Graduate Computational Genomics

02-710 / 10-810 & MSCBIO2070

Introduction to motif finding

Takis Benos

Lecture #10, February 15, 2007

Reading: handouts & papers

Outline

- The problem
- Motif representation
- Motif discovery
  - A greedy approach
  - Expectation-Maximization
  - Gibbs sampling
  - Self-organizing maps

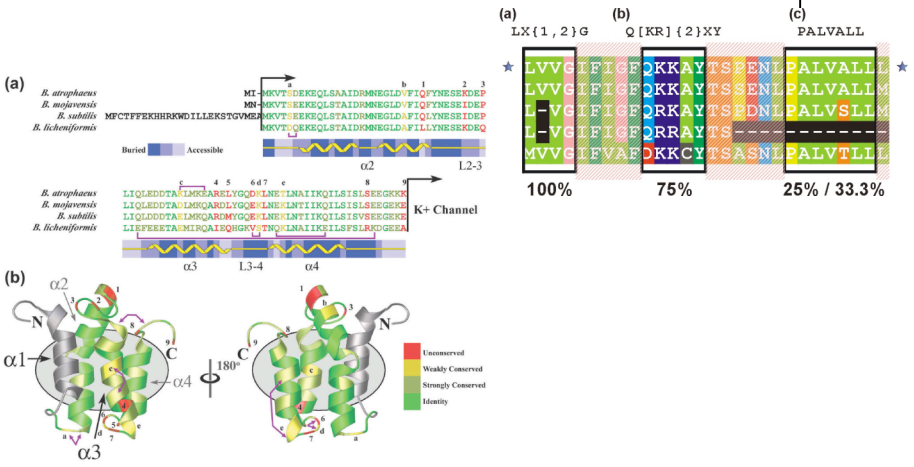
The problem

- Motif is a pattern that nature has preserved in biological sequences, usually for a reason.
- Motifs can be viewed as degenerate sequences (strings).
- Motifs are typically detected either by examining one gene in multiple species or by examining many genes in one species.

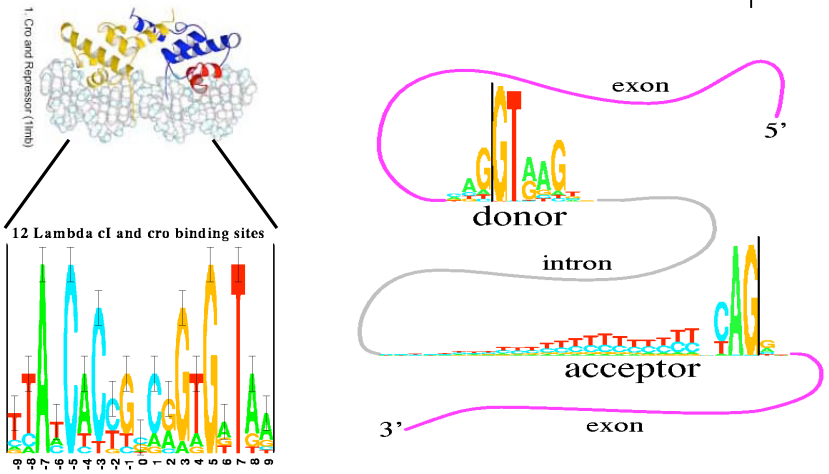
Why find motifs?

- In proteins - may be an important component
  - Find similarities to known proteins
  - Find important areas of a protein family
- In DNA - may be a *regulatory element*
  - Discover how the gene expression is regulated

Protein motifs



DNA motifs



Pattern representation

- Majority rule
- String representation
  - Consensus sequence
  - Regular expressions

```
C-26    AGGATATT
C-57    AGGATATT
C-25    CATATTTT
7       AGAGTTTT
67      AGCATTTT
        AGNATWTT
```

C - X(2,4) - C - X(3) - [LIVMFYWC] - X(8) - H - X(3,5) - H

1SP1:
KKFA- C - PE - C - PKR - F - MRSDHLSK - H - IKT - H - QNKK

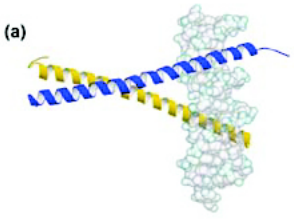
Pattern representation

- Majority rule
- String representation
  - Consensus sequence
  - Regular expressions
- Probabilistic modeling
  - Weight matrices (PSSM models)
  - ✓ HMMs

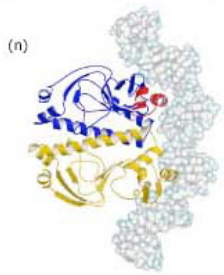
Transcription factors

- *Function:* they recognise a specific DNA target sequence (typically in the proximity of the gene) and facilitate transcription initiation and/or repression.
- Preferential DNA-binding.
- Tolerant in (some) base substitutions at the most preferred target sequence.
- Moderate change in affinity between "good" targets.
- TF binding patterns are generally more difficult to detect than protein patterns. *(why?)*

Transcription factors (cntd)



21. Leucine zipper (2dgc)



14. Catabolic gene activator (2cgp)

Source: Luscombe et al. (2000) Genome Biol. 1(1):REVIEWS001

Consensus representation

Sequences:

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

Putative new targets:

```
A G A G T T G T
G G A G T T T T
```


Position Specific Scoring Matrices (PSSM)

A	9	214	63	142	118	8
C	22	7	26	31	52	13
G	18	2	29	38	29	5
T	193	19	124	31	43	216

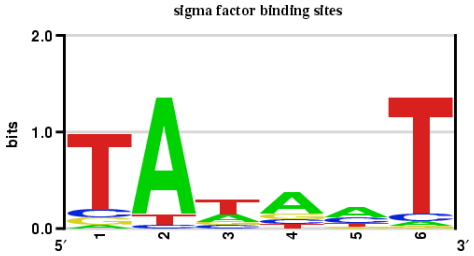
Counts from 242 known σ^{70} sites

A	.04	.88	.26	.59	.49	.03
C	.09	.03	.11	.13	.21	.05
G	.07	.01	.12	.16	.12	.02
T	.80	.08	.51	.13	.18	.89

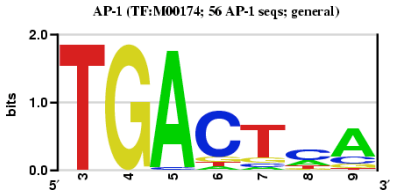
Relative frequencies $f_b(i)$

A	-38	19	1	12	10	-48
C	-15	-38	-8	-10	-3	-32
G	-13	-48	-6	-7	-10	-40
T	17	-32	8	-9	-6	19

$10 \times \log_2 f_b(i)/f_{REF}$



Sequence LOGO



$$Height_{col}(x) = 2 + \underbrace{\sum_{b=A}^{T} f(b,x) \cdot \log_2 f(b,x)}_{\text{Shannon's entropy } (H)}$$

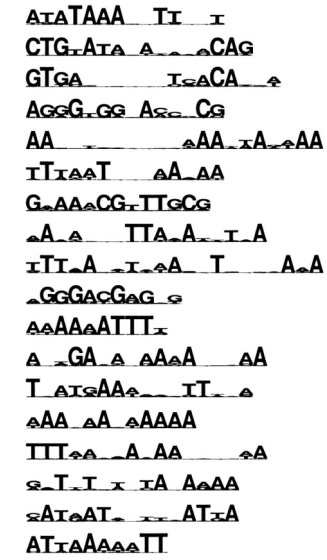
- A motif is interesting if it differs from the background. So…

$$RH(\text{mod}el) = \sum_{x=1}^{L} \sum_{b=A}^{T} f(b,x) \cdot \ln \frac{f(b,x)}{f_{REF}(x)}$$

Relative Entropy (Information Content)

not a proper metric!!

Characteristics of Regulatory Motifs



- Tiny
- Highly Variable
- ~Constant Size
  - Because a constant-size transcription factor binds
- Often repeated
- Low-complexity-ish

A little more about PSSMs

- The PSSM model is a convenient representation of a transcription factor binding preferences.
- Scoring a sequence against a PSSM model calculates the log-likelihood ration of the sequence coming from the model *vs.* the background model.
- PSSM models assume position independency, meaning that the nucleotide distributions between two positions are independent.
- PSSM scores are related to interaction energy *via* the Boltzmann distribution.

1.- PSSM scoring

	C	T	A	T	A	A	T	C
A	-38	19	1	12	10	-48		
C	-15	-38	-8	-10	-3	-32		
G	-13	-48	-6	-7	-10	-40		
T	17	-32	8	-9	-6	19		

-93

	C	T	A	T	A	A	T	C
A	-38	19	1	12	10	-48		
C	-15	-38	-8	-10	-3	-32		
G	-13	-48	-6	-7	-10	-40		
T	17	-32	8	-9	-6	19		

+85

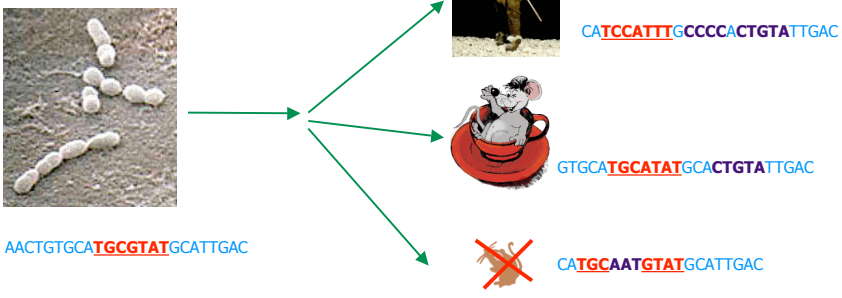
	C	T	A	T	A	A	T	C
A	-38	19	1	12	10	-48		
C	-15	-38	-8	-10	-3	-32		
G	-13	-48	-6	-7	-10	-40		
T	17	-32	8	-9	-6	19		

-95

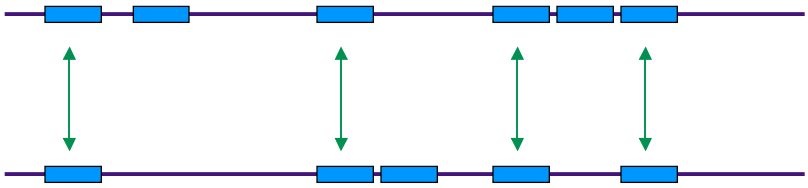
More on PSSM scoring

- PSSM scanning can be used to identify sites that match the profile in new promoters.
- Setting up a threshold is somewhat arbitrary.
  - Controlling FDR
  - Controlling false positives
- Even if the threshold is perfect, still some sequences will turn up as "positive" (why?)
- Solution: ask nature's help!

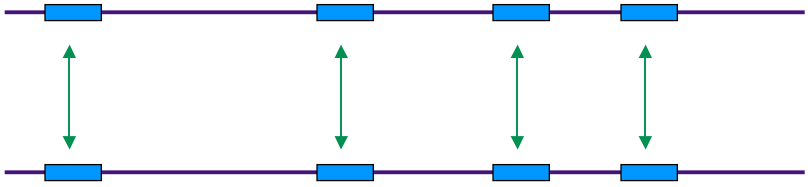
Evolution of regulatory regions



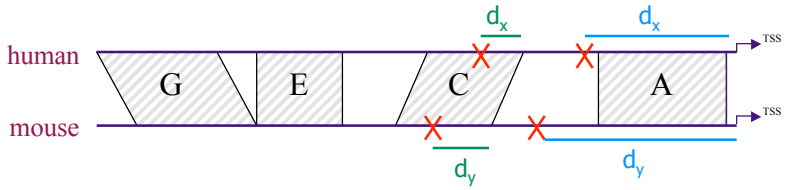
Phylogenetic footprinting



Phylogenetic footprinting



FOOTER scoring



$$PF_D = P(d_{XY} \le d_1) = \frac{1}{D} + \sum_{k=1}^{d_1} \frac{2 \cdot (D-k)}{D^2}$$

$$PF_S = P((S+T) \le (s+t) \mid M_1, M_2)$$

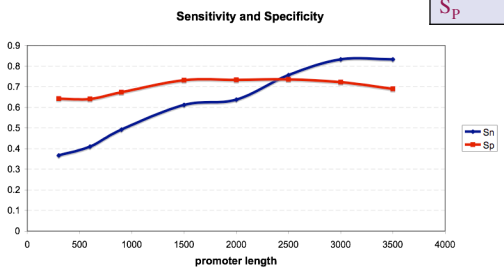
$$PF = -w_D \cdot \ln(PF_D) - w_S \cdot \ln(PF_S)$$

FOOTER performance

$$S_N = \frac{TP}{TP + FN}$$

$$S_P = \frac{TP}{TP + FP}$$

	Footer	*ConSite* (def)	*ConSite* (adj)	*rVista* v1.0
# of sites	**72**	49	49	69
TP	**60**	23	34	54
FP	**23**	15	28	189
S_N	**83.3%**	46.9%	69.4%	78.2%
S_P	**72.3%**	60.5%	54.8%	22.2%



PSSM to Binding Energy

- Assuming equilibrium…

$$P(d \mid p) = \frac{P_{REF}(d) \cdot e^{-E(p,d)/k_B T}}{\sum_{x_i} P_{REF}(x_i) \cdot e^{-E(p,x_i)/k_B T}} \Rightarrow \frac{P(d \mid p)}{P_{REF}(d)} = \frac{e^{-E(p,d)/k_B T}}{Z} \Rightarrow$$

$$\Rightarrow \ln \frac{P(D \mid P)}{P_{REF}(D)} = -E(p,d)/k_B T - c$$

Ludwig Boltzmann (1844-1906)

- The PSSM log-likelihood score is equal to the binding energy (in $k_B T$ units) minus the average background energy.
- The absolute energy is not important.
  - Try adding/subtracting a number from all energy values

Acknowledgements

Some of the slides used in this lecture are adapted or modified slides from lectures of:

- Serafim Batzoglou, Stanford University
- Eric Xing, Carnegie-Mellon University

Theory and examples from the following:

- http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html