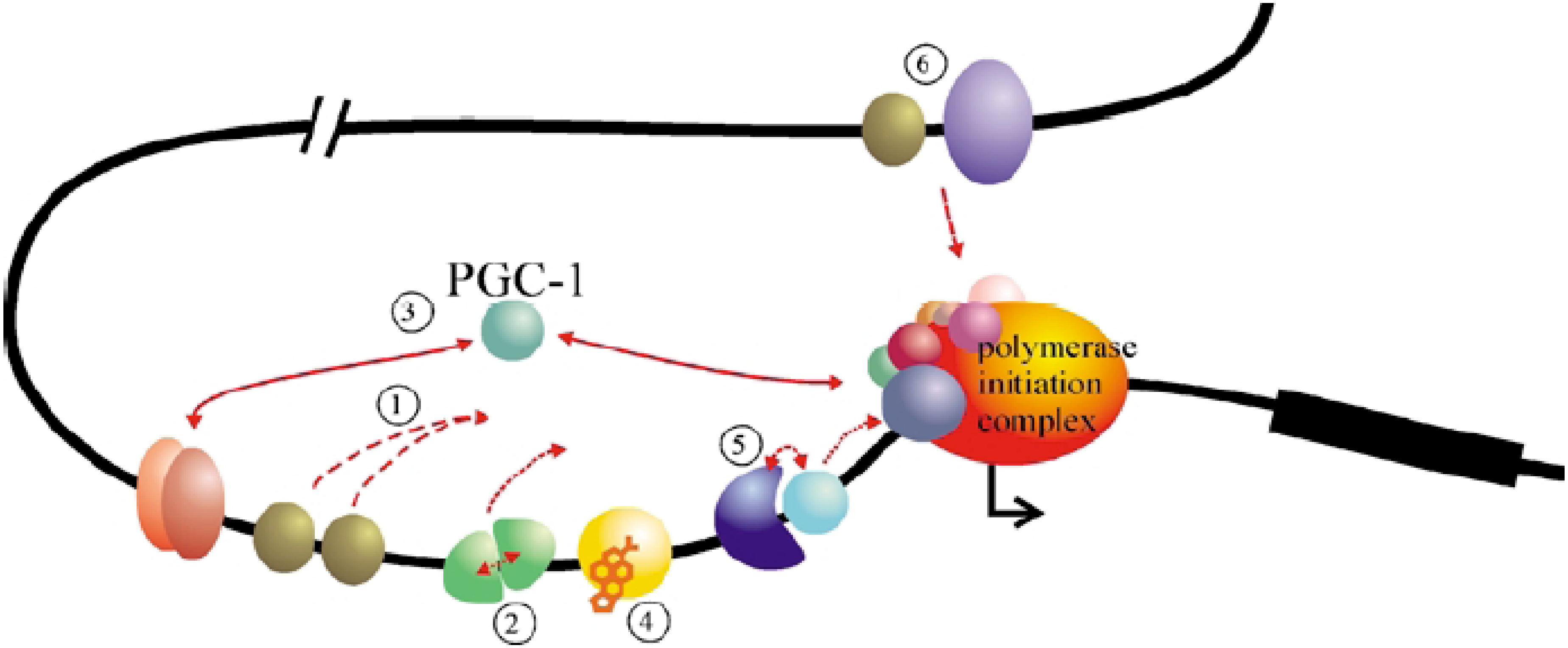


TF Transcription Factor

🟥 TF binding site

🟩 Gene

Transcription Regulation



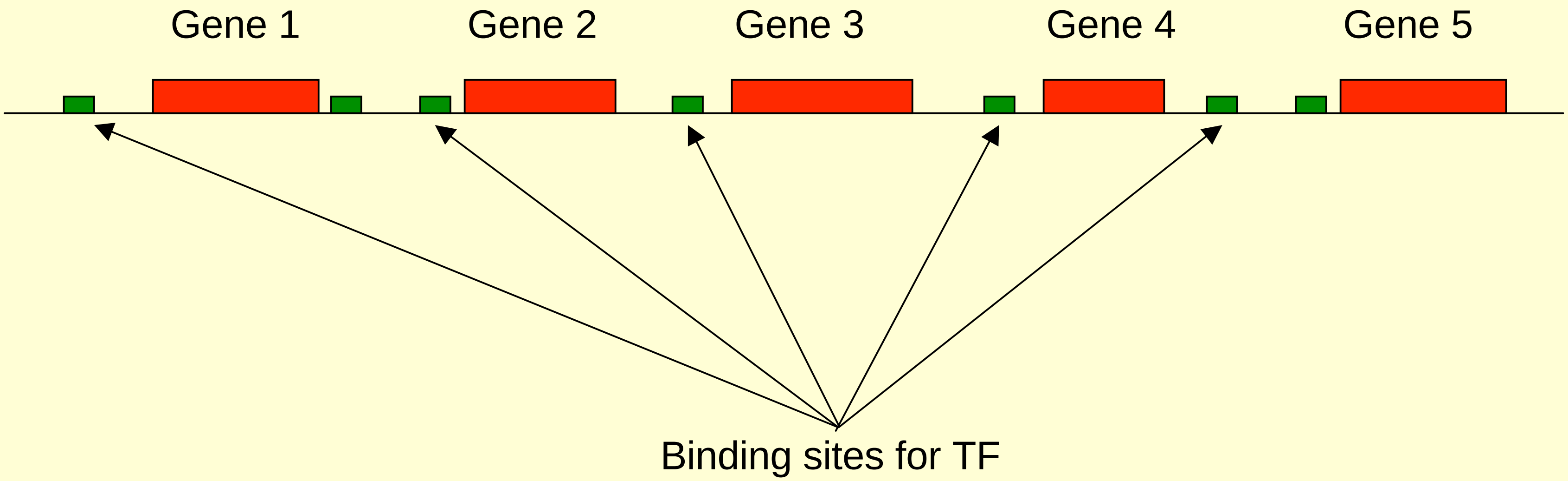
[Goffart *et al. Exp. Physiology* (2003)]

Motif-prediction: Whole genome

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

Motif-prediction: Whole genome

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]
- ❑ …

EM Method: Model Parameters

- ❑ Input: upstream sequences
  - ➢ $X = \{X_1, X_2, \dots, X_n\}$,
- ❑ Motif profile: 4°k matrix ¶›= (¶»$_{rp}$),
  - ● $r \in \{A,C,G,T\}$
  - ● $1 \le p \le k$
  - ● ¶»$_{rp}$ = Pr(residue r in position p of motif)
- ❑ Background distribution:
  - ➢ ¶»$_{r0}$ = Pr(residue r in background)

EM Method: Hidden Information

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

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

- Iterate over probabilistic models that could generate X and Z, trying to converge on this solution

BIORG
BioInformatics Research Group

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:

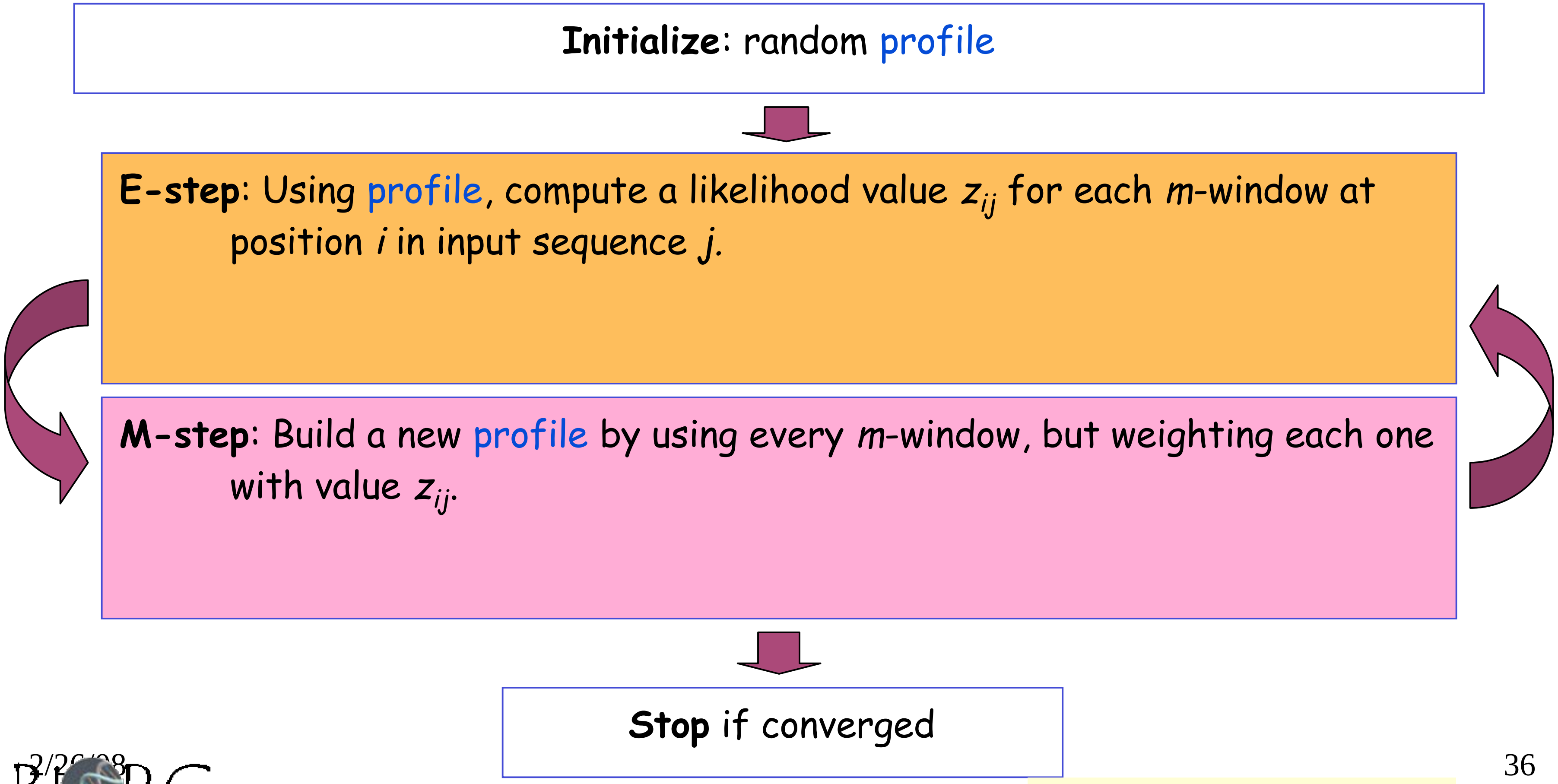
$$L(w) = \frac{\Pr(w \mid M)}{\Pr(w \mid Bg)} = \prod_{j=1}^{m} \frac{m_{i,j}}{b_i}$$

Counts

Frequencies

EM Algorithm

Goal: Find ¶¸»Z that maximize Pr (X, Z | ¶»)



Initialize: random profile

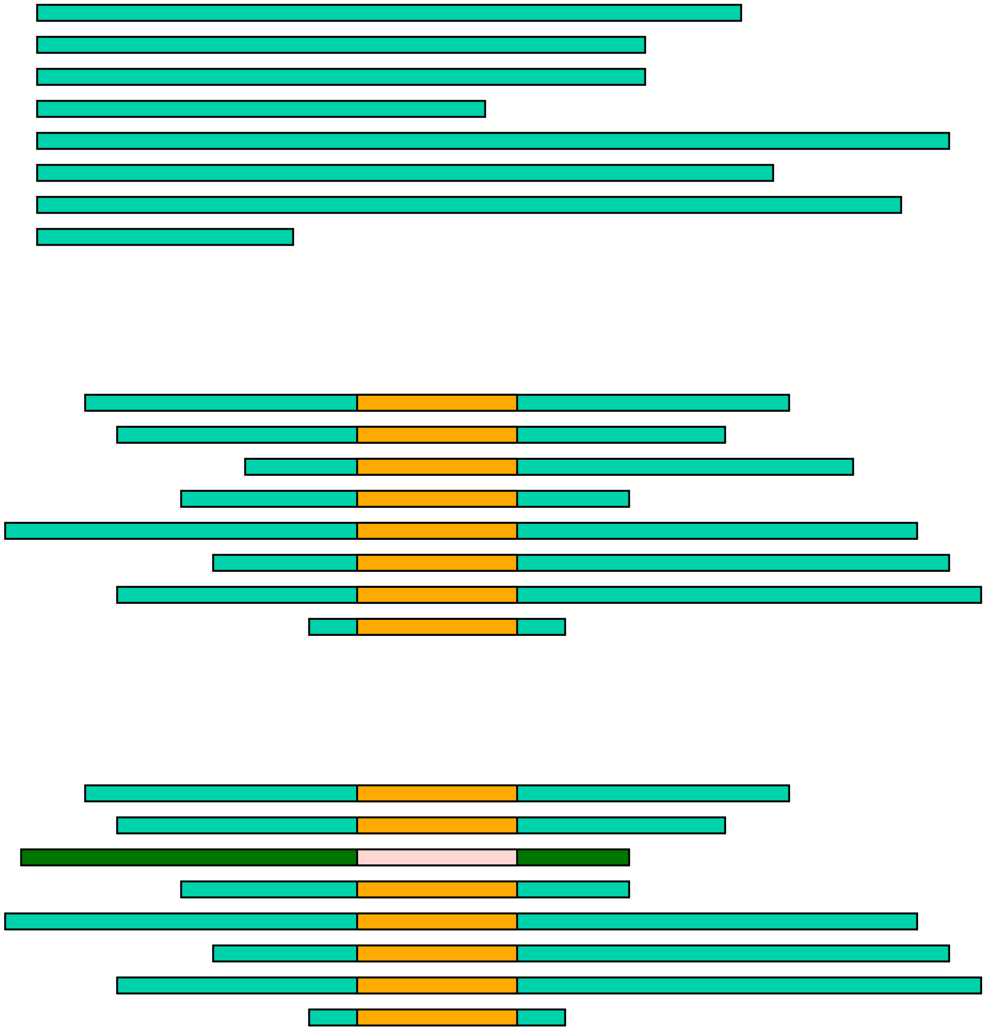
E-step: Using profile, compute a likelihood value z_{ij} for each *m*-window at position *i* in input sequence *j*.

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

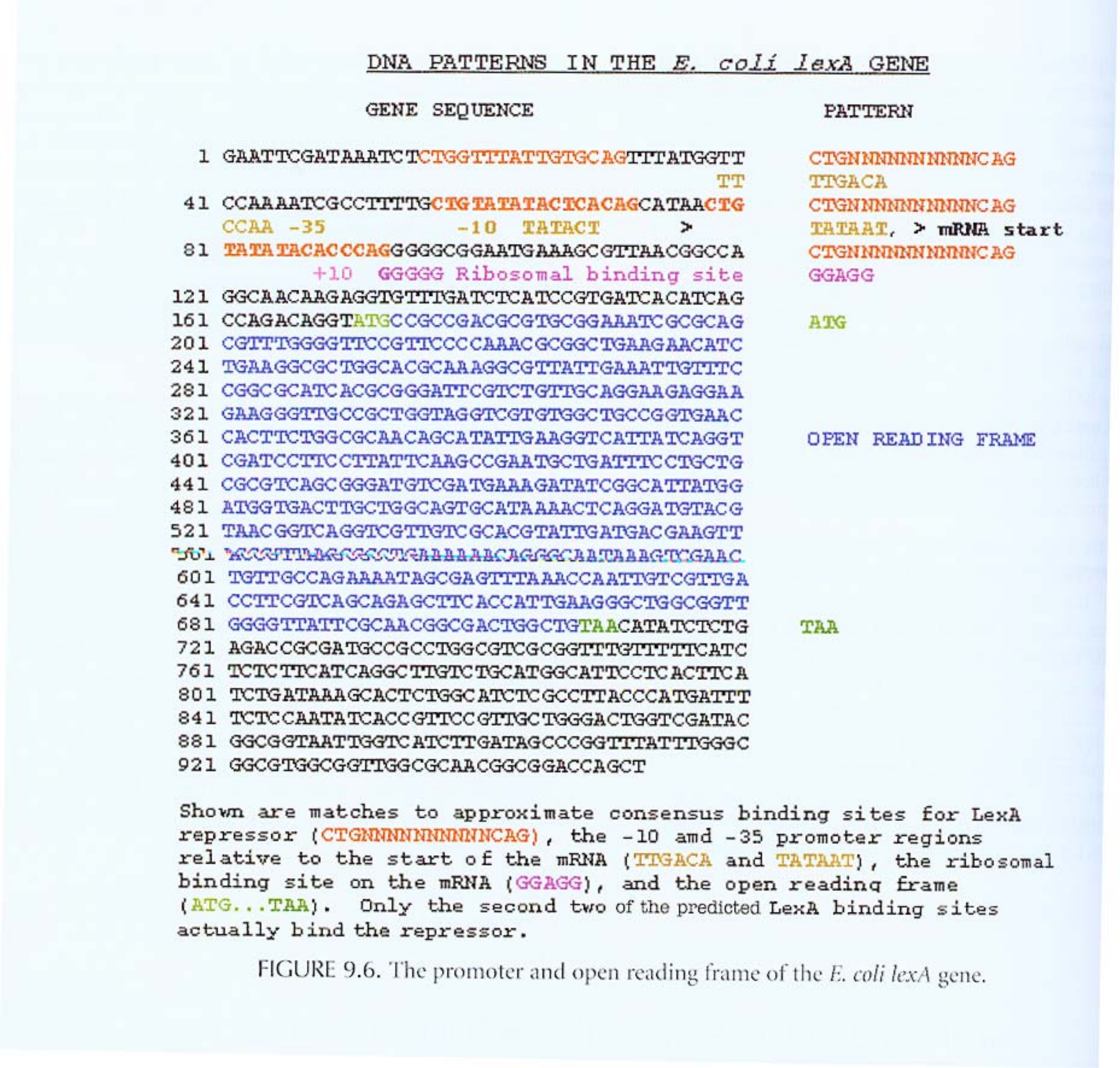
Stop if converged

MEME [Bailey, Elkan 1994]

Gibbs Sampling for Motif Detection



Prokaryotic Gene Characteristics



DNA PATTERNS IN THE *E. coli lexA* GENE

```
            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 ACCGTTAAGCGCCTGAAAAAACAGGGCAATAAAGTCGAAC
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.

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.

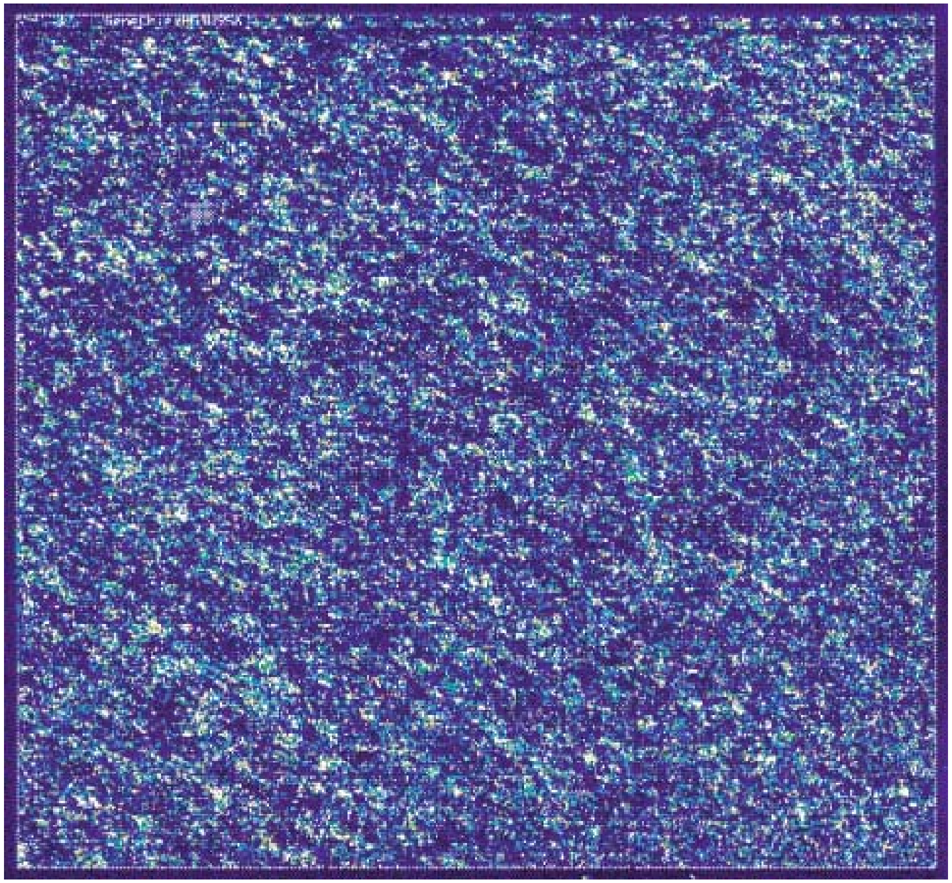
Microarray/DNA chip technology

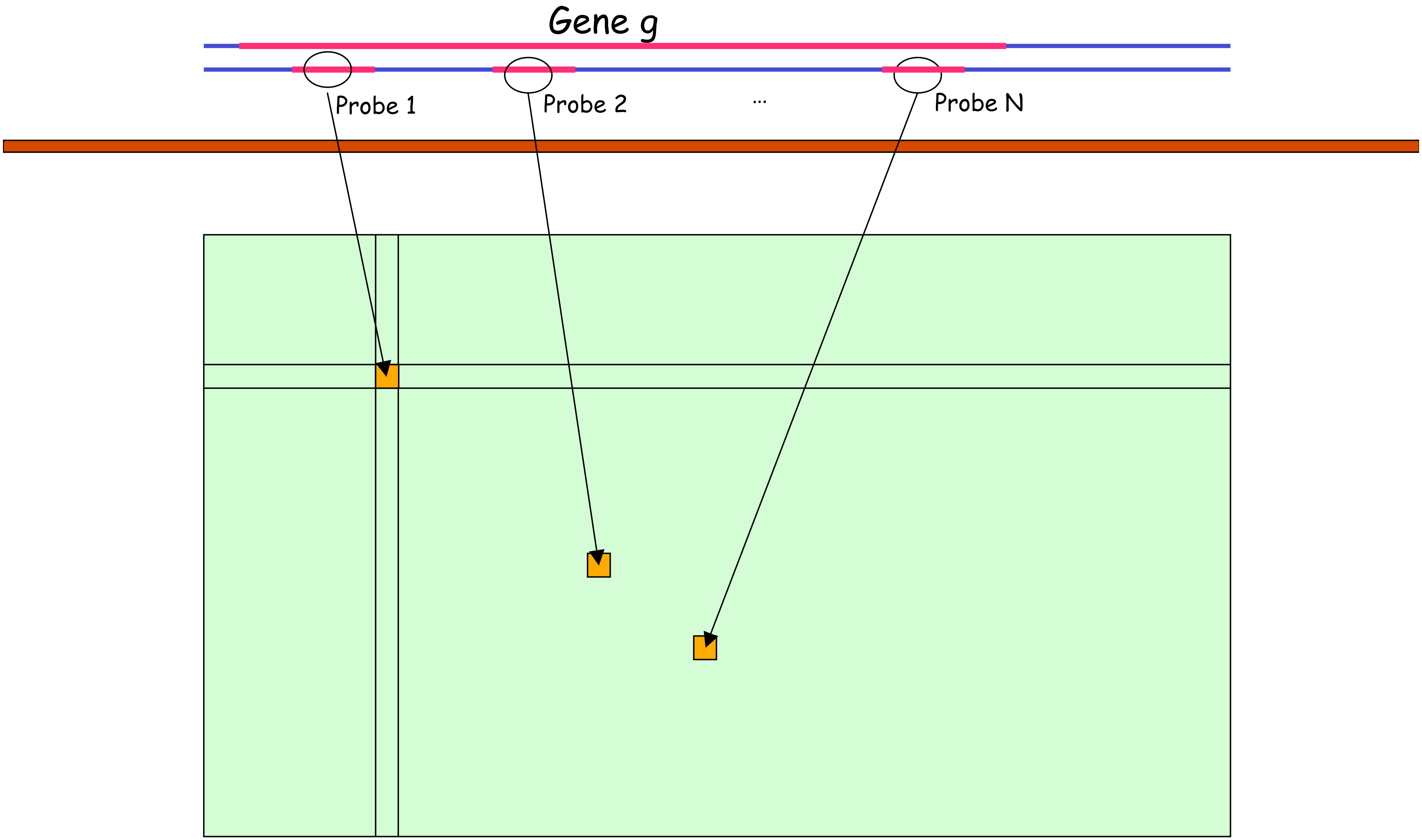
- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between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

Microarray Data

Gene	Expression Level
Gene1	
Gene2	
Gene3	
…	

Gene Chips

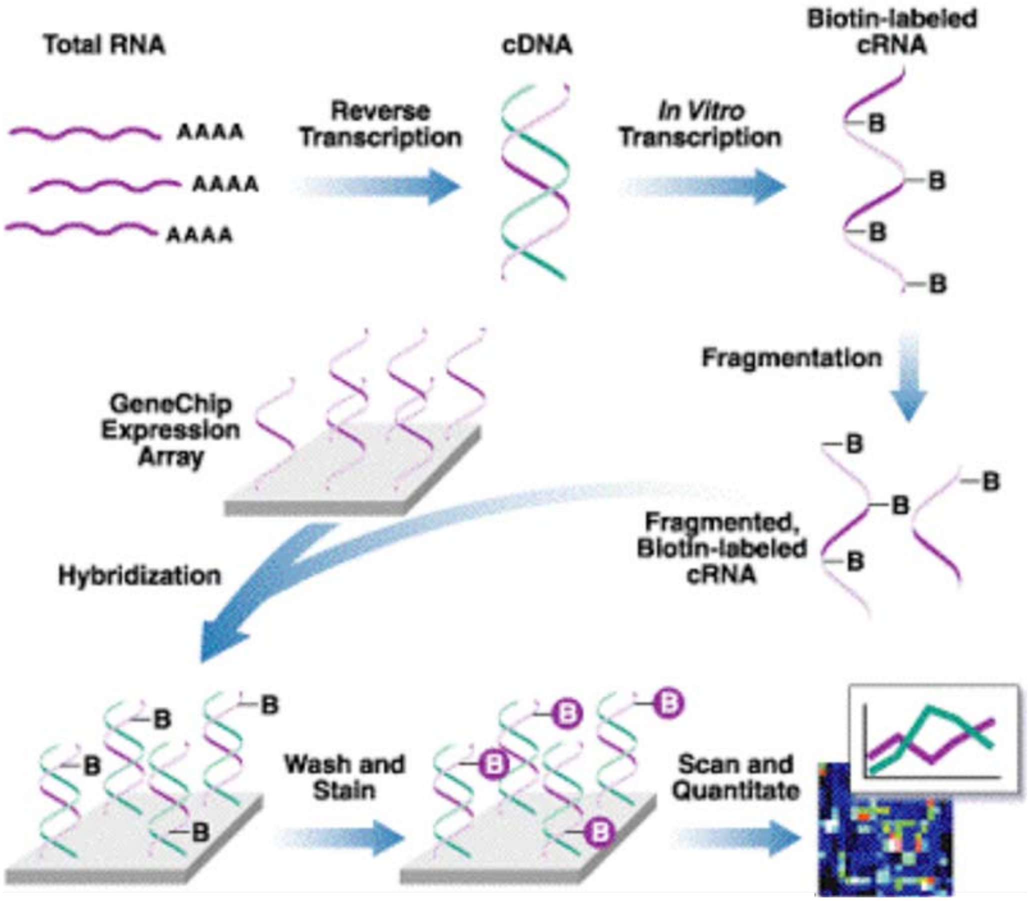



Gene g
Probe 1
Probe 2
…
Probe N

Microarray/DNA chips (Simplified)

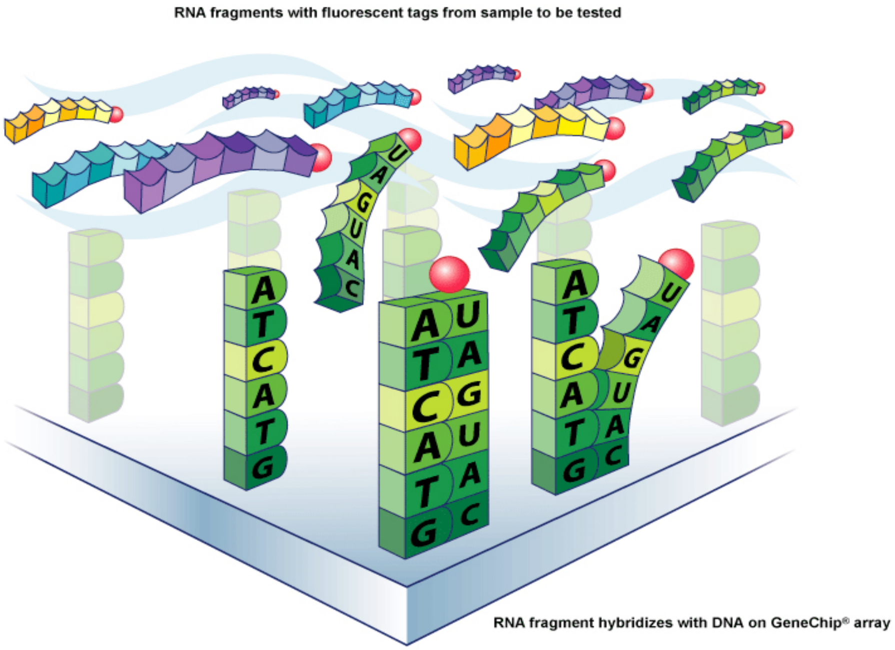
- Construct probes corresponding to reverse complements of genes of interest.
- Microscopic quantities of probes placed on solid surfaces at defined spots on the chip.
- Extract mRNA from sample cells and label them.
- Apply labeled sample (mRNA extracted from cells) to every spot, and allow hybridization.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

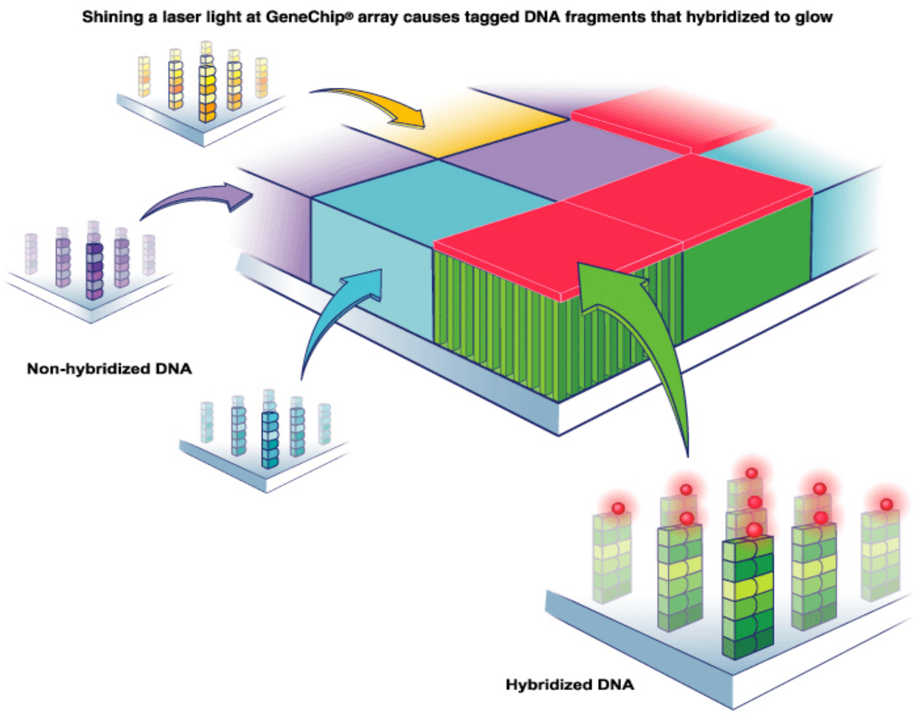
Affymetrix DNA chip schematic



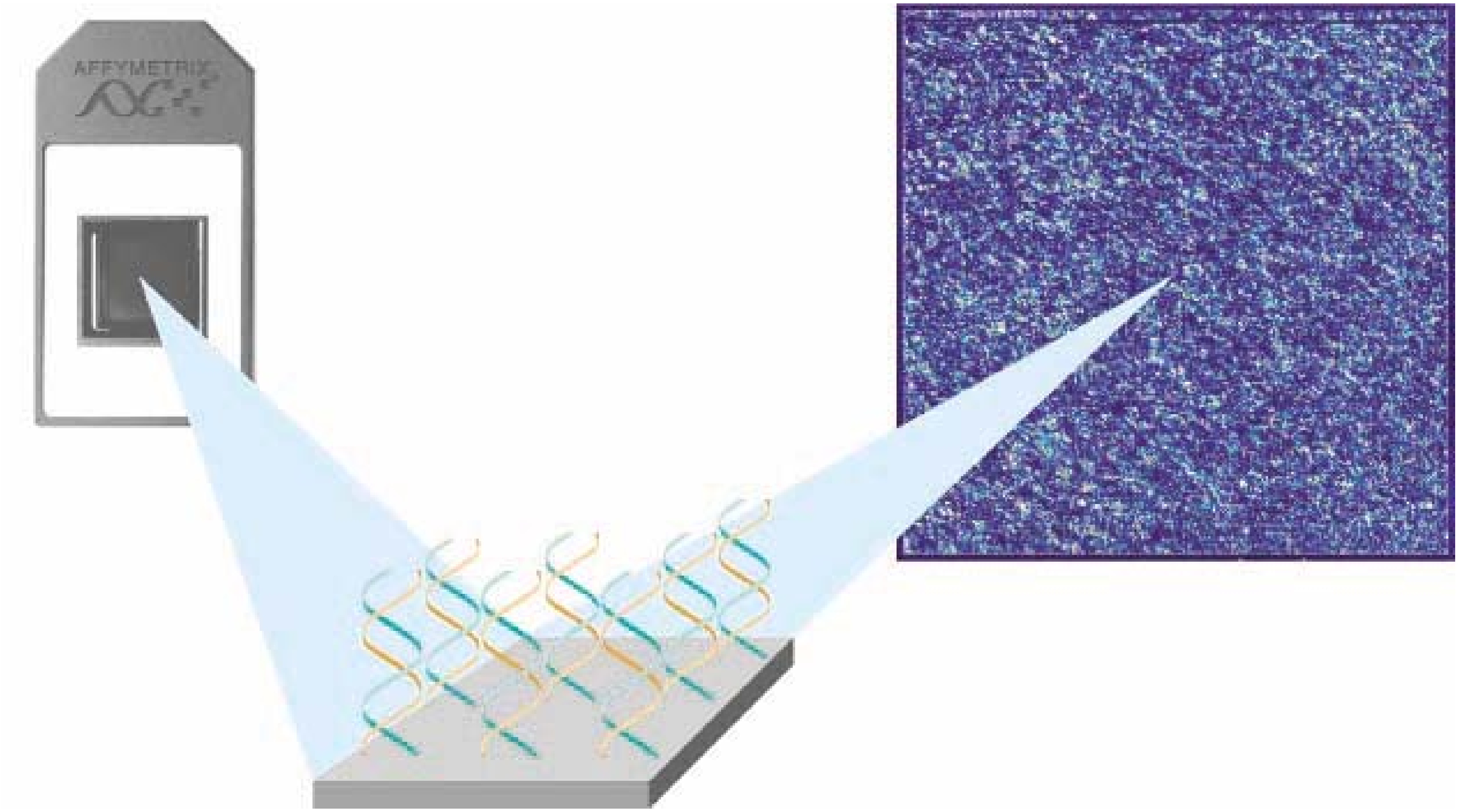
www.affymetrix.com

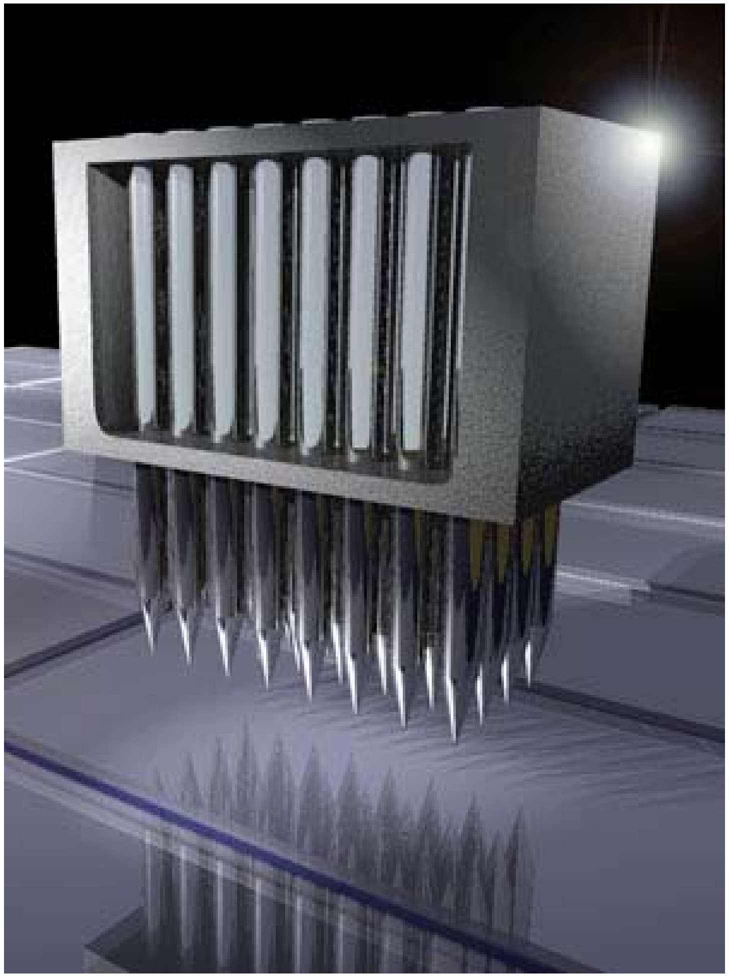
What’s on the slide?


RNA fragments with fluorescent tags from sample to be tested
A T C A T G
U A G U A C
A T C A T G
RNA fragment hybridizes with DNA on GeneChip® array


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