

# Elements of Bioinformatics

## Autumn 2010



VELI MÄKINEN

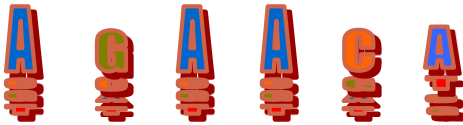
[HTTP://WWW.CS.HELSINKI.FI/EN/COURSES/  
582606/2010/S/K/1](http://www.cs.helsinki.fi/en/courses/582606/2010/S/K/1)

# Lecture Thu 18.11.



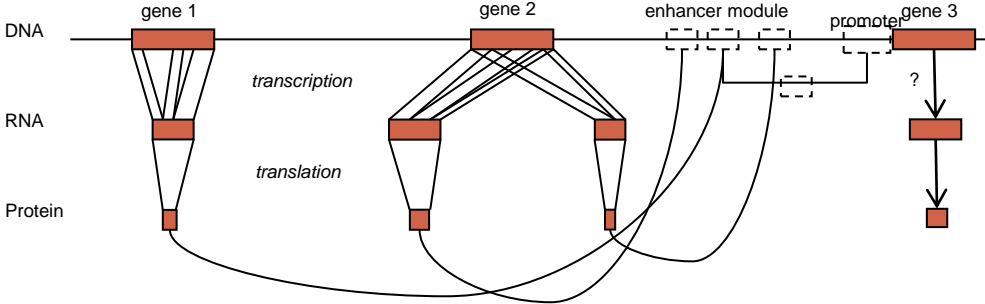
## SIGNALS IN DNA

1.54	-1.46	1.54	1.54	-1.46	1.35
-1.46	-0.46	-1.46	-1.46	1.35	-1.46
-1.46	1.35	-1.46	-1.46	-0.46	-1.46
-1.46	-1.46	-1.46	-1.46	-1.46	-0.46



```

12854400 tcaaaagttaagttagataaaacatgatacattcacaggtccagatgttttaaaaaaaaaaatcatttgggttaacctcacatgtagaacaactcagaaattcctc
12854200 tggactaccagaatttgaagtacctagctactctcattctcattttaccctaaagtcctaaataaataaaagctacctagccctctcgttttatgattcctc
12854000 taggaaaagtttaagtttaggcccaactcatttttaaacggcccaaacacatataataggtccaaatatacattttccctagaaatattcacaact
12853800 attgcaactcaaaaagtgacaaaaggaggtgtataaaggagaccactaactgaectatttagagctaggatcagacagatgatttttgcaataactc
12853600 ctgttaaagttatccacatttcactcccaagaaaaatagactgtagaagaatataatcagatatgcaaggccgtgctgtttaggtttagctaaccttaca
12853400 aggtttagggtctcagataaaacacaaagcagatagagaagcgaaccactcacaatcgagcagcagcagcagcagcagcagcagcagcagcagcagcagc
12853200 cctttttctccacgccttcccaaccttccacaccttccacccttattggttcaggttctgttttlttgccttttacacacacagactccacacac
12853000 tcaacttattgggtttcttcaatttgaacacagactttccattgggactcattggaaagaaagaggaggtttctcaattctctccacacactccattgagc
12852800 acataggcaagcgggaaatcacctgacttttggcctctctccctcctcaactcgttctcctccgtagaggtcactttctgttactctttagaacattcga
12852600 taccacattttgactcagcgaattcttactcctcagcttacttaccggaaagctatagcattcaaacagctcccaaataggctcagagggcggaactga
12852400 aatccttaattcaagcgtgaaatcaaaagaaapaagaactttggctgatatatgattgattcaaccacagacactcagagaaagacgattaaatgcgata
12852200 accggagagagattttgtagggagccggcagactcagacccttaaccctagagcttcagaaagagctgcgcaatggatgaattggcttcaaacctgaaatcg
12852000 gatctccatgggttggagatttgattatgtttgaggttatgcatctccaccaccaaatatatacgttcaggttaaatcacaatgatctcaaatatagac
12851800 aacagattagatatagaacaataggttggataattatattactatagatataatagatatacaggttgaatgaggttatttactactatttagat
12851600 ataaagaaatcaagttcaattcaagaaatcaagaaatcaagaaatcaagaaatcaagaaatcaagaaatcaagaaatcaagaaatcaagaaatcaagaaatca
12851400 accgatttttccgctgggtggaaaatggcagatataaagtagcggagggagacgggaaactagactatgatcagaacagcactccgctcggctcacaacg
12851200 tggctcagagagcgggtgggtgggtgcttggacagcttggattcaccaccacaagagactcttaccagcttgcctcaaaagctgagcttgcagactaaag
12851000 actcggagggaaaacactggtaggaaatcagccggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggcggc
12850800 tgataaagcttggacttggataactttatpacttactcaccagaaactccttggcatttgaagtacattttagtagtagctatactataaactac
12850600 aactcttggtagatatttttggtagtctttaaataaacactttggcagatctcaagaatgatatataatgaaatataaaatgataaagaatcacataa
12850400 atgggattggttagcacaacagctctgtaactaaagcggcagggcagctctcctctccttggctatgattatgataaattaccagaaattggaccaaa
12850200 gcaagatcgggaacacttggctctcaattttgcttagcttattccagcccttgaacttctcctcctgggagaaagaacaaacgcaaatcgaactcga
12850000 agaaaagtcaatcgaatgattctctcctttaaataaacactttggcagatctcaagaatgatatataatgaaatataaaatgataaagaatcacataa
12849800 caaaaagcactagtgggttaaaagatacaactcagtagaagaaagagtcgaagtgaagaggtcgaacttgaagaatagatggaaagttc
12849600 catgcttggtttggtagcaactaatttatataatggcgcactggttaagattggagccactcaactcaagatattgattcttaacca
12849400 cacaactctgccaatcagaagcaatataattattgtagaagaagaaaaaagattggtgggaagtgggaacagttagacaggttaattcgaataaa
    
```



### ChIP-seq: welcome to the new frontier

High-resolution sequencing technology combined with chromatin immunoprecipitation (ChIP) is revolutionizing the study of gene regulation. This combination of techniques is providing a genome-wide look at transcription factor binding sites (TFBS) and other regulatory elements in the genome.

ChIP-seq is a powerful tool for studying the interactions between proteins and DNA. It allows researchers to identify the binding sites of transcription factors and other proteins on the genome. This information is crucial for understanding gene regulation and the development of diseases.

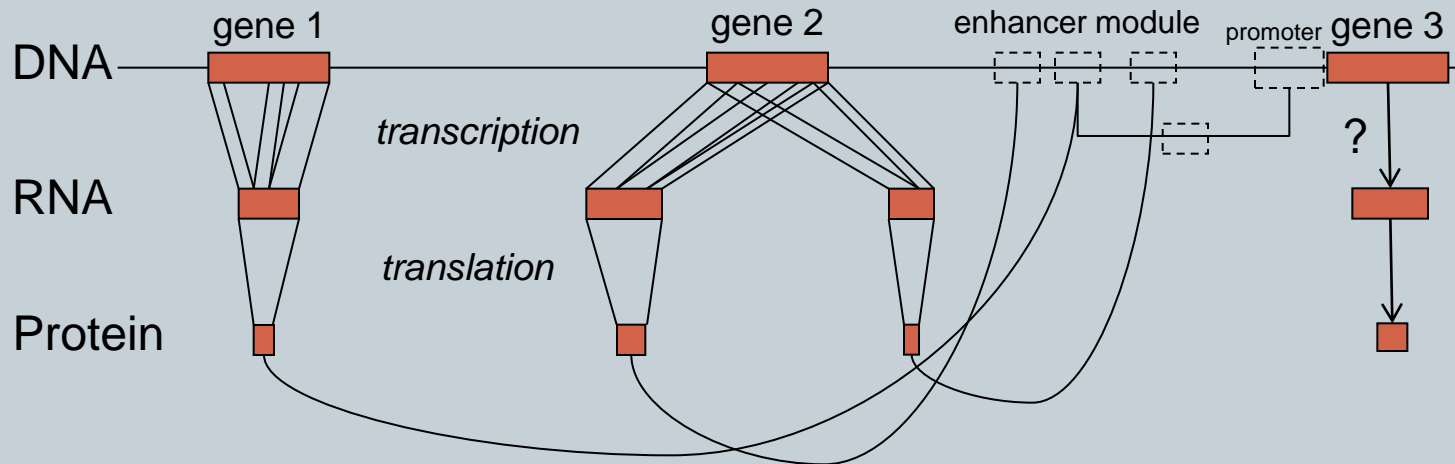
The diagram illustrates the ChIP-seq workflow, from chromatin immunoprecipitation to sequencing and data analysis. It highlights the importance of high-resolution sequencing and the use of bioinformatics tools to analyze the resulting data.

© 2010 Cold Spring Harbor Laboratory Press

# Signals in DNA



- Genes
- Promoter regions
- Binding sites for regulatory proteins (*transcription factors, enhancer modules, motifs*)



# Typical gene

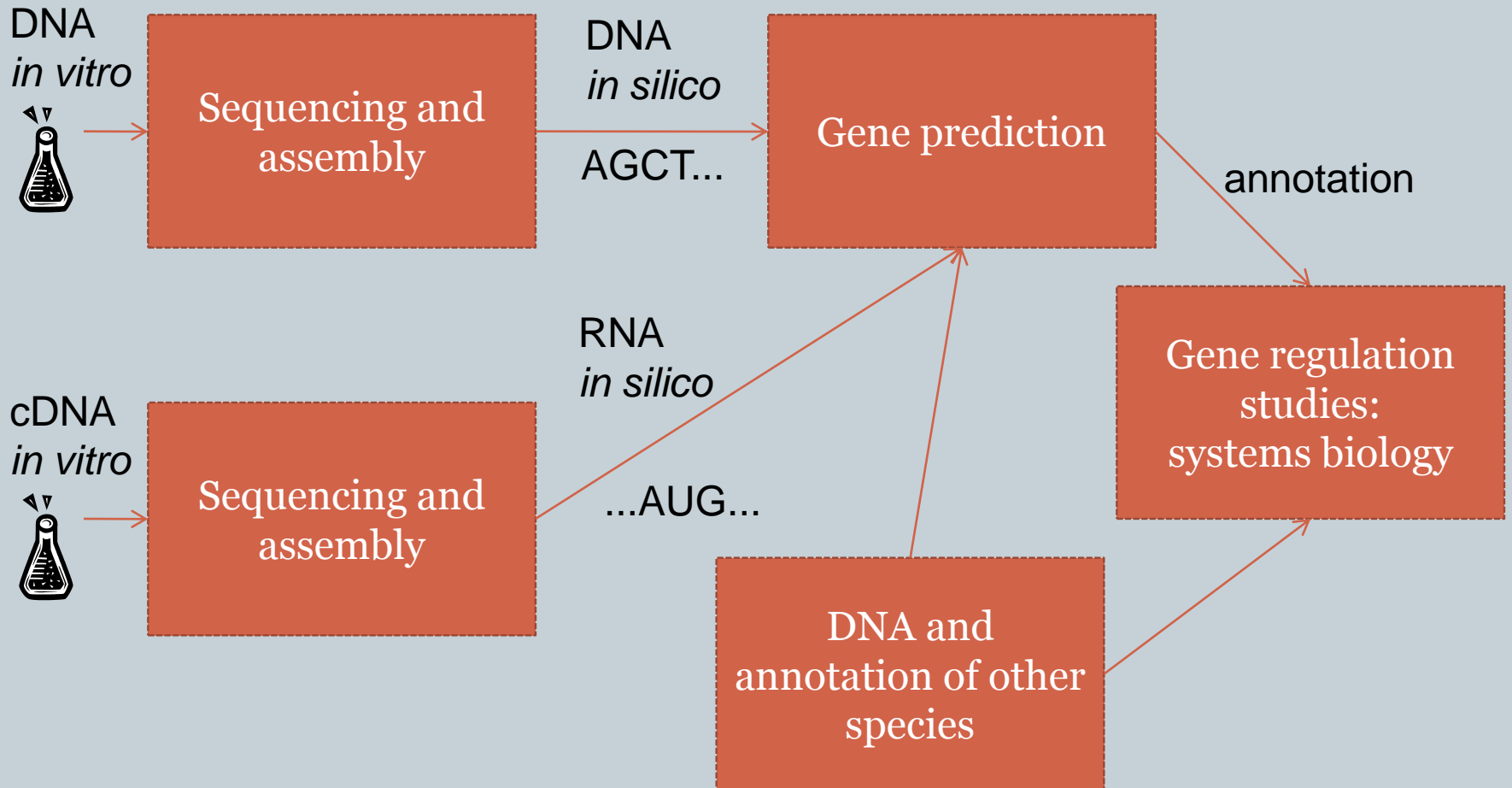


```
12854400 tcaaagtaagttagataaacaatgatcattcacaggtcagatgtttaaaaaaaatcattatgggtgtacatcacatgtagacaatacttcagaattcatc
12854200 tggactaccagaattgagttacctagtagtacttctcaattctatctttaccctaacgtctaataaataacaagtagtcttagcctcttcgtttatgattcctc
12854000 taggaaaagttaatggtacggccaatcacttttttaacagcccaacaacataatattagctccaaatcattttttcccctagaatattctcaacct
12853800 attgtccactcaaaacgtgacaaatggaggtctaaagggagaccatacttagtactcatttttagagcttaggatcagacagagtagattttttggcataactc
12853600 cttgtaaatgtattcacatttcattcccaagaaaaatagactgatgaagaaatataatcagatatgacaaggccgtgctcgttttaggttacgtaactctaca
12853400 aggtttagggtctcaatataaacacacaaaagcagatagaagaagcaaacattcacaaatcagacaATGACATCTCTCCATACGTTACTCTTCTCTCTCT
12853200 TCTTTTCTTCATCGTCTTTCCAACCTTCACGTTTTCCTCCACCTTATTGTTTCAGgttcgctcttttagttttgcttctttacatacacagactctacacac
12853000 tcacttattgggtttcctttcaattgtgaaacagAGTTTCAATTGGGAGTCATGGAAGAAAGAAGGAGGATTCTACAATTCTCTCCACAACCTCCATTGACG
12852800 ACATAGCCAACGCTGGAATCACTCATCTTTGGCTTCTCCTCCTTCTCAATCCGTTGCTCCTGAAGgttccatttctgctttactctttacacattcaca
12852600 taccaatcttgttactcagcaatcttcattcctcagGTTACTTACCGGGAAAGCTATACGATCTAAACAGCTCCAAATACGGTTCAGAGGCGGAACTGA
12852400 AATCGTTAATCAAAGCGTTGAATCAAAAAGGAATAAAAAGCTTTGGCTGATATAGTGATTAACCACAGAACAGCTGAGAGGAAAGACGATAAAATGTGGATA
12852200 CTGTTAATTTGGAAGGTGGGACTTCCGATGATCGTCTTGATTGGGATCCTTCTTTGTCGCCGCAATGACCCTAAATTTCCCGGTACCGGAAACCTCGAC
12852000 ACCGGAGGAGATTTTGATGGAGCGCCCGACATCGACCACCTTAACCCTAGAGTTCAGAAAGAGTTGTCCGAATGGATGAATTTGGCTTAAAACCTGAAATCG
12851800 GATTCCATGGTTGGAGATTTGATTATGTTTCGAGGTTATGCATCTTCCATCACCAAAATTATACGTTTCAGgtaaatcacatatgaattctcaaatatcagac
12851600 aacagtattagttatataagaacaataggttgagataattatctactattagttatataagaatcataggttgatagggttatttactactatcttagtat
12851400 ataagaacaataagtcattgcaatcaataagaataatataagaagttcactactgattatgtgataaattcctctgtttttggatacacagAATACATC
12851200 ACCGGATTTTTCGGTGGGTGAGAAATGGGACGATATGAAGTACGGAGGAGACGGGAAACTAGACTATGATCAGAACGAGCATCGGTTCGGGTCTCAAACAG
12851000 TGGATCGAGGAAGCGGGTGGTGGTGTGTTGACAGCTTTTTGATTTACCACCAAAGGGATCTTACAGTCTGTGTCAAAGGTGAGCTTTGGAGACTAAAGG
12850800 ACTCGCAGGGAAAACCGCTGGTATGATAGGAATCATGCCCGGAAACGCTGTCACATTCATAGATAAACCATGATACATTCAGAACGTGGGTTTTCCCTTC
12850600 TGATAAAGTCTTGCTTGATACGTTTATATACTTACTCATCCAGGAACCTCTTGCAATTgtaagtatcatttttagttatgtagctatactatttacaactac
12850400 aatcttggttgatatggtatttttggttgcagTTTTATAATCATTACATAGAATGGGGACTAAAAGAGAGCATCTCAAAGCTGGTGGCTATCAGGAACAAAA
12850200 ATGGGATTTGGTAGCACAACTGTGTAACAGATAAAAAGCCGAGAGCGGATCTCTACTTGGCTATGTTGATGATAAAGTTATCATGAAGATTGGACAAA
12850000 GCAAGATGTGGGAACACTTGTTCCTTAATTTGCTTATTAGCTTATTAGGCTTGCATCTTGGCTGTCTGGGAGAAGAATAAcgcataactcgaatcaca
12849800 agaaaagtaaatcgaatgtatcttcttcttctttaaataaaacatctttggcagtatctaagatatgtataatgaaatataaaatgataaagaatacctaaa
12849600 taaaaagagcactagtgggtgtaaaaggatacaactccagtgaaagaaaagagttcaagtgaagaagtgtcaactttagtagaataagttggaaagtttc
12849400 catcgttttgttttgttgcatacaactaatatattatatttgccgactcgtataagatttggagccctactaaaatcagaattatgatgtcttaacca
12849200 cacaatactgccaaaatcagaacgaattatattattgtagaagaagaaaaaaaagtaggtgggaagtggaacagttagacaggttaattcgaataaa
```

A  
T  
4  
G  
5  
0  
0  
.  
1  
v

<http://en.wikipedia.org/wiki/File:AMY1gene.png>

# Genome analysis pipeline



# Gene regulation



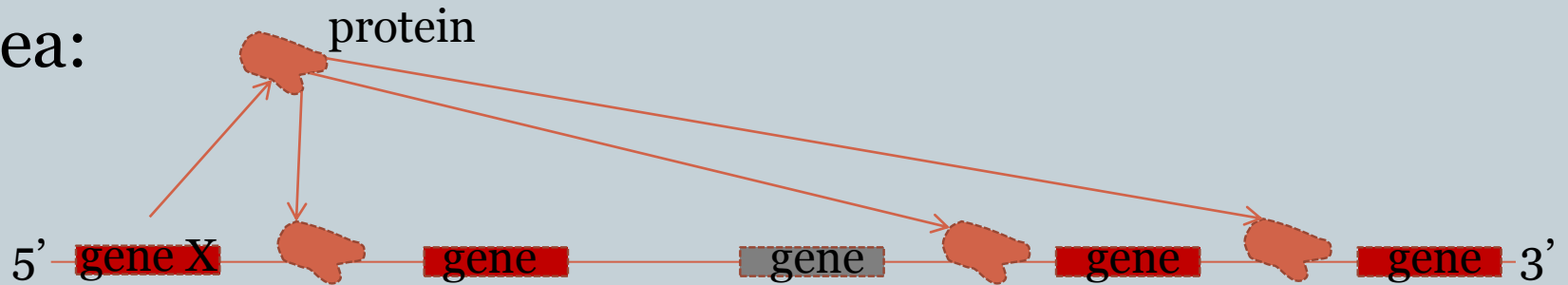
- Let us assume that gene prediction is done (covered later today & next week).
- We are interested in signals that influence gene regulation:
  - How much mRNA is transcribed, how much protein is translated?
  - How to measure those?
    - ✦ 2D gel electrophoresis (traditional technique to measure protein expression)
    - ✦ Microarrays (the standard technique to measure RNA expression)
    - ✦ RNA-sequencing (a new technique to measure RNA expression, useful for many other purposes as well, including gene prediction)

# Microarrays and gene expression

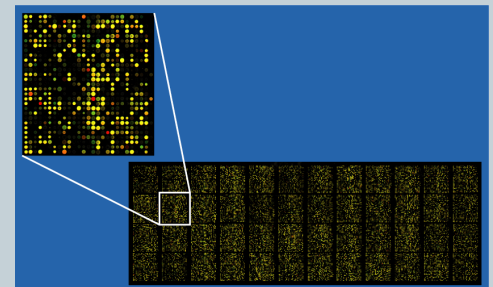
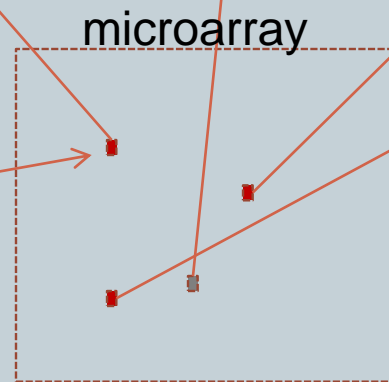


- Will be covered in detail in the Spring course High-throughput Bioinformatics (Aalto University).

- Idea:



a probe specific to the gene:  
e.g. complement of short unique fragment of cDNA



<http://en.wikipedia.org/wiki/File:Microarray2.gif>

# Time series expression profiling



- It is possible to make a series of microarray experiments to obtain a time series expression profile for each gene.



- *Cluster* similarly behaving genes.



# Analysis of clustered genes



- Similarly expressing genes may share a common transcription factor located upstream of the gene sequence.
  - Extract those sequences from the clustered genes and search for a common motif sequence.
  - Some basic techniques for Motif discovery covered in the Algorithms for Bioinformatics course.
- We concentrate now on the structure of upstream region, representation of motifs, and the simple tasks of locating the occurrences of already known motifs.

# Promoter sequences



- Immediately before the gene.
- Clear structure in prokaryotes, more complex in eukaryotes.
- An example from *E coli* is shown in next slide (taken from the course book).

# Promoter example



**Table 9.2.** A sample of *E. coli* promoter sequences. These sequences have been aligned relative to the transcriptional start site at position +1 (boldface large letter). Sequences from -40 to +11 are shown. Close matches to consensus -35 and -10 hexamers are underlined. See also Appendix C.3 for additional examples and sources of the data.

	-35	-10	-1
ORF83P1			
	CTCTGCTGGCATT <u>CACA</u> AATGCGCAGGGGT <u>AAAA</u> CGTTTC <b>C</b> TGTAGCACCG		
<i>ada</i>	GTTGGTTTTTCGGTGATGGTGACCGGGCAGCCTAAAGGCTATCCTTAACCA		
<i>amnP4</i>	TTCACATTTCTGTGACA TACTATCGGATGTGCGGTAATTGTATGGAACAGG		
<i>araFGH</i>	CTCTCCTATGGAGAATTAATTTCTCGCT <u>AAAA</u> CTATGTCAACACAGTCACT		
<i>aroG</i>	CCCCGTTTACACATTCTGACGGAAGATATAGATTGGAAGTATTGCATTAC		
<i>atpI</i>	TATTGTTT <u>GAAA</u> TCACGGGGCGCACCGTATAATTGACCGCTTTTGTATG		
<i>caiT</i>	AATCACAGAATACAGCTTATTGAATACCCATTATGAGTTAGCCATTAACGC		
<i>clpAP1</i>	TTATTGACGTGTTACAAAAATTCTTTTCTATGATGTAGAACGTGCAACGC		
<i>errP2-I</i>	GTGGTGAGCTTGCTGCCGATGAACGTGCT <u>TACACT</u> TCTGTTGCTGGGGATGG		

# Representing signals in DNA



- Consensus sequence:
  - -10 site in E coli: TATAAT
  - GRE half-site consensus: AGAACA
- Simple regular expression:
  - $A(C/G)AA(C/G)(A/T)$
- Positional weight matrix (PWM):

A	1.00	0.00	1.00	1.00	0.00	0.86
C	0.00	0.14	0.00	0.00	0.86	0.00
G	0.00	0.86	0.00	0.00	0.14	0.00
T	0.00	0.00	0.00	0.00	0.00	0.14

GRE half-sites:

AGAACA

ACAACA

AGAACA

AGAAGA

AGAACA

AGAACT

AGAACA

consensus: AGAACA

# Position-specific scoring matrix (PSSM)



- PSSM is a log-odds normalized version of PWM. <sup>1</sup>
- Calculated by  $\log(p_{ai}/q_a)$ , where
  - $p_{ai}$  is the frequency of **a** at column **i** in the samples.
  - $q_a$  is the probability of **a** in the whole organism (or in some region of interest).
- Problematic when some values  $p_{ai}$  are zero.
- Solution is to use pseudocounts:
  - add **1** to all the sample counts where the frequencies are calculated.

<sup>1</sup> In the following **log** denotes base **2** logarithm.

# PWM versus PSSM



counts

$$\begin{bmatrix} 7 & 0 & 7 & 7 & 0 & 6 \\ 0 & 1 & 0 & 0 & 6 & 0 \\ 0 & 6 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

PWM



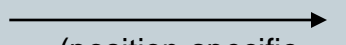
$$\begin{bmatrix} 1.00 & 0.00 & 1.00 & 1.00 & 0.00 & 0.86 \\ 0.00 & 0.14 & 0.00 & 0.00 & 0.86 & 0.00 \\ 0.00 & 0.86 & 0.00 & 0.00 & 0.14 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.14 \end{bmatrix}$$

pseudocounts



$$\begin{bmatrix} 8 & 1 & 8 & 8 & 1 & 7 \\ 1 & 2 & 1 & 1 & 7 & 1 \\ 1 & 7 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 & 2 \end{bmatrix}$$

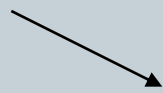
PSSM



(position-specific scoring matrix)

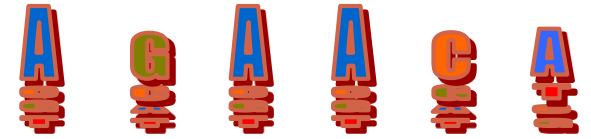
$$\begin{bmatrix} 1.54 & -1.46 & 1.54 & 1.54 & -1.46 & 1.35 \\ -1.46 & -0.46 & -1.46 & -1.46 & 1.35 & -1.46 \\ -1.46 & 1.35 & -1.46 & -1.46 & -0.46 & -1.46 \\ -1.46 & -1.46 & -1.46 & -1.46 & -1.46 & -0.46 \end{bmatrix}$$

- $\log((8/11)/(1/4))$
- $\log((1/11)/(1/4))$
- $\log((2/11)/(1/4))$
- $\log((7/11)/(1/4))$



← assuming  $q_a=0.25$  for all  $a$

# Sequence logos



- Many known transcription factor binding site PWM:s can be found from JASPAR database (<http://jaspar.cgb.ki.se/>).
- PWM:s are visualized as *sequence logos*, where the height of each nucleotide equals its proportion of the relative entropy (expected log-odds score) in that column.

- $$E(S_i) = \sum_a p_{ai} \log(p_{ai} / q_a)$$

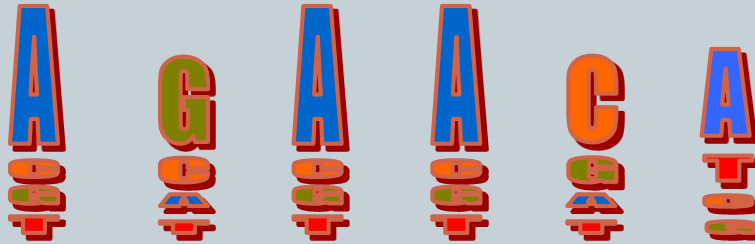
- Height of **a** at column **i** is  $p_{ai}E(S_i)$

# Example sequence logo



1.54	-1.46	1.54	1.54	-1.46	1.35
-1.46	-0.46	-1.46	-1.46	1.35	-1.46
-1.46	1.35	-1.46	-1.46	-0.46	-1.46
-1.46	-1.46	-1.46	-1.46	-1.46	-0.46

2 bits





# Searching PSSMs



- As easy as naive exact text search (see next slide).
- Much faster methods exist. For example, one can apply branch-and-bound technique on top of suffix tree (see [http://sysdb.cs.helsinki.fi/~tkt\\_suds/gb/](http://sysdb.cs.helsinki.fi/~tkt_suds/gb/) for demonstration).
- **Warning:**
  - Good hits for any PSSM are too easy to find!
  - Search domain must be limited by other means to find anything statistically meaningful with PSSMs only.
    - ✦ Typically used on upstream regions of genes clustered by gene expression profiling.

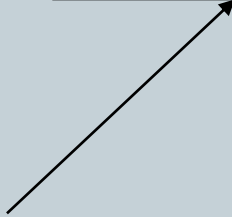
```

#!/usr/bin/env python
import sys
import time
# naive PSSM search
matrix = {'A':[1.54,-1.46,1.54,1.54,-1.46,1.35],
          'C':[-1.46,-0.46,-1.46,-1.45,1.35,-1.46],
          'G':[-1.46,1.35,-1.46,-1.46,-0.46,-1.46],
          'T':[-1.46,-1.46,-1.46,-1.46,-1.46,-0.46]}
count = {'A':0,'C':0,'G':0,'T':0}
textf = open(sys.argv[1],'r')
text = textf.read()
m=len(matrix['A'])
bestscore = -m*2.0
t1 = time.time()
for i in range(len(text)-m+1):
    score = 0.0
    for j in range(m):
        if text[i+j] in matrix:
            score = score + matrix[text[i+j]][j]
            count[text[i+j]] = count[text[i+j]]+1
        else:
            score = -m*2.0
    if score > bestscore:
        bestscore = score
        bestindex = i
t2 = time.time()
totalcount = count['A']+count['C']+count['G']+count['T']
expectednumberofhits = 1.0*(len(text)-m+1)
for j in range(m):
    expectednumberofhits = expectednumberofhits*float(count[text[bestindex+j]])/float(totalcount)
print 'best score ' + str(bestscore) + ' at index ' +str(bestindex)
print 'best hit: ' + text[bestindex:bestindex+m]
print 'computation took ' + str(t2-t1) + ' seconds'
print 'expected number of hits: ' + str(expectednumberofhits)

```

pssm.py hs\_ref\_chrY\_nolinebreaks.fa  
best score 8.67 at index 397  
best hit: AGAACA  
computation took 440.56187582 seconds  
expected number of hits: 18144.7627936

no sense in  
this search!



# Refined motifs



- Our example PSSM (GRE half-site) represents only half of the actual motif: the complete motif is a palindrome with consensus:

- AGAACAnnnTGTTCT

```
pssmpalindrome.py hs_ref_chrY_nolinebreaks.fa  
best score 17.34 at index 17441483  
best hit: AGAACAGGCTGTTCT  
computation took 1011.4800241 seconds  
expected number of hits: 5.98440033042  
total number of maximum score hits: 2
```

- Exercise: modify pssm.py into pssmpalindrome.py  
... or learn biopython to do the same in few lines of code

# Discovering motifs



- **Principle:** discover over-represented motifs from the promotor / enhancer regions of co-expressing genes.
- How to define a motif?
  - Consensus, PWM, PSSM, palindrome PSSM, co-occurrence of several motifs (enhancer modules),...
  - Abstractions of protein-DNA chemical binding.
- **Computational challenge in motif discovery:**
  - Almost as hard as (local) multiple alignment.
  - Exhaustive methods too slow.
  - Lots of specialized pruning mechanisms exist.
- **New sequencing technologies will help (ChIP-seq).**
  - Covered in the Spring course Biological Sequence Analysis.

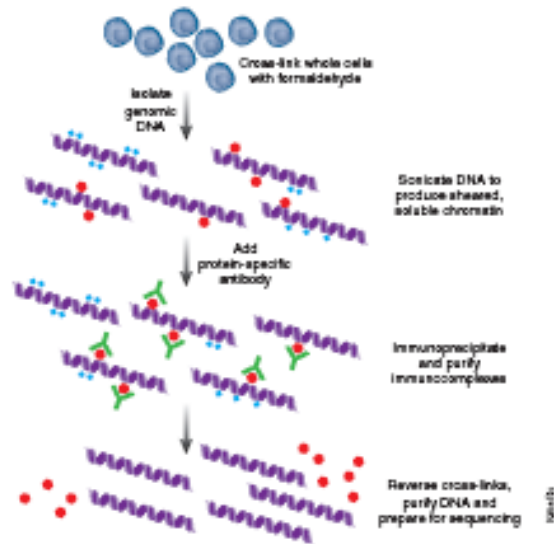
## ChIP-seq: welcome to the new frontier

Haine R Mards

Next-generation sequencing technology combines with chromatin immunoprecipitation to provide a genome-wide look at transcription-factor binding.

Next-generation sequencing technologies, capable of producing tens of millions of sequence reads during each instrument run, are quickly being applied in a myriad of creative ways to answer genome-wide

questions. In this issue, Robertson and colleagues describe such an application, comparing chromatin immunoprecipitation (ChIP) of the Stat1 transcription factor using a next-generation sequencing



**Figure 1** | Workflow of ChIP-seq. DNA and proteins are cross-linked and purified; then bound DNA is analyzed by massively parallel, short-read sequencing.

Haine R. Mards at the Washington University School of Medicine, St. Louis, Missouri © 1998, USA.  
e-mail: amards@wustl.edu

platform and a conventional microarray-based platform<sup>3</sup>.

This report provides an elegant example of the power of next-generation sequencing platforms to expand what once was a focused assay to a genome-wide scope. In the process, our ability to characterize and understand phenomena such as alterations in transcription-factor binding in response to environmental stimuli can be evaluated for the entire genome in a single experiment. Hence, the ramifications for the pace of biological inquiry and the functional annotation of genomes are profound.

ChIP, first described by Varshavsky and colleagues<sup>2</sup> as a method to study protein-DNA interactions, comprises three basic steps. First, covalent cross-links between proteins and DNA are formed, typically by treating cells with formaldehyde or another chemical reagent. In the second step, an antibody specific to the protein of interest is used to selectively coimmunoprecipitate the protein-bound DNA fragments that were covalently cross-linked. Finally, the immunoprecipitated protein-DNA links are reversed and the recovered DNA is assayed to determine the sequences bound by that protein (Fig. 1). Because random protein-DNA cross-linking can occur, and nonspecific DNA can be pulled down in the immunoprecipitation step, the ChIP-selected DNA is typically compared to a mock sample of DNA collected without antibody addition during the immunoprecipitation step. Typically, these two DNA populations are differentially labeled and compared by hybridization to a genomic microarray ('ChIP-chip'), as initially reported by Ren and colleagues<sup>3</sup> in yeast.

Although ChIP-chip approaches have greatly expanded our understanding of genome-wide protein-DNA associations, the substitution of next-generation sequencing technology to analyze the DNA fragments released after ChIP ('ChIP-seq') has distinct advantages over microarray hybridization. As shown in Robertson et al.<sup>4</sup>, the Solexa sequencing technology<sup>4</sup> provided short read length sequences of ~30 base pairs that were ideal for characterizing ChIP-derived fragments. The

ChIP, first described by Varshavsky and colleagues<sup>2</sup> as a method to study protein-DNA interactions, comprises three basic steps. First, covalent cross-links between proteins and DNA are formed, typically by treating cells with formaldehyde or another chemical reagent. In the second step, an antibody specific to the protein of interest is used to selectively coimmunoprecipitate the protein-bound DNA fragments that were covalently cross-linked. Finally, the immunoprecipitated protein-DNA links are reversed and the recovered DNA is assayed to determine the sequences bound by that protein (Fig. 1). Because random protein-

<http://www.nature.com/nmeth/journal/v4/n8/pdf/nmeth0807-613.pdf>

# Demos



- Faster PSSM search  
(<http://sysdb.cs.helsinki.fi/~tktsuds/gb/>)
  - Check also simulation of descending suffix walk covered previous week
- JASPAR (<http://jaspar.genereg.net/>)