

Clustering: a bed-time story

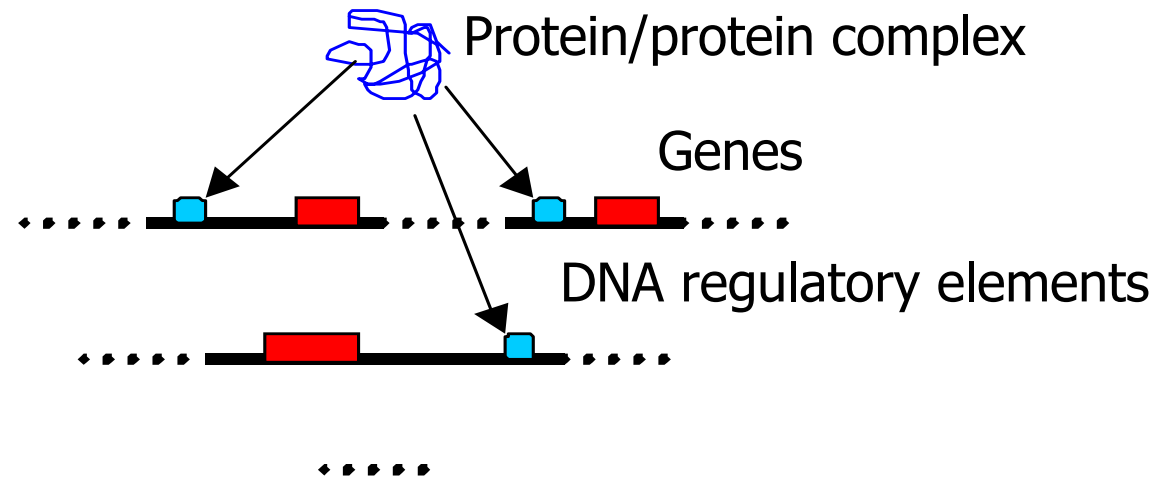
Perry Moerland
Bioinformatics Lab, KEBB, AMC

`p.d.moerland@amc.uva.nl`

Unsupervised learning: clustering

Clustering: grouping together similar objects. Microarray data:

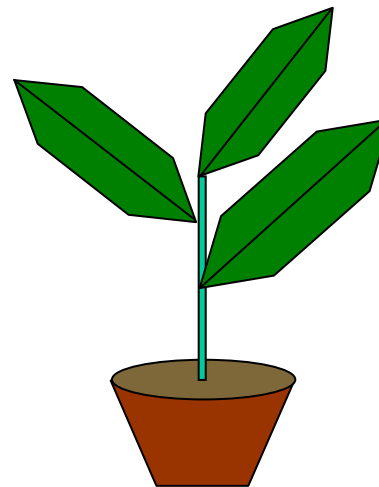
Genes: similar \sim co-expression \sim co-regulation \sim same pathway / same function



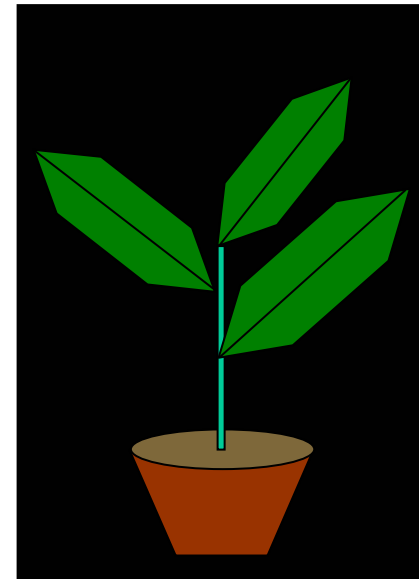
Samples: similar \sim same type of tissue

Used for discovery of new subclasses in a form of cancer

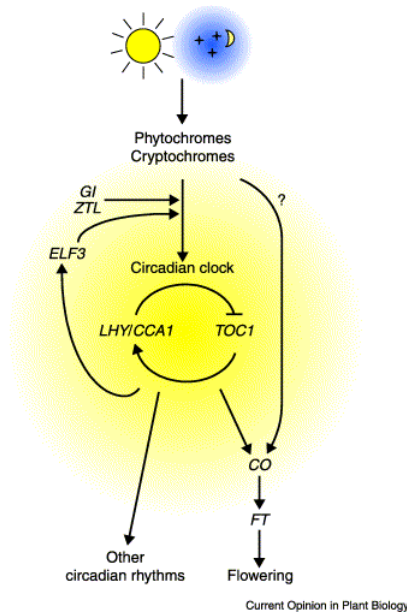
Case study: circadian clock in Arabidopsis



Light



Dark



Experimental design:

μ arrays at 0h=8am, 4h, 8h, ..., 48h

Harmer et al, Science, 290, pp. 2110 – 2113 (2000)

Circadian clock: clusters

- Clock-regulated genes: 437 (6% of the genes on the chip) – correlation with cosine
- Clustered in three groups according to the temporal axis in which the experiment was done
- Cluster 1: genes whose expression peaks in phases 0, 4
- **Cluster 2**: phases 8, 12 – with 191 genes
- Cluster 3: phases 16, 20
- Promoter: 1000 bases upstream
- Motifs: short, recurring patterns in DNA that are presumed to have a biological function

Sites

Site: short sequence containing
some signal

A	C	A	A	T	G
T	C	A	A	T	C
A	C	A	A	G	C
A	G	A	A	T	C
A	C	C	A	T	C

Examples: intron splice sites, transcription start site,
transcription factor binding sites

Goals: - give a mathematical description (**model**) of a site
- find possible sites in a long sequence

Consensus sequence

majority vote:

A C A A T C

A	C	A	A	T	G
T	C	A	A	T	C
A	C	A	A	G	C
A	G	A	A	T	C
A	C	C	A	T	C

W S M A K S

from IUPAC code:

M	A/C
R	A/G
W	A/T
S	C/G
Y	C/T
K	G/T
B	C/G/T
D	A/G/T
H	A/C/T
V	A/C/G
N	A/C/G/T

Regular expressions

[ab]: union {a,b}
ab : concatenation {ab}
 ϵ : empty string
a* : Kleene star { ϵ ,a,aa,aaa, ...}

A	C	A	A	T	G
T	C	A	A	T	C
A	C	A	A	G	C
A	G	A	A	T	C
A	C	C	A	T	C

[AT][CG][AC]A[TG][GC]

A C A A T C , but also T G C A G G

See also <http://au.expasy.org/prosite/>

Weight matrices

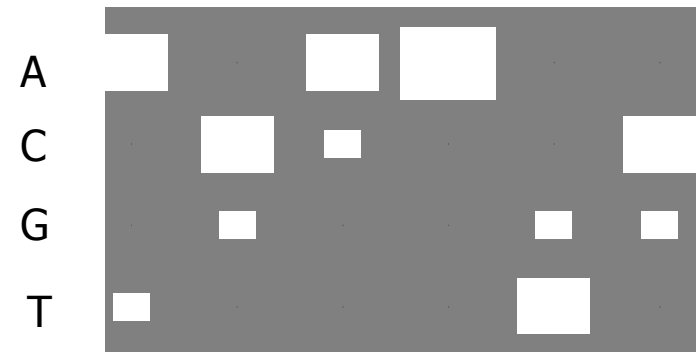
$$\begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ A & \left(\begin{array}{cccccc} 4 & 0 & 4 & 5 & 0 & 0 \\ 0 & 4 & 1 & 0 & 0 & 4 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 4 & 0 \end{array} \right) \\ C & & & & & & \\ G & & & & & & \\ T & & & & & & \end{matrix}$$

counts

A	C	A	A	T	G
T	C	A	A	T	C
A	C	A	A	G	C
A	G	A	A	T	C
A	C	C	A	T	C

probabilities

$$W = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ A & \left(\begin{array}{cccccc} 0.8 & 0.0 & 0.8 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.8 & 0.2 & 0.0 & 0.0 & 0.8 \\ 0.0 & 0.2 & 0.0 & 0.0 & 0.2 & 0.2 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.8 & 0.0 \end{array} \right) \\ C & & & & & & \\ G & & & & & & \\ T & & & & & & \end{matrix}$$



aka position specific score matrix

Weight matrices

Sequence: $x = x_1x_2\dots x_N$

independence

$$P(x_1x_2\dots x_N | W) = \prod_{i=1}^N w_{x_i,i} = \prod_{i=1}^N P_i(x_i | W)$$

$$P(\text{ACAATC} | W) = P_1(\text{A})P_2(\text{C})P_3(\text{A})P_4(\text{A})P_5(\text{T})P_6(\text{C})$$
$$= 0.8 \times 0.8 \times 0.8 \times 1 \times 0.8 \times 0.8 = 0.33$$

	1	2	3	4	5	6
A	0.8	0.0	0.8	1.0	0.0	0.0
C	0.0	0.8	0.2	0.0	0.0	0.8
G	0.0	0.2	0.0	0.0	0.2	0.2
T	0.2	0.0	0.0	0.0	0.8	0.0

Weight matrices

Sequence: $x = x_1x_2\dots x_N$

independence

$$P(x_1x_2\dots x_N | W) = \prod_{i=1}^N w_{x_i,i} = \prod_{i=1}^N P_i(x_i | W)$$

$$P(\text{CCAATC} | W) = P_1(\text{C})P_2(\text{C})P_3(\text{A})P_4(\text{A})P_5(\text{T})P_6(\text{C})$$
$$= 0 \times 0.8 \times 0.8 \times 1 \times 0.8 \times 0.8 = 0$$

	1	2	3	4	5	6
A	0.8	0.0	0.8	1.0	0.0	0.0
C	0.0	0.8	0.2	0.0	0.0	0.8
G	0.0	0.2	0.0	0.0	0.2	0.2
T	0.2	0.0	0.0	0.0	0.8	0.0

Weight matrices: pseudocounts

$$P(x) = \frac{\#x + 1}{\sum_i (\#i + 1)}$$

pseudocount (Laplace)

A	C	A	A	T	G
T	C	A	A	T	C
A	C	A	A	G	C
A	G	A	A	T	C
A	C	C	A	T	C

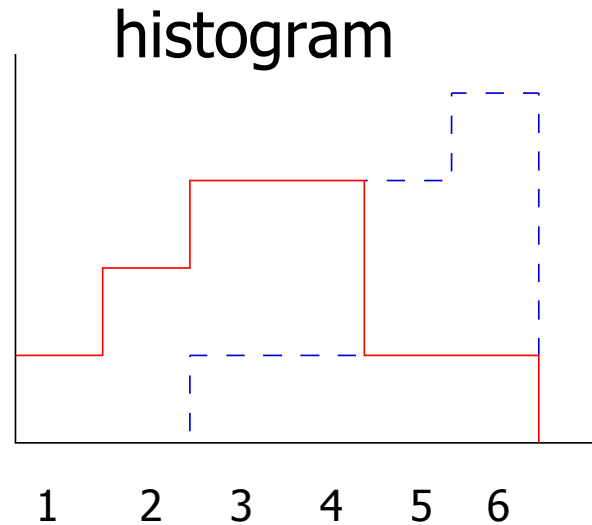
$$\begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{pmatrix} 5 & 1 & 5 & 6 & 1 & 1 \\ 1 & 5 & 2 & 1 & 1 & 5 \\ 1 & 2 & 1 & 1 & 2 & 2 \\ 2 & 1 & 1 & 1 & 5 & 1 \end{pmatrix}$$

$$W' = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{pmatrix} 0.56 & 0.11 & 0.56 & 0.67 & 0.11 & 0.11 \\ 0.11 & 0.56 & 0.22 & 0.11 & 0.11 & 0.56 \\ 0.11 & 0.22 & 0.11 & 0.11 & 0.22 & 0.22 \\ 0.22 & 0.11 & 0.11 & 0.11 & 0.56 & 0.11 \end{pmatrix}$$

$$P(\text{ACAATC} | W') = P_1(A)P_2(C)P_3(A)P_4(A)P_5(T)P_6(C) = 0.56^5 \times 0.67 = 0.037$$

$$P(\text{CCAATC} | W') = P_1(C)P_2(C)P_3(A)P_4(A)P_5(T)P_6(C) = 0.11 \times 0.56^4 \times 0.67 = 0.0072$$

Bayes' decision rule



•	• •	• • •	• • •	•	•
		•	•	• • •	• • • •

— die 1
 - - - - die 2

A x is thrown, from which die does it come?

x is assigned to die 1 $\Leftrightarrow P(\text{die 1} | x) > P(\text{die 2} | x)$

$\Leftrightarrow P(x | \text{die 1})P(\text{die 1}) > P(x | \text{die 2})P(\text{die 2})$

$\Leftrightarrow P(x, \text{die 1}) > P(x, \text{die 2})$

Bayes' decision rule: odds ratio

class A: sites

class B: non-sites

$$x \text{ is assigned to class } A \Leftrightarrow \frac{P(x | \text{class } A)P(A)}{P(x)} > \frac{P(x | \text{class } B)P(B)}{P(x)}$$

$$\Leftrightarrow \frac{P(x | \text{class } A)}{P(x | \text{class } B)} > \frac{P(B)}{P(A)} \rightarrow \text{priors}$$

equal priors:

$$\frac{P(x | \text{class } A)}{P(x | \text{class } B)} > 1 \Leftrightarrow \log \left(\frac{P(x | \text{class } A)}{P(x | \text{class } B)} \right) > 0$$

odds ratio **log-odds ratio**

unequal priors, e.g.:

$$\log \frac{P(B)}{P(A)} = \log \frac{0.7}{0.3} = 1.22$$

Weight matrices: odds ratio

W : weight matrix, R : background model (independent of position)

$$\frac{P(x_1 x_2 \dots x_N | W)}{P(x_1 x_2 \dots x_N | R)} = \frac{\prod_{i=1}^N P_i(x_i | W)}{\prod_{i=1}^N P(x_i | R)}$$

$$\log_2 \left(\frac{P(x_1 x_2 \dots x_N | W)}{P(x_1 x_2 \dots x_N | R)} \right) = \log_2 \left(\frac{\prod_{i=1}^N P_i(x_i | W)}{\prod_{i=1}^N P(x_i | R)} \right) = \sum_{i=1}^N \log_2 \left(\frac{P_i(x_i | W)}{P(x_i | R)} \right)$$

log-odds ratio

Weight matrices: log-odds ratio

R **uniform**: $P(A|R) = P(C|R) = P(G|R) = P(T|R) = 0.25$

$$\begin{matrix} A \\ C \\ G \\ T \end{matrix} \left(\begin{array}{cccccc} \boxed{1.16} & -1.17 & \boxed{1.16} & \boxed{1.42} & -1.17 & -1.17 \\ -1.17 & \boxed{1.16} & -0.17 & -1.17 & -1.17 & \boxed{1.16} \\ -1.17 & -0.17 & -1.17 & -1.17 & -0.17 & -0.17 \\ -0.17 & -1.17 & -1.17 & -1.17 & \boxed{1.16} & -1.17 \end{array} \right) \longrightarrow \log(0.56 / 0.25)$$

$$\log\text{-odds(ACAATC)} = 1.16 + 1.16 + 1.16 + 1.42 + 1.16 + 1.16 = 7.22$$

$$\log\text{-odds(TGCAGG)} = -0.17 - 0.17 - 0.17 + 1.42 - 0.17 - 0.17 = 0.57$$

$$\log\text{-odds(CTTGAT)} = 6 \times -1.17 = -7.02$$

Sequence logo

$$W = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} 0.8 & 0.0 & 0.8 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.8 & 0.2 & 0.0 & 0.0 & 0.8 \\ 0.0 & 0.2 & 0.0 & 0.0 & 0.2 & 0.2 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.8 & 0.0 \end{pmatrix} \end{matrix}$$



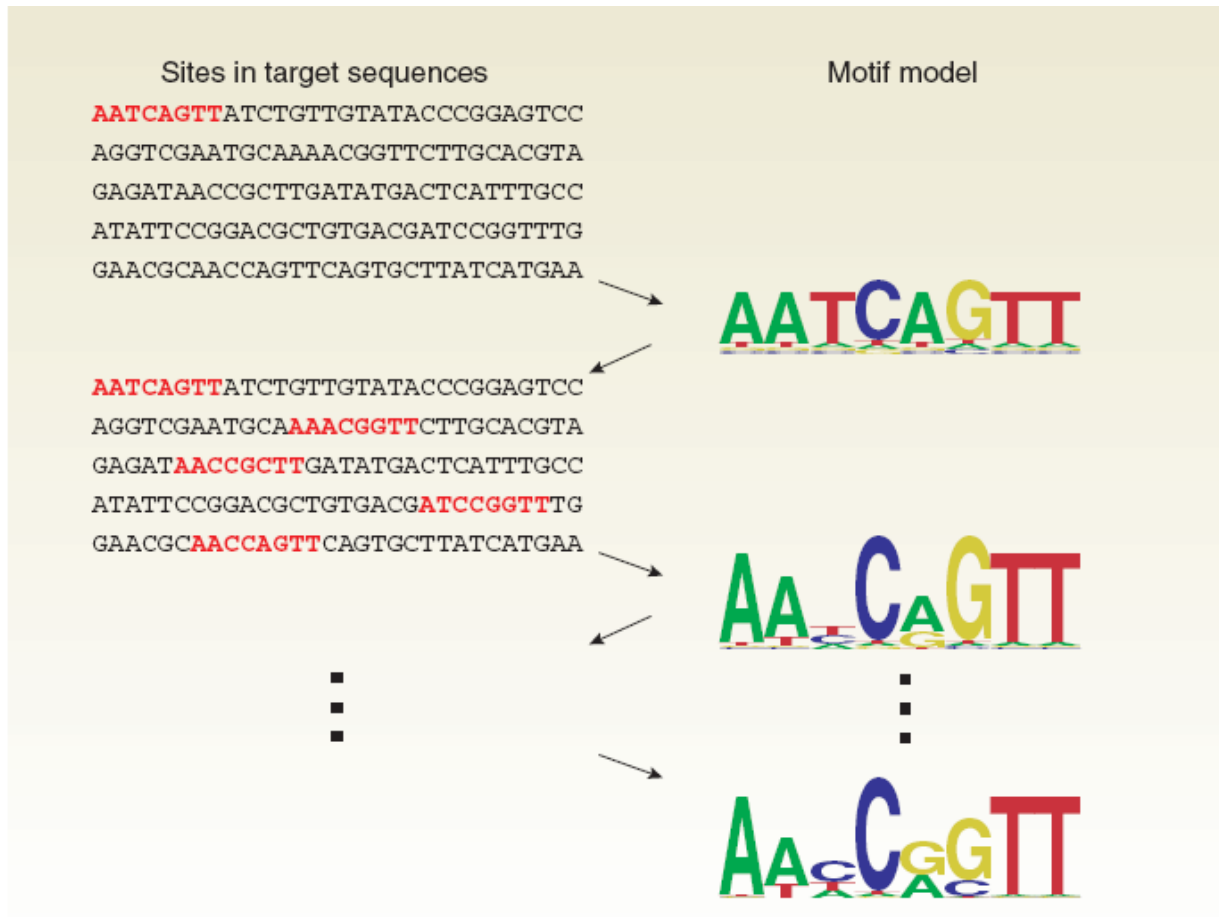
<http://weblogo.berkeley.edu/>

height of stack:
$$H(j) = 2 - \sum_{i \in \{A,C,G,T\}} P_j(i) \log_2 \frac{1}{P_j(i)}$$

Extreme cases: full consensus = 2 uniform = 0

$$H(1) = 2 - 0.8 \log_2 \left(\frac{1}{0.8} \right) - 2 \times 0 \log_2 \left(\frac{1}{0} \right) - 0.2 \log_2 \left(\frac{1}{0.2} \right) = 1.28$$

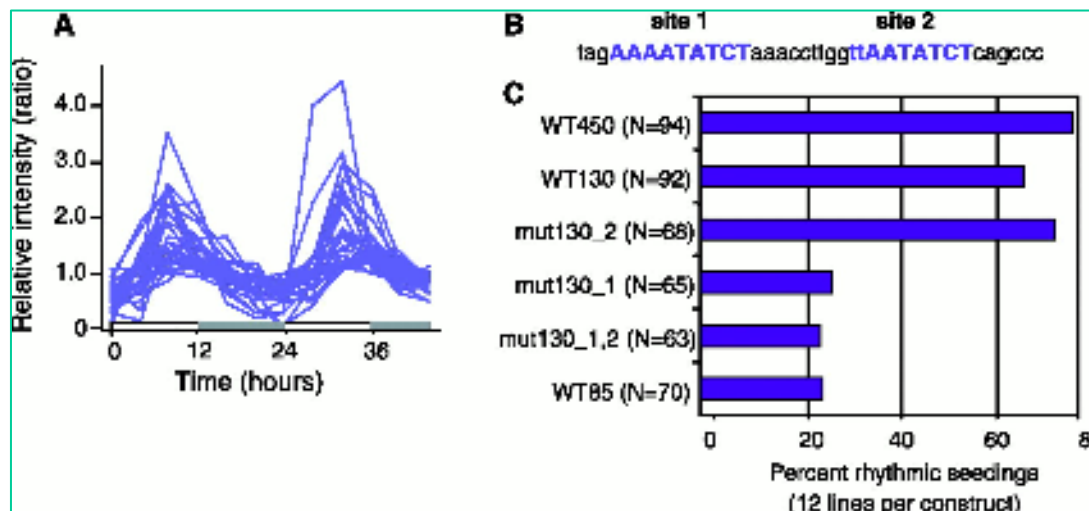
MEME



init: choose site
for (i in 1:n){
• fit motif model
• choose sites
}

D'Haeseleer, Nature Biotechnology, 24, pp959-961, (2006)

Evening element in Arabidopsis



Science 290:2110-3 (2000)

Harmer, et al. discovered an “evening element” in the promoter that was common between many genes that followed the same pattern of expression in Cluster 2

Evening element: 9-mers

Margin: |frequency in cluster 2 – frequency in other clusters|

- **AAAAAAAAA** TTTTTTTTT
- AAAATATCT AGATATTTT
- **CTCTCTCTC** GAGAGAGAG
- **AGAGAGAGA** TCTCTCTCT
- **AAAAAAAAAC** GTTTTTTTTT
- AAATATCTT AAGATATTT
- AAAAATATC GATATTTTT
- AAATAAAAT ATTTTATTT
- AAAATATAA TTATATTTT
- **TAAAAAAAA** TTTTTTTTA

Filter repeats and near-repeats

Motifs

- AAAATATCT AGATATTTT
- AAATATCTT AAGATATTT
- AAAAATATC GATATTTT
- AAATAAAAT ATTTTATTT
- AAAATATAA TTATATTTT

Significance: randomization

Validation: fuse CCR2 promoter to a luciferase reporter gene