

Course schedule

- Project group meetings on Thursday
 - Please sign up
- One more homework
 - Incremental & team assignment
- Motif finding
 - EM
 - Gibbs sampling
- Phylogenetic trees

Motif finding

Bioinformatics, Fall 2011 1

Lecture outline

- Why find motifs
 - Discover regulatory elements
 - Promoters, repressors, enhancers, activators
 - Deduce network structure
 - Gene structure determination
- Examples

Motif finding

Bioinformatics, Fall 2011 2

Expression of Genes in Cells

• To produce a protein, a gene (DNA) has to be converted to an intermediary molecule called **RNA**, in a process called **transcription**.

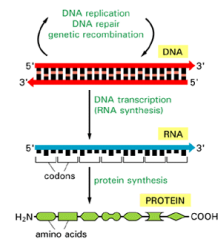
- Each cell contains the same genome. Different cells have a different set of genes which are turned on (**expressed**) by allowing the genes to be transcribed.
- Different cells have different **mixtures** of gene regulatory proteins to turn genes on or off.

Motif finding

Bioinformatics, Fall 2011 3

Terminology

- **Genome** – entire genetic material of an individual
- **Transcriptome** – set of transcribed sequences
- **Proteome** – set of proteins encoded by the genome



Motif finding

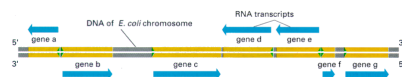
Bioinformatics, Fall 2011 4

Terminology

- Only one strand of DNA serves as a template for transcription.

(5') CGCTATAGCGTTT(3') DNA nontemplate (coding) strand
 (3') GCGATATCGCAA(5') DNA template strand
 (5') CGCUAUAGCGUUU(3') RNA transcript

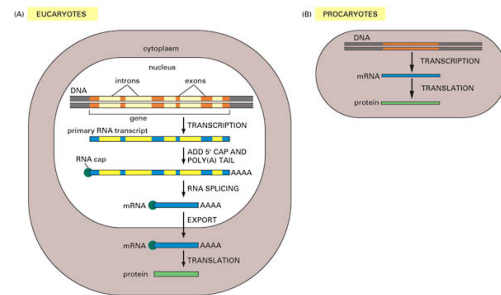
- Different genes are transcribed from different strands



Motif finding

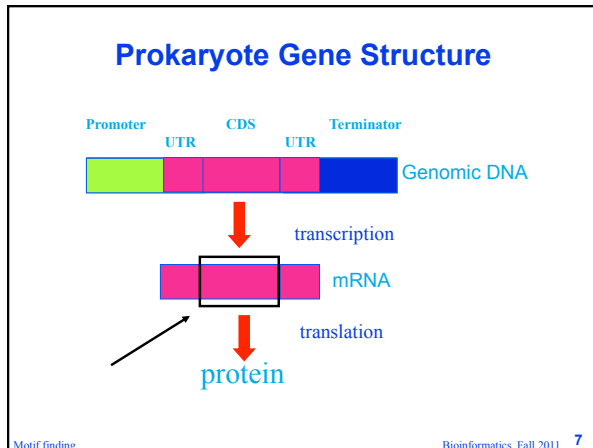
Bioinformatics, Fall 2011 5

Gene Structure



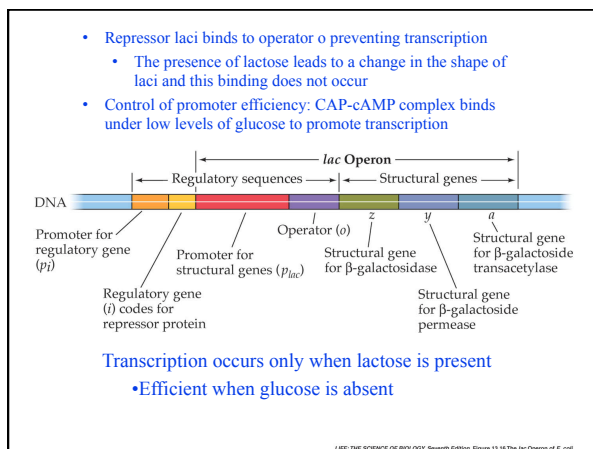
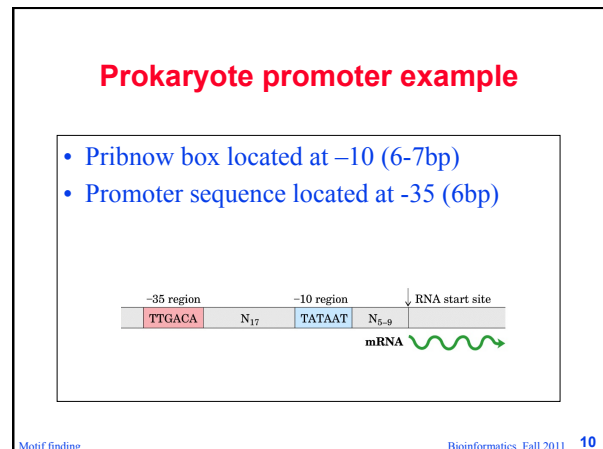
Motif finding

Bioinformatics, Fall 2011 6

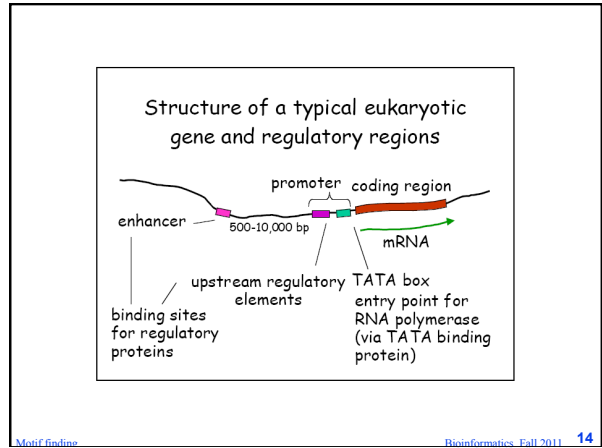
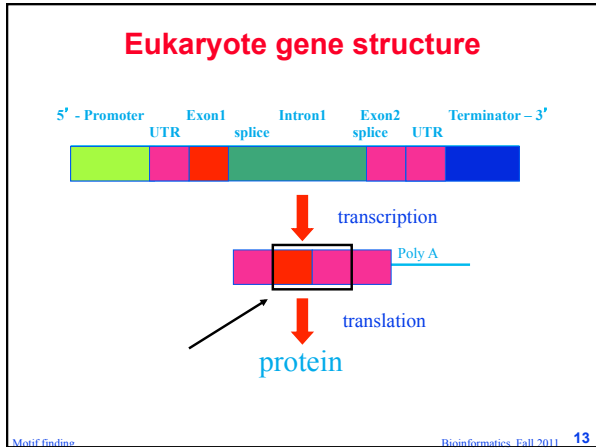


- ### Prokaryote gene structure
- Promoter : RNA polymerase binding site consisting of
 - minus 10 site:
 - Pribnow box (TATAAT)
 - Minus 35 site
 - TTGACA
 - Sigma-specific (sigma factors are part of the RNA polymerase complex)
 - Other units
 - Transcription start site
 - Coding region (ORF): aa sequence in protein
 - Translational start site (AUG)
 - Translational stop site (UAA, UAG, UGA)
 - Transcription stop site
- Motif findline Bioinformatics, Fall 2011 8

- ### Prokaryotic promoters and regulators
- Promoter determines:
 - Which strand will serve as a template.
 - Transcription starting point.
 - Strength of polymerase binding.
 - RNA polymerase subunit for promoter recognition is called sigma-factor
 - Different variations (7 for *E. coli*)
 - Consensus binding sequences
 - Operons for co-transcription
 - Regulators affect the binding of RNA polymerase to DNA (positive and negative)
- Motif findline Bioinformatics, Fall 2011 9

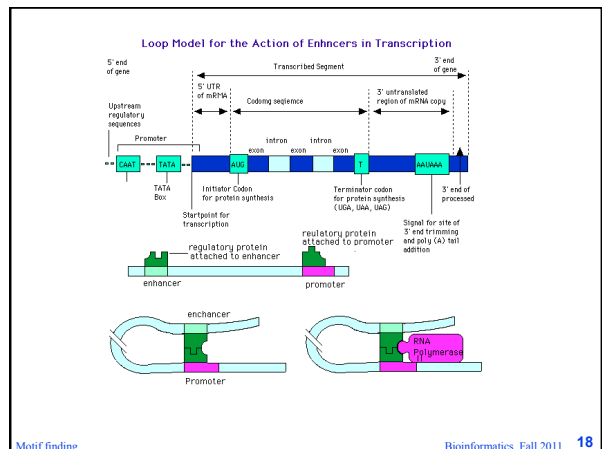
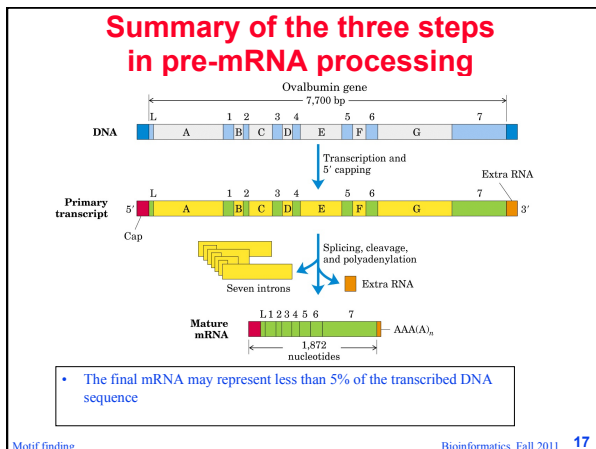


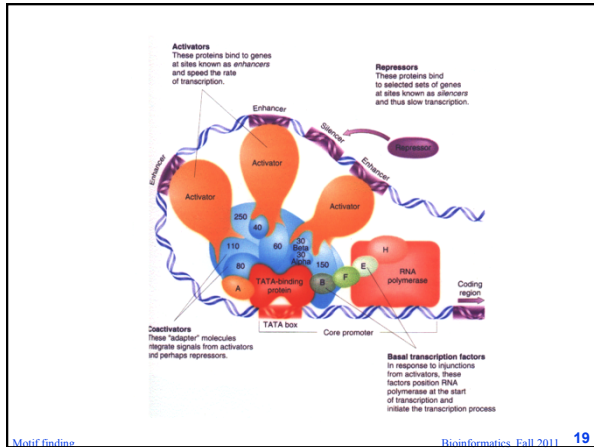
- ### Consensus sequences
- Promoters sequences can vary tremendously.
 - RNA polymerase recognizes hundreds of different promoters
-
- The diagram shows a sequence alignment of various *E. coli* promoters. The consensus sequence is highlighted in green. The sequence is: TCTCAACGTAAACACTTTACAGGGGGCG••CGTCATTGATATGATGC•GCCCCCTTCCCGATAAGGGGATCAAAAATAATCTGTGCAAAAA••TTGGATCCCTATAATGGCCCTCCCTTGAGACGCAACCGATGCTGCAATTTTCGCGCTTCTTCCCTGA••GCCGACTCCCTATAATGGCCCTCCCTGACACGGCGGATCCTGAAATTCGAGGCTTGACTCTGAAA••GAGAAAGCCTATAATAC•GCCACTGGCCAGCTGACCTGCAATTTCTATATGGGCTTGGC••GAGAACTCCCTATAATGGCCCTCCCTGACACGGCGGATTTTTAAATTTCCCTGTGTGAGGGCGGG••AATAACTCCCTATAATGGCCGACCCCTGAGAGGGAAACAAACAAAAATAATGCTTGAAGCTGTAG••CGGGAAGGCGTATTATGC•ACACC•CGGCGCTGAGAAAPeTAACCCGTGGCTGTGACTATTTTA•CCCTCGGGGTGATAATGG••TTGGCTGTACTAAGGAGGTATCTCTGGGGTGTGACATATAATA•CGACTGGGGTGTACTGA••GCACTCAGGAGGACGACGTGAAACAAACCGTTGACACATGA•AGTAAACACGGTACGATGT•ACGACTGAAACGACAGTGAATCAAAAAGAGTATGAGTAAAGT•CTAACCTATAGGATCTTA•CAGCCCTGGAGAGGACACGTTAGACGAAAACAGBTATGAGACATGAGTAAACATGAGATGATGAC•AAATCTCTAGGTACACTAGGATACAAATCCCGTTGACTTTGTT••TCGGCTGGTATAATCG•CTGGGGTCAAGATGAGTG
- Motif findline Bioinformatics, Fall 2011 12



- ### Eukaryote gene structure
- TATA box located around -25
 - TATA(A/T)A(A/T)
 - Recognized by TATA-binding protein
 - Initiator sequence at +1
 - YYCARR; Y is C/T, R is G/A
 - +1 is usually the A
 - Transcription factors bind to promoters
 - Possible distant regions acting as enhancers or silencers (even more than 50 kb).
 - More complex mechanism than prokaryotes
- Motif findline Bioinformatics, Fall 2011 15

- ### Eukaryote vs. prokaryote gene structure
- No operons
 - Capping at 5' end and polyadenylation at 3' end
 - Transport of mRNA out of nucleus
 - Effects stability and efficiency of translation
 - Introns
 - Alternative splicing
 - CpG islands around promoter regions
 - CpG tends to methylate and mutate
 - Conservation implies function
- Motif findline Bioinformatics, Fall 2011 16





Promoter...

(a) Strong *E. coli* promoters

```

tyr tRNA AACACTTTACAGCGGCG • CGTCATTTGATATGATGC • GCCCGCTTCCCGA
rm D1 AATACTTGTGCAAAAAA • TTGGGATCCCTATAATGCGCTCCGTGAGACG
rm X1 TCCGCTTGTCTTCTCTGA • GCCGACTCCCTATAATGCGCTCCATCGACACG
rm (DXE)2 CAGGGTTGACTCTGAAA • GAGGAAAGCGTAATATAC • GCCACCTCGCGACA
rm E1 TTCTATTGCGGCTGCG • GAGAACTCCCTATAATGCGCTCCATCGACACG
rm A1 TCCTTTGTAGGCGCG • AATAACTCCCTATAATGCGCCACCGTACACG
rm A2 AATGCTTGACTCTGTAG • CGGGAAGCGTATTATGC • ACACCCGCGCCGC
λ Pg GCGTGTGACTATTTTA • CCTCTGGCGGTGATAATGG • TTGCATGTACTAA
λ Pl CGGTGTTGACATAAATA • CCACTGGCGGTGACTACTGA • GCACATCAGCAGG
T7 A3 AACGGTTGACAACATGA • AGTAACACCGGTACGATGT • ACCACATGAAACGA
T7 A1 GAGTATTGACTTAAAGT • CTAACCTATAGGACTCTTA • CAGCCATCGAGAGG
T7 A2 AGGTATTGACAACATGAAGTAACTGCAGTAAGATAC • AAATCGTAGGTAA
fd VIII CTCCTTGTACTTTGTT • TCGCGCTTGGTATAATCG • CTGGGGTCAAAGA
-35 -10 +1
  
```

(b) Consensus sequences of promoters

-35 region 15-17 bp -10 region

TTG**CAT** **TAT****AAT**

(Griffiths et al. 1996)

Motif findline Bioinformatics, Fall 2011 20

Difficulty of Finding Regulatory Elements

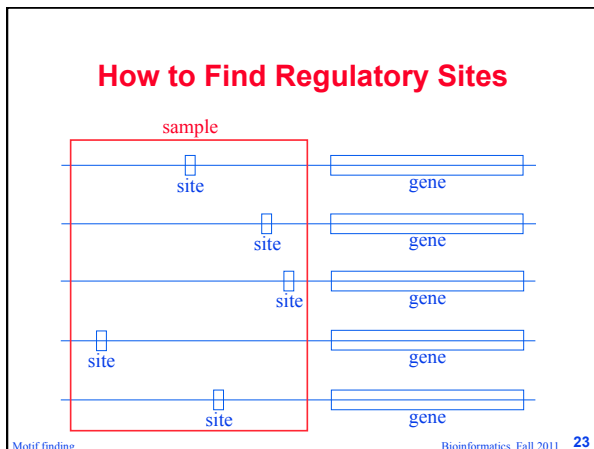
- Regulatory sites are short (up to 30 nucleotides).
- Non-coding regions are very long (includes all regions which are not translated into proteins).
- Experiments to find regulatory sites are tedious and time-consuming. One approach is to mutate different combinations of nucleotides until functionality changes.

Motif findline Bioinformatics, Fall 2011 21

Computational Approach

- Identify a set of genes believed to be controlled by the same regulatory mechanism (**co-regulated genes**).
- Extract regulatory regions of the genes (usually upstream sequences) to form a sample of sequences.
- Find some way to identify **conserved** elements in these sequences, resulting in a list of potential regulatory sites.

Motif findline Bioinformatics, Fall 2011 22



Motif Finding Approach

- Given a set of sequences, find a common **motif** shared by these sequences.
- Steps:
 - Construct a model of what we mean by common motif.
 - Solve the problem within the model on simulated samples.
 - Evaluate performance on biological samples.

Motif findline Bioinformatics, Fall 2011 24

Motif Finding Problem

• Given a sample of sequences and an unknown pattern (motif) that appears at different unknown positions in each sequence, can we find the unknown pattern?

• **Input:** a set of sequences, each one with an unknown pattern at an unknown position.

• **Output:** a set of starting positions of the pattern in each sequence.

• Naïve technique: $O((n-m+1)^k \cdot f)$

– n = length of each sequence

– m = size of motif

– k = number of sequences

– f = cost of computation of distance between motif and background

Motif.findline

Bioinformatics, Fall 2011 25

Sample with Motif AAAAAAAGGGGGG

```

atgaccggatactgatCAAAAAGGGGGGggcgtacacattagataaacgtatgaagctgtagactcggccgcgcg
accctattttttgacagatttagtgcctggaaaaaaatttagtacaacactttccgaataCAAAAAGGGGGG
tgagtaccctgggatgacttCAAAAAGGGGGGgctctcccattttgaatgtaggatcattccagggtccga
gctgagaattggtgCAAAAAGGGGGGccacgcaatcgcgaaccaacgagcccaaggaagcagcagataaaggaga
tccttttgcggttaagtgcgggaggtggttagcaggaagcccaacgacttaattCAAAAAGGGGGGttatag
gtcaatcatgtctgtgaatgattCAAAAAGGGGGGaccctgtggcaccacaattcagtggtggcgagcgcaa
cggtttgcctctgttagagcccccgtCAAAAAGGGGGGcaattatgagagactaattatcgctgctgttcat
aacttgattCAAAAAGGGGGGtggggcacatacagaggagtctcttcatcagtaagtgtgtgacactatgta
ttggccattgctaaaagcccaactgacaatggaatgagaatccttgcattCAAAAAGGGGGGccgaagggaag
ctgtgagcaacgacagattcttagctgacttagctcctcgggatctaatagacgaagcttCAAAAAGGGGGG
    
```

Motif.findline

Bioinformatics, Fall 2011 26

Sample with (15,4)-motif AAAAAAAGGGGGG

```

atgaccggatactgatCAAAAAGGGGGGggcgtacacattagataaacgtatgaagctgtagactcggccgcgcg
accctattttttgacagatttagtgcctggaaaaaaatttagtacaacactttccgaataCAAAAAGGGGGG
tgagtaccctgggatgacttCAAAAAGGGGGGgctctcccattttgaatgtaggatcattccagggtccga
gctgagaattggtgCAAAAAGGGGGGccacgcaatcgcgaaccaacgagcccaaggaagcagcagataaaggaga
tccttttgcggttaagtgcgggaggtggttagcaggaagcccaacgacttaattCAAAAAGGGGGGttatag
gtcaatcatgtctgtgaatgattCAAAAAGGGGGGaccctgtggcaccacaattcagtggtggcgagcgcaa
cggtttgcctctgttagagcccccgtCAAAAAGGGGGGcaattatgagagactaattatcgctgctgttcat
aacttgattCAAAAAGGGGGGtggggcacatacagaggagtctcttcatcagtaagtgtgtgacactatgta
ttggccattgctaaaagcccaactgacaatggaatgagaatccttgcattCAAAAAGGGGGGccgaagggaag
ctgtgagcaacgacagattcttagctgacttagctcctcgggatctaatagacgaagcttCAAAAAGGGGGG
    
```

Motif.findline

Bioinformatics, Fall 2011 27

Difficult Motif Finding Problem

• Find a motif in a sample of sequences, each 600 nucleotides long and each containing a pattern of length 15 with 4 mismatches ((15,4)-motif):

- (15,4)-motif is not too infrequent (appears once in 3,000 positions on average).
- Any two instances of the (15,4)-motif may differ in as many as 8 positions, a very large number.

Motif.findline

Bioinformatics, Fall 2011 28

Why Not Use Multiple Alignment?

• The motif is **short** and may appear at different location in different sequences. Most other areas are random.

• The problem is made more complicated since **not** every sequence contains a motif, due to:

- The upstream region used may not be long enough to include a regulatory site in every sequence.
- Usually, potential co-regulated genes are used to construct the sample; unclear whether all these genes are really co-regulated.

Motif.findline

Bioinformatics, Fall 2011 29

Computational Approaches

- **CONSENSUS** – Use a greedy algorithm to iteratively build up motifs by adding more and more pattern instances.
- **MEME** – Use the expectation maximization (EM) algorithm. Popular program for motif finding.
- **Gibbs sampler** – Start from a random initial solution, use the Gibbs sampling approach to make a series of local moves, trying to get to the solution with the best score.

Motif.findline

Bioinformatics, Fall 2011 30

CONSENSUS Algorithm

- **CONSENSUS** uses an iterative procedure to add more and more patterns to form potential motifs:
 - Initialize each l -mer in sequence 1 as a single-pattern motif.
 - Add each l -mer in sequence 2 to each single-pattern motif, forming motifs consisting of 2 patterns. Keep only the top n motifs.
 - Repeat the process by adding each l -mer in sequence 3 to the top n motifs from the last round, forming motifs consisting of 3 patterns, and so on until the last sequence. Only the top n motifs are kept each time.

Motif finding

Bioinformatics, Fall 2011 31

More Details of CONSENSUS

- CONSENSUS uses the entropy score for scoring a motif as a set of ungapped patterns.
- Instead of following the sequence order as given in the input sequence set, a randomized ordering is used to avoid dependence on the input set.

Motif finding

Bioinformatics, Fall 2011 32

Problems of Combinatorial Approaches

- Most combinatorial approaches do not consider the statistical significance of motifs.
- Although these new approaches confirm to have better performance over older approaches on simulated samples, it has not been shown that they have significant advantages on real biological samples.

Motif finding

Bioinformatics, Fall 2011 33

Summary of Statistical Approaches

- Most of these approaches use an iterative improvement procedure. In most cases, random starting points are used and the iteration often ends in a local optimal solution which is not the global optimal but a suboptimal solution.
- However, in most practical cases, very good solutions are obtained.

Motif finding

Bioinformatics, Fall 2011 34

Biological Considerations

- In practice, motif finding algorithms have to take into account **characteristics** of real input samples. These include:
 - Motifs with unknown length.
 - Samples with biased nucleotide composition.
 - Corrupted samples (not every sequence contains a motif).
 - Regulatory sites can lie on either DNA strand.

Motif finding

Bioinformatics, Fall 2011 35

Motif Finding in Real Life

- In practice, the motif finding process is a **three-step** process:
 - Assemble a sample of upstream sequences of (potentially) co-regulated genes.
 - Use motif finding software to find potential regulatory sites.
 - Verify the predictions experimentally.

Motif finding

Bioinformatics, Fall 2011 36

Higher-Order Motif Finding Problem

• Usually more than one motif is involved in regulation. Also, there are many regulatory proteins that control the expression of a gene, and the set of regulatory proteins involved is different under different situations.

References (Statistical Approaches)

- Stormo G.D. and Hartzell G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA*, **86**, 1183-1187.
- Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F. and Wootton J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208-214.
- Bailey T.L. and Elkan C.P. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, **21**, 51-80.

References (Combinatorial Approaches)

- Pevzner P.A. and Sze S.-H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. of 8th Int. Conf. on Intelligent Systems for Mol. Biol. (ISMB'2000)*, 269-278.
- Marsan L. and Sagot M.-F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comp. Biol.*, **7**, 345-362.
- Buhler J. and Tompa M. (2002) Finding motifs using random projections. *J. Comp. Biol.*, **9**, 225-242.
- Sze S.-H., Lu S. and Chen J. (2004) Integrating sample-driven and pattern-driven approaches in motif finding. *Lecture Notes in Computer Science/Lecture Notes in Bioinformatics (WABI'2004)*, 438-449.

References (Combinatorial Approaches)

- Keich U. and Pevzner P.A. (2002) Finding motifs in the twilight zone. *Bioinformatics*, **18**, 1374-1381.
- Price A., Ramabhadran S. and Pevzner P.A. (2003) Finding subtle motifs by branching from sample strings. *Bioinformatics*, **19**, S1149-155.
- Eskin E. (2004) From profiles to patterns and back again: a branch and bound algorithm for finding near optimal motif profiles. *Proc. of the 8th Ann. Int. Conf. on Comp. Mol. Biol. (RECOMB'2004)*, 115-124.