# A Biophysical Approach to Transcription Factor Binding Site Discovery

Anirvan Sengupta
Dept. of Physics and Astronomy
and
BioMaPS Institute
Rutgers University

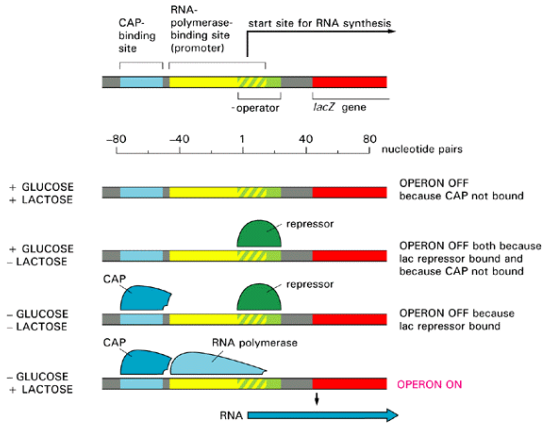
## Plan of the Talk

- Global Regulators and DNA-binding Specificity
- Predicting Binding Sites: Bioinformatics
- Experimental Efforts

# Examples of Global Regulators in *E. Coli*

| Name of Protein | Function |
|---|---|
| Crp/CAP | Hunger (cAMP) sensor |
| LexA | SOS response |
| Fnr | Oxygen sensing |
| Lrp | Leucine?? response |
| Various sigma factors | Response to different stresses |

# Textbook Example: Lac Operon



# Some Known Targets of lacI and of crp

lacI binding sites:

AATTGTGAGCGGATAACAATT
AAATGTGAGCGAGTAACAACC
GGCAGTGAGCGCAACGCAATT

some crp binding sites:

.....
TAATGTGACGTCCTTTGCATAC
GAAGGCGACCTGGGTCATGCGA
GGTG TTAAATTGATCACGTTTC
GATG CGAGGCGGATCGAAAAA
AAA TTCAATATTCATCACACTT
.....
TTTTGCGATCAAAATAACACTT
AAACGTGATCAACCCCTCAATT
TAATGTGAGTTAGCTCACTCAT
AATTGTGAGCGGATAACAATTT
.....

Consensus:

AAATGTGATCTAGATCACATTT

# Describing Fuzzy Motifs

IUPAC way:
A, C, G, T
R (A or G) from puRine,
Y (C or T) pYrimidine,
W (A or T) Weak,
S (G or C) Strong,
M (A or C) aMino,
K (G or T) Keto,
B (C or G or T) not-A
D (A or G or T) not-C
H (A or C or T) not-G
V (A or C or G) not-T
N aNy

For CRP, we have to look for
WWNTGTGANNNNNNTCTCANWW
or something less stringent
WWNYKYDVNNNNNNBHRMRNWW

# Weight Matrix

[Berg, vonHippel,Studen,Stormo,…]

**Given a set of known factor binding motifs, like,**

TAATGTGACGTCCTTTGCATAC
GAAGGCGACCTGGGTCATGCTG
CGATGCGAGGCGGATCGAAAAA
……..
……
ATTTGAACCAGATCGCATTA
AAATGTAAGCTGTGCCACGTTT

**construct a frequency matrix $n_{ib}$**

| Position | 1 | 2 | 3 | …….. | 22 |
|---|---|---|---|---|---|
| A | **3** | **4** | **5** | **……..** | **3** |
| C | **2** | **1** | **1** | **……..** | **2** |
| G | **2** | **2** | **2** | **……..** | **2** |
| T | **2** | **2** | **2** | **……..** | **2** |

# Weight Matrix Continued…

Calculate weights by taking logarithm : $w_{ib} = \log(n_{ib} / n_s)$

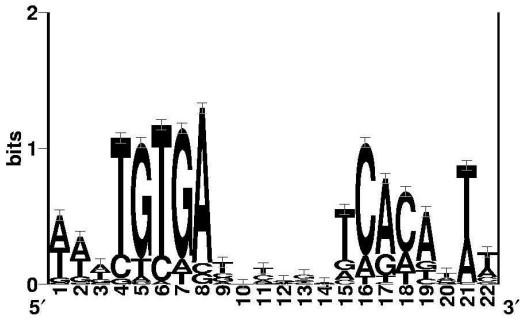
For any sequence S, the score W is given by: $W = \sum_{ib} w_{ib} S_{ib}$

For example:
$W(TTAGCA…..) = w_{1T} + w_{2T} + w_{3A} + w_{4G} + w_{5C} + w_{6A} + \cdots\cdots$

Sequences with higher $W$ are better binders.

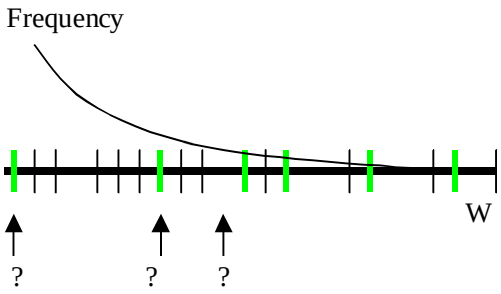
Precise relationship with binding energy in certain limits.

# Compromise

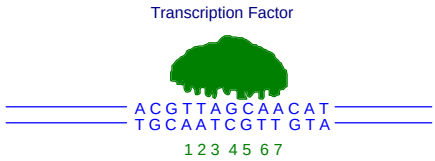



Robison, McGuire, Church

# Problem of Threshold Selection



# A Model for Transcription Factor Binding DNA



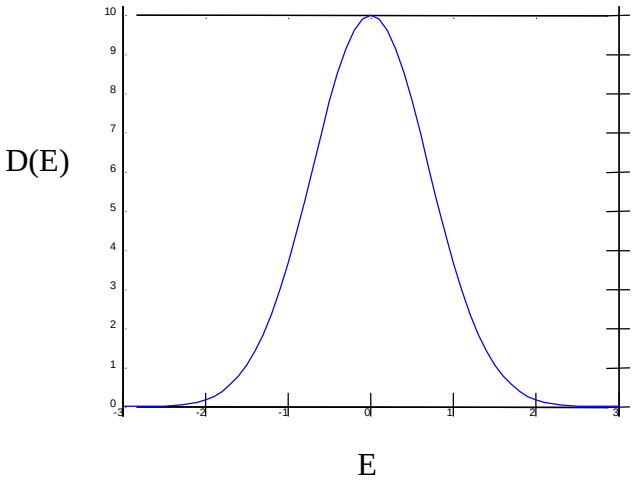

Independent Nucleotide Model for Binding Energy

(Berg, vonHippel, Stormo, …)

The energy is the sum of independent contributions from bases.

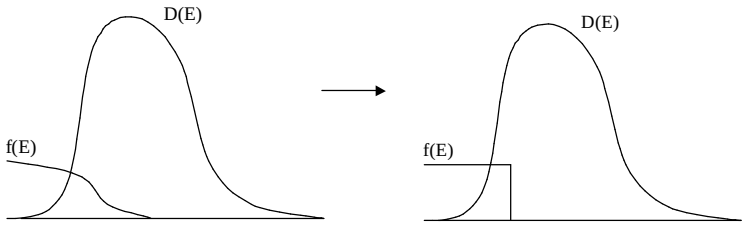
$E(TTAGCAA) = \varepsilon_{1T} + \varepsilon_{2T} + \varepsilon_{3A} + \varepsilon_{4G} + \varepsilon_{5C} + \varepsilon_{6A} + \varepsilon_{7A} = \sum_{ib} \varepsilon_{ib} S_{ib} = \varepsilon . S$

# Distribution of Energies



# Binding Probability

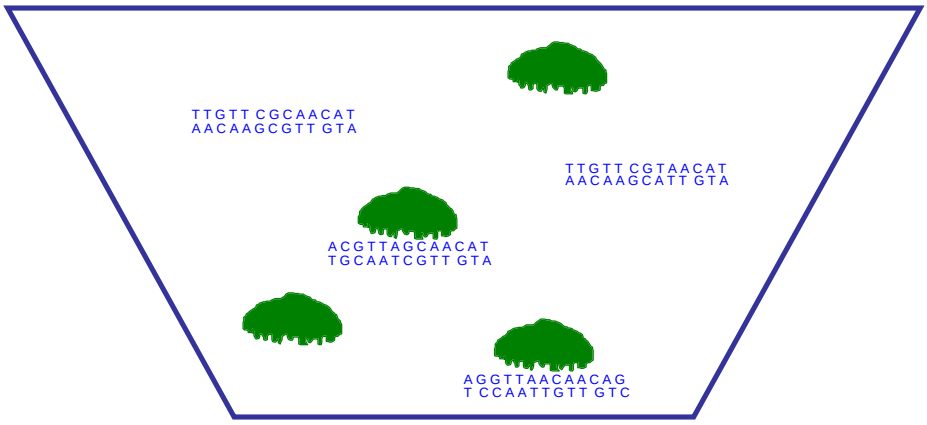
$$f(E(s)) = \frac{1}{\exp(\beta(E(S)-\mu))+1} = \frac{K\exp(-\beta E(S))n}{1+K\exp(-\beta E(S))n}$$




Finite Temperature

Zero Temperature

# The Probability Model for Data: Low Stringency SELEX



# From Sequences to Energies

Maximum Likelihood Method

$$e^{L} = \prod_{S \in O} (pf(E(S))) \prod_{S' \notin O} (1 - pf(E(S')))$$

Leads to minimization of

$$-L \approx -n_s \ln(p) - \sum_{S \in O} \ln(f(E(S))) + p \int dE D(E) f(E)$$

Simplifies in the 'low temperature' limit

# Temperature<< Variation of Energies

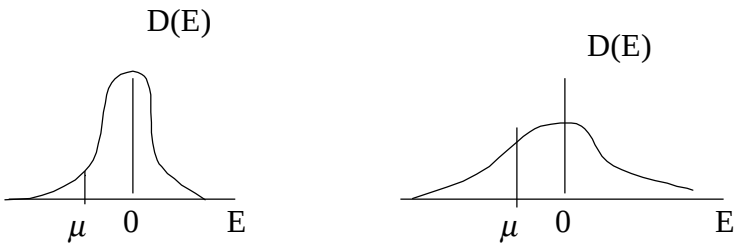
Leads to Minimization of

$$-\sum_{S \in O} \ln(\Theta(\mu - E(S))) + p \int_{-\infty}^{\mu} dE D(E)$$

The first term forces binding energy of sample sequences to be less than the chemical potential.

The second term forces the number of random strings that bind to be minimum.

## Increasing Width of D(E) increases number of 'False Positives'/ Random Background




Minimize the Variance!

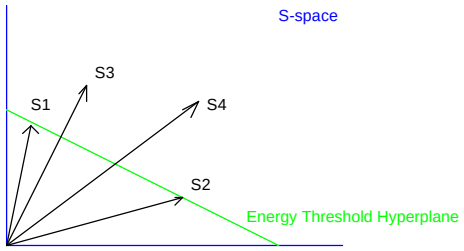
## Quadratic Programming Method for Energy Parameter Estimation

Minimize variance $\varepsilon^2$
Subject to constraints
$E(S_a) = \varepsilon$. $S_a < \mu = -1$
for each example $a$.

Solvable by Quadratic Programming.

Similar to Support Vector Machine (SVM) pattern finder.

# Support Vectors Machines




S1 and S2 supports the separating hyperplane.

# Low Concentration Limit: Weight Matrix Method

$\mu \to -\infty$, $f(E(S)) \to e^{\beta\mu}\, e^{-\beta E(S)}$

Maximum likelihood estimate of parameters leads to weight matrix formula.

# Comparison to Weight Matrix based Search

Training set for 55 E. Coli. transcription factors from DPInteract Database (G. Church, Harvard Medical School).

Get additional sites from RegulonDB (Collado-Vides, UNAM)

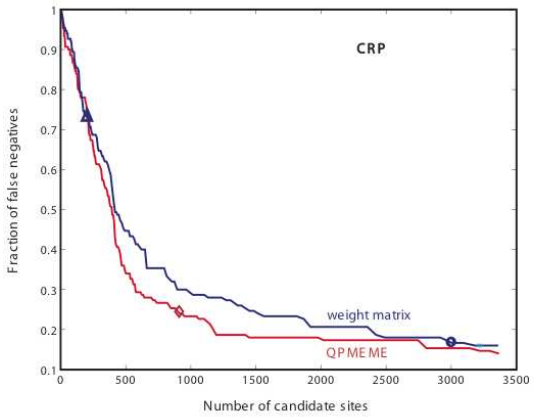
Compare the success of our method with

that of Weight Matrix Method:

Problem: No natural threshold choice for weight matrix.

Resolution: Compare for more than one thresholds.

# False Negatives vs List Size



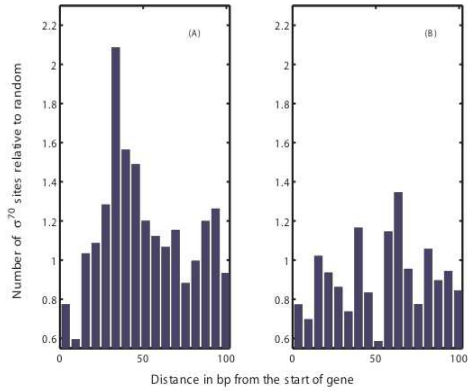
# Statistical Significance

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.

Table I: Statistical summary of *E. coli* search results (see website for details)

| Name | Length | Number of examples | Weight matrix 'hits' | QPMEME 'hits' | Significance |
|---|---|---|---|---|---|
| ArcA | 15 | 14 | 391 | 52 | 6.3 |
| ArgR | 18 | 17 | 320 | 79 | 8.9 |
| CRP | 22 | 49 | 3093 | 796 | 27.2 |
| CytR | 18 | 5 | 745 | 42 | 4.3 |
| DnaA | 15 | 8 | 98748 | 461 | 0.5 |
| FadR | 17 | 7 | 28 | 10 | 9.0 |
| FarR | 10 | 4 | 1893 | 241 | 3.7 |
| FIS | 35 | 19 | 7687 | 255 | 4.1 |
| Fnr | 22 | 14 | 174 | 36 | 13.9 |
| FruR | 16 | 12 | 31 | 23 | 14.8 |
| GalR | 16 | 7 | 10 | 9 | >9 |
| GcvA | 20 | 4 | 15 | 5 | >5 |
| GlpR | 20 | 13 | 9132 | 192 | 1.6 |
| H-NS | 11 | 15 | 14619 | 2340 | 2.7 |
| IHF | 48 | 26 | 82494 | 359 | 13.6 |
| LexA | 20 | 19 | 39 | 39 | >39 |
| LRP | 25 | 12 | 90676 | 4087 | 32.9 |
| MalR | 10 | 10 | 96 | 61 | 8.7 |
| MetJ | 16 | 15 | 404 | 42 | 1.6 |
| MetR | 15 | 8 | 344 | 26 | 3.2 |
| NagC | 23 | 6 | 72 | 8 | 7.0 |
| NarL | 16 | 11 | 2090 | 19 | 7.5 |
| OmpR | 20 | 9 | 4890 | 93 | 2.6 |
| PhoB | 22 | 15 | 258 | 23 | 14.8 |
| PurR | 26 | 22 | 47 | 28 | 27.0 |
| $\sigma^{70}$ (15) | 27 | 27 | 11517 | 635 | 2.2 |
| $\sigma^{70}$ (16) | 28 | 48 | 15867 | 912 | 2.6 |
| $\sigma^{70}$ (17) | 29 | 116 | 41488 | 3923 | 0.6 |
| $\sigma^{70}$ (18) | 30 | 34 | 10133 | 381 | 0.6 |
| $\sigma^{70}$ (19) | 31 | 25 | 15086 | 391 | 0.4 |
| $\sigma^{54}$ | 16 | 6 | 16 | 7 | 6.0 |
| $\sigma^{S}$ | 29 | 15 | 10669 | 245 | 2.0 |
| SoxS | 35 | 14 | 3963 | 49 | 2.2 |
| TyrR | 22 | 17 | 3843 | 73 | 6.2 |

## Statistics of $\sigma^{70}$ Hits: Orientation Dependence




## False Positives: Experiments

- Checking examples by gel-retardation experiments
- Effort to do genome-wide location analysis for some pleiotropic factors in *E. Coli*.
- Low stringency SELEX experiments.

# Some of Predicted Binding Sites

**AAGTGTGACCCGGTTCACGTAG 1.23 1402612**
**in between**
**1401279 1402604 -1;**
**1402765 1403673 +1;**

**TCCTGCGCCTTTGCTCACAATC 1.05 1846106**
**in between**
**1844989 1846032 -1;**
**1846149 1846700 -1;**

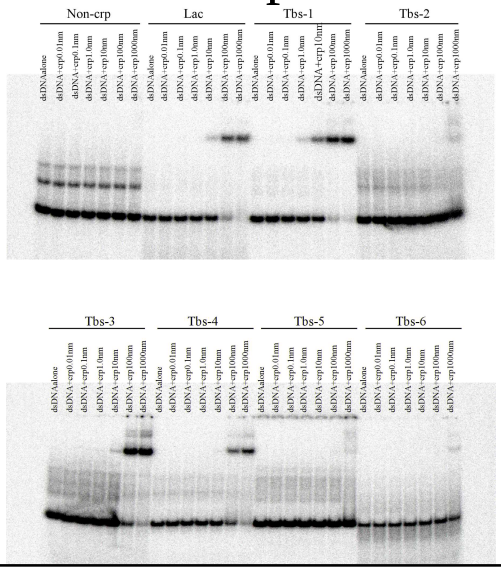
**TATCGAGATAACGATCACAAAA 1.17 2175284**
**in between**
**2174370 2175230 -1;**
**2175532 2176656 -1;**

**TTATGGAAGAGATATCACATTT 1.01 3003939**
**in between**
**3002030 3003808 -1;**
**3004356 3005447 +1;**

**TAACGCGATTCCGCTCAAAAAT 1.16 3874615**
**in between**
**3873768 3874580 -1;**
**3874695 3875102 -1;**

**CAATTTGATCTACATCTCTTTA 1.04 4131285**
**in between**
**4130196 4131086 +1;**
**4131415 4133595 +1**

# Gel-shift Experiments



## Summary

- Low DNA-binding specificity for global regulators and need to quantify variability.
- New bio-informatic tool for binding site prediction with a built-in threshold.
- Preliminary experimental results encouraging.

## Collaborators

Computation: Boris Shraiman (Rutgers)
Marco Djordjevic (Columbia)

Experiments: Viji Nagaraj (Rutgers)
Boris Shraiman (Rutgers)
Richard Ebright (Rutgers)