The evening element

Kay and coworkers in 2000 discover the key element in the promoter region of 500 genes of *Arabidopsis thaliana* involved in the circadian behaviour

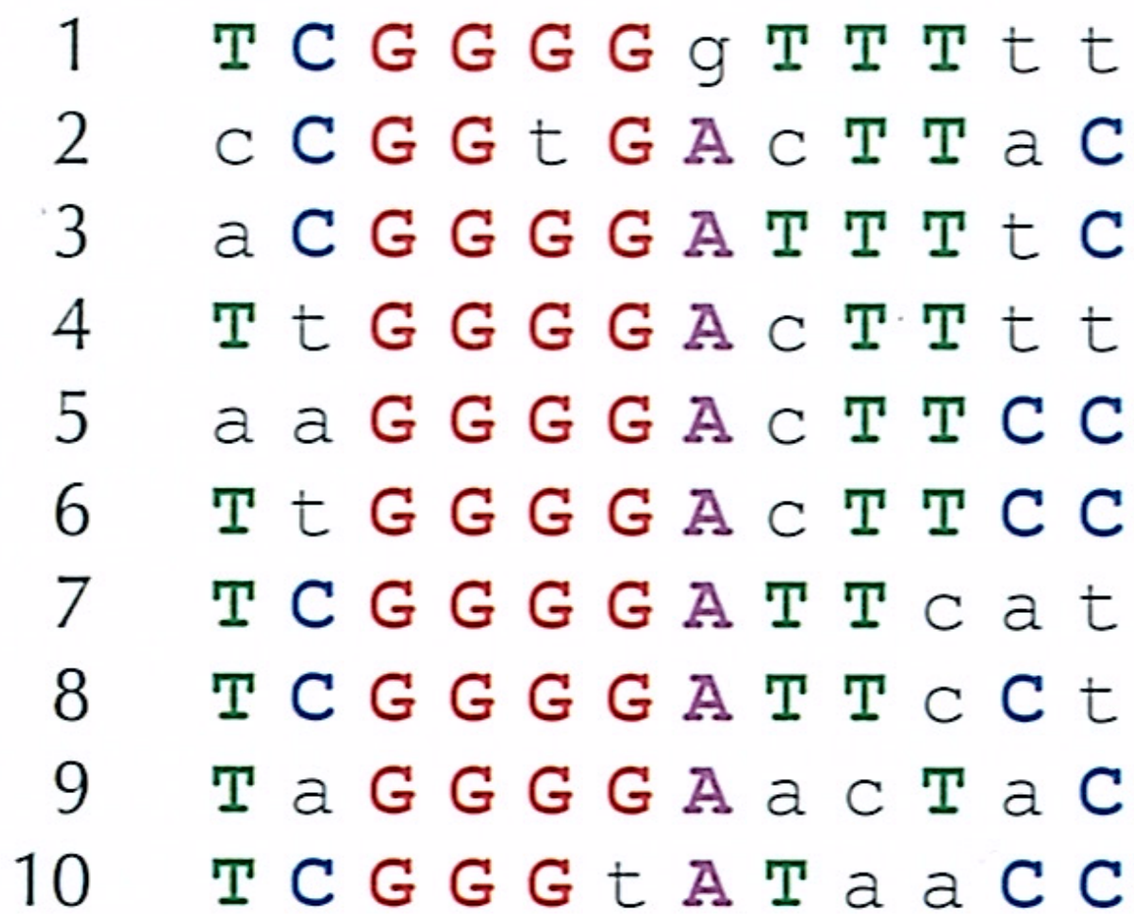
Very conserved

A A A A A T A T C T

Easy to find

The element is very conserved and a mutation of this elements in the upstream region of one gene leads it to no longer exhibited circadian behaviour

Immunity genes



Infected fly with a bacterium, the fly switch on the immunity genes to fight the infection

NF-xB

No well conserved among the immunity genes

Upstream regions of ten genes

1 atgaccgggatactgataaaaaaaaggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
2 acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaactttccgaataaaaaaaaggggggga
3 tgagtatccctgggatgacttaaaaaaaaggggggtgctctcccgattttgaatatgtaggatcattcgccagggtccga
4 gctgagaattggatgaaaaaaaaggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
5 tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaaggggggcttatag
6 gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
7 cggttttggcccttgttagaggcccccgtaaaaaaaaggggggcaattatgagagagctaatctatcgcgtgcgtgttcat
8 aacttgagttaaaaaaaaggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
9 ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaaggggggggaccgaaagggaag
10 ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga

Can you find the implanted hidden message (no mismatches)?

Frequent Words Problem

find the most frequent k-mers in a string

input: A string Text and a integer k

output: All most frequent k-mers in the Text

```
FREQUENTWORDS(Text, k)
    FrequentPatterns ← an empty set
    for i ← 0 to |Text| − k
        Pattern ← the k-mer Text(i, k)
        COUNT(i) ← PATTERNCOUNT(Text, Pattern)
    maxCount ← maximum value in array COUNT
    for i ← 0 to |Text| − k
        if COUNT(i) = maxCount
            add Text(i, k) to FrequentPatterns
    remove duplicates from FrequentPatterns
    return FrequentPatterns
```

```
PATTERNCOUNT(Text, Pattern)
    count ← 0
    for i ← 0 to |Text| − |Pattern|
        if Text(i, |Pattern|) = Pattern
            count ← count + 1
    return count
```

Text	A	C	T	G	A	C	T	C	C	C	A	C	C	C	C
Count	2	1	1	1	2	1	1	3	1	1	1	3	3		

Upstream regions of ten genes

```
1  atgaccgggatactgataaaaaaaaggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
2  acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaactttccgaataaaaaaaaggggggga
3  tgagtatccctgggatgacttaaaaaaaaggggggtgctctcccgattttgaatatgtaggatcattcgccagggtccga
4  gctgagaattggatgaaaaaaaaggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
5  tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaaggggggcttatag
6  gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
7  cggttttggcccttgttagaggcccccgtaaaaaaaaggggggcaattatgagagagctaatctatcgcgtgcgtgttcat
8  aacttgagttaaaaaaaaggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
9  ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaaggggggggaccgaaagggaag
10 ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga
```

Can you find the implanted hidden message (no mismatches)?

```
1  atgaccgggatactgatAAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
2  acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaactttccgaataAAAAAAAAGGGGGGGa
3  tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgattttgaatatgtaggatcattcgccagggtccga
4  gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
5  tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag
6  gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
7  cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat
8  aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
9  ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag
10 ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa
```

The number of mismatches between two strings is called the ***Hamming distance***

Frequent Words Problem considering mismatches

compute the Hamming distance between two strings

input: Two string of equal length

output: The Hamming distance between these strings

```
FREQUENTWORDS(Text, k)
    FrequentPatterns ← an empty set
    for i ← 0 to |Text| − k
        Pattern ← the k-mer Text(i, k)
        COUNT(i) ← PATTERNCOUNT(Text, Pattern)
    maxCount ← maximum value in array COUNT
    for i ← 0 to |Text| − k
        if COUNT(i) = maxCount
            add Text(i, k) to FrequentPatterns
    remove duplicates from FrequentPatterns
    return FrequentPatterns
```

```
APPROXIMATEPATTERNCOUNT(Text, Pattern, d)
    count ← 0
    for i ← 0 to |Text| − |Pattern|
        Pattern′ ← Text(i, |Pattern|)
        if HAMMINGDISTANCE(Pattern, Pattern′) ≤ d
            count ← count + 1
    return count
```

Example with muted pattern

```
1   atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg
2   acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa
3   tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgattttgaatatgtaggatcattcgccagggtccga
4   gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga
5   tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag
6   gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa
7   cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat
8   aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta
9   ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag
10  ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa
```

Brute force algorithm for motif finding (inspirited from The gold bug problem)

Brute force search is a general problem-solving technique that explores ALL possible candidate solutions and checked whether each candidate solves the problem

Implanted Motif Problem

find all (k,d)-motifs in a collection of strings

input: A collection of string dna, and integers k and d

output: All (k,d)-motifs in dna

From motifs to profile matrices and consensus strings

Motifs

T	C	G	G	G	G	g	T	T	T	t	t
c	C	G	G	t	G	A	c	T	T	a	C
a	C	G	G	G	G	A	T	T	T	t	C
T	t	G	G	G	G	A	c	T	T	t	t
a	a	G	G	G	G	A	c	T	T	C	C
T	t	G	G	G	G	A	c	T	T	C	C
T	C	G	G	G	G	A	T	T	c	a	t
T	C	G	G	G	G	A	T	T	c	C	t
T	a	G	G	G	G	A	a	c	T	a	C
T	C	G	G	G	t	A	T	a	a	C	C

SCORE(*Motifs*) 3 + 4 + 0 + 0 + 1 + 1 + 1 + 5 + 2 + 3 + 6 + 4 = 30

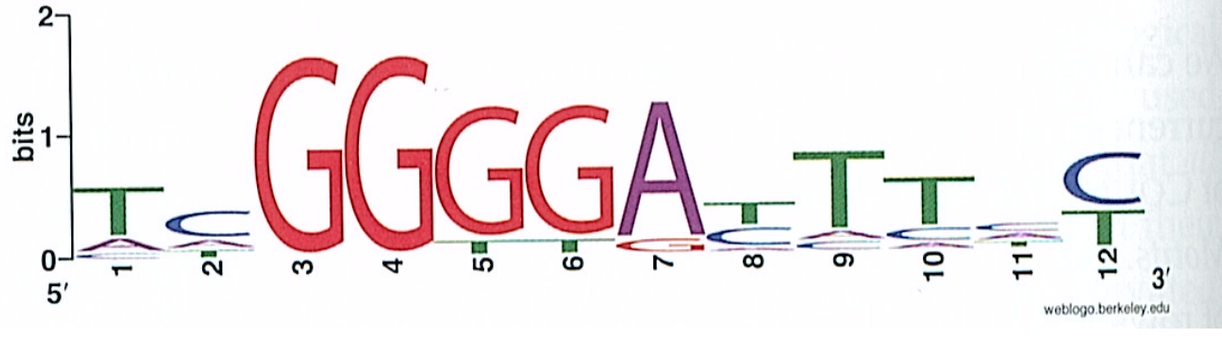
COUNT(*Motifs*)

A:	2	2	0	0	0	0	**9**	1	1	1	3	0
C:	1	**6**	0	0	0	0	0	4	1	2	4	**6**
G:	0	0	**10**	**10**	**9**	**9**	1	0	0	0	0	0
T:	**7**	2	0	0	1	1	0	**5**	**8**	**7**	3	4

PROFILE(*Motifs*)

A:	.2	.2	0	0	0	0	**.9**	.1	.1	.1	.3	0
C:	.1	**.6**	0	0	0	0	0	.4	.1	.2	.4	**.6**
G:	0	0	**1**	**1**	**.9**	**.9**	.1	0	0	0	0	0
T:	**.7**	.2	0	0	.1	.1	0	**.5**	**.8**	**.7**	.3	.4

CONSENSUS(*Motifs*) T C G G G G A T T T C C



TFBSs representation

Bussemaker at al. give a new interpretation of the information encoded by PWM. In their view PWMs contain two kinds of knowledge:

1. thermodynamics interactions between Transcription Factor and DNA,
2. evolutionary selection

The underlying assumptions are:

- natural selection gives rise to a certain level of sequence specificity for each TF
- sequences that give rise to the same physically binding affinity are equally likely to be selected.

TFBSs representation

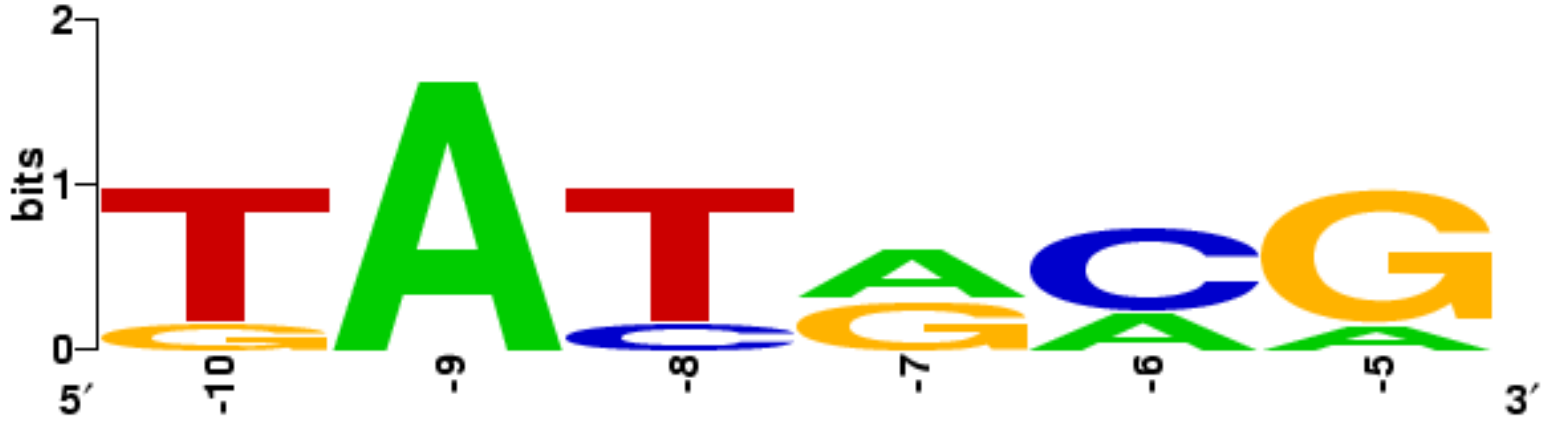
According with the additivity assumption each position contributes independently to the total binding energy, there is some matrix $H(b, i)$ that contains those binding energy contributions as its elements. Given any particular sequence S_α its total binding energy is since given by $H(b, i) \cdot S_\alpha$.

The measure of significance for one position in the PWM, compared with the frequency in the genome, is commonly given by the *Information Content* (IC) defined as:

$$I_i = 2 + \sum_{b=A}^{T} f_{b,i} \log_2 f_{b,i} \tag{1}$$

where i is the position in the site, b refers to each of the possible nucleotides, and $f_{b,i}$ is the observed frequency of each base at i^{th} position.

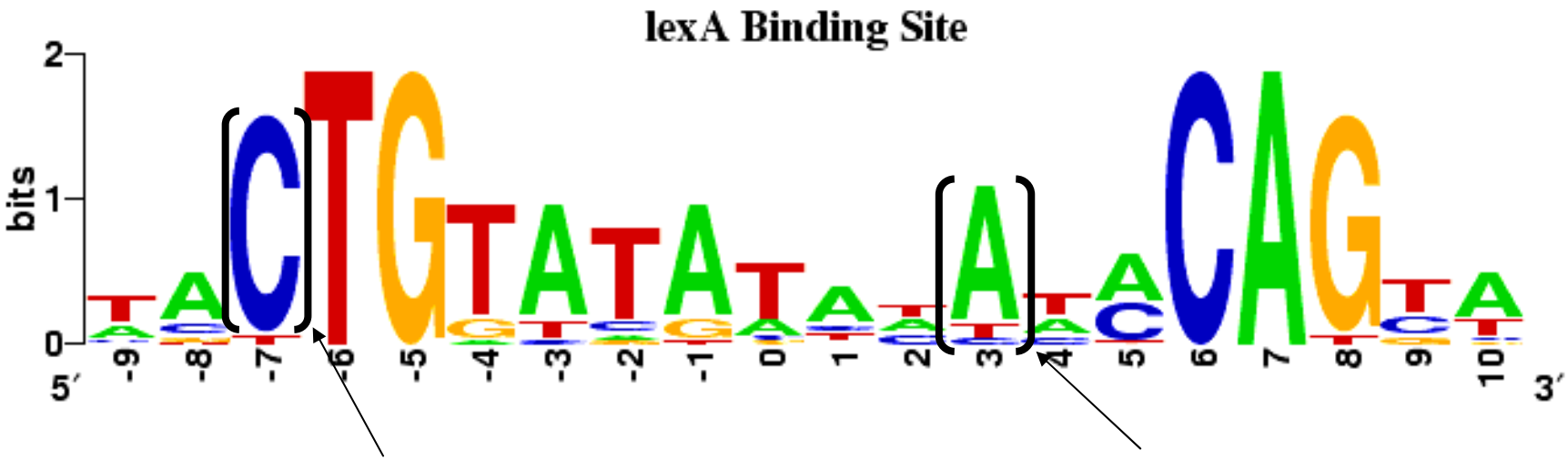
TFBSs representation



The I values are between 0, for positions that are 25% of each base, and 2 *bits* for positions completely conserved.

Visualizing Motifs – Motif Logos

Represent both base frequency and conservation at each position



Height of letter proportional to frequency of base at that position

Height of stack proportional to conservation at that position

TFBSs representation

This formula provides a good approximation only in genomes with a perfect balance distribution of frequency among the four nucleotides (25% for each bases). Berg et al. shows that the logarithms of base frequencies should be proportional to the binding energy contribution of the bases:

$$I_{seq}(i) = \sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b} \tag{2}$$

Limitations:

- the positions in site contribute additively to the total activity

Two comprehensive and annotated databases that contain information on TFs binding site profiles: JASPAR and TRANSFAC.

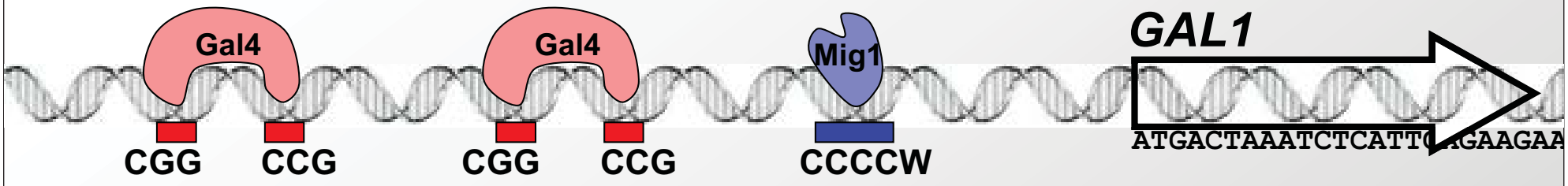
TFBSs identification

The interest in promoter analysis received a great improvement due to the identification of co-regulated groups of gene. A basic assumption is that these profiles reflect a similar structure of the regions involved in transcription regulation.

Transcription modules are *self-consistent regulatory units*: a set of genes are co-regulated, responding to different conditions that alter expression of all genes in the module.

Another type of orthogonal data are functional sequences that are preferentially conserved over the course of evolution by selective pressure.

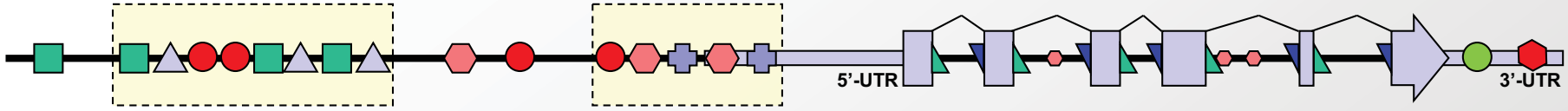
Regulatory motif discovery



- Regulatory motifs
  - Genes are turned on / off in response to changing environments
  - No direct addressing: subroutines (genes) contain sequence tags (motifs)
  - Specialized proteins (transcription factors) recognize these tags

- What makes motif discovery hard?
  - Motifs are short (6-8 bp), sometimes degenerate
  - Can contain any set of nucleotides (no ATG or other rules)
  - Act at variable distances upstream (or downstream) of target gene

The regulatory code: All about regulatory motifs



Enhancer regions — **Where in the body?**

Promoter motifs — **When in time?**

Splicing signals — **Which variants?**

Motifs at RNA level — **Which subsets?**

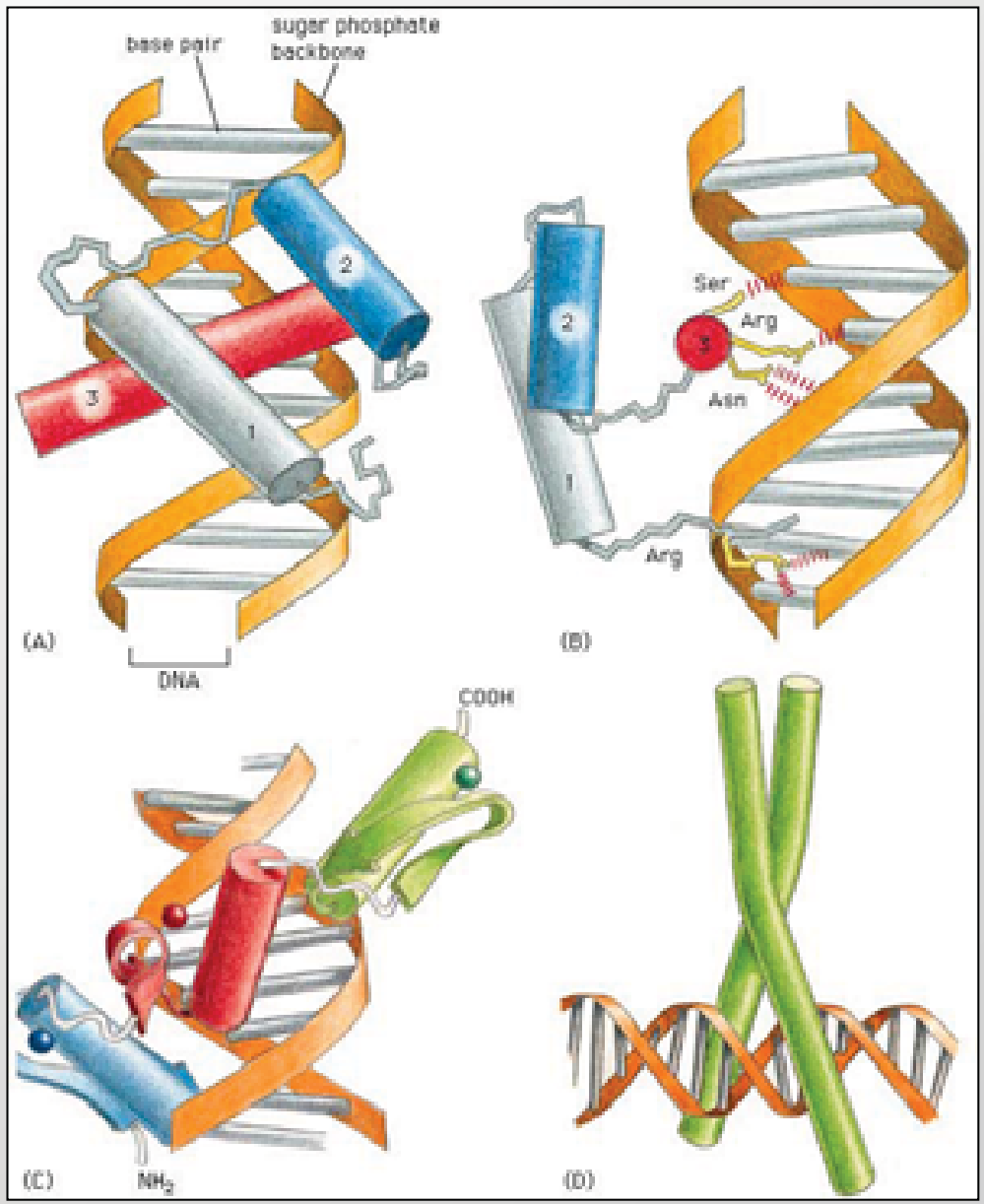
- The parts list: ~20-30k genes
  - Protein-coding genes, RNA genes (tRNA, microRNA, snRNA)
- The circuitry: constructs controlling gene usage
  - Enhancers, promoters, splicing, post-transcriptional motifs
- The regulatory code, complications:
  - Combinatorial coding of 'unique tags'
    - Data-centric encoding of addresses
  - Overlaid with 'memory' marks
    - Large-scale on/off states
  - Modulation of the large-scale coding
    - Post-transcriptional and post-translational information
- Today: discovering motifs in co-regulated promoters and *de novo* motif discovery & target identification

Motifs are not limited to DNA sequences

- Splicing Signals at the RNA level
  - Splice junctions
  - Exonic Splicing Enhancers (ESE)
  - Exonic Splicing Surpressors (ESS)
- Domains and epitopes at the Protein level
  - Glycosylation sites
  - Kinase targets
  - Targetting signals
  - MHC binding specificities
- Recurring patterns at the physiological level
  - Expression patterns during the cell cycle
  - Heart beat patterns predicting cardiac arrest
    - Final project in previous year, now used in Boston hospitals!
  - Any probabilistic recurring pattern

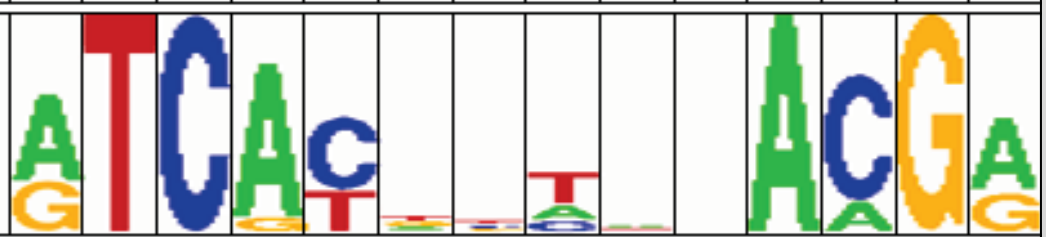
How Transcription Factors actually recognize motifs

- Proteins 'feel' DNA
  - Read chemical properties of bases
  - Do NOT open DNA (no base complementarity)
- 3D Topology dictates specificity
  - Fully constrained positions: ➔ every atom matters
  - "Ambiguous / degenerate" positions ➔ loosely contacted
- Other types of recognition
  - MicroRNAs: complementarity
  - Nucleosomes: GC content
  - RNAs: structure/seqn combination



Motifs summarize TF sequence specificity

Target genes bound by ABF1 regulator		Coordinates		Genome sequence at bound site
ACS1	acetyl CoA synthetase	-491	-479	|ATCATTCTGGACG|
ACS1	acetyl CoA synthetase	-433	-421	|ATCATCTCGGACG|
ACS1	acetyl CoA synthetase	-311	-299	|ATCATTTGCCACG|
CHA1	catabolic L-serine dehydratase	-260	-254	A|ATCACCGCGAACG|GA
ENO2	Enolase	-470	-461	ggcgttat|GTCACTAACGACG|tgcacca
HMR	silencer	-256	-283	ATCAATAC|ATCATAAAATACG|AACGATC
LPD1	lipoamide dehydrogenase	-288	-300	gat|ATCAAAATTAACG|tag
LPD1	lipoamide dehydrogenase	-301	-313	gat|ATCACCGTTGACG|tca
PGK	phosphoglycerate kinase	-523	-496	CAAACAA|ATCACGAGCGACG|GTAATTTC
RPC160	RNA pol III/C 160 kDa subunit	-385	-349	|ATCACTATATACG|TGAA
RPC40	RNA pol III/C 40 kDa subunit	-137	-116	|GTCACTATAAACG|
rpL2	ribosomal protein L2	-185	-167	TAAT|aTCAcgtcACACG|AC
SPR3	CDC3/10/11/12 family homolog	-315	-303	|ATCACTAAATACG|
YPT1	TUB2	-193	-172	CCTAG|GTCACTGTACACG|TATA

Position		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Position Weight Matrix (PWM)	A	**56**	4	4	**81**	4	23	15	27	31	31	**89**	23	4	**58**
	G	**32**	4	4	12	4	31	23	4	19	23	4	4	**89**	**35**
	C	4	4	**89**	4	**58**	12	23	19	19	23	4	**69**	4	4
	T	4	**89**	4	4	**35**	35	39	50	31	23	4	4	4	4
Motif Logo															
Consensus		**R**	**T**	**C**	**A**	**Y**	N	N	H	N	N	**A**	**C**	**G**	**R**

- Summarize information
- Integrate many positions
- Measure of information
- Distinguish motif vs. motif instance
- Assumptions:
  - Independence
  - Fixed spacing

Uncertainty and probability

Uncertainty is related to our surprise at an event

"The sun will rise tomorrow" **Not surprising (p~1)**

"The sun will *not* rise tomorrow" ***Very*** **surprising (p<<1)**

Uncertainty is inversely related to probability of event

Average Uncertainty

Two possible outcomes for sun rising

A "The sun will rise tomorrow" **$P(A)=p_1$**

B "The sun will *not* rise tomorrow" **$P(B)=p_2$**

What is our *average uncertainty* about the sun rising

$$= P(A)\text{Uncertainty(A)} + P(B)\text{Uncertainty(B)}$$

$$= -p_1 \log p_1 - p_2 \log p_2$$

$$= -\sum p_i \log p_i$$ **= Entropy**

Entropy

Entropy measures average uncertainty

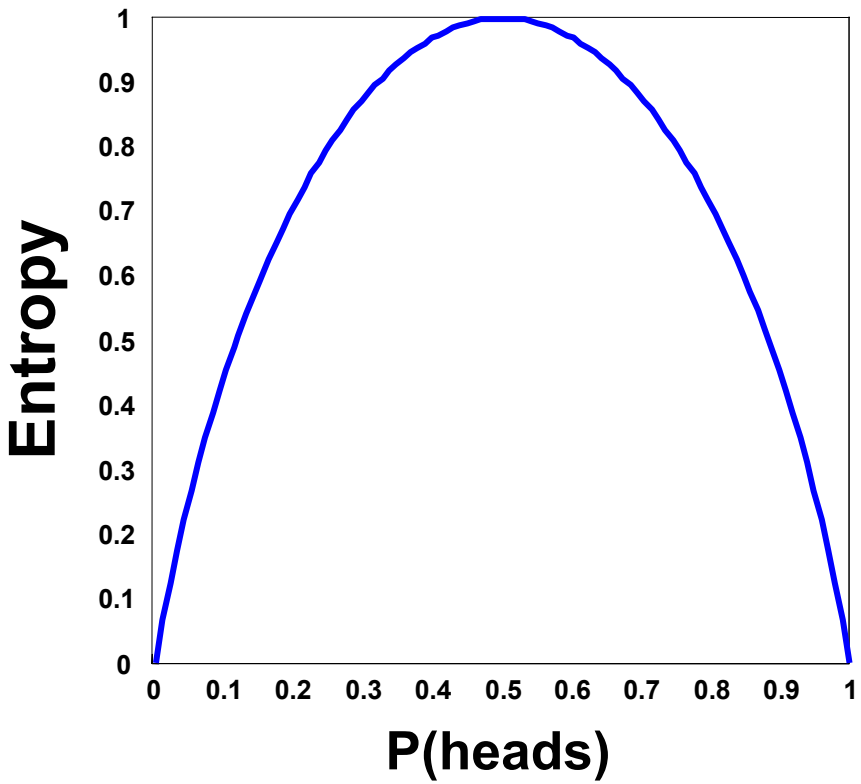
Entropy measures randomness

$$H(X) = -\sum_i p_i \log_2 p_i$$

If log is base 2, then the units are called bits

Entropy versus randomness

Entropy is maximum at maximum randomness

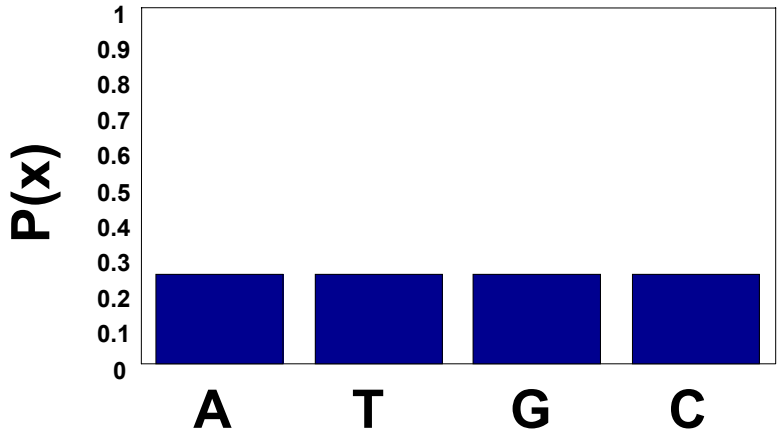


Example: Coin Toss

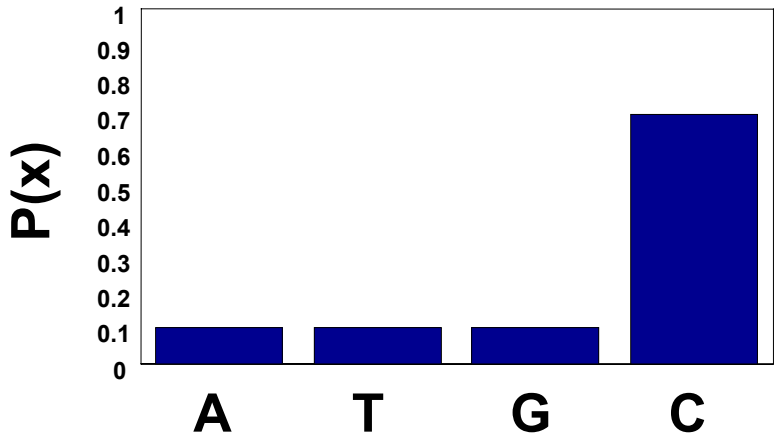
P(heads)=0.1 Not very random
H(X)=0.47 bits

P(heads)=0.5 Completely random
H(X)=1 bits

Entropy Examples



$$H(X) = -[0.25\log(0.25) + 0.25\log(0.25) + 0.25\log(0.25) + 0.25\log(0.25)]$$
$$= 2 \text{ bits}$$



$$H(X) = -[0.1\log(0.1) + 0.1\log(0.1) + 0.1\log(0.1) + 0.75\log(0.75)]$$
$$= 0.63 \text{ bits}$$

Information Content

Information is a *decrease in uncertainty*

Once I tell you the sun will rise, your uncertainty about the event decreases

$$\textbf{Information} = H_{before}(X) - H_{after}(X)$$

Information is difference in entropy after receiving information

Motif Information

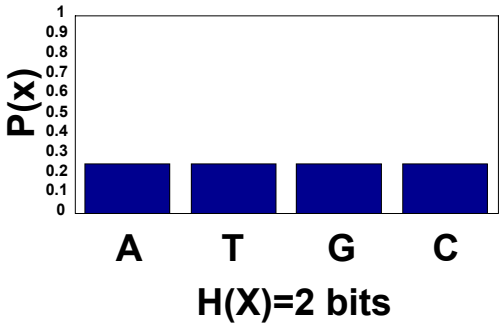
Motif Position Information = $2 - \sum_{b=\{A,T,G,C\}} -p_b \log p_b$

$H_{background}(X)$

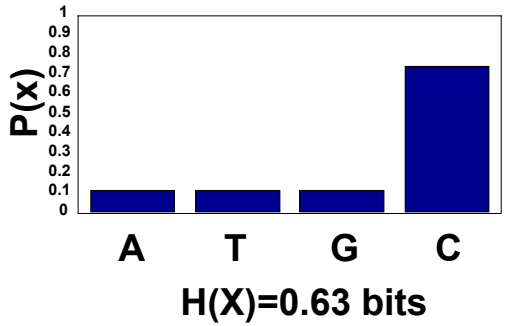
$H_{motif\_i}(X)$

Prior uncertainty about nucleotide

Uncertainty after learning it is position i in a motif



H(X)=2 bits

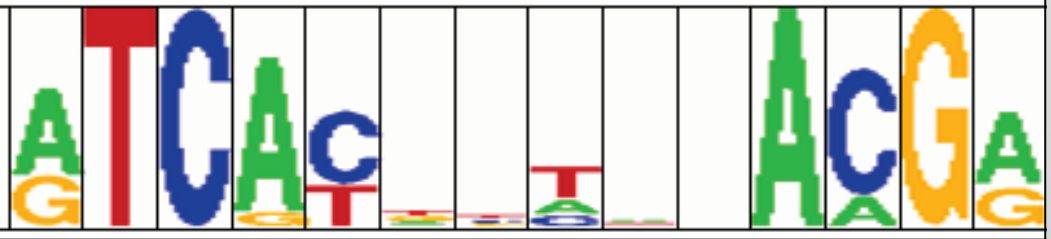


H(X)=0.63 bits

Uncertainty at this position has been reduced by 0.37 bits

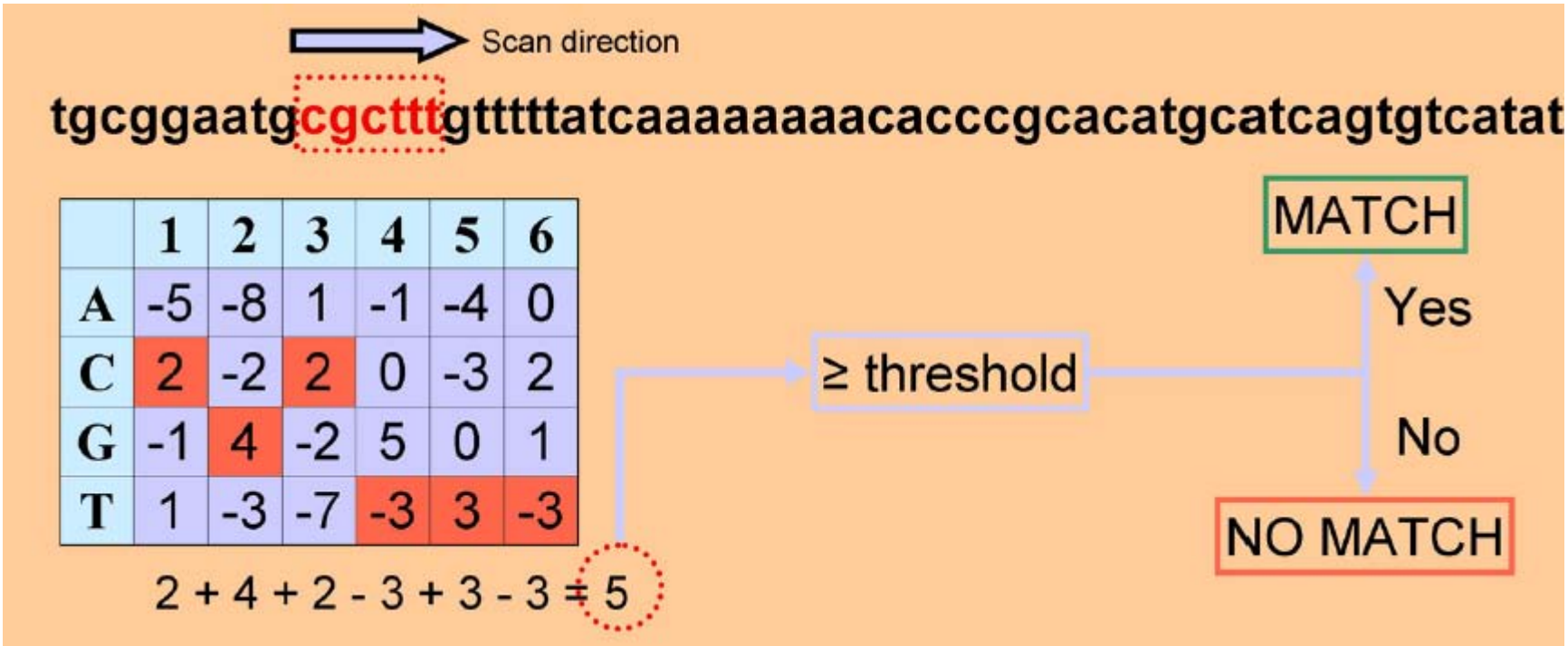
Motifs summarize TF sequence specificity

Target genes bound by ABF1 regulator		Coordinates		Genome sequence at bound site
ACS1	acetyl CoA synthetase	-491	-479	| ATCATTCTGGACG |
ACS1	acetyl CoA synthetase	-433	-421	| ATCATCTCGGACG |
ACS1	acetyl CoA synthetase	-311	-299	| ATCATTTGCCACG |
CHA1	catabolic L-serine dehydratase	-260	-254	A | ATCACCGCGAACG | GA
ENO2	Enolase	-470	-461	ggcgttat | GTCACTAACGACG | tgcaccca
HMR	silencer	-256	-283	ATCAATAC | ATCATAAAATACG | AACGATC
LPD1	lipoamide dehydrogenase	-288	-300	gat | ATCAAAATTAACG | tag
LPD1	lipoamide dehydrogenase	-301	-313	gat | ATCACCGTTGACG | tca
PGK	phosphoglycerate kinase	-523	-496	CAAACAA | ATCACGAGCGACG | GTAATTTC
RPC160	RNA pol III/C 160 kDa subunit	-385	-349	| ATCACTATATACG | TGAA
RPC40	RNA pol III/C 40 kDa subunit	-137	-116	| GTCACTATAAACG |
rpL2	ribosomal protein L2	-185	-167	TAAT | aTCAcgtcACACG | AC
SPR3	CDC3/10/11/12 family homolog	-315	-303	| ATCACTAAATACG |
YPT1	TUB2	-193	-172	CCTAG | GTCACTGTACACG | TATA

Position		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Position Weight Matrix (PWM)	A	**56**	4	4	**81**	4	23	15	27	31	31	**89**	23	4	**58**
	G	**32**	4	4	12	4	31	23	4	19	23	4	4	**89**	**35**
	C	4	4	**89**	4	**58**	12	23	19	19	23	4	**69**	4	4
	T	4	**89**	4	4	**35**	35	39	50	31	23	4	4	4	4
Motif Logo															
Consensus		**R**	**T**	**C**	**A**	**Y**	N	N	H	N	N	**A**	**C**	**G**	**R**

- Summarize information
- Integrate many positions
- Measure of information
- Distinguish motif vs. motif instance
- Assumptions:
  - Independence
  - Fixed spacing

Scoring a Sequence



Common threshold = 60% of maximum score

MacIsaac & Fraenkel (2006) PLoS Comp Bio