# Gene Regulation and Computation

Voichita D. Marinescu, PhD
Children’s Hospital Informatics Program

6.872/HST.950 April 15th, 2004

---

## Topics for today

- The transcriptional machinery and its regulation
- The role of transcriptional regulation in development
- Representations of transcriptional regulation components from a computational point of view
- Popular tools and methods available
- Examples

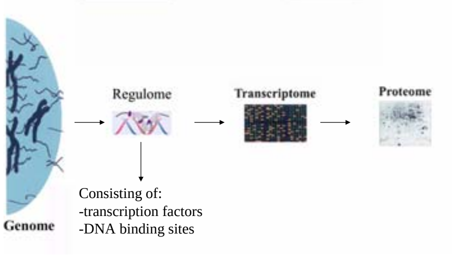
---

## Definition

- Gene regulation: a complex mechanism that allows the cell to vary the amount of mRNA produced for a gene
- Consequences:
  - mediating the transition from genotype to phenotype
  - ensuring adaptation to environmental changes
  - making differentiation possible

---

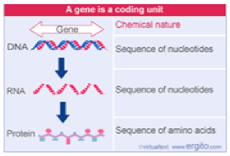
## The regulome

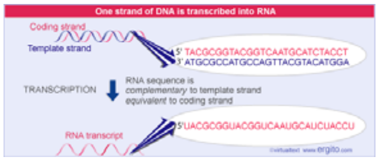


Consisting of:
-transcription factors
-DNA binding sites

---

## The transcriptional mechanism

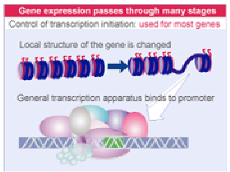


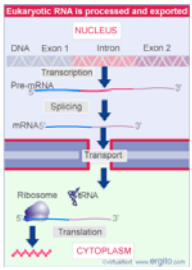


(From Lewin, *Genes VIII*, 2004)

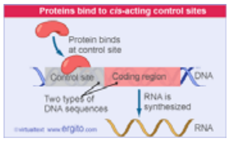
---

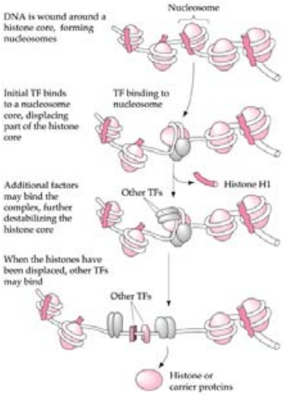
## Stages and compartmentalization



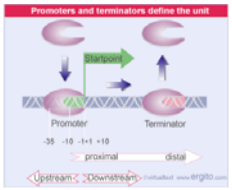


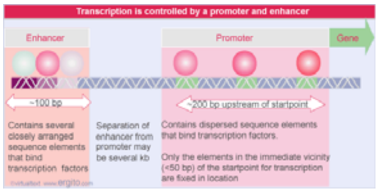
## Key players: transcription factors



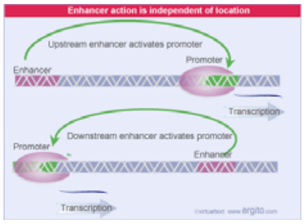


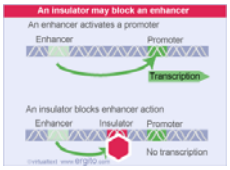
## Transcription factor binding sites and modules



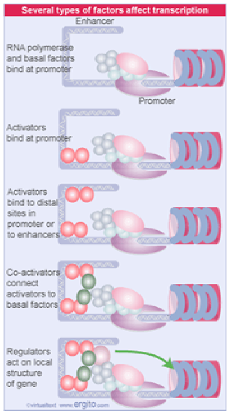


## Enhancers and Insulators

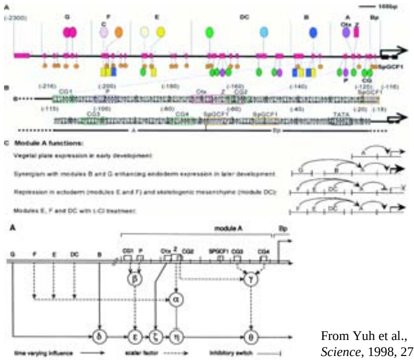




## Coactivators



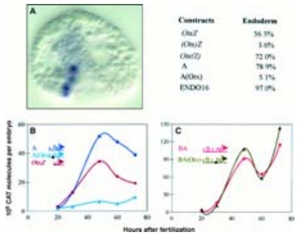
## The complexity of gene regulation



From Yuh et al.,
*Science*, 1998, 279(5358)

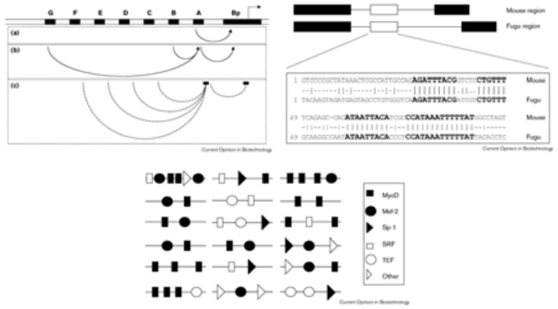
## The importance of gene regulation

Temporal and spatial gene expression patterns control:
- developmental processes
- morphogenesis
- cell differentiation
- tissue specificity
- stress responses



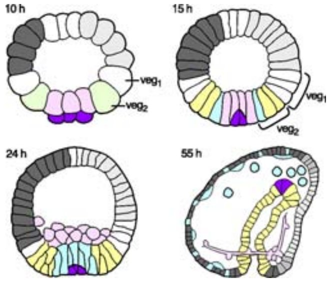
From Yuh et al.,
*Science*, 1998, 279(5358)

## The combinatorial control of eukaryotic transcription



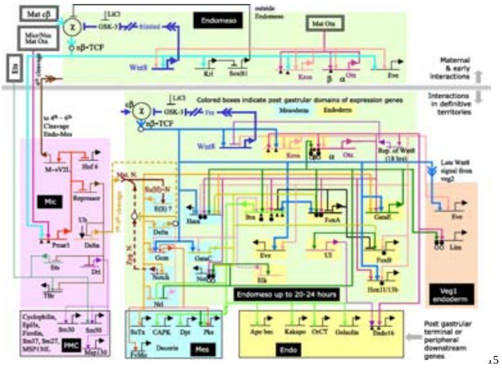
From Fickett and Wasserman, *Curr.Opin.Biotechnol.*, 2000, 11(1)

## The sea urchin model for a genomic regulatory network



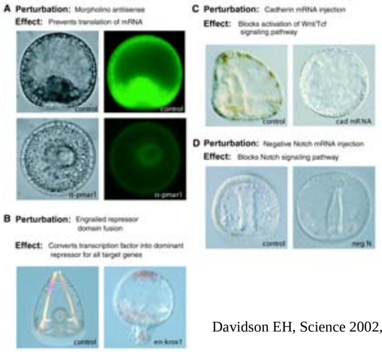
Davidson EH, Science 2002, 295:1669

## The regulatory network for endomesoderm specification



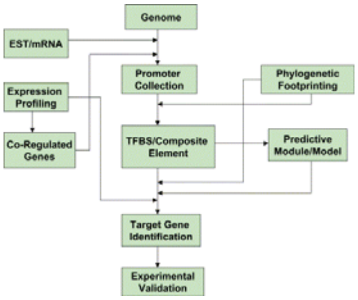
Davidson EH, Science 2002, 295:1669

## Perturbations used in network analysis



Davidson EH, Science 2002, 295:1669

## How can we get there for other organisms?



Genome

EST/mRNA

Promoter Collection

Phylogenetic Footprinting

Expression Profiling

Co-Regulated Genes

TFBS/Composite Element

Predictive Module/Model

Target Gene Identification

Experimental Validation

(From Qiu, *Biochem. Biophys. Res. Commun*, 2003, 495-501)

## Functional genomics insights

- Assumption: co-regulation = same cluster + common regulatory motifs
- Cluster analysis of combined yeast data sets:
- Can expression clusters be predicted?
- Can clusters be used to predict structurally related genes?

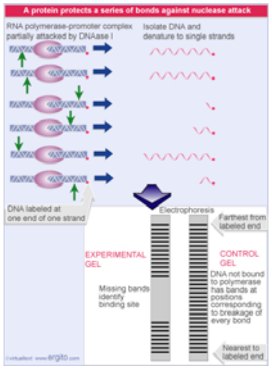
From Eisen et al., *PNAS*, 1998, 95(25)

## The simple scenario

- Perform genome-wide screening for identifying regulatory elements
- Group the genes in classes sharing common motifs
- Group the genes in clusters based on the expression profile
- Look for the intersection of the two sets

## Finding putative regulatory elements (e.g. TFBSs)

- TFBS = a short sequence of DNA (6-20bp)
- Conserved across the target genes
- Usually identified experimentally by molecular biology techniques
- Use this information to abstract a pattern



## Deterministic TFBS representations

– Exact string matching (e.g. “CAGTTCA”, “CATATGA”, “CAACTCA”, “CACGTGA”)

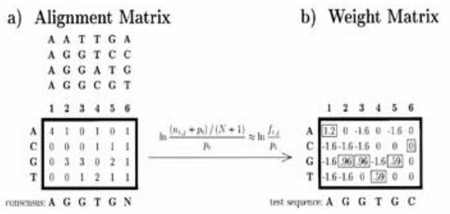
– Regular expressions (e.g. “CA-$X_2$-T-(C/G)-A”)

## Probabilistic TFBS representations

- Position weight matrices (PMWs)
- oligonucleotide frequency analysis:
  - word counting algorithm
  - overrepresentation
- dyad analysis:
  - many regulatory sites: XXX-$(Y)_n$-XXX
  - count number of occurrences
- HMMs

## Position weight matrices

By far the most popular approach



N – total number of sequences
$n_{i,j}$ – number of times nucleotide *i* was observed in position *j*
$p_i$ – *a priori* probability of letter *i*
$f_{i,j}$ – frequency of letter *i* at position *j*

## PWMs – major disadvantages

- Assumes that the positions are independent
- Overemphasize a core matrix (3-4 bp in length)
- Search methods based on them are prone to a very high number of false positives

## Popular tools available
## I. Databases: TRANSFAC

- http://www.gene-regulation.com/index.html
- Large and frequently updated database containing information on TF and their binding sites in target genes
- Manually currated
- Based on experimental data

---

## TRANSFAC data files

MATRICES

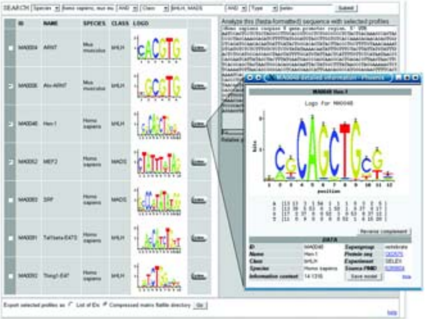
```
AC  M00761
ID  V$P53_DECAMER_Q2
NA  p53 decamer
XX
BF  T00671; p53; Species: human, Homo sapiens.
BF  T01806; p53; Species: mouse, Mus musculus.
BF  T04997; p53; Species: rat, Rattus norvegicus.
XX
P0      A      C      G      T
01     25      3     16      2      R
02     14      0     32      0      G
03     25      0     21      0      R
04      2     39      4      1      C
05     32      2      4      8      A
06     23      2      2     19      W
07      3      0     43      0      G
08      9     15      5     17      N
09      2     28      9      7      C
10      5     22      5     14      Y
XX
BA  46 compiled sequences
XX
BS  R04373; Start: 1; Length: 10; Gaps:; Orientation: p.
BS  R04373; Start: 11; Length: 10; Gaps:; Orientation: p.
-----------------------------------------------------
BS  R11629; Start: 21; Length: 10; Gaps:; Orientation: n.
BS  R11629; Start: 31; Length: 10; Gaps:; Orientation: p.
XX
//
```

FACTORS

```
AC  T00505
FA  MEF-2
SY  MEF-2; MEF-2A.
OS  mouse, Mus musculus
-------------------------
XX
MX  M00006; V$MEF2_01.
MX  M00231; V$MEF2_02.
MX  M00232; V$MEF2_03.
MX  M00233; V$MEF2_04.
MX  M00405; V$MMEF2_Q6.
MX  M00405; V$MMEF2_Q6.
XX
BS  R03583; AS$MEF2_01; Quality: 6.
BS  R03586; CHICK$CTNT_05; Quality: 6; cTNT, G000057; chick
BS  R09144; HS$DSMN_01; Quality: 6; DSMN, G001827; human
BS  R04212; HS$PGAMM_01; Quality: 4; PGAM-M, G000361; human
BS  R02201; MEF2$CONS_01; Quality: 6.
BS  R04530; MEF2$CONS_02; Quality: 6.
BS  R04804; MEF2$CONS_03; Quality: 6.
BS  R04805; MEF2$CONS_04; Quality: 6.
BS  R04806; MEF2$CONS_05; Quality: 6.
BS  R00244; MOUSE$MCK_05; Quality: 1; mck, G000557; mouse
BS  R03585; MOUSE$MCK_07; Quality: 6; mck, G000557; mouse
BS  R03587; RAT$AMHC_03; Quality: 6; MHC-A, G000694; rat
BS  R03588; RAT$AMHC_04; Quality: 6; MHC-A, G000694; rat
BS  R03589; RAT$MHC_02; Quality: 6; MHC, G000770; rat
BS  R03584; RAT$MLCC_01; Quality: 6; MLC2, G000771; rat
XX
---------------------------------
DR  SwissProt: Q60929; MEFA_MOUSE
XX
RN  [1]
RX  MEDLINE; 96104596.
---------------------------------
XX
//
```

---

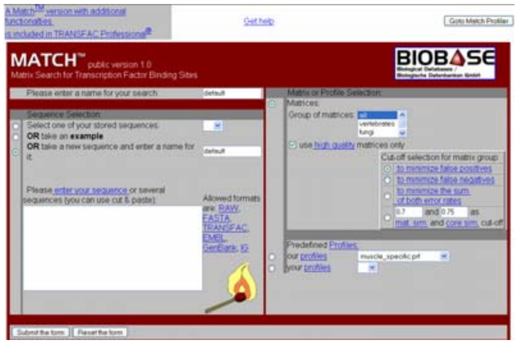
## Databases (contd): JASPAR



http://jaspar.cgb.ki.se/

---

## II. Search engines: Match



Please enter a name for your search: default

OR take a new sequence and enter a name for it: default

Group of matrices: all

Cut-off selection for matrix group: 0.7 and 0.75 as mat. sim. and core sim. cut-off

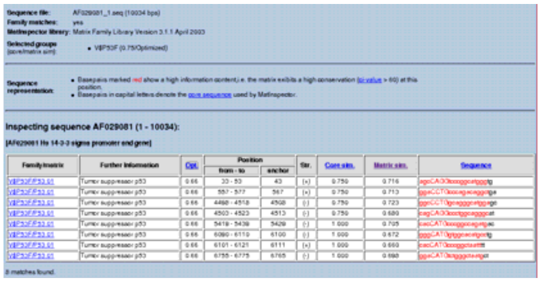
our profiles: muscle_specific.prf

http://www.gene-regulation.com/pub/programs.html

---

## Search engines (contd.): MatInspector

**MatInspector output for sequence AF029081**



**Sequence file:** AF029081_1.seq (10034 bps)
**Family matches:** yes
**MatInspector library:** Matrix Family Library Version 3.1.1 April 2003
**Selected groups (core/matrix sim):**
- V$P53F (0.75/Optimized)

**Sequence representation:**
- Basepairs marked red show a high information content,i.e. the matrix exibits a high conservation (ci-value > 60) at this position.
- Basepairs in capital letters denote the core sequence used by MatInspector.

**Inspecting sequence AF029081 (1 - 10034):**

**[AF029081 Hs 14-3-3 sigma promoter and gene]**

| Family/matrix | Further Information | Opt. | Position |  | Str. | Core sim. | Matrix sim. | Sequence |
|---|---|---|---|---|---|---|---|---|
|  |  |  | from - to | anchor |  |  |  |  |
| V$P53F/P53.01 | Tumor suppressor p53 | 0.66 | 33 - 53 | 43 | (+) | 0.750 | 0.716 | [illegible] |
| V$P53F/P53.01 | Tumor suppressor p53 | 0.66 | 557 - 577 | 567 | (+) | 0.750 | 0.713 | [illegible] |
| V$P53F/P53.01 | Tumor suppressor p53 | 0.66 | 4498 - 4518 | 4508 | (-) | 0.750 | 0.723 | [illegible] |
| V$P53F/P53.01 | Tumor suppressor p53 | 0.66 | 4503 - 4523 | 4513 | (-) | 0.750 | 0.680 | [illegible] |
| V$P53F/P53.01 | Tumor suppressor p53 | 0.66 | 5418 - 5438 | 5428 | (-) | 1.000 | 0.705 | [illegible] |
| V$P53F/P53.01 | Tumor suppressor p53 | 0.66 | 6090 - 6110 | 6100 | (-) | 1.000 | 0.672 | [illegible] |
| V$P53F/P53.01 | Tumor suppressor p53 | 0.66 | 6101 - 6121 | 6111 | (+) | 1.000 | 0.660 | [illegible] |
| V$P53F/P53.01 | Tumor suppressor p53 | 0.66 | 6755 - 6775 | 6765 | (-) | 1.000 | 0.698 | [illegible] |

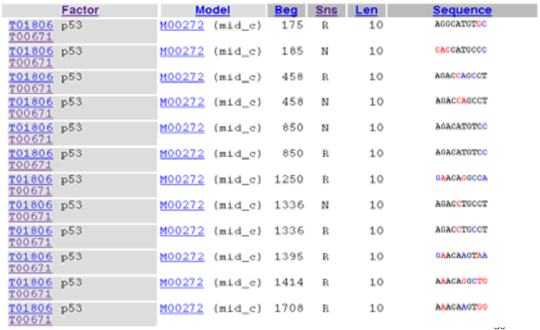
8 matches found.

http://www.gene-regulation.com/pub/programs.html

---

## Search engines (contd.):TESS

**TESS output for nucleotides 6000-8000 of sequence AF029081**



| Factor | Model | Beg | Sns | Len | Sequence |
|---|---|---|---|---|---|
| T01806 p53<br>T00671 | M00272 (mid_c) | 175 | R | 10 | AGGCATGTGC |
| T01806 p53<br>T00671 | M00272 (mid_c) | 185 | N | 10 | CACCATGCCC |
| T01806 p53<br>T00671 | M00272 (mid_c) | 458 | R | 10 | AGACCAGCCT |
| T01806 p53<br>T00671 | M00272 (mid_c) | 458 | N | 10 | AGACCAGCCT |
| T01806 p53<br>T00671 | M00272 (mid_c) | 850 | N | 10 | AGACATGTCC |
| T01806 p53<br>T00671 | M00272 (mid_c) | 850 | R | 10 | AGACATGTCC |
| T01806 p53<br>T00671 | M00272 (mid_c) | 1250 | R | 10 | GAACAGGCCA |
| T01806 p53<br>T00671 | M00272 (mid_c) | 1336 | N | 10 | AGACCTGCCT |
| T01806 p53<br>T00671 | M00272 (mid_c) | 1336 | R | 10 | AGACCTGCCT |
| T01806 p53<br>T00671 | M00272 (mid_c) | 1395 | R | 10 | GAACAAGTAA |
| T01806 p53<br>T00671 | M00272 (mid_c) | 1414 | R | 10 | AAACAGGCTG |
| T01806 p53<br>T00671 | M00272 (mid_c) | 1708 | R | 10 | AAACAAGTGG |

http://www.cbil.upenn.edu/tess/

## Search engines (contd.)

# MDscan

## A Fast and Accurate Motif Finding Algorithm With Applications To Chromatin Immunoprecipitation Microarray Experiments

Xiaole Shirley Liu, Douglas L. Brutlag, Jun S. Liu
Stanford Medical Informatics, Stanford University

While chromatin immunoprecipitation followed by cDNA microarray hybridization (ChIP-array) has become a popular procedure for studying genome-wide protein-DNA interactions and transcription regulation, it can only map the probable protein-DNA interaction loci within 1-2kb resolution. To pinpoint the interaction sites down to the base pair level, we introduce a novel computational method, Motif Discovery scan (MDscan), that examines the ChIP-array selected sequences and searches for DNA sequence motifs representing the protein-DNA interaction sites. MDscan combines the advantages of two widely adopted motif search strategies, word enumeration and position-specific weight matrix updating, and incorporates the ChIP-array ranking information to accelerate the search and enhance its success rate. MDscan correctly identified all experimentally verified motifs from published ChIP-array experiments in yeast, i.e., STE12, GAL4, RAP1 and all the known cell-cycle related motifs SCB, MCB, MCM1, SFF and SWI5, and predicted two motif patterns for the differential binding of Rap1 protein in telomere regions. The method showed both speed and accuracy advantages in our studies compared to several established motif-finding algorithms. MDscan can be used not only with the ChIP-array experiments, but also to find DNA motifs in other experiments in which a subgroup of the sequences can be inferred to contain relatively abundant motif sites.

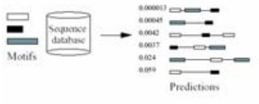
The intuition of MDscan is to first search for groups of similar words in the sequences more likely to contain the motif (high-confidence sequences) because these sequences have higher signal to noise ratio. Words in each similarity group can initialize a position specific motif matrix and the motif can be updated and refined with the whole input sequences.

**Obtaining a local copy of MDscan:**
MDscan is free-of-charge to academia. Please check out Brutlag Bioinformatics Group Software Download and Academic License Instructions for details.

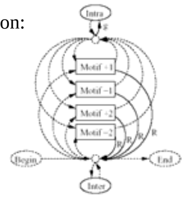
http://bioprospector.stanford.edu/MDscan/

---

## The next step: finding regulatory modules

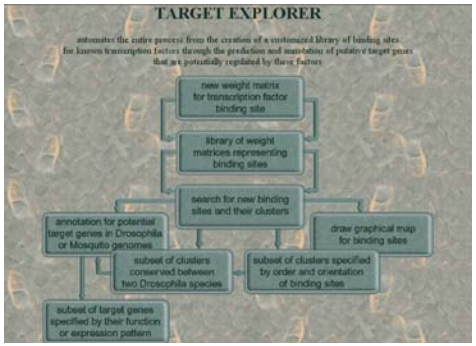


Module representation:



---

## III. Searching for *cis*-regulatory modules



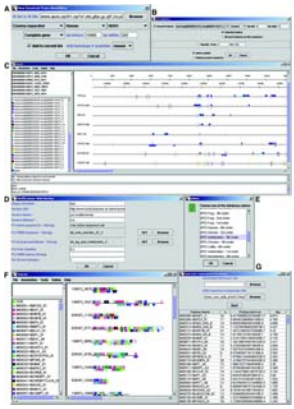
**TARGET EXPLORER**

automates the entire process from the creation of a customized library of binding sites for known transcription factors through the prediction and annotation of putative target genes that are potentially regulated by these factors

http://trantor.bioc.columbia.edu/Target_Explorer/

---

## Searching for modules (contd): Toucan



http://www.esat.kuleuven.ac.be/~saerts/software/toucan.php

---

## Searching for modules (contd.): FlyEnhancer

**Fly Enhancer**

Select a help icon to get help on any part of this form.

Site A
Site B
Site C
Site D
Site E
Site F
Site G
Site H
Site I
Site J

Chromosome Arm: all
Cluster Size: at least 2 binding sites
Window Size: within 1000 base pairs
Boolean Condition:

Please be patient. It takes about one minute to scan the entire genome.

If you have any questions or comments about this service, we'd like to hear from you. Please write to [illegible]

http://opengenomics.org/

---

## Searching for modules: Ahab

**AHAB WEB INTERFACE VERSION 1.0**

**Step 1: DNA Sequence**

You may EITHER upload OR paste your sequence below. Select the corresponding button and enter your sequence. Warning: Please make sure that the total sequence length of each sequence combined is no longer than 1 Megabase.

● upload
○ paste
☑ mark microsatellites

**Step 2: Select transcription factor binding motifs**

You may upload, choose from a selection of motif weight matrices, or do both. Check the boxes next to the one(s) you do.

☐ upload
☑ choose — Note that if you come back to this page by clicking the Browser's Back button, your motif selections will be saved.

**Step 3: Parameters**

Specify the parameters. The current values are recommended default values. Ahab slides a search window of width W nucleotides in steps of S nucleotides along the sequence.

Window Shift S: 30 — In general, S should be at least a third of the window width W. However, the smaller S, the longer the computation time. S smaller than 10 (scale of binding sites) is superfluous.

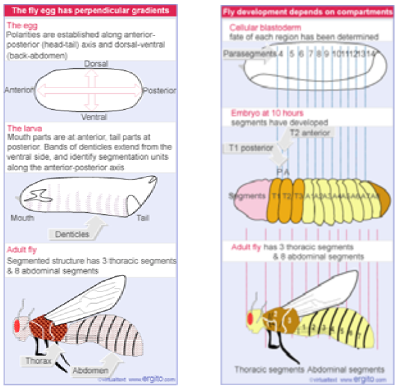
Most eukaryotic minimal enhancer elements are on the scale of 500 nucleotides. We

http://gaspard.bio.nyu.edu/ahab_beta_version.html

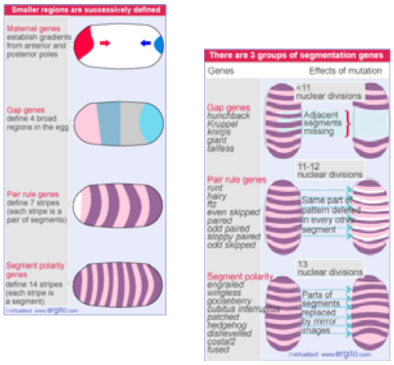
# An example

- Finding cis-regulatory modules involved in body patterning in early Drosophila development
- Tool: Ahab
- Prior knowledge: TFs represented in the modules known, more than 20 modules determined experimentally
- Reference: Rajewsky et al., BMC Bioinformatics, 3:30

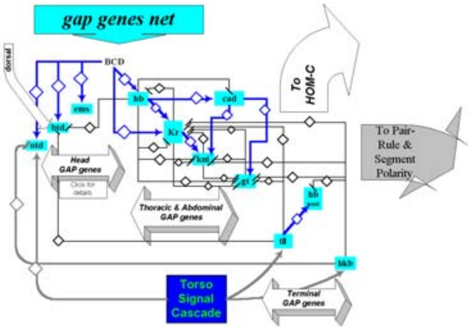
# Segmentation in *Drosophila*



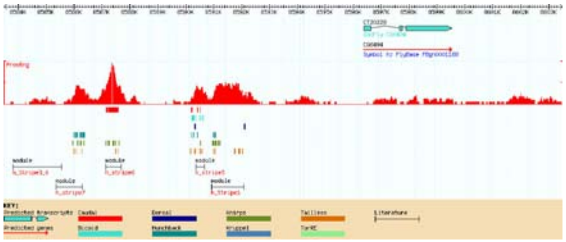
# Segmentation genes in *Drosophila*



# Gap genes network

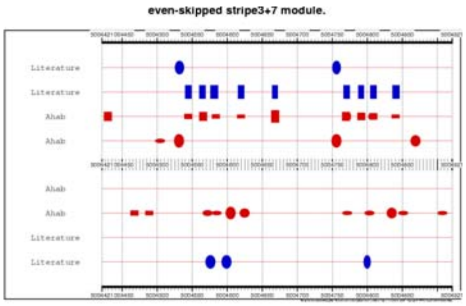


# Ahab prediction for the *hairy* locus



(From Rajewsky et al., *BCM Bioinformatics*, 2002, 3(1):30)

# Ahab prediction (contd.)

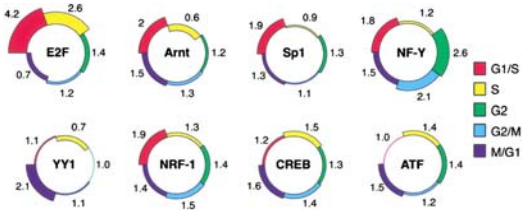


(From Rajewsky et al., *BCM Bioinformatics*, 2002, 3(1):30)

# A second example

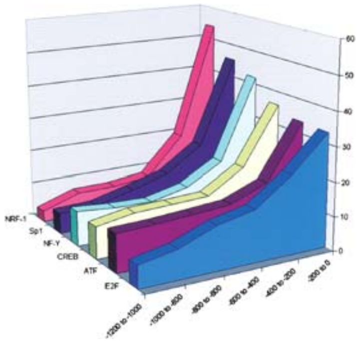
- Correlating overexpression of particular TFs with gene co-regulation
- Looking for synergistic TFs
- Also looking for overrepresentation of a given TF among the genes expressed at a given phase of the cell cycle.
- Reference: Elkon et al., Genome Research, 2003, 13:773-780

# TFBS representation in the cell cycle phase clusters



(From Elkon et al., *Genome Research*, 2003, 13:773-780)

# Distribution of the locations of the putative TFBSs



(From Elkon et al., *Genome Research*, 2003, 13:773-780)

# Conclusions

- Transcriptional regulation is critical for normal development and differentiation
- Known regulatory elements can be represented by powerful models
- The models can be used for scanning for putative regulatory elements and modules
- The combination of computational prediction and experimental validation can provide very valuable insights into the mechanisms of development and pathogenesis.