Finding biological sequence motifs

November 9th 2006

Ackn.: CPSC 545/445; CPSC 536A, 2001/2002; CS527, 2000]

How?

Range of the problem: identification of long functional regions such as genes (see lecture Pierre Rouze), as well as shorter functional regions such as signals.

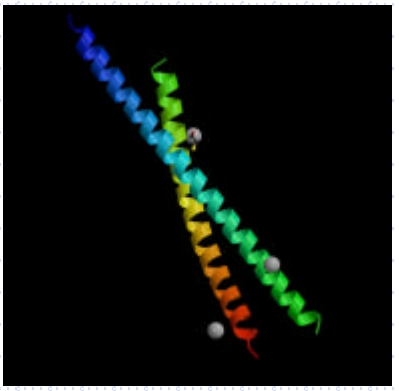
We can subdivide the problem further into:

- finding instances of a known site
- finding instances of unknown sites

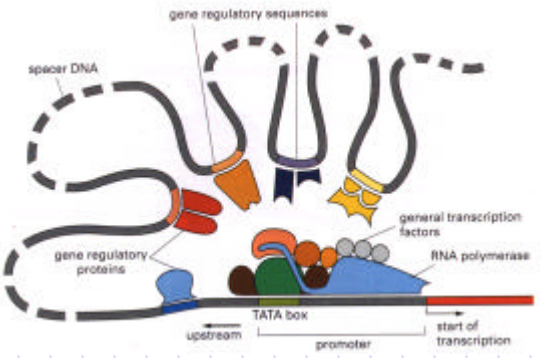
For this discussion, we will concentrate on the detection of shorter functional regions such as regulatory sequences in DNA

Motifs in Protein Sequences

- The leucine zipper may explain how some eukaryotic gene regulatory proteins work.
- L-x(6)-L-x(6)-L-x(6)-L
- The leucine side chains extending from one alpha-helix interact with those from a similar alpha helix of a second polypeptide, facilitating dimerization

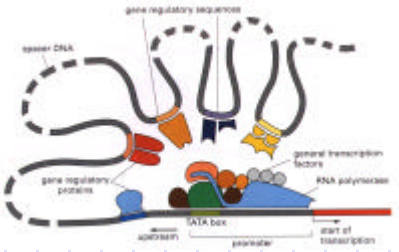


Motifs in DNA Sequences

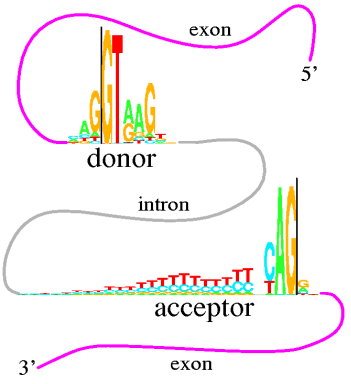


Motifs in DNA Sequences

- Promoter regions, e.g. TATA box
- Transcription factor binding sites, e.g. Eve in *Drosophila:*
  - G-G-T-C-C-T-G-G
- Cis-Regulatory regions



Motifs in RNA sequences

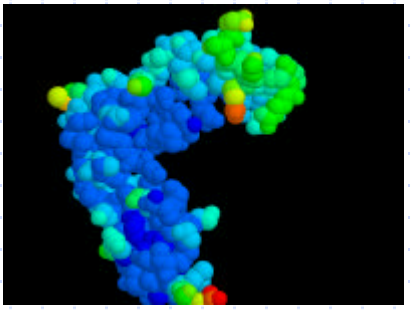


Human RNA-splice junctions sequence matrix

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

Motifs in Protein Structures

- Protein structure patterns can encode information about protein function.
- Structure motifs can be used to improve multiple alignments of protein sequences.



Regulation of Expression

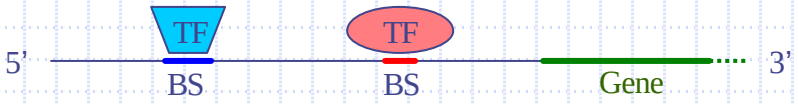
- Each cell contains a copy of the whole genome
- BUT utilizes only a subset of the genes
- Most genes are highly regulated –
  their expression is limited to specific tissues, developmental stages, physiological condition

How is the expression of genes regulated?

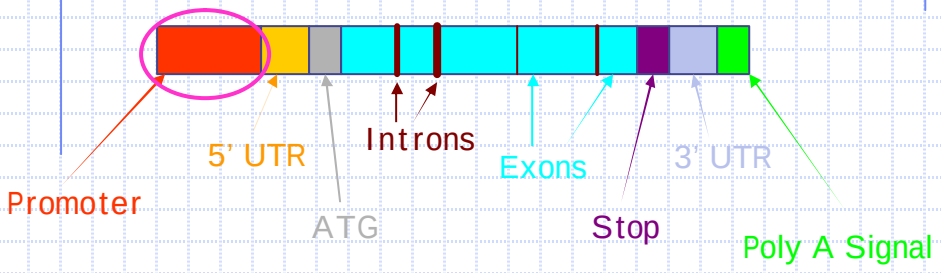
One way is through *transcriptional regulation*

Regulation of Transcription

- The conditions in which a gene is transcribed are mainly encoded in the DNA in a region called "promoter"
- Each promoter contains several short DNA subsequences, called "binding sites" (BSs) that are bound by specific proteins called "transcription factors" (TFs)



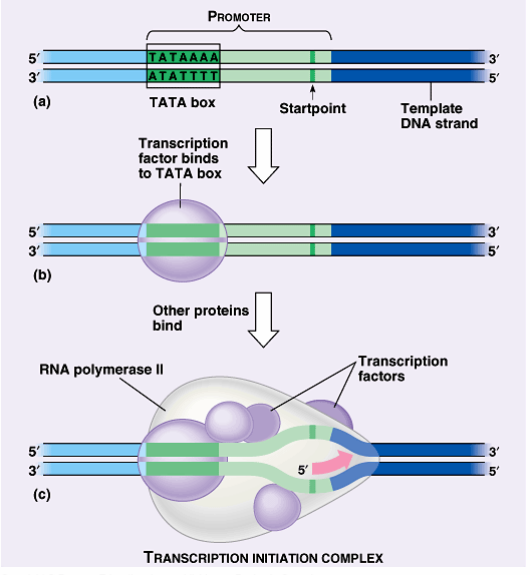
Gene Structure



Transcription Factors

- Proteins involved in the regulation of gene expression that bind to the promoter elements upstream of transcription initiation sites
- Composed of two essential functional regions: a DNA-binding domain and an activator domain.

Transcription – A quick review

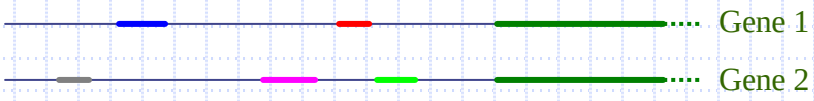


Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

http://www.msu.edu/course/lbs/145/smith/s02/graphics/campbell_17.7.gif

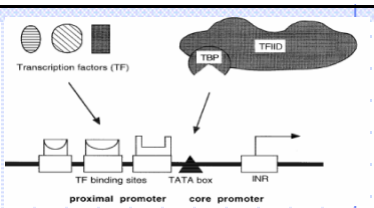
Regulation of Transcription

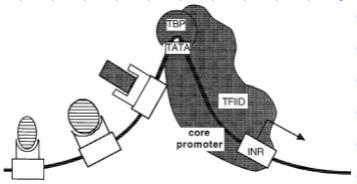
- By binding to a gene's promoter, TFs can either *promote* or *repress* the recruitment of the transcription machinery
- The conditions in which a gene is transcribed are determined by the specific combination of BSs in its promoter

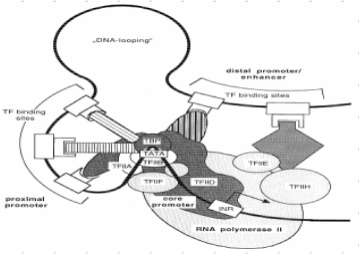


Key events in transcriptional initiation

- Transcription factors (TFs) bind to upstream promoter sequences to form a multiprotein complex.
- Recruits a pol II and some GTFs to the transcription start site.

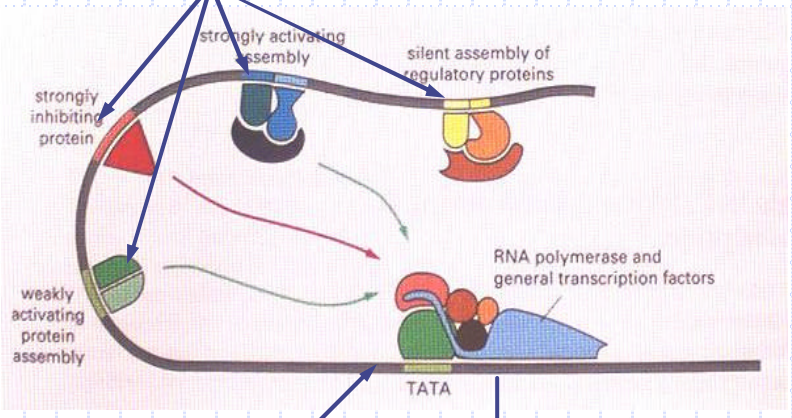






Regulation of Transcription

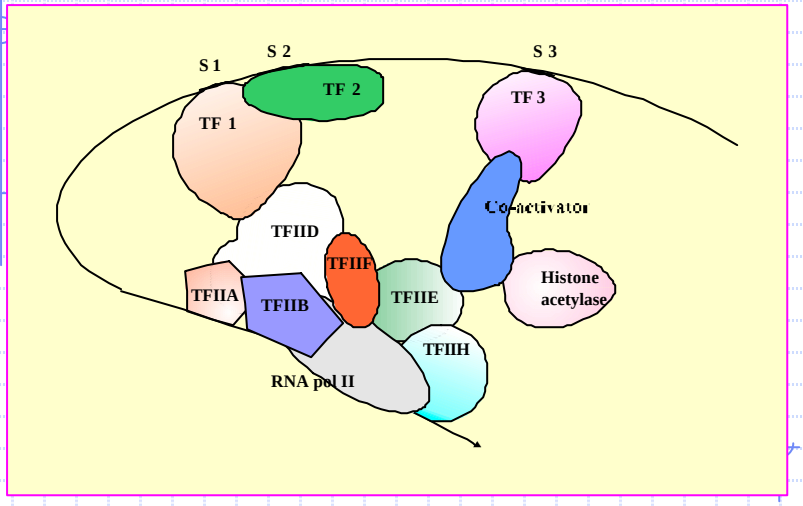
TFs bound to their BSs

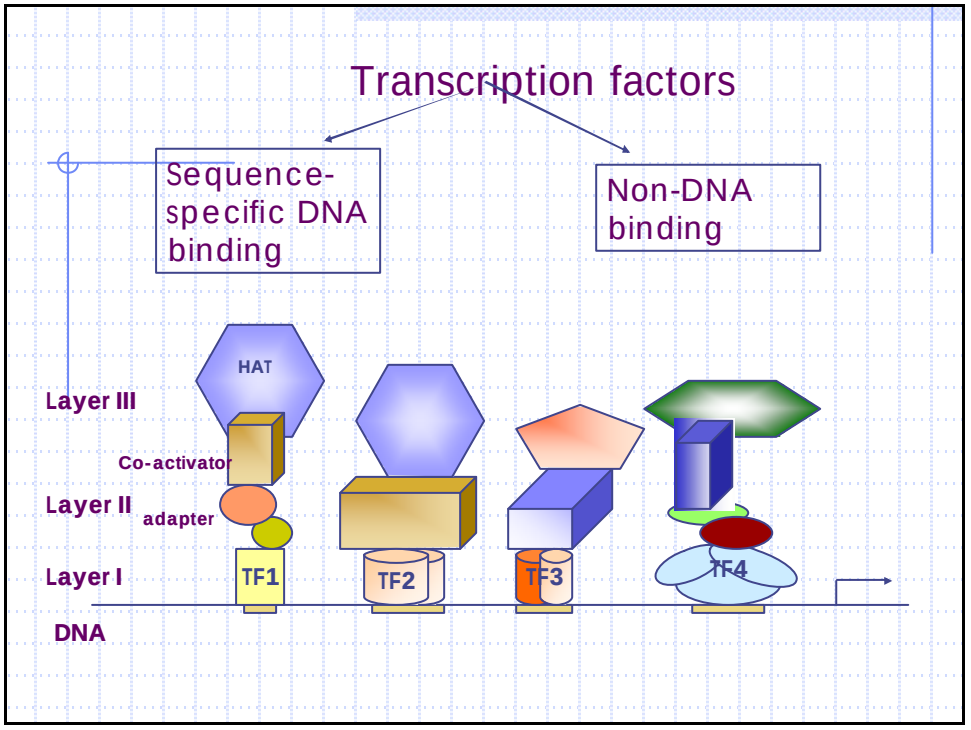


Transcription machinery

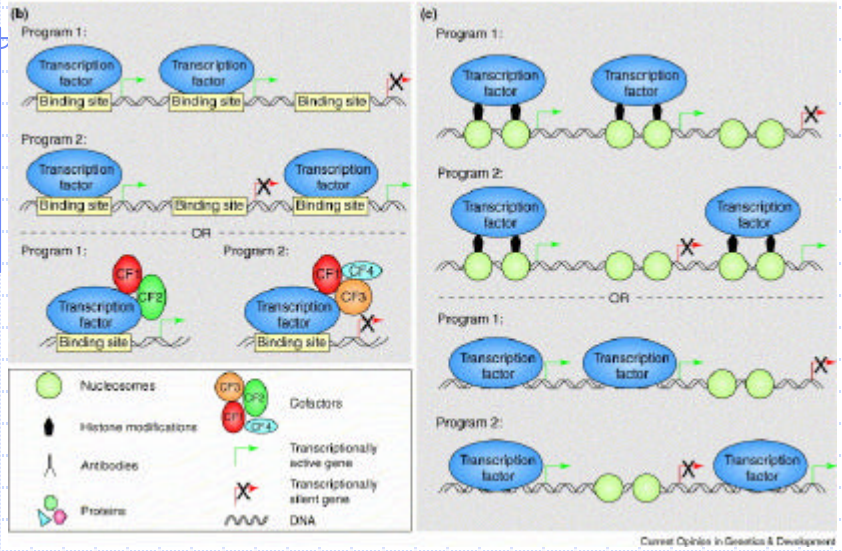
Gene start

Protein-DNA and protein-protein interactions in gene transcriptional regulation.





Single TF-Multiple Responses



Hanlon and Lieb (2004) Curr. Opin. Gen. & Dev. 14:697-705

What is a promoter?

- A sequence that is used to initiate and regulate transcription of a gene.
- The minimum region of DNA allowing formation of a functional initiation complex
- Most genes in higher eukaryotes are transcribed from polymerase II dependent promoters.

Two major class of mammalian promoter

- TATA-box containing promoter
  - Minority
  - Tissue-specific
  - High conservation
  - Exonic promoter activity
  - More constrained
- CpG island-associated promoter
  - Majority
  - Rapidly evolving
  - Bidirectional promoters
  - Epigenetic control of transcriptional activity

Significance of promoter study

- Regulation mechanisms study
- Tissue-specific promoter identification
- Gene therapy targeting
- Variation origin of some phenotypic traits

Promoters identification

- Very difficult
- No good tools yet

Why is promoter prediction so difficult?!

- Not one single type of core promoter
- Promoters are dependent on additional regulatory elements
- Transcription may be activated, enhanced or repressed by regulatory proteins/protein complex
- Cis-activation factor is short, but the recognize sites are highly similar.
- Transcriptional activators and repressors act very specifically both in terms of the cell type and time in the cell cycle
- Many regulatory factors have not been characterized yet

Problems to be solved

- No well defined “Core promoter”
- Promoter control depends on regions both upstream and downstream of the promoter region
- The transcriptional machinery is capable of recognize Promoters in contrast with present statistical data that suggest that the regulatory elements do not contain sufficient information to do so

Experimental Methods for promoter analysis

- High-throughput
  - CSGE: Cap analysis of gene expression
  - CHIP & CpG island microarray analysis
  - Genome Sequencing & Bioinformatics prediction
- Experiments
  - CHIP
  - Expression & Reporter gene
  - EMSA (gel shift)

Regulation of Transcription

Assumption:

Co-expression

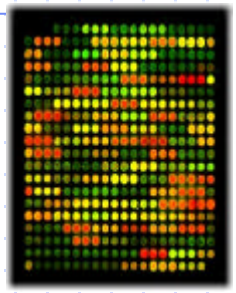
?

Transcriptional co-regulation

?

Common BSs

DNA chips



?

Data analysis (normalization, clustering)

?

Co-expression

Human Genome



?

Promoter analysis (find common BSs)

?

Co-regulation

Promoter Region

What is the promoter region?

- Upstream Transcription Start Site (TSS)
  - Too short ? miss many real BSs (false negatives)
  - Too long ? lots of wrong hits (false positives)
  - Length is species dependent (e.g., yeast ~600bp, thousands in human)
  - Common practice: ~ 500-2000bp
- Mask-out repetitive sequences?
  - Common practice: Yes
- Consider both strands?
  - Common practice: Yes

The What? question

Computational tasks:

- New BSs of known TFs
- New motifs (BSs of unknown TFs)
- Modules = combinations of TFs

BSs Models

(a) Exact string(s)

Example:

BS = TACACC , TACGGC

CAATGCAGGATACACCGATCGGTA
GGAGTACGGCAAGTCCCCATGTGA
AGGCTGGACCAGACTCTACACCTA

BSs Models (II)

(b) String with mismatches

Example:

BS = TACACC + 1 mismatch

CAATGCAGGATTCACCGATCGGTA
GGAGTACAGCAAGTCCCCATGTGA
AGGCTGGACCAGACTCTACACCTA

BSs Models (III)

(c) Degenerate string

Example:

BS = TASDAC (S={C,G} D={A,G,T})

CAATGCAGGATACAACGATCGGTA
GGAGTAGTACAAGTCCCCATGTGA
AGGCTGGACCAGACTCTACGACTA

BSs Models (IV)

(d) Position Weight Matrix (PWM)

Example: BS =

A	0.1	0.8	0	0.7	0.2	0
C	0	0.1	0.5	0.1	0.4	0.6
G	0	0	0.5	0.1	0.4	0.1
T	0.9	0.1	0	0.1	0	0.3

Need to set score threshold

ATGCAGGATACACCGATCGGTA 0.0605

GGAGTAGAGCAAGTCCCGTGA 0.0605

AAGACTCTACAATTATGGCGT 0.0151

BSs Models (V)

(e) More complex models

- PWM with spacers (e.g., for p53)
- Markov model (dependency between adjacent columns of PWM)
- Hybrid models, e.g., mixture of two PWMs
- ...

... And we also need to model the non-BSs sequences in the promoters...

Motif Representations

1. Consensus

CGGCGCACTCTCGCCCG
CGGGGCAGACTATTCCG
CGGCGGCTTCTAATCCG
. . .
CGGGGCAGACTATTCCG

→ **CGGNGCACANTCNTCCG**

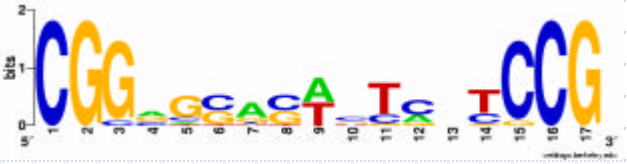
2. Frequency Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	0	0	0	$\frac{7}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{9}{14}$	0	$\frac{6}{14}$	$\frac{1}{14}$	0	$\frac{3}{14}$	$\frac{1}{14}$	$\frac{2}{14}$	0	0	0
C	$\frac{14}{14}$	0	$\frac{1}{14}$	$\frac{3}{14}$	$\frac{3}{14}$	$\frac{6}{14}$	$\frac{3}{14}$	$\frac{8}{14}$	0	$\frac{5}{14}$	$\frac{3}{14}$	$\frac{7}{14}$	$\frac{5}{14}$	$\frac{3}{14}$	$\frac{12}{14}$	$\frac{14}{14}$	0
G	0	$\frac{14}{14}$	$\frac{13}{14}$	$\frac{4}{14}$	$\frac{9}{14}$	$\frac{6}{14}$	$\frac{1}{14}$	$\frac{5}{14}$	0	$\frac{6}{14}$	$\frac{1}{14}$	$\frac{2}{14}$	$\frac{6}{14}$	$\frac{1}{14}$	$\frac{2}{14}$	0	$\frac{14}{14}$
T	0	0	0	0	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{8}{14}$	$\frac{2}{14}$	$\frac{10}{14}$	$\frac{2}{14}$	$\frac{2}{14}$	$\frac{8}{14}$	0	0	0

3. Logo

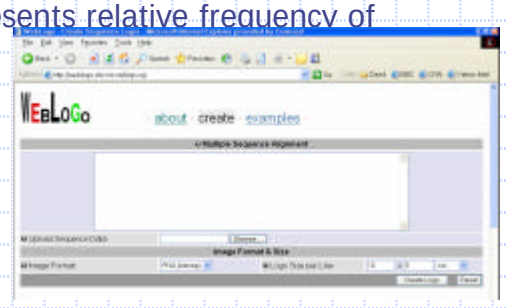


Logos



- Graphical representation of nucleotide base (or amino acid) conservation in a motif (or alignment)
- Information theory $2 + \sum_{b=\{A,C,G,T\}} p(b)\log_2 p(b)$
- Height of letters represents relative frequency of nucleotide bases

http://weblogo.berkeley.edu/



How to find novel motifs

Degenerate string:

- YMF - Sinha & Tompa '02

String with mismatches:

- WINNOWER – Pevzner & Sze '00
- Random Projections – Buhler & Tompa '02
- MULTIPROFILER – Keich & Pevzner '02

PWM:

- MEME – Bailey & Elkan '95
- AlignACE – Hughes et al. '98
- CONSENSUS - Hertz & Stormo '99

How to find TF modules

- **BioProspector** – Liu et al. '01
- **Co-Bind** – GuhaThakurta & Stormo '01
- **MITRA** – Eskin & Pevzner '02
- **CREME** – Sharan et al. '03
- **MCAST** – Bailey & Noble '03

Novel Motif Prediction

- Goal: Characterize and predict locations of novel motif in sequences
- Challenges:
  - Short (6-20 bases)
  - Degenerate
  - Locations not fixed
  - Signal to noise
    - *eg.*, yeast 600-800bps

Motif-finding Methods

- Methods:
  - Word enumeration method
  - Gibbs sampling
  - Random projection
  - Phylogenetic footprinting
  - Reducer

Algorithms

- Pattern-Driven
  - TRANSFAC
  - rVISTA
- Sequence-Driven
  - FootPrinter
  - MEME
  - BioProspector
  - AlignACE

How to summarize known sites?

Given:

A: large sample of length n sites

B: large sample of length n nonsites

s: sequence of length n ($s_1 s_2 ... s_n$)

Asked:

Is s more likely to be a site or a nonsite?

Positions 3–9 (out of the 22 sequence positions) from 23 CRP Binding Sites

CRP: cyclic AMP receptor protein (E. coli)

TTGTGGC
TTTTGAT
AAGTGTC
ATTTGCA
CTGTGAG
ATGCAAA
GTGTTAA
ATTTGAA
TTGTGAT
ATTTATT
ACGTGAT
ATGTGAG
TTGTGAG
CTGTAAC
CTGTGAA
TTGTGAC
GCCTGAC
TTGTGAT
TTGTGAT
GTGTGAA
CTGTGAC
ATGAGAC
TTGTGAG

Describing Motifs using Frequency Matrices

Definition:

For a motif of length n using an alphabet of c characters, a frequency matrix A is a c by n matrix in which each element contains the frequency at which a given member of the alphabet is observed at a given position in an aligned set of sequences containing the motif

Profile for the 23 CRP sites:

A	0.35	0.043	0	0.043	0.13	0.83	0.26
C	0.17	0.087	0.043	0.043	0	0.043	0.3
G	0.13	0	0.78	0	0.83	0.043	0.17
T	0.35	0.87	0.17	0.91	0.043	0.087	0.26

Features:

- 4 x 7 matrix
- The profile shows the distribution of residues in each of the n positions

Simplifying assumptions:

- consider only motifs with same length
- do not allow gaps
- consider DNA sequences

Using Probabilities to Test for Sites

Given:

t: randomly and uniformly chosen from A ($t=t_1 t_2 \ldots t_n$)

Then:

$$A_{r,j} = \Pr(t_j = r \mid t \in A)$$

"$A_{r,j}$ is the probability that the j-th residue of t is the residue r, given that t is chosen randomly from A."

The Independence Assumption

"which residue occurs at a certain position is independent of the residues occurring at other positions.
In other words, residues at any two different positions are uncorrelated."

Justification:

1. It keeps the model and resulting analysis simple.
2. its predictive power in some (but admittedly not all) situations.

Independent Events

Definition:

Two probabilistic events E and F are said to be independent if the probability that they both occur is the product of their individual probabilities:

$$\Pr(E \cap F) = \Pr(E).\Pr(F)$$

Probability of a site having a specified sequence

the probability that a randomly chosen site has a specified sequence $r_1, r_2, \ldots r_n$ is determined by:

$$\Pr(t = r_1 r_2 \ldots r_n \mid t \in A) = \Pr(t_1 = r_1 \cap t_2 = r_2 \ldots t_n = r_n \mid t \in A)$$

$$\Leftrightarrow \prod_{j=1}^{n} \Pr(t_j = r_j \mid t \in A)$$

$$\Leftrightarrow \prod_{j=1}^{n} A_{r_j, j}$$

What is the probability that a randomly chosen CRP binding site will be TTGTGAC?

Given:

A	0.35	0.043	0	0.043	0.13	0.83	0.26
C	0.17	0.087	0.043	0.043	0	0.043	0.3
G	0.13	0	0.78	0	0.83	0.043	0.17
T	0.35	0.87	0.17	0.91	0.043	0.087	0.26

Pr(t = TTGTGAC | t is a site) = (.35)(.87)(.78)(.91)(.83)(.83)(.3) = 0.045

Likelihood ratio

Using the same approach, we can form the c x n profile from the sample B of nonsites.

Given:

$s = s_1 s_2 \ldots s_n$

The likelihood ratio, LR(A,B,s) is then defined to be:

$$\frac{\Pr(t = s | t \in A)}{\Pr(t = s | t \in B)} = \frac{\prod_{j=1}^{n} A_{s_j, j}}{\prod_{j=1}^{n} B_{s_j, j}} = \prod_{j=1}^{n} \frac{A_{s_j, j}}{B_{s_j, j}}$$

Likelihood Ratio - Example

Given:

$B = \{A,C,G,T\}^7$ (the set of all length 7 sequences)

$B_{r,j} = 0.25$ for all r and j

s = TTGTGAC

Calculate LR(A,B,s)

$$LR(A,B,s) = \frac{\prod_{j=1}^{n} A_{s_j, j}}{\prod_{j=1}^{n} B_{s_j, j}} = \frac{0.045}{(0.25)^7} = 732$$

The need for a "cutoff" value

To test a sequence s, compare LR(A,B,s) to a prespecified constant "cutoff" L, and declare s more likely to be a site if:

$$LR(A,B,s) \geq L$$

The Log Likelihood Ratio

Given the sequence $s=s_1s_2...s_n$, the log likelihood ratio (LLR(A,B,s)) is defined to be:

$$LLR(A,B,s) = \log_2 \prod_{j=1}^{n} \frac{A_{s_j,j}}{B_{s_j,j}} = \sum_{j=1}^{n} \log_2 \frac{A_{s_j,j}}{B_{s_j,j}}$$

The corresponding test of s that s is more likely to be a site becomes:

$$LLR(A,B,s) \geq \log_2 L$$

Log Likelihood Weight Matrix for CRP Binding Sites

A	0.48	−2.5	$-\infty$	−2.5	−0.94	1.7	0.061
C	−0.52	−1.5	−2.5	−2.5	$-\infty$	−2.5	0.28
G	−0.94	$-\infty$	1.6	$-\infty$	1.7	−2.5	−0.52
T	0.48	1.8	−0.52	1.9	−2.5	−1.5	0.061

Practical:

- Create a scoring matrix W whose entries are the log likelihood ratios:

$$W_{r,j} = \log_2 \frac{A_{r,j}}{B_{r,j}}$$

- In order to compute LLR(A,B,s), add the corresponding scores from W:

$$LLR(A,B,s) = \sum_{j=1}^{n} W_{s_j,j}$$

Small sample correction

When $A_{r,j}=0$ then $W_{r,j}$ becomes –infinity!

Solutions:

1. Increase the sample A of sites
2. Replace $A_{r,j}$ by a small, positive number

Weight Matrices

Definition:

A *weight matrix* is any c x n matrix W that assigns a score to each sequence $s=s_1s_2...s_n$ according to the formula:

$$\sum_{j=1}^{n} W_{s_j,j}$$

A	0.48	−2.5	$-\infty$	−2.5	−0.94	1.7	0.061
C	−0.52	−1.5	−2.5	−2.5	$-\infty$	−2.5	0.28
G	−0.94	$-\infty$	1.6	$-\infty$	1.7	−2.5	−0.52
T	0.48	1.8	−0.52	1.9	−2.5	−1.5	0.061

How Informative is a Weight Matrix?

How good is it in distinguishing between sites and nonsites?

Some Definitions:

Sample Space:

A sample space S is the set of all possible values of some random variable s.

Probability Distribution:

A probability distribution P for a sample space S assigns a probability P(s) to every s from S, satisfying:

1. $0 \leq P(s) \leq 1$

2. $\sum_{s \in S} P(s) = 1$

For us:
Sample space = set of all length n sequences
The site profile A induces a probability distribution on this sample space as does the nonsite profile B

Some Definitions (cont.)

Relative Entropy:

Let P and Q be probability distributions on the same sample space S. The relative entropy (or "information content", or "Kullback-Leibler meisure" of P with respect to Q is defined as follows:

$$D_b(P\|Q) = \sum_{s\in S} P(s)\log_b \frac{P(s)}{Q(s)}$$

The RE corresponds to a weighted average of the LLR with weights P(s).

Some Definitions (cont.)

Expected value:

The expected value of a function f(s) with respect to a probability distribution P on sample space S is:

$$E(f(s)) = \sum_{s\in S} P(s) f(s)$$

Thus, the relative entropy is the expected value of the LLR(P,Q,s)

In general, the RE measures how different the two distributions P and Q are.

In our case, we want RE to be large and we will use it as a measure of how informative the LLR test is.

The Background Distribution

$$LLR(A,B,s) = \sum_{j=1}^{n} \log_2 \frac{A_{s_j,j}}{B_{s_j,j}}$$

$B_{sj,j}$: (often) the "background" distribution of residue s_j in the entire genome, or a large portion of the genome.

$B_{sj,j}$ is not always 0.25 in the case of nucleotides!!!

Example: *Methanococcoccus jannaschii*
$B_{A,j}=B_{T,j}=0.34$
$B_{C,j}=B_{G,j}=0.16$

A Translation Start Site example

Given:

1. A uniform background distribution
   $B_{r,j}=0.25$
2. 8 (hypothetical) TSSs:

ATG
ATG
ATG
ATG
ATG
GTG
GTG
TTG

Analysis

A	0.625	0	0
C	0	0	0
G	0.25	0	1
T	0.125	1	0

(a)

Profile

A	1.32	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	0	$-\infty$	2
T	-1	2	$-\infty$

(b)

Log Likelihood Weight Matrix

	0.701	2	2

(c)

Positional Relative Entropies

The Role of the Background Distribution (I)

A	0.625	0	0
C	0	0	0
G	0.25	0	1
T	0.125	1	0

(a)

A	1.32	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	0	$-\infty$	2
T	-1	2	$-\infty$

(b)

	0.701	2	2

(c)

Pos.2:

- A,C,G do not contribute to RE (c)
- T contributes $1.W_{T,2} = 2$

$\Rightarrow$ 2 bits of information in pos. 2

Pos.3: similar to pos.2

Pos.1:

- RE is 0.7 => more similar to the background distribution than columns 2 and 3.

Total RE = 4.7

The Role of the Background Distribution (II) - nonuniform

Given:

$B_{A,j} = B_{T,j} = 0.375$

$B_{C,j} = B_{G,j} = 0.125$

The site profile:

A	0.625	0	0
C	0	0	0
G	0.25	0	1
T	0.125	1	0

Calculate the log likelihood weight matrix and the total and positional RE.

Result

A	0.737	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	1	$-\infty$	3
T	−1.58	1.42	$-\infty$
	0.512	1.42	3

1. RE of each position has changed: the last 2 columns no longer have equal entropy
2. RE of pos.2 is now closer to the background distribution (G is rarer in the background distribution)
3. RE=3 => G is $2^3 = 8$ times more likely to occur in the third position of a site than a nonsite
4. The total RE is 4.93

Exercise: Calculate the positional relative entropy for our CRP sites:

Given:

The Profile:

A	0.35	0.043	0	0.043	0.13	0.83	0.26
C	0.17	0.087	0.043	0.043	0	0.043	0.3
G	0.13	0	0.78	0	0.83	0.043	0.17
T	0.35	0.87	0.17	0.91	0.043	0.087	0.26

The Weight Matrix:

A	0.48	-2.5	$-\infty$	-2.5	-0.94	1.7	0.061
C	-0.52	-1.5	-2.5	-2.5	$-\infty$	-2.5	0.28
G	-0.94	$-\infty$	1.6	$-\infty$	1.7	-2.5	-0.52
T	0.48	1.8	-0.52	1.9	-2.5	-1.5	0.061

Result:

0.12	1.3	1.1	1.5	1.2	1.1	0.027

Recap

- Problem 1: Given a motif, finding its instances
- Problem 2: Finding motif ab initio.
  - Paradigm: look for over-represented motifs
  - Gibbs sampling

Finding Instances of Unknown Sites

- Problem: given a set of biological sequences, find instances of a short site that occur more often than you would expect by chance, with no *a priori* knowledge about the site.
- Given a collection of such instances, this induces a profile A. From the background, we compute a profile B. From A and B, we compute the RE and use this as a measure of how good the collection is.
- Goal: Find a collection that maximizes RE
- Computationally stated: take as inputs k sequences and an integer n, and output one length n substring from each input sequence, such that the resulting relative entropy is maximized.
- This the *relative entropy site selection proble*m.
- Unfortunately, this problem is likely to be computationally intractable (Akutsu, 1998).

Greedy Algorithm

Greedy algorithms pick the locally best choice at each step, without concern for the impact on future choices.

$\Rightarrow$may result in solutions that are far from optimal

Greedy Algorithm

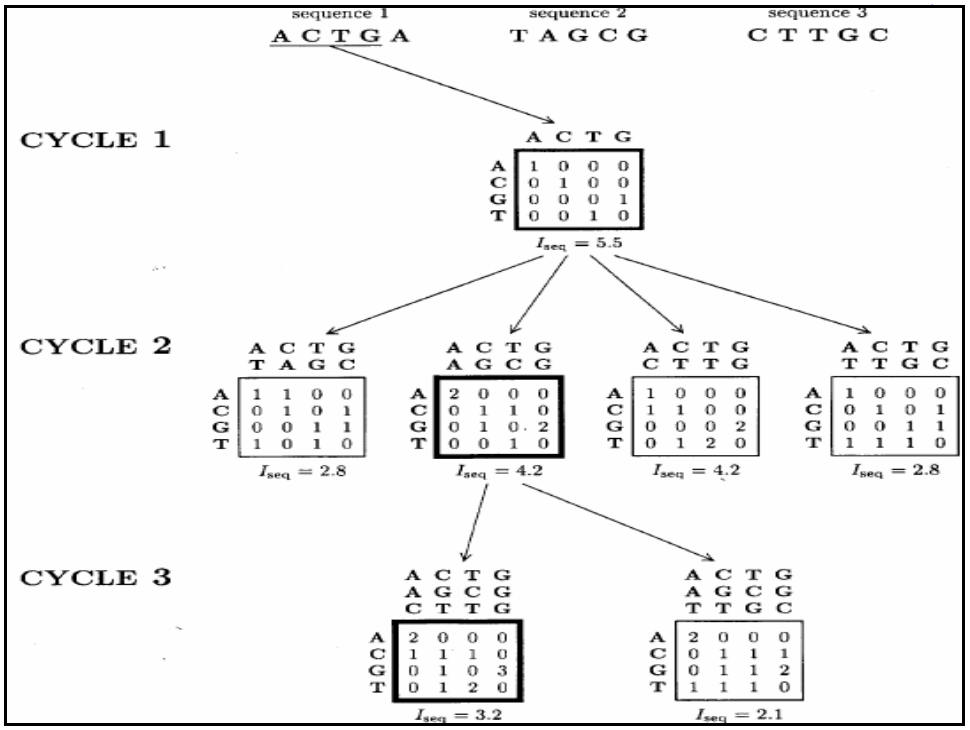
INPUT:

- sequences $s_1, s_2, \ldots, s_k$
- the length n of sites
- the maximum number d of profiles to retain

ALGORITHM:

1. Create a singleton set (i.e., only one member) for each possible length n substring of each of the k input sequences.
2. For each set S retained, add each possible length n substring from an input sequence s_i not yet present in S:
   1. Compute the Profile
   2. Compute the RE

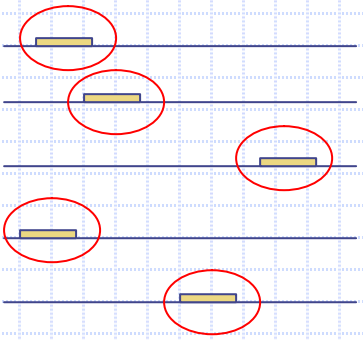
   => Retain the d sets with the highest RE
3. Repeat step 2 until each set has k members



Ab initio motif finding: Gibbs sampling

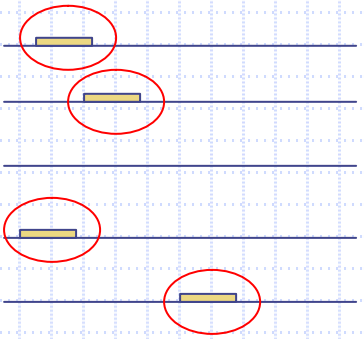
- Popular algorithm for motif discovery
- Motif model: Position Weight Matrix
- Local search algorithm

Gibbs sampling: basic idea



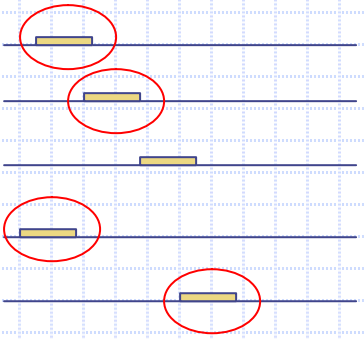
Current motif = PWM formed by circled substrings

Gibbs sampling: basic idea



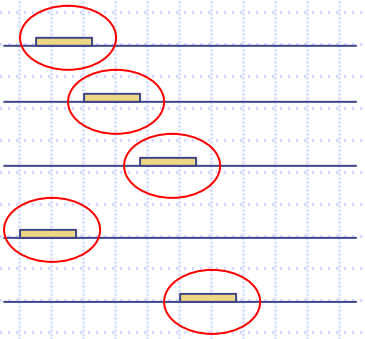
Delete one substring

Gibbs sampling: basic idea



Try a replacement:
Compute its score,
Accept the replacement depending on the score.

Gibbs sampling: basic idea



Ab initio motif finding: Expectation Maximization

- Popular algorithm for motif discovery
- Motif model: Position Weight Matrix
- Local search algorithm
  - Move from current choice of motif to a new similar motif, so as to improve the score
  - Keep doing this until no more improvement is obtained : Convergence to local optima

How is a motif evaluated ?

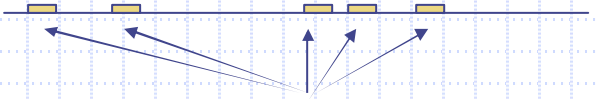
- Let W be a PWM. Let S be the input sequence.
- Imagine a process that randomly picks different strings matching W, and threads them together, interspersed with random sequence
- Let Pr(S|W) be the probability of such a process ending up generating S.

How is a motif evaluated ?

- Find W so as to maximize Pr(S|W)
- Difficult optimization
- Special technique called “Expectation-Maximization” or E-M.
- Iteratively finds a new motif W that improves Pr(S|W)

Basic idea of iteration

1. PWM  Current motif
2. Scan sequence for good matches to the current motif.



3. Build a new PWM out of these matches, and make it the new motif

Guarantees

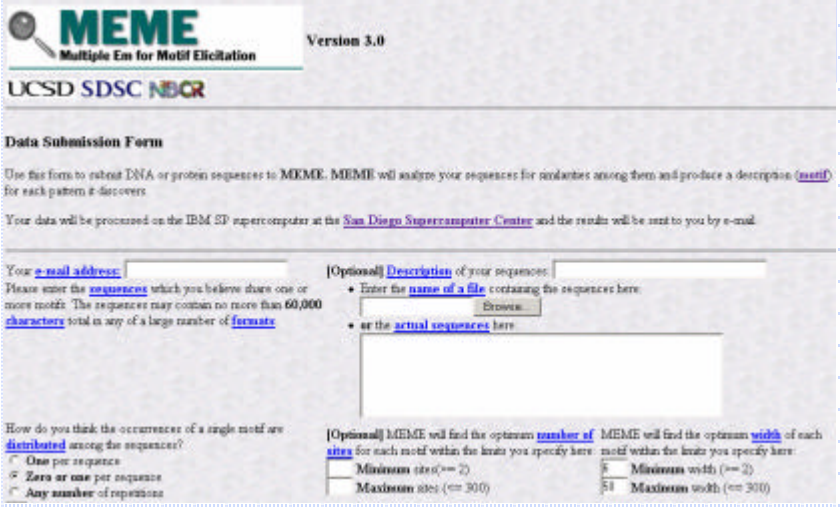
- The basic idea can be formalized in the language of probability
- Provides a formula for updating W, that guarantees an improvement in Pr(S|W)

MEME

- Popular motif finding program that uses Expectation-Maximization
- Web site

http://meme.sdsc.edu/meme/website/meme.html

PRACTICAL:
http://meme.sdsc.edu/meme/website/meme.html



Consensus sequences

- A **consensus sequence** is a sequence that summarizes or approximates the pattern observed in a group of aligned sequences containing a sequence feature
- Consensus sequences are **regular expressions**

Representation of Sequences

- characters
  - simplest
  - easy to read, edit, etc.
- bit-coding
  - more compact, both on disk and in memory
  - comparisons more efficient

Character representation of sequences

- DNA or RNA
  - use 1-letter codes (e.g., A,C,G,T)
- protein
  - use 1-letter codes
    - can convert to/from 3-letter codes

Representing uncertainty in nucleotide sequences

- It is often the case that we would like to represent uncertainty in a nucleotide sequence, i.e., that more than one base is "possible" at a given position
  - to express ambiguity during sequencing
  - to express variation at a position in a gene during evolution
  - to express ability of an enzyme to tolerate more than one base at a given position of a recognition site

Representing uncertainty in nucleotide sequences

- To do this for nucleotides, we use a set of single character codes that represent all possible combinations of bases
- This set was proposed and adopted by the International Union of Biochemistry and is referred to as the **I.U.B. code**

The I.U.B. Code

- A, C, G, T, U
- R = A, G (puRine)
- Y = C, T (pYrimidine)
- S = G, C (Strong hydrogen bonds)
- W = A, T (Weak hydrogen bonds)
- M = A, C (aMino group)
- K = G, T (Keto group)
- B = C, G, T (not A)
- D = A, G, T (not C)
- H = A, C, T (not G)
- V = A, C, G (not T/U)
- N = A, C, G, T/U (iNdeterminate) X or - are sometimes used

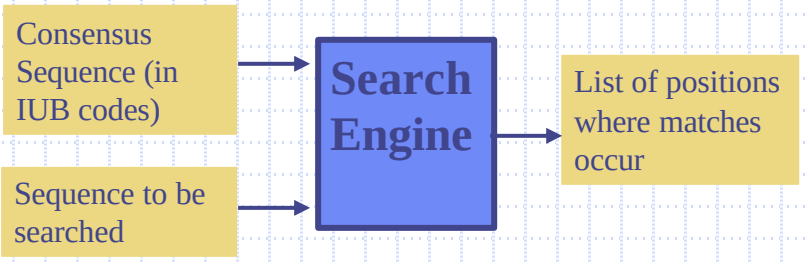
Representing uncertainty in protein sequences

- Given the size of the amino acid "alphabet", it is not practical to design a set of codes for ambiguity in protein sequences
- Fortunately, ambiguity is less common in protein sequences than in nucleic acid sequences
- Could use bit-coding as for nucleic acids but rarely done

Finding occurrences of consensus sequences

- Example: recognition site for a restriction enzyme
  - **EcoRI** recognizes **GAATTC**
  - **AccI** recognizes **GTMKAC**
- Basic Algorithm
  - Start with first character of sequence to be searched
  - See if enzyme site matches starting at that position
  - Advance to next character of sequence to be searched
  - Repeat Steps 2 and 3 until all positions have been tested

Block Diagram for Search with a Consensus Sequence



Sequence Analysis Tasks

Oligonucleotide frequency analysis ~ word counting

Statistical significance

Given:

1. A number N of unaligned sequences of length L
2. A pattern with width w and length k
3. The background frequencies of the nucleotides

Asked:

the probability to observe s or more occurrences of w

Does this situation involves a binomial random variable ?

Binomial properties:

1. The experiment consists of a fixed number of Bernouilli trials, resulting in either a success or a failure
2. The trials are identical and independent and therefore the probability of a success, p, remains the same from trial to trial
3. The random variable X denotes the number of successes obtained in T trials

Analysis

The total number of possible matching positions of a given word, T (trials), within a window is:

$$T = N \times (L - w + 1)$$

The expected frequency of a oligomer w of length k can be calculated, based on word composition and the background frequency of the nucleotides, w_i (i=1..k):

$$Freq_E(w) = \prod_{i=1}^{k} w_i$$

Analysis

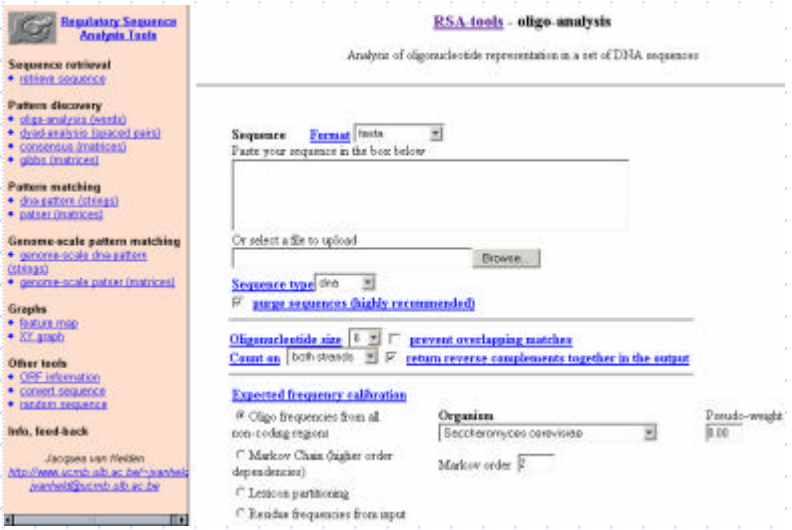
Let X denote the number of occurrences found of the oligomer w. The probability to observe exactly s occurrences of this oligomer can then be found by the binomial formula:

$$P(X = s) = \binom{T}{s} (Freq_E(w))^s (1 - Freq_E(w))^{(T-s)}$$

Finally, the probability to observe s or more occurrences of w is given by:

$$P(X \geq s) = 1 - \sum_{j=0}^{s-1} P(x = j)$$

Web Application:
http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/



Motif-finding Programs

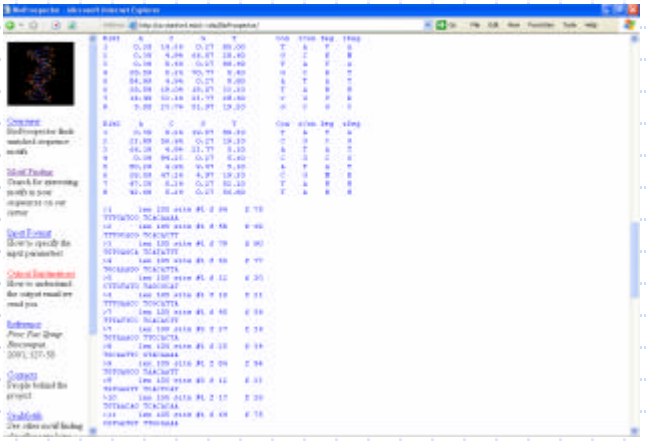
- TRANSFAC http://www.gene-regulation.com/
- RSA tools http://rsat.ulb.ac.be/rsat/
- BioProspector
  http://bioprospector.stanford.edu
- Alignace http://atlas.med.harvard.edu
- MEME
  http://meme.sdsc.edu/meme/website/intro.html
- FootPrinter
  http://wingless.cs.washington.edu/htbin-post/unrestricted/FootPrinterWeb/FootPrinterInput2.pl
- rVista http://rvista.dcode.org/

BioProspector Webserver

http://bioprospector.Stanford.edu

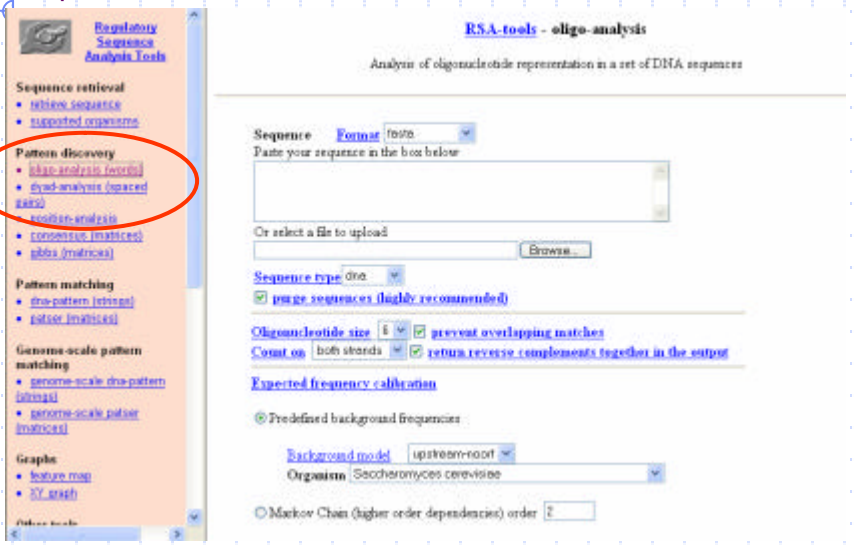


BioProspector Output



RSA Tools

http://rsat.ulb.ac.be/rsat/



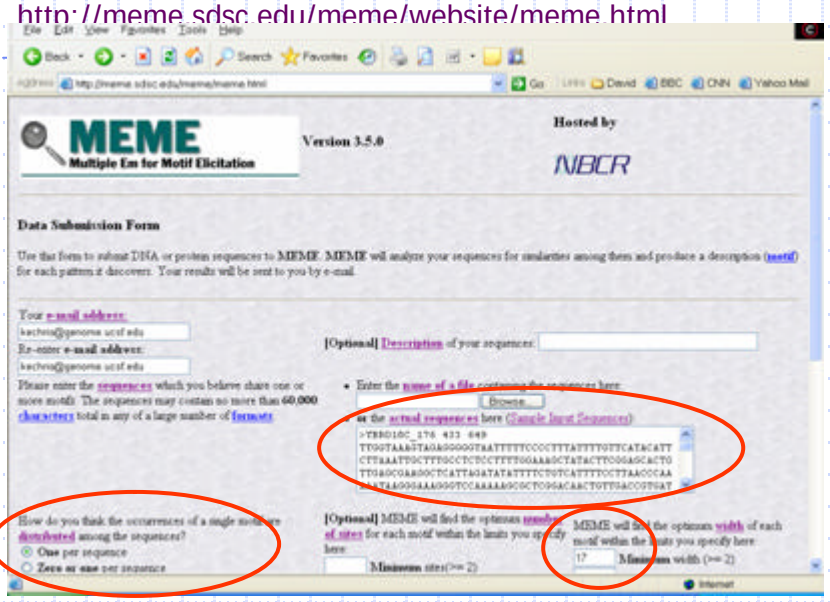
Gibbs Motif Sampler

http://bayesweb.wadsworth.org/gibbs/gibbs.html



MEME

http://meme.sdsc.edu/meme/website/meme.html



MEME Output

