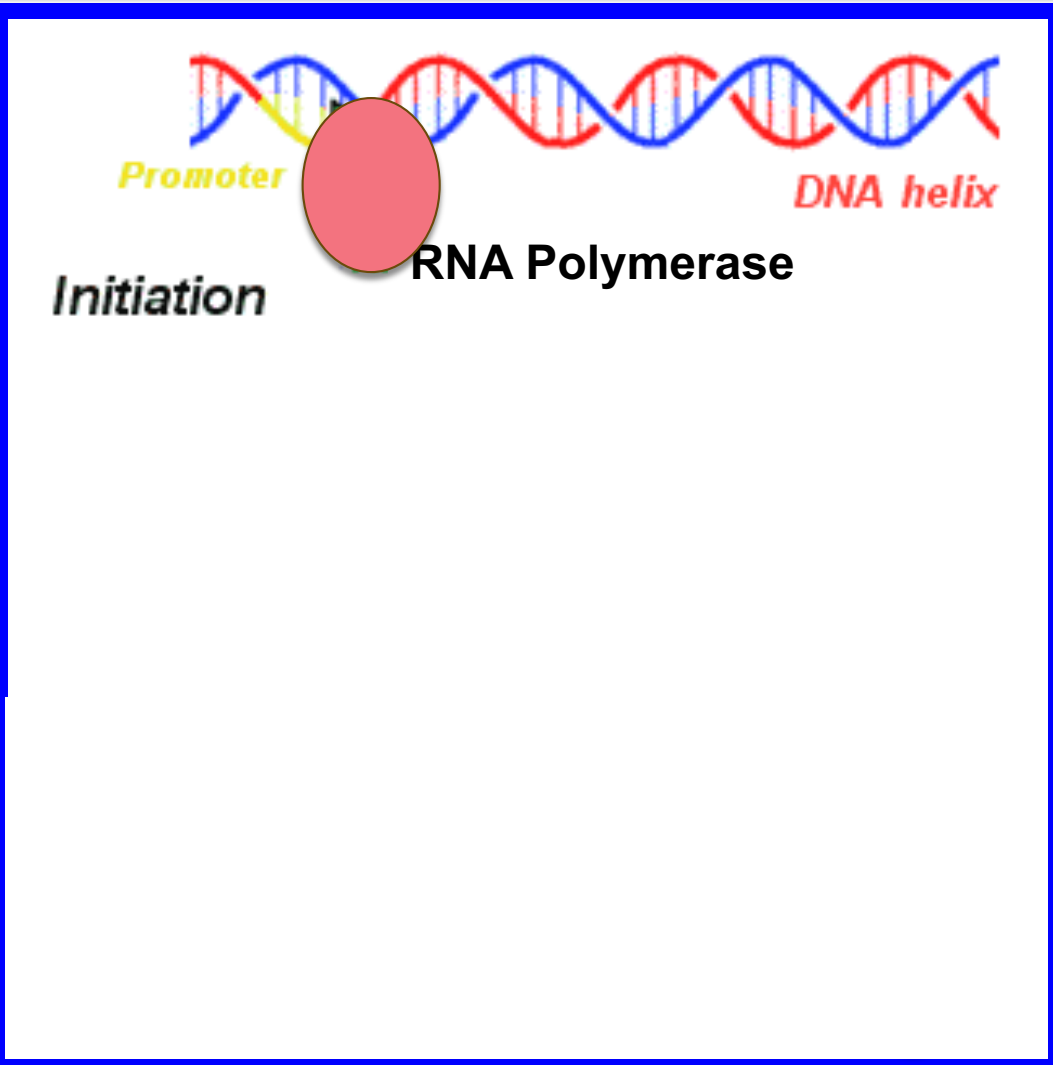
REGULATORY GENOMICS

Saurabh Sinha, Dept. of Computer Science & Carl R. Woese Institute of Genomic Biology, University of Illinois.

Introduction …

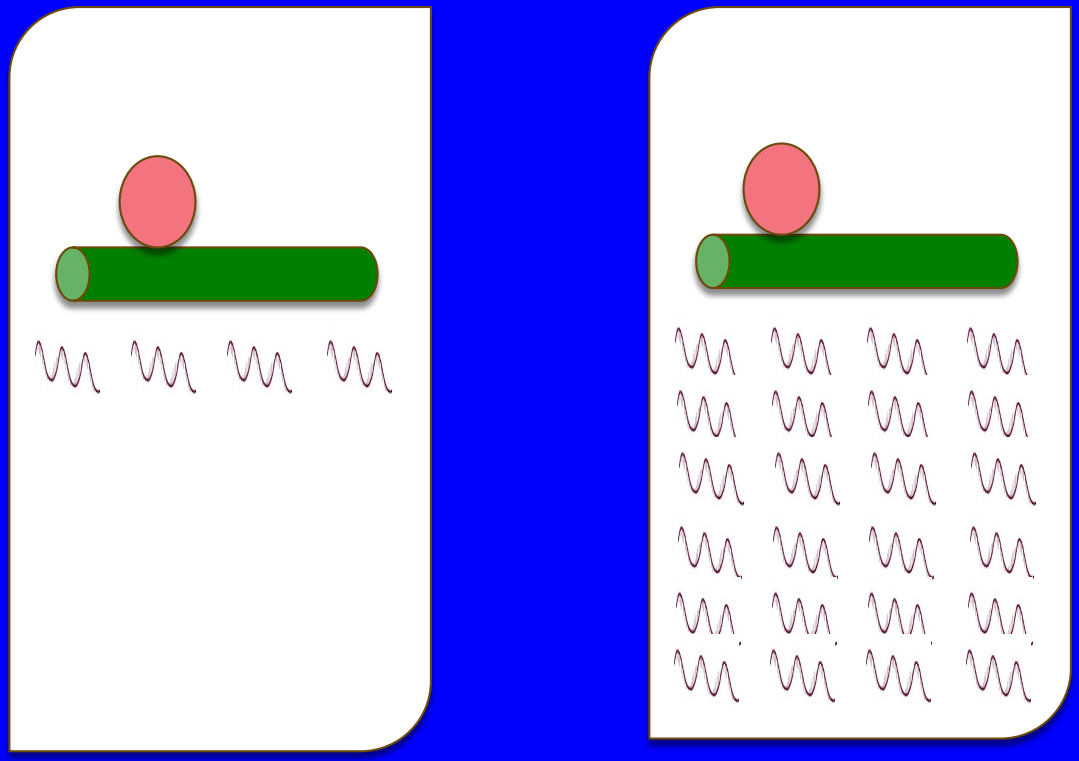
Genes to proteins: “transcription”



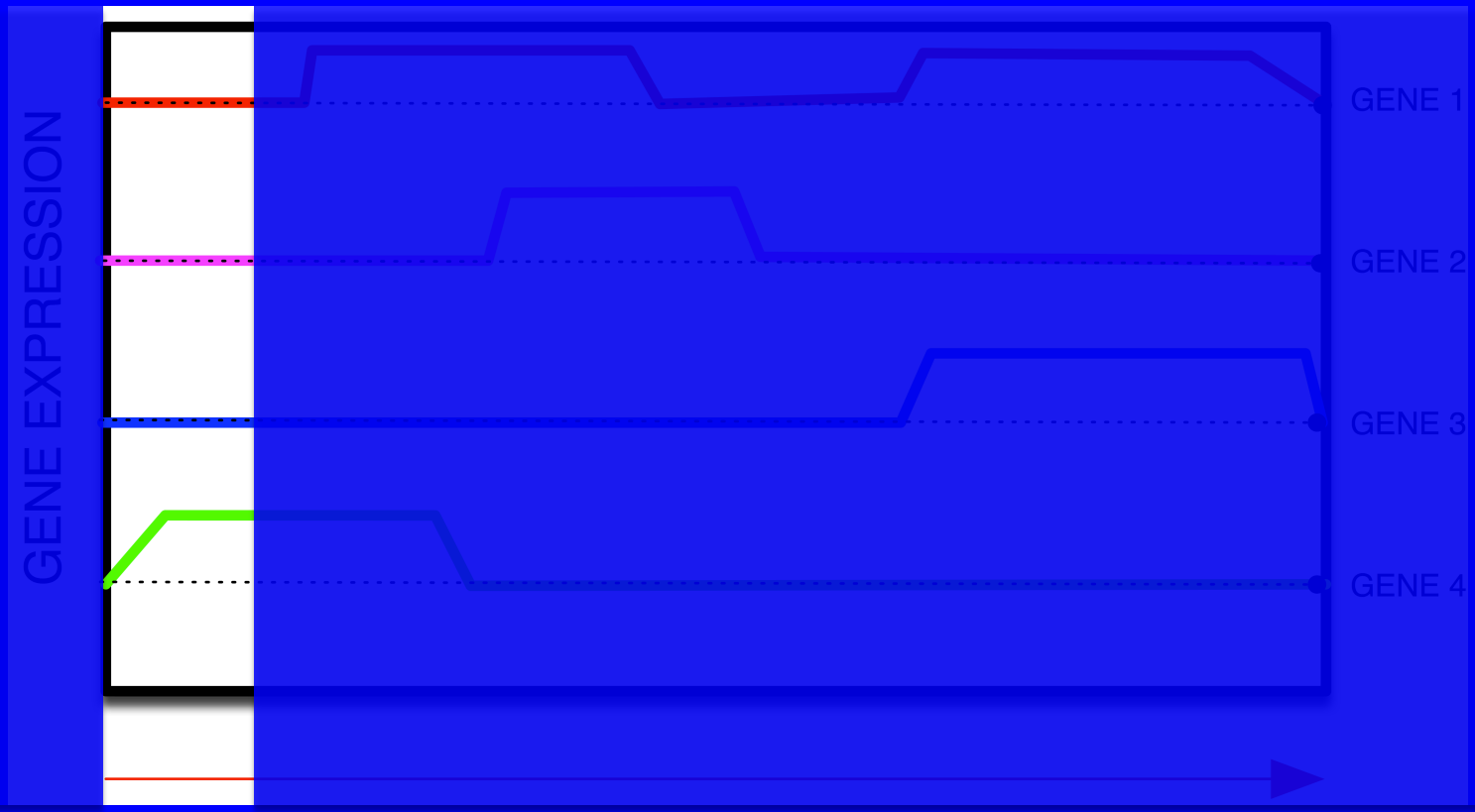
http://instruct.westvalley.edu/svensson/CellsandGenes/Transcription%5B1%

Regulation of gene expression

- More frequent transcription => more mRNA => more protein.

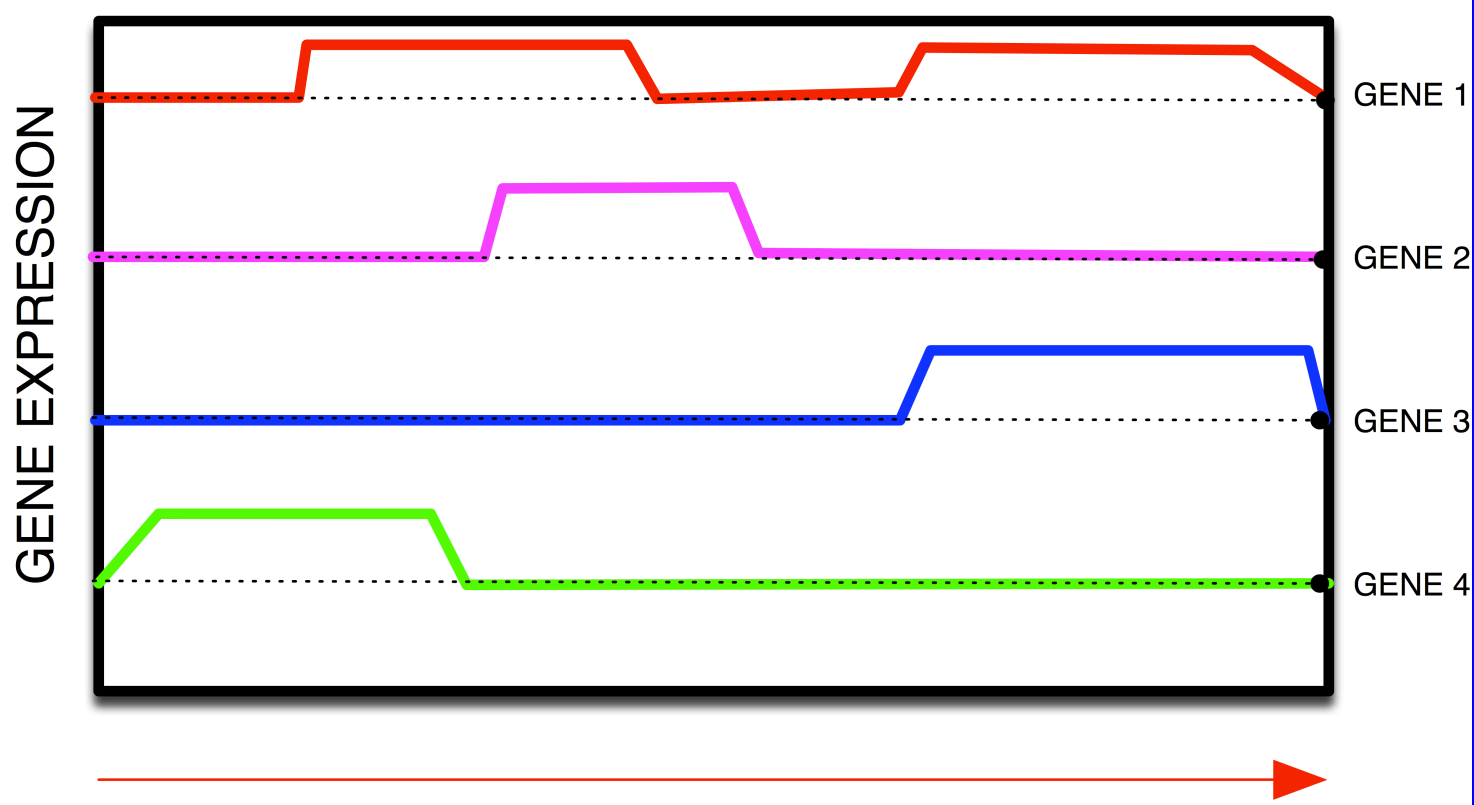


Cell states defined by gene expression



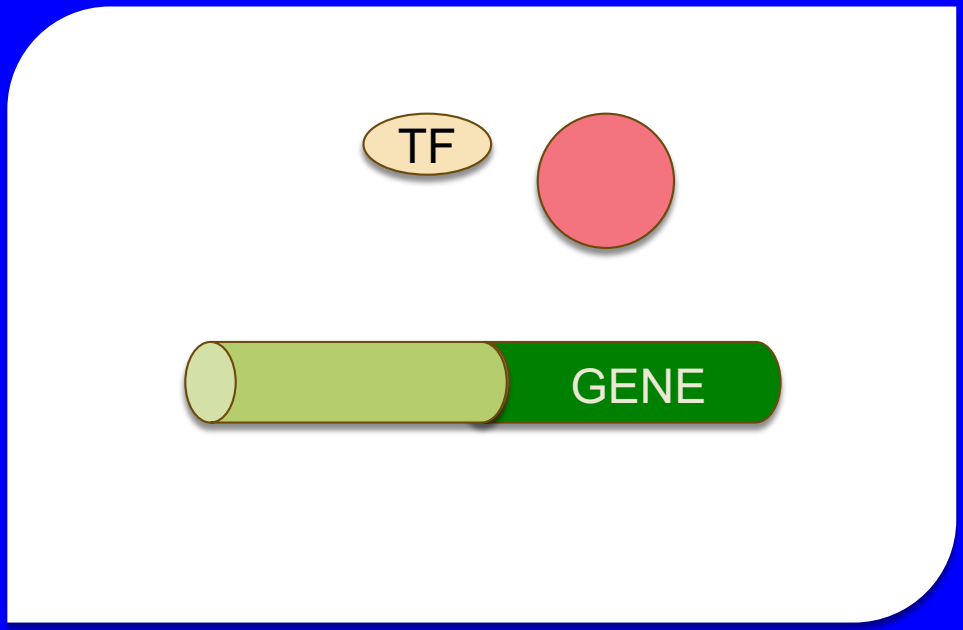
Gene 4 Genes 1&4 Genes 1&2 Genes 1&3

Cell states defined by gene expression



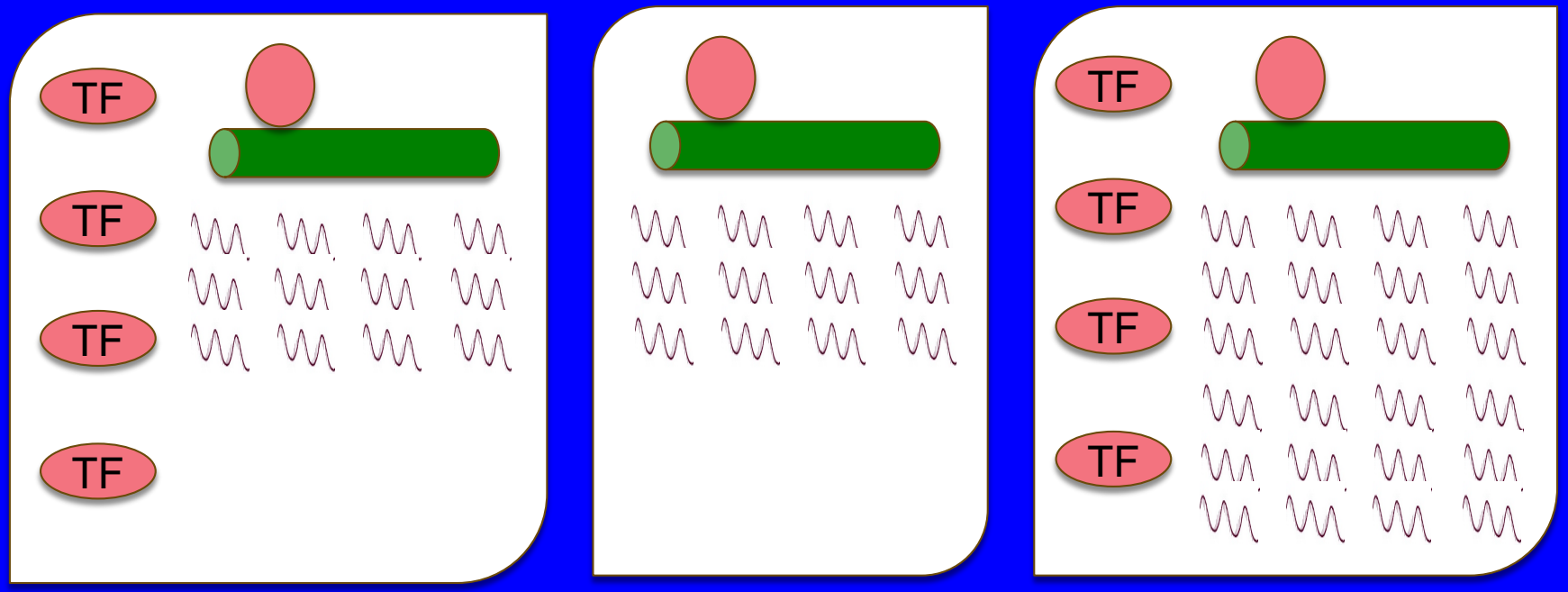
- *How is the cell state specified ?*
- *In other words, how is a specific set of genes turned on at a precise time and cell ?*

Gene regulation by “transcription factors”

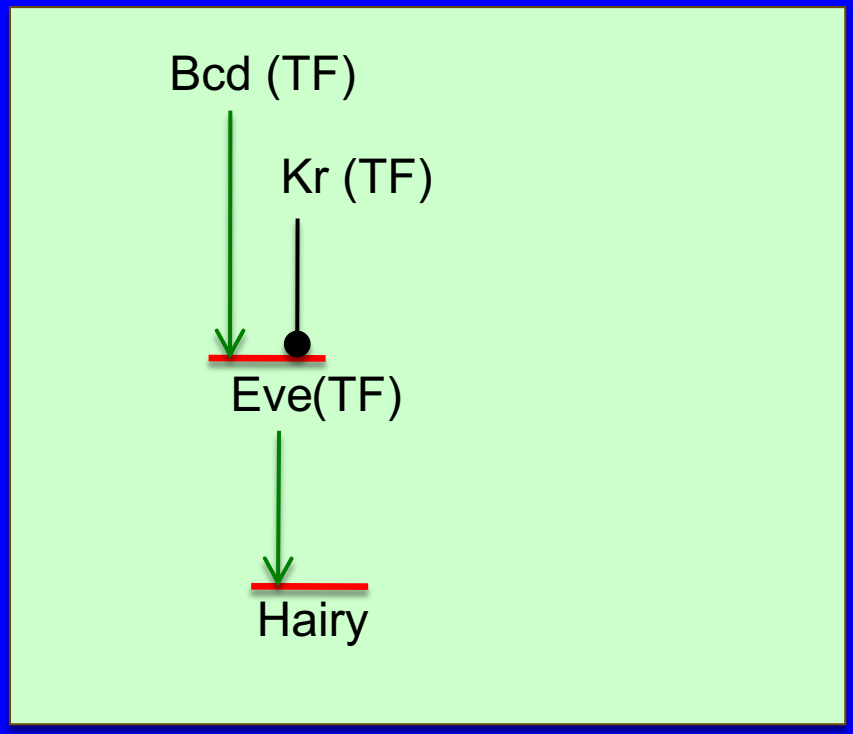


Transcription factors

- may activate …
- or repress

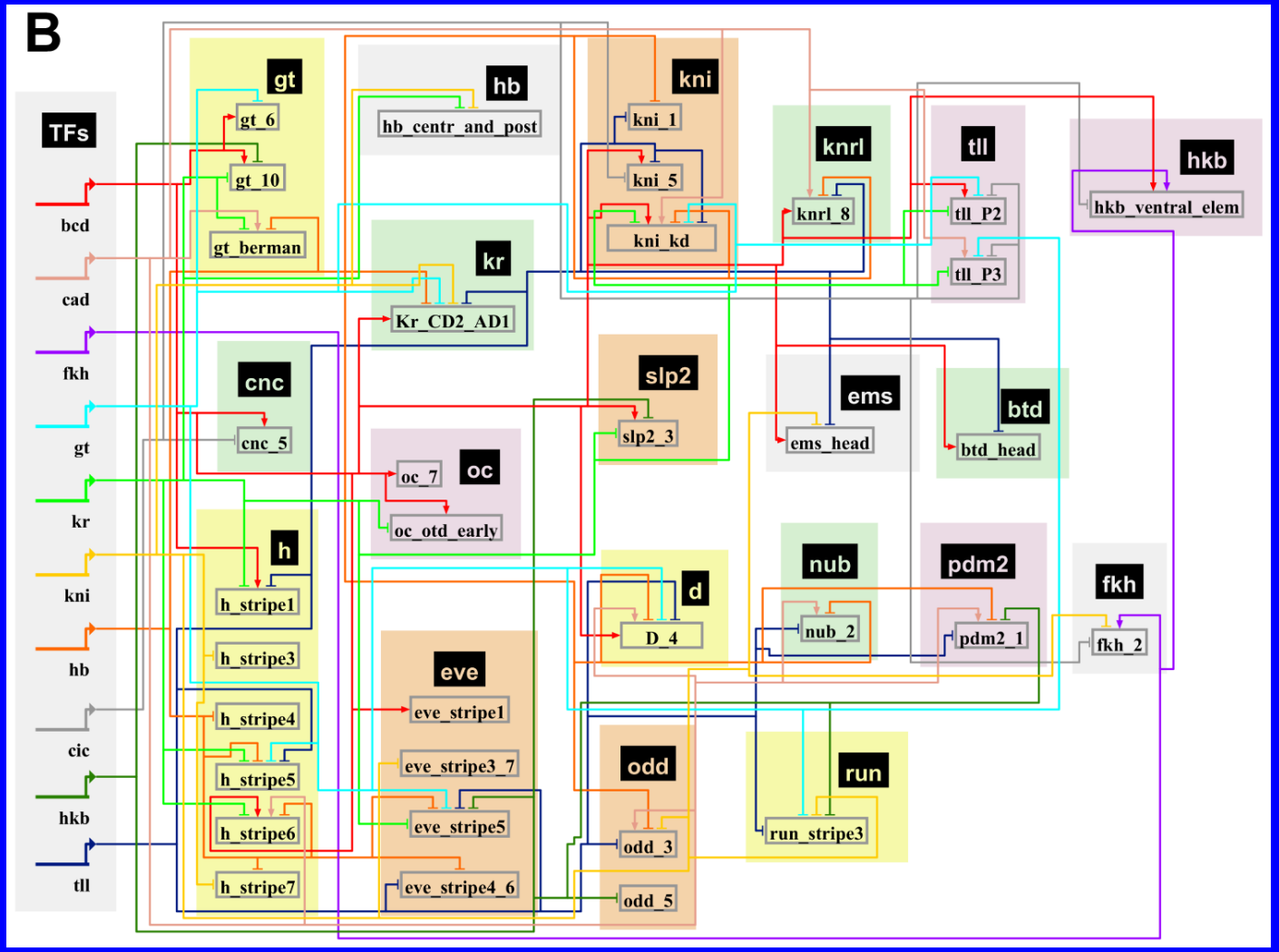


Gene regulation by transcription factors determines cell state



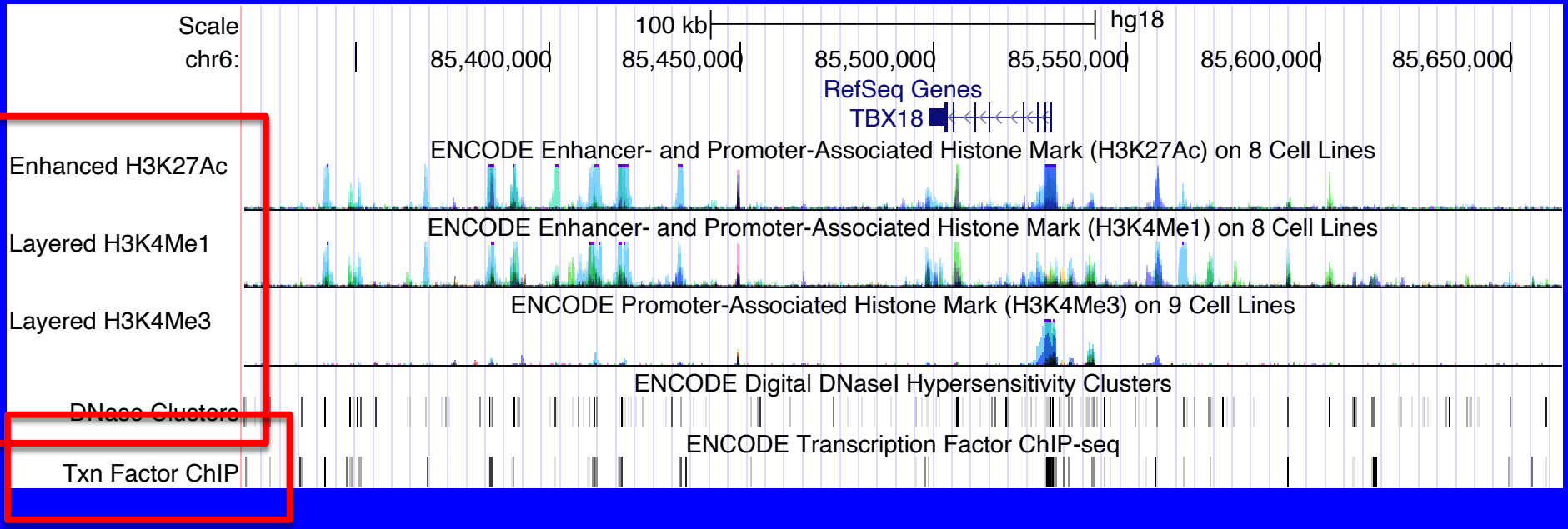
Gene Eve is “on” if “Bcd present AND Kr NOT present”.

Gene regulatory networks



- *Goal: discover the gene regulatory network*
- *Sub-goal: discover the genes regulated by a transcription factor*

Genome-wide assays



One experiment per cell type ... tells us where the regulatory signals may lie

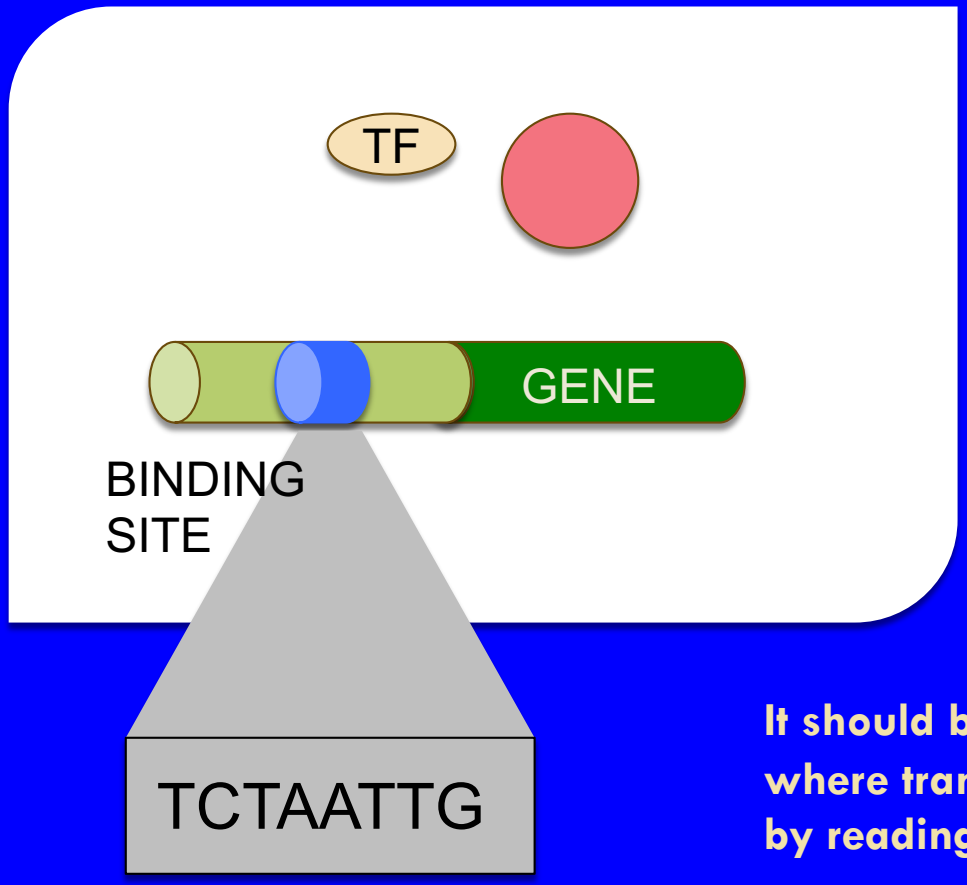
One experiment per cell type AND PER TF

... tells us which TF might regulate a gene of interest

Expensive !

- *Goal: discover the gene regulatory network*
- *Sub-goal: discover the genes regulated by a transcription factor*
- *… by DNA sequence analysis*

The regulatory network is encoded in the DNA



It should be possible to predict where transcription factors bind, by reading the DNA sequence

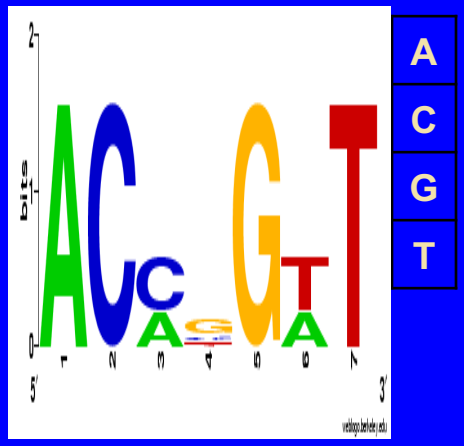
Motifs and DNA sequence analysis

Finding TF targets

- Step 1. Determine the binding specificity of a TF
- Step 2. Find motif matches in DNA
- Step 3. Designate nearby genes as TF targets

Step 1. Determine the binding specificity of a TF

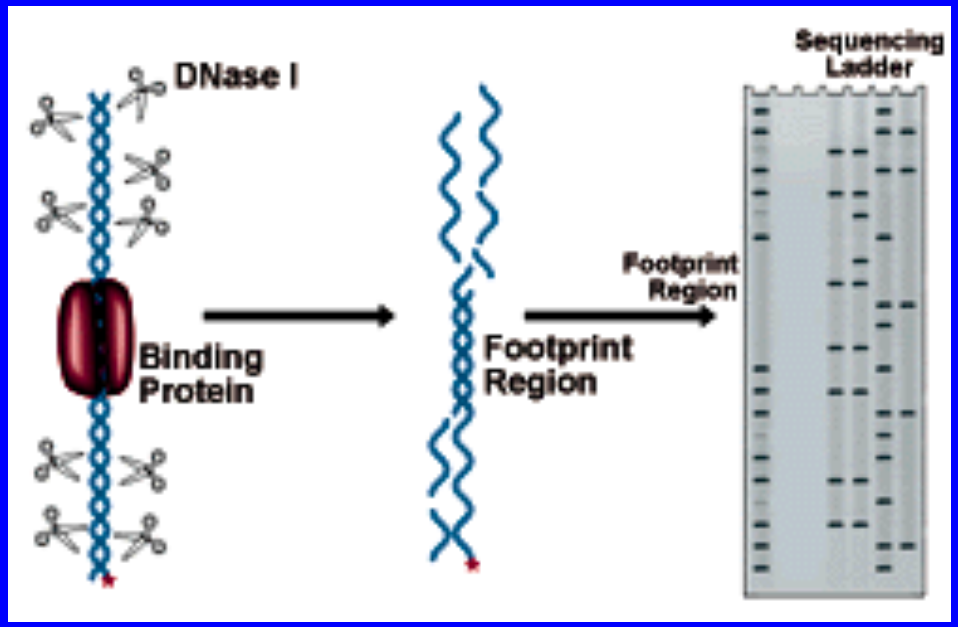
ACCCGTT
ACCGGTT
ACAGGAT
ACCGGTT
ACATGAT



"MOTIF"

How?

- 1. DNase I footprinting

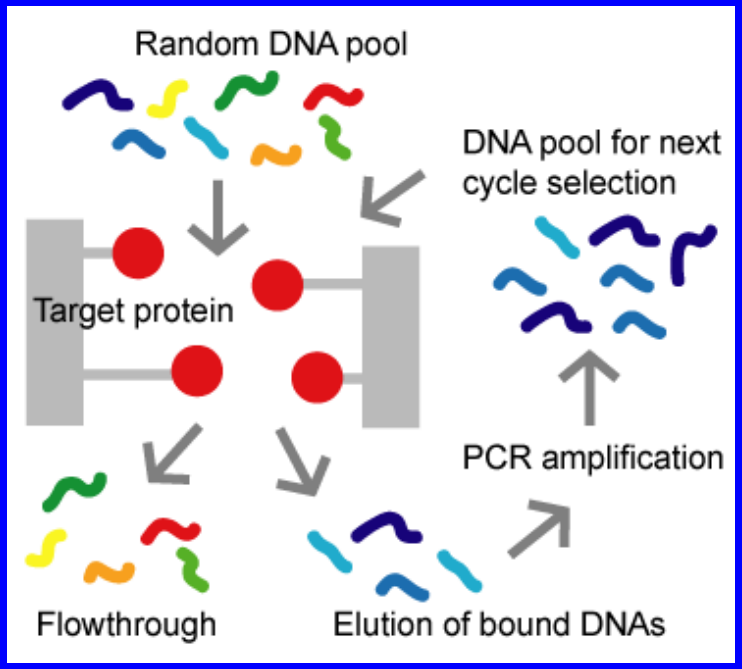


TAACCCGTTC
GTACCGGTTG
ACACAGGATT
AACCGGTTA
GGACATGAT

http://nationaldiagnostics.com/article_info.php/articles_id/31

How?

- 2. SELEX

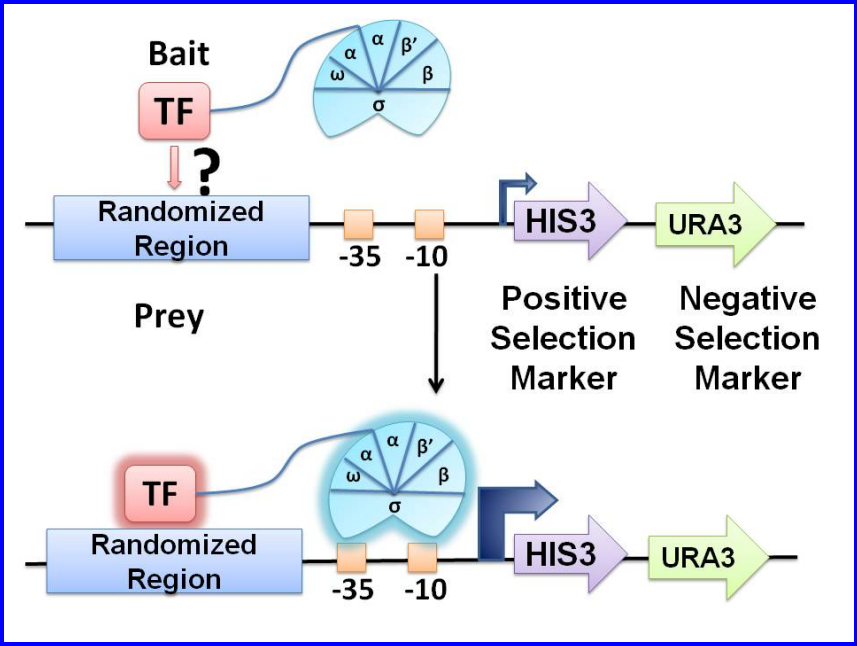


http://altair.sci.hokudai.ac.jp/g6/Projects/Selex-e.html

TAACCCGTTC
GTACCGGTTG
ACACAGGATT
AACCGGTTA
GGACATGAT

How?

- 3. Bacterial 1-hybrid

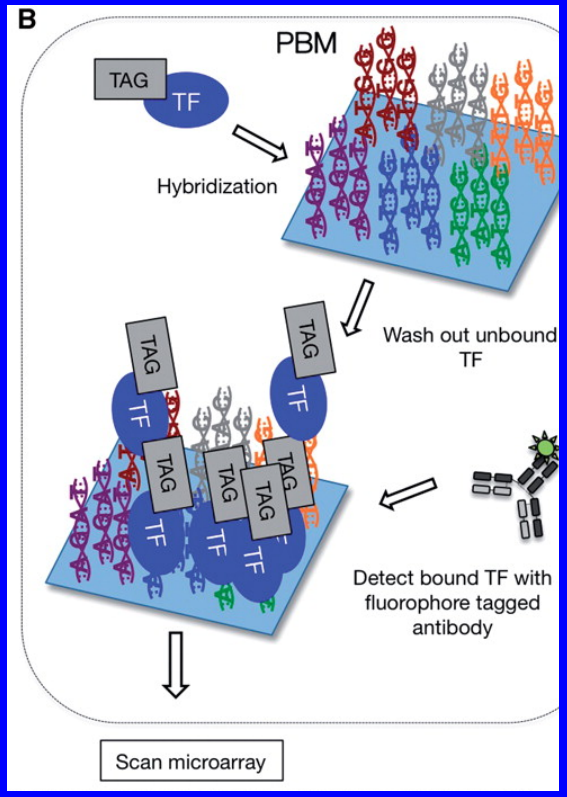


TAACCCGTTC
GTACCGGTTG
ACACAGGATT
AACCGGTTA
GGACATGAT

http://upload.wikimedia.org/wikipedia/commons/5/56/FigureB1H.JPG

How?

- 4. Protein binding microarrays



TAACCCGTTC
GTACCGGTTG
ACACAGGATT
AACCGGTTA
GGACATGAT

http://bfg.oxfordjournals.org/content/9/5-6/362/F2.large.jpg

Motif finding tools

How did this happen?

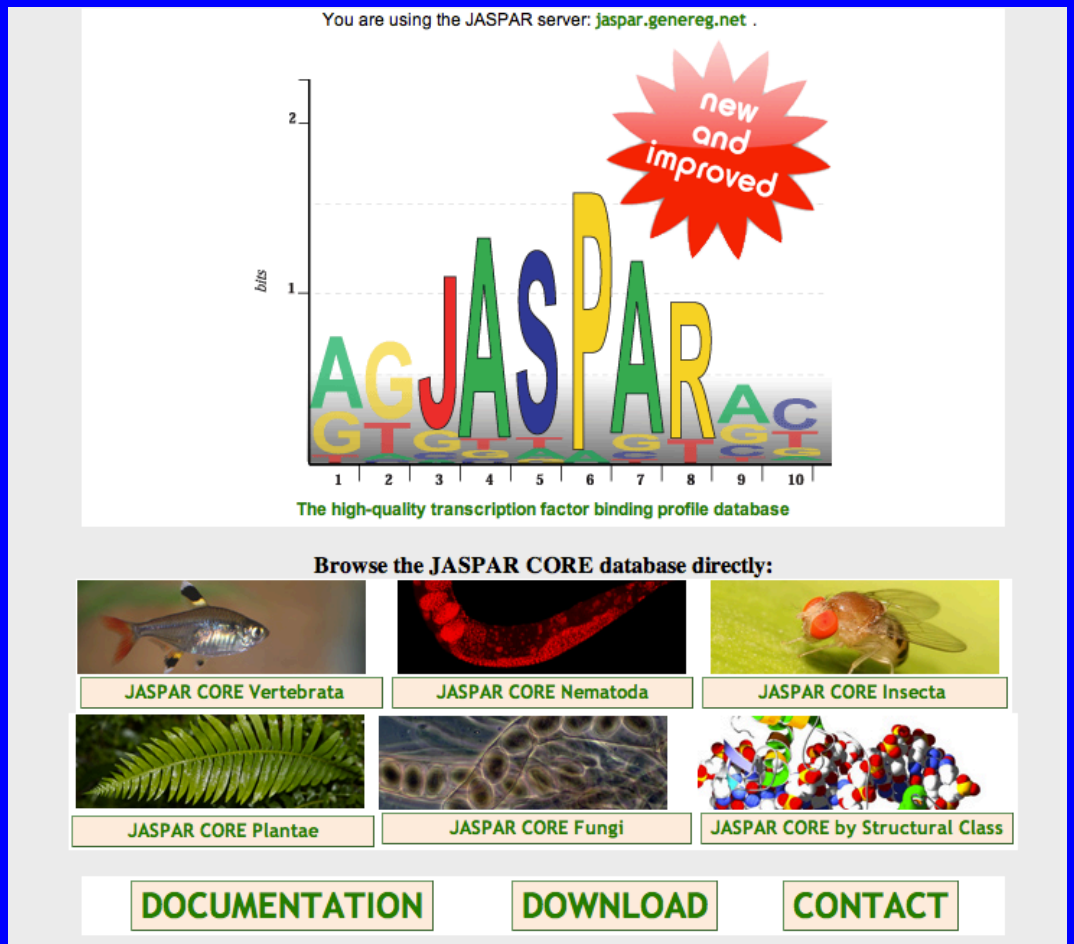
```
TAACCCGTTC
GTACCGGTTG
ACACAGGATT
 AACCGGTTA
GGACATGAT
```

Run a motif finding tool (e.g., "MEME") on the collection of experimentally determined binding sites, which could be of variable lengths, and in either orientation.

We'll come back to motif finding tools later.

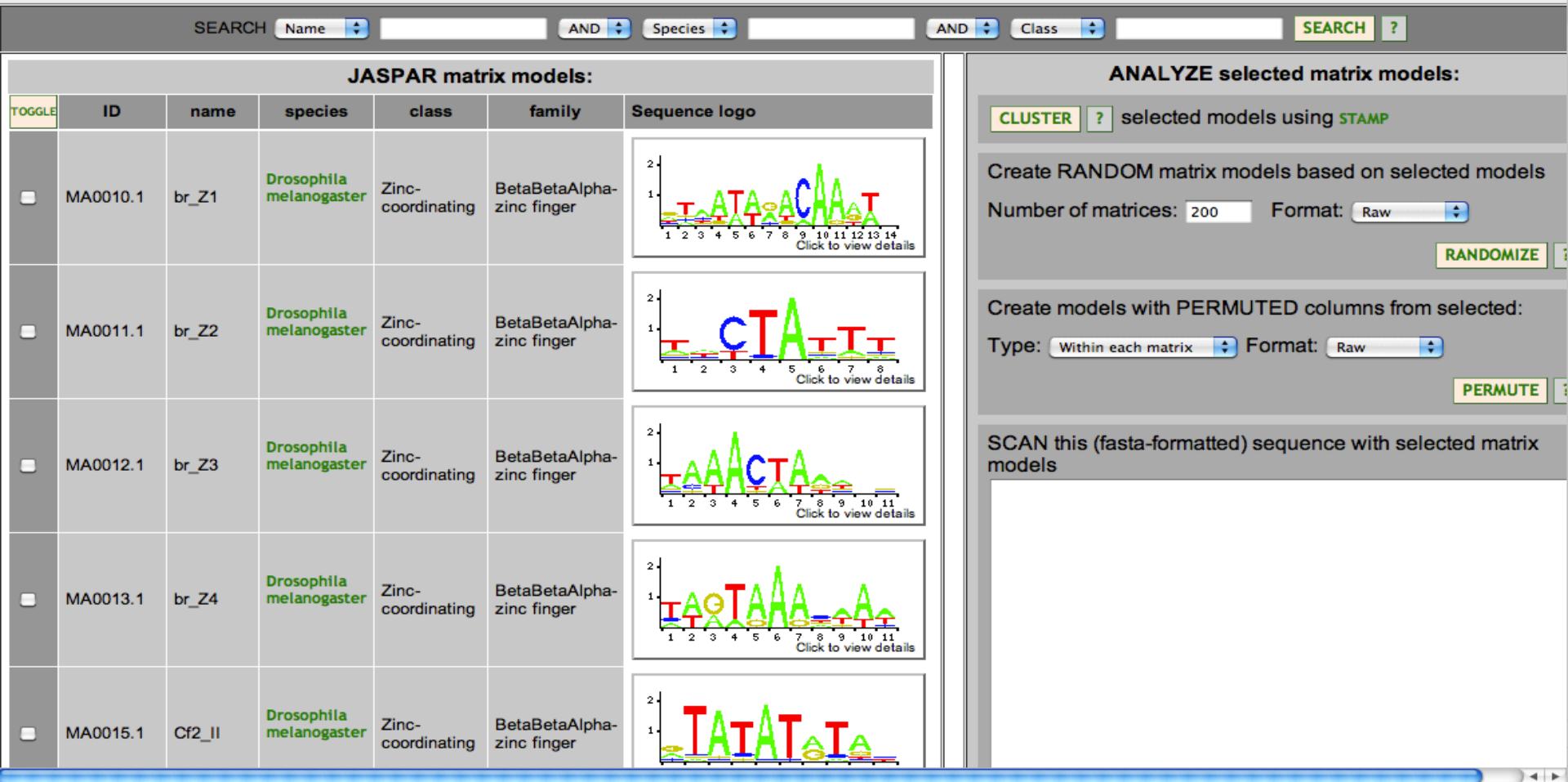
Motif Databases

- JASPAR: http://jaspar.genereg.net/



Motif Databases

- JASPAR: http://jaspar.genereg.net/



Motif Databases

- TRANSFAC
  - Public version and License version
- Hocomoco: http://hocomoco.autosome.ru/
  - Human and mouse motifs
- UniProbe: http://thebrain.bwh.harvard.edu/uniprobe/
  - variety of organisms, mostly mouse and human

Motif Databases

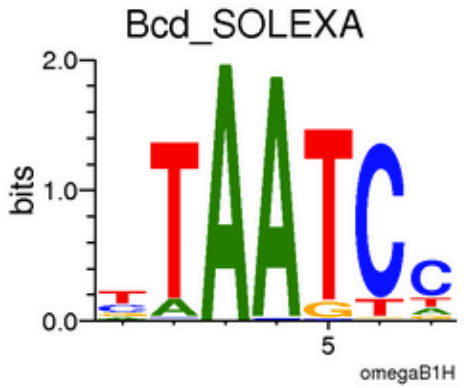
- Fly Factor Survey: http://pgfe.umassmed.edu/TFDBS/
  - Drosophila specific
  - In analyzing insect genomes, motifs from this database can be used, with some additional checks.

Sample entry in Fly Factor Survey

Summary page for FlybaseID: FBgn0000166 from the database of Drosophila TF DNA-binding Specificities

Gene Name	bicoid
Gene Symbol	bcd
Secondary ID	CG1034
Synonyms	BG:DS00276.7 Bcd CG1034 bcd bic mum prd4
Protein ID	BCD_DROME
Unipro ID	P09081
FlyMine	Link to FlyMine
Primary DNAbindingDomain	Homeobox
Secondary DNAbindingDomain	
Pfam Domain	Homeobox

Motif



Reverse Complement

Frequency Matrix

Vertical Count
Horizontal Count
Horizontal PSPM (Probability)
Horizontal PSSM (logodds) ?
Aligned Sequence
Unique Raw Sequence

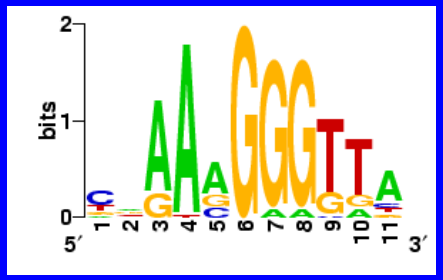
Other Information

Motif frequency in Drosophila genome	Genome Surveyor
Source	B1H
Sequence Method	SOLEXA
PubmedID	18585360
Vector	omegaUV2zf
Inhibitor Concentration	5 mM
Inducer Concentration	10 µM
AA sequence of fragment	PRRTRTTFTSSQIAELEQHFLQGRYLTAPRLADLSAKLALGTAQVKIWF KNRRRRHKIQSDQHKDQSYEG

Step 2. Finding motif matches in DNA

- Basic idea:

Motif:



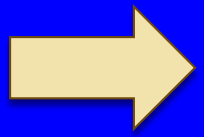
Match: CAAAAGGGTTA
Apprx. Match: CAAAAGGGGTA

- To score a single site s for match to a motif **W**, we use

$$\Pr(s \mid W)$$

What is Pr (s | W)?

5	0	2	0	0	2	0	**A**
0	5	3	1	0	0	0	**C**
0	0	0	3	5	0	0	**G**
0	0	0	1	0	3	5	**T**



1	0	0.4	0	0	0.4	0	**A**
0	1	0.6	0.2	0	0	0	**C**
0	0	0	0.6	1	0	0	**G**
0	0	0	0.2	0	0.6	1	**T**

Now, say s =ACCGGTT (consensus)

$Pr(s|W) = 1 \times 1 \times 0.6 \times 0.6 \times 1 \times 0.6 \times 1 = 0.216.$

Then, say s = ACACGTT (two mismatches from consensus)

$Pr(s|W) = 1 \times 1 \times 0.4 \times 0.2 \times 1 \times 0.6 \times 1 = 0.048.$

Scoring motif matches with "LLR"

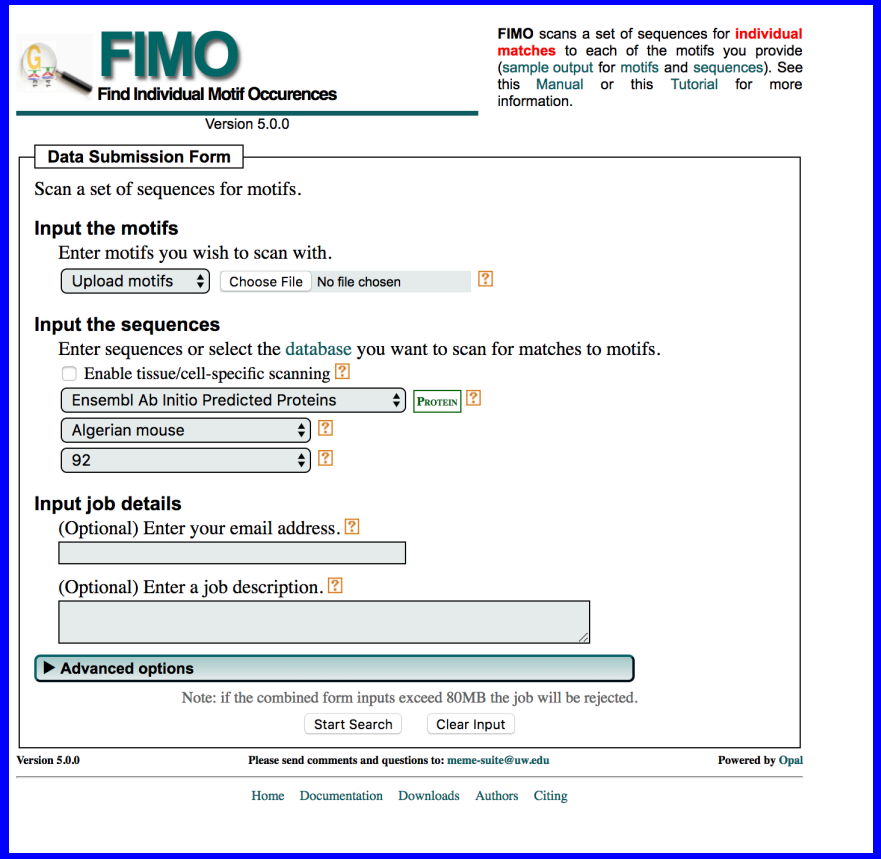
- Pr (s | W) is the key idea.
- However, some statistical massaging is done on this.
- Given a motif W, background nucleotide frequencies W_b and a site s,
- LLR score of *s* =

$$\log \frac{\Pr(s \mid W)}{\Pr(s \mid W_b)}$$

- Good scores > 0. Bad scores < 0.

The FIMO program

- Grant, Bailey, Noble; *Bioinformatics* 2011.

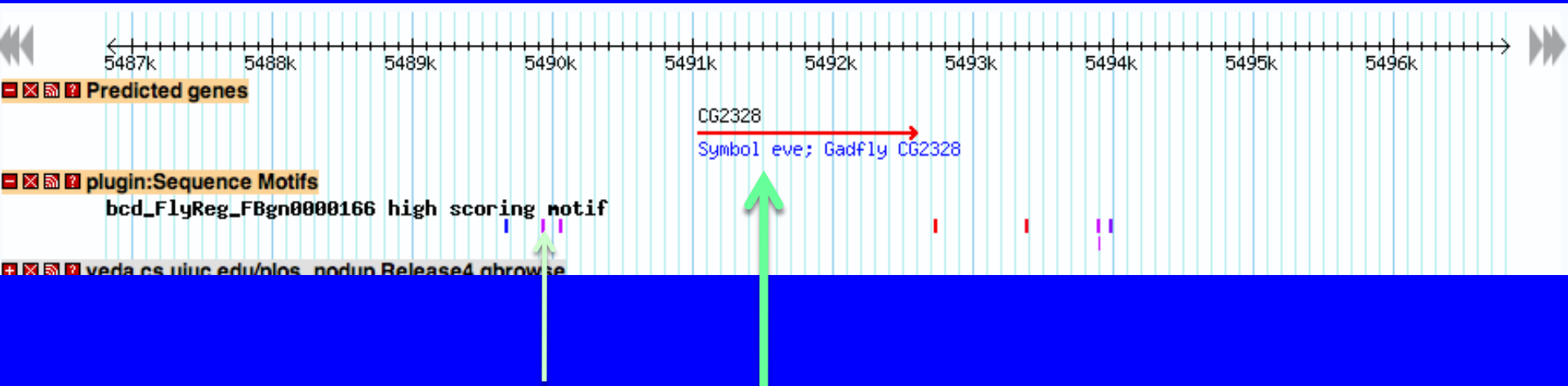


- Takes motif W, background W_b and a sequence *S*.
- Scans every site *s* in *S*, and computes its LLR score.
- Uses sound statistics to deduce an appropriate (p-value) threshold on the LLR score. All sites above threshold are predicted as binding sites.

Finding TF targets

- Step 1. Determine the binding specificity of a TF
- Step 2. Find motif matches in DNA
- Step 3. Designate nearby genes as TF targets

Step 3: Designating genes as targets



Predicted binding sites for motif of TF called “bcd”

Designate this gene as a target of the TF

Sub-goal: discover the genes regulated by a transcription factor … by DNA sequence analysis

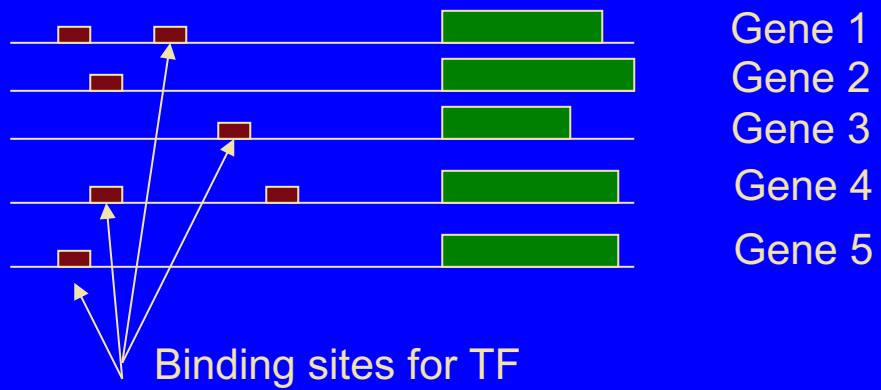
Computational motif discovery

Why?

- We assumed that we have experimental characterization of a transcription factor's binding specificity (motif)
- What if we don't?
- There's a couple of options …

Option 1

- Suppose a transcription factor (TF) regulates five different genes
- Each of the five genes should have binding sites for TF in their promoter region



Option 1

- Now suppose we are given the promoter regions of the five genes G1, G2, … G5
- Can we find the binding sites of TF, without knowing about them *a priori* ?
- This is the computational motif finding problem
- To find a motif that represents binding sites of an unknown TF

Option 2

- Suppose we have ChIP-chip or ChIP-Seq data on binding locations of a transcription factor.

- Collect sequences at the peaks
- Computationally find the motif from these sequences
- This is another version of the motif finding problem

Motif finding algorithms

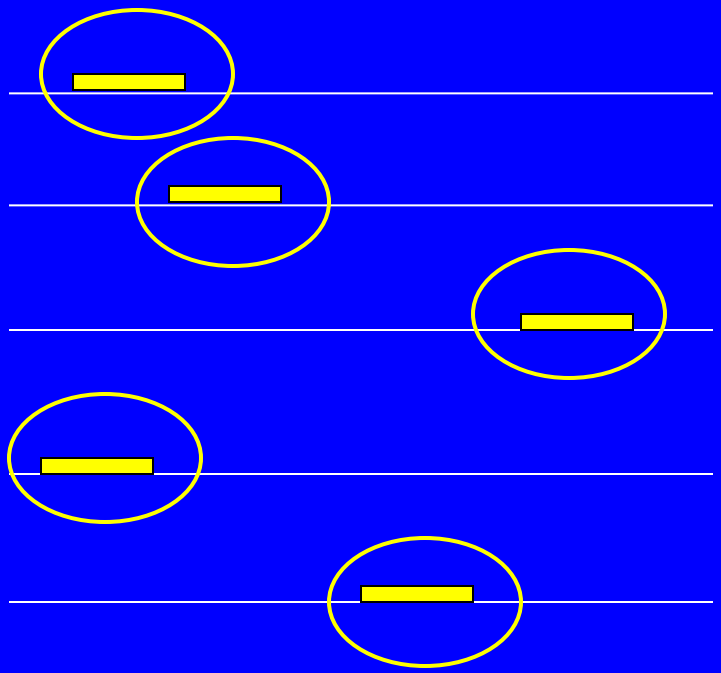
- Version 1: Given promoter regions of co-regulated genes, find the motif
- Version 2: Given bound sequences (ChIP peaks) of a transcription factor, find the motif
- Idea: Find a motif with many (surprisingly many) matches in the given sequences

Motif finding algorithms

- Gibbs sampling (MCMC) : Lawrence et al. 1993
- MEME (Expectation-Maximization) : Bailey & Elkan 94. (Very popular, visited in today's lab.)
- CONSENSUS (Greedy search) : Stormo lab.
- Priority (Gibbs sampling, but allows for additional prior information to be incorporated): Hartemink lab.
- Many many others …

Examining one such algorithm

The "CONSENSUS" algorithm

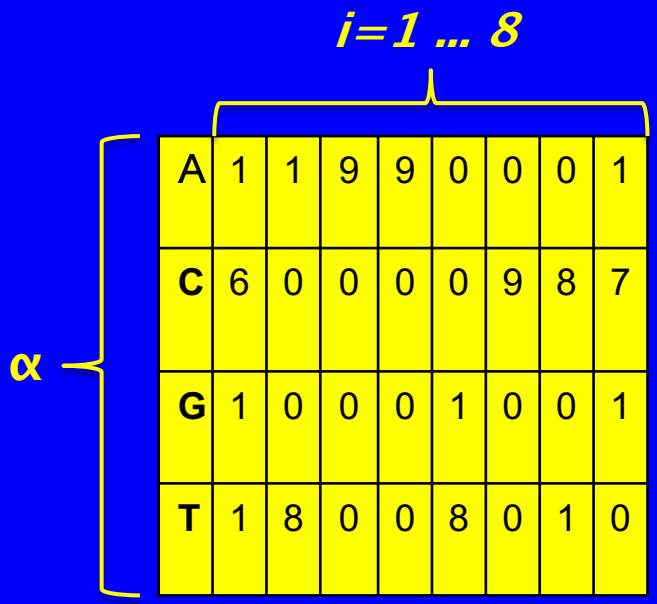


Final goal: Find a set of substrings, one in each input sequence

Set of substrings define a motif. Goal: This motif should have high "information content".

High information content means that the motif "stands out".

The "CONSENSUS" algorithm



Compute information content of motif:

For each column,

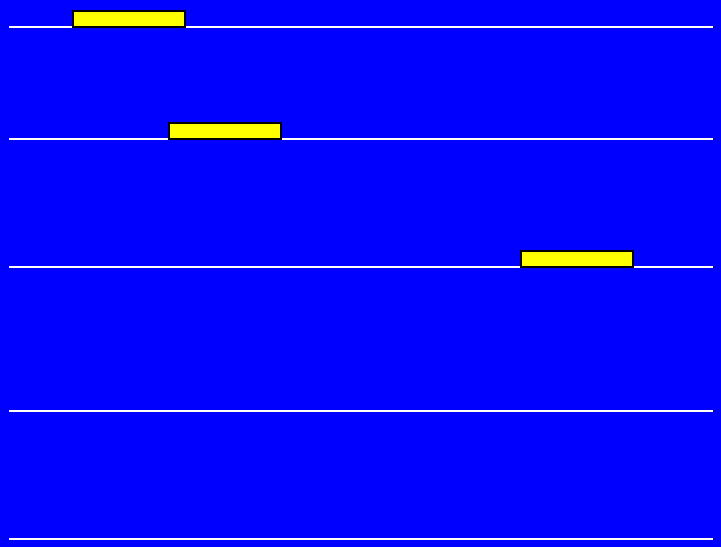
Compute information content of column

$$\sum_{\alpha} W_{i\alpha} \log \frac{W_{i\alpha}}{0.25}$$

Sum over all columns

High information content means that the motif "stands out".

The “CONSENSUS” algorithm

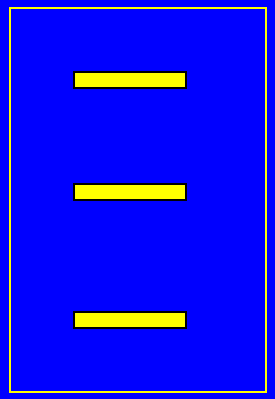


Start with a substring in one input sequence

Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

The "CONSENSUS" algorithm



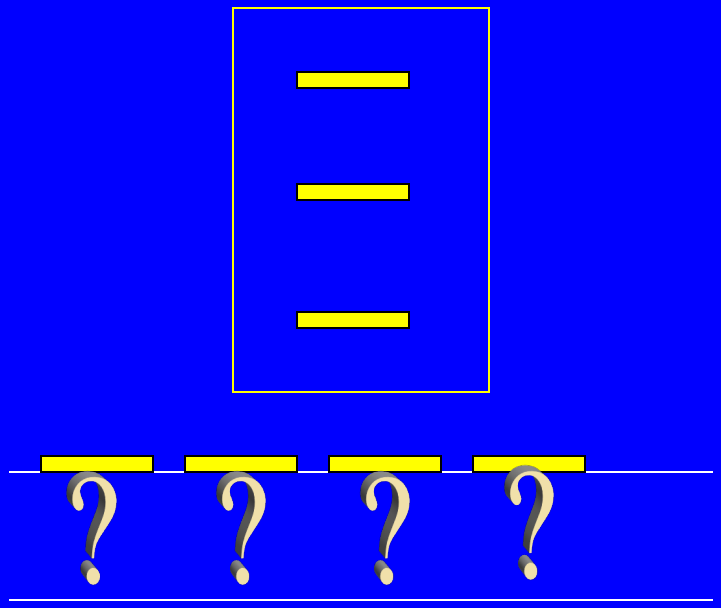
Start with a substring in one input sequence

Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

The current motif.

The "CONSENSUS" algorithm



Start with a substring in one input sequence

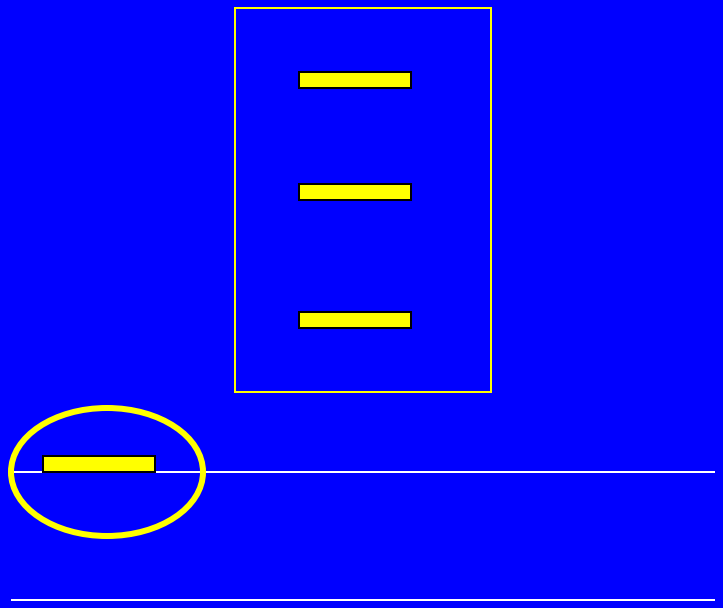
Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

The current motif.

Consider every substring in the next sequence, try adding it to current motif and scoring resulting motif

The "CONSENSUS" algorithm



Start with a substring in one input sequence

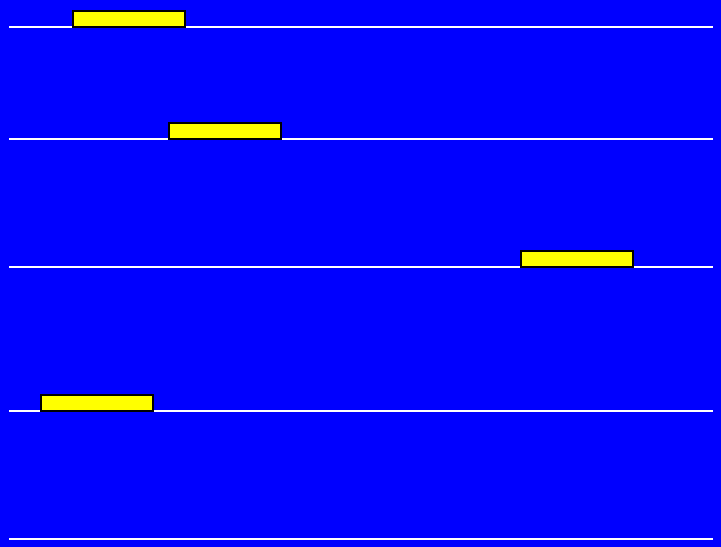
Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

The current motif.

Pick the best one ….

The "CONSENSUS" algorithm



Start with a substring in one input sequence

Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

The current motif.

Pick the best one ….

… and repeat

Summary so far

- To find genes regulated by a TF
  - Determine its motif experimentally
  - Scan genome for matches (e.g., with FIMO & the LLR score)
- Motif can also be determined computationally
  - From promoters of co-expressed genes
  - From TF-bound sequences determined by ChIP assays
  - MEME, CONSENSUS, etc.

Further reading

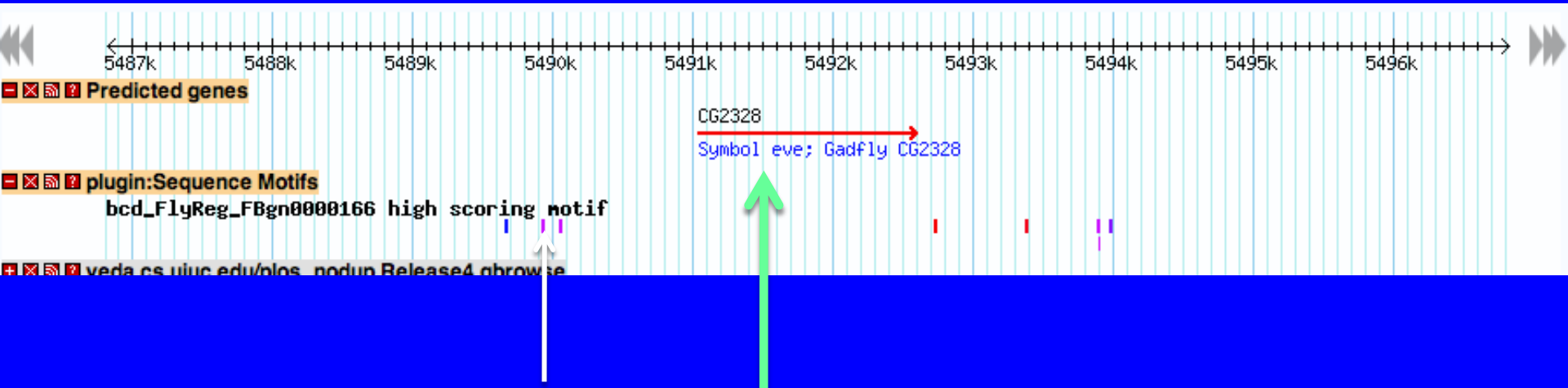
- Introduction to theory of motif finding
  - Moses & Sinha: http://www.moseslab.csb.utoronto.ca/Moses_Sinha_Bioinf_Tools_apps_2009.pdf
  - Das & Dai: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2099490/pdf/1471-2105-8-S7-S21.pdf

Motif finding tools

- MEME: http://meme-suite.org/
- Weeder: http://159.149.160.88/modtools/
- CisFinder: http://lgsun.grc.nia.nih.gov/CisFinder/
- RSAT: http://rsat.sb-roscoff.fr/

Motif scanning

Recap Step 3: Designating genes as targets

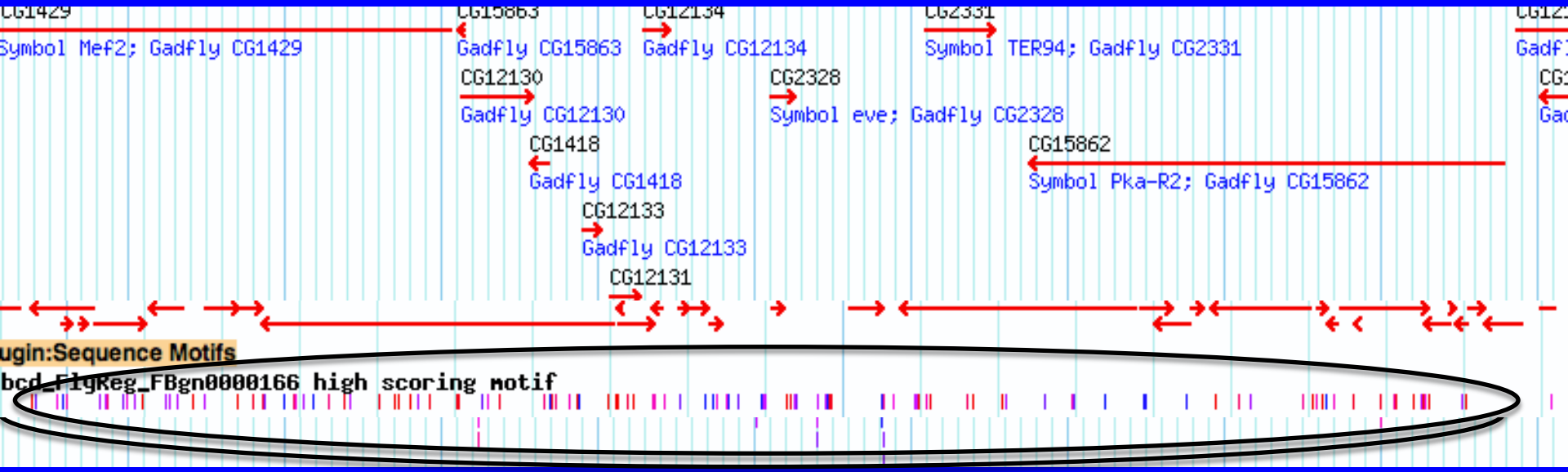


Predicted binding sites for motif of TF called “bcd”

Designate this gene as a target of the TF

But there is a problem …

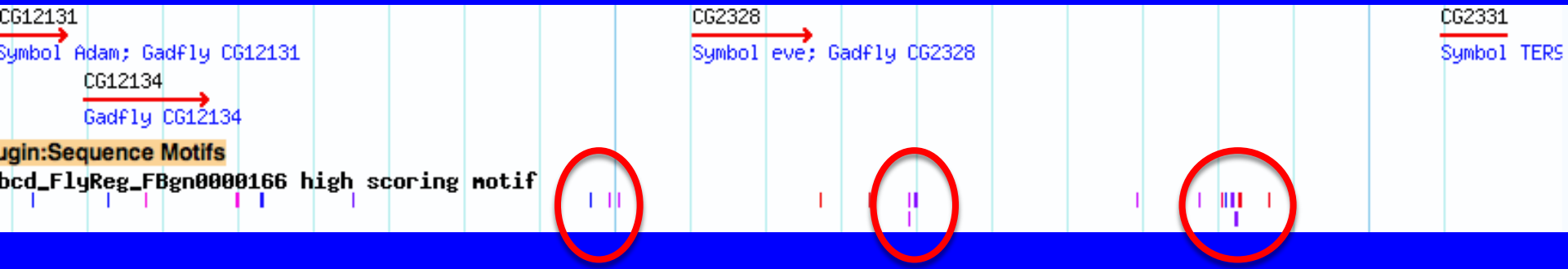
Too many predicted sites !



- An idea: look for clusters of sites (motif matches)

Why clusters of sites?

- Because functional sites often do occur in clusters.



- Because this makes biophysical sense. Multiple sites increase the chances of the TF binding there.

On looking for "clusters of matches"

- To score a single site ***s*** for match to a motif ***W***, we use

length ~10

$$\Pr(s \mid W)$$

- To score a sequence ***S*** for a cluster of matches to motif ***W***, we use

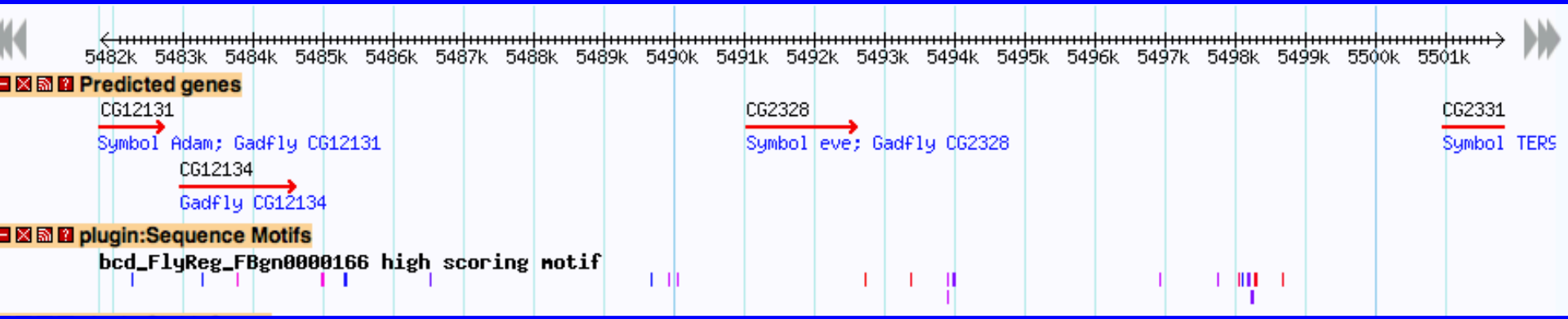
length ~1000

$$\Pr(S \mid W)$$

- But what is this probability?
- A popular approach is to use "Hidden Markov Models" (HMM)

The HMM score profile

- This is what we had, using LLR score:



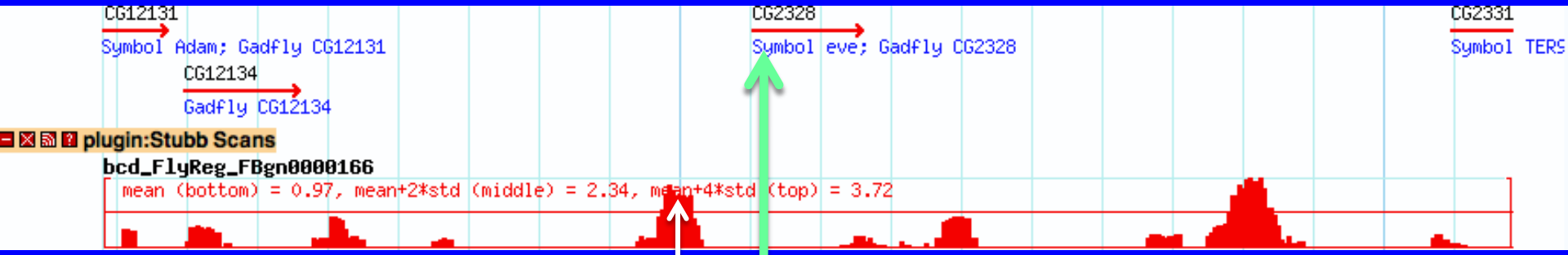
- This is what we have, with the
- These are the functional targets of the TF

Nucleic Acids Res. 2015 Apr 30; 43(8): 3998–4012. PMCID: PMC4417154
Published online 2015 Mar 19. doi: 10.1093/nar/gkv195 PMID: 25791631
NAR Breakthrough Article

Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism

Charles Blatti,[1] Majid Kazemian,[2] Scot Wolfe,[3,4] Michael Brodsky,[3,5] and Saurabh Sinha[1,6,*]

Step 3: Designating genes as targets

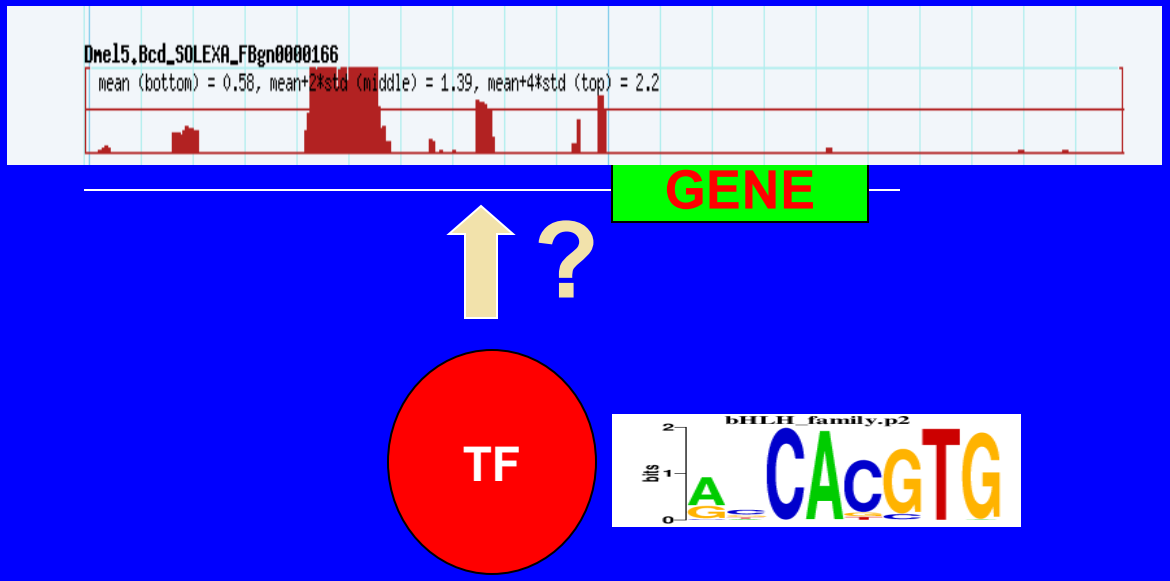


HMM Score for binding site presence
In 1000 bp window centered here.
Designate this gene as a target of the TF

Sub-goal: discover the genes regulated by a transcription factor … by DNA sequence analysis

Integrating sequence analysis and expression data

1. Predict regulatory targets of a TF

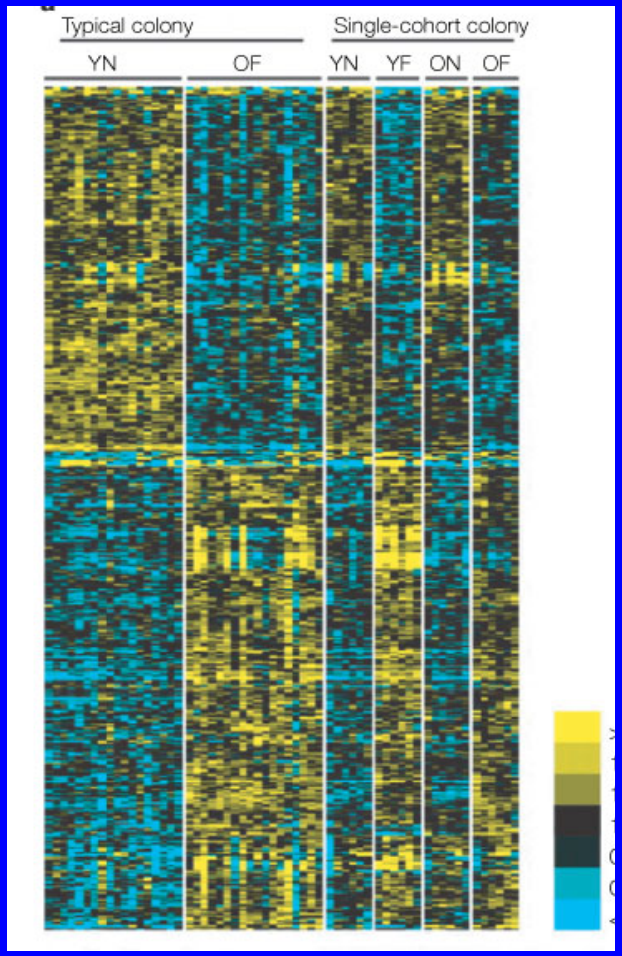


Motif module: a set of genes predicted to be regulated by a TF (motif)

2. Identify dysregulated genes in phenotype of interest

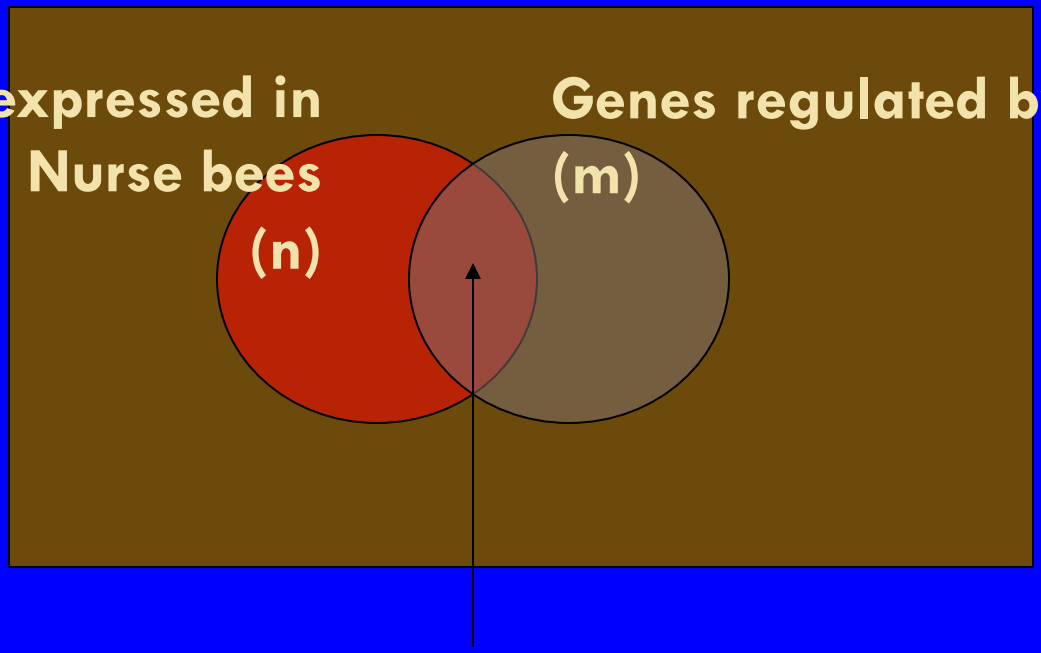
- Honeybee whole brain transcriptomic study

Genes up-regulated in nurses vs foragers



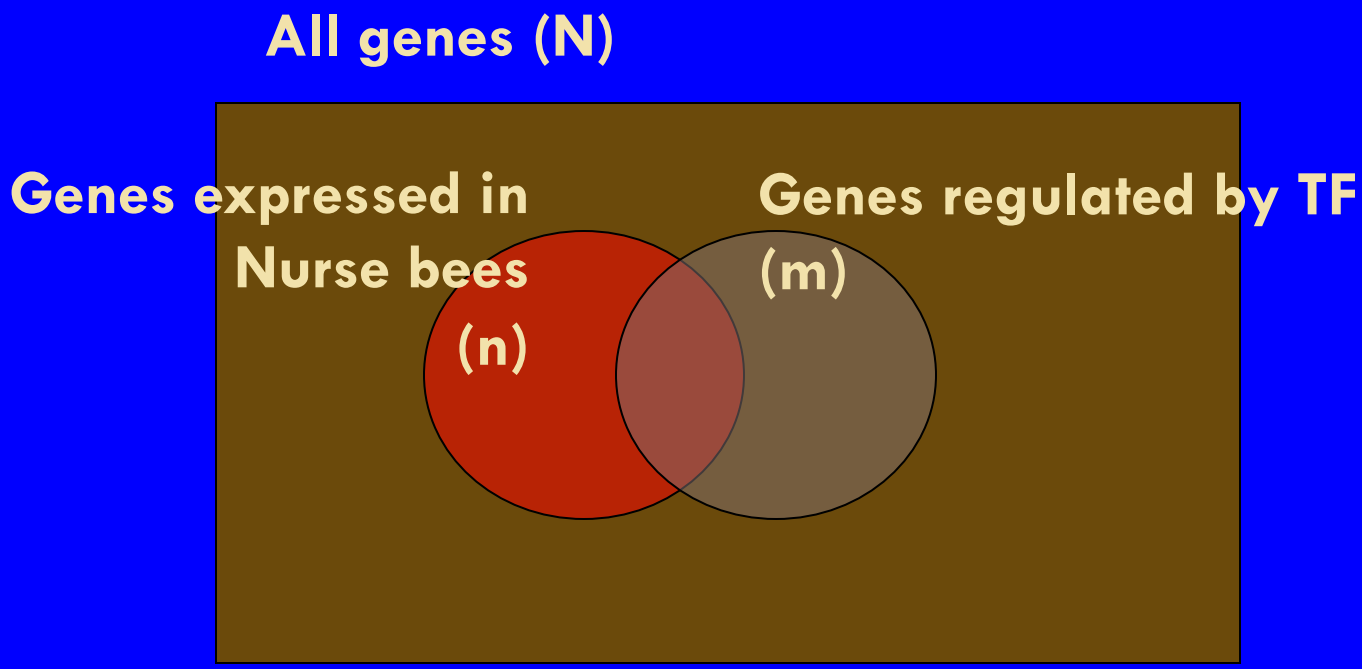
Robinson et al 2005

3. Combine motif analysis and gene expression data



Is the intersection (size "k") significantly large, given N, m, n?

3. Combine motif analysis and gene expression data



Infer: TF may be regulating "Nurse" genes.
An "association" between motif and condition

Hypergeometric Distribution

Given that **n** of **N** genes are labeled "Nurse high".
If we picked a random sample of **m** genes, how likely is an intersection **equal to k** ?

$$f(k; N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

Hypergeometric Distribution

Given that **n** of **N** genes are labeled "Nurse high".
If we picked a random sample of **m** genes, how likely is an intersection **equal to or greater than k** ?

$$P = \sum_{j \geq k} f(j; N, m, n)$$

Further reading

- Motif scanning and its applications
  - Kim et al. 2010. PMID: 20126523
    - http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000652
  - Sinha et al. 2008. PMID: 18256240. http://genome.cshlp.org/content/18/3/477.long

Useful tools

- GREAT: http://bejerano.stanford.edu/great/public/html/
  - Input a set of genomic segments (e.g., ChIP peaks)
  - Obtain what annotations enriched in nearby genes
  - only for human, mouse and zebrafish

- DAVID: https://david.ncifcrf.gov/
  - Input a set of genes
  - Obtain what annotations enriched in those genes
  - Many different species

Questions ?