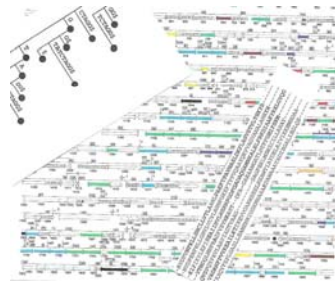


Lecture 9: Motifs and Motifs finding

Principles of Computational Biology

Instructor: Teresa Przytycka, PhD

Igor Rogozin, PhD



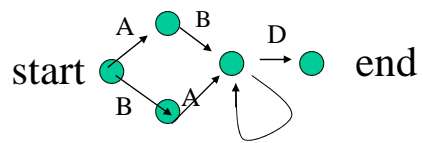
Motifs

- Motif is a region (a subsequence) of protein or DNA sequence that has a specific structure
- Motifs are candidates for functionally important sites
- Presence of a motif may be used as a base of protein classification

Representation of motifs

- Profile or sequence logos
- Regular expression

Describing patterns using regular expressions

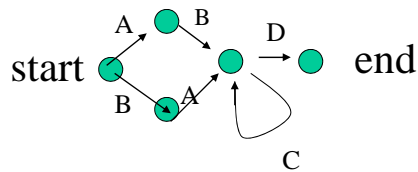


A graph like one above is called in CS literature a **finite automaton** can be used to describe a sequence family (CS literature such a set of sequences is called a language):

Take any path from “start” to “end” and as you go print the letters that label the edges you used. Any sequence that can be printed in this way will be called **generated** (CS term: accepted) by the automaton.

E.g.: ABCCCCD; BACCD;.....

Regular expressions



A finite automaton can be translated to so called regular expressions:

Notation:

[choice1, choice2,...] = a set of choices in a brunching point ,

- = “followed by”

* = repeat 0 or more times

E.g. The regular expression describing automaton above:

[A-B , B-A]-C*-D

PROSITE

- A data base of regular expression that describe protein motifs
- Developed since 1988
- 1999 – authors recognize that some protein families are characterized by profiles than regular expression and extended the data base to contain profiles
- Profiles are generated from multiple sequence alignments

PROSITE patterns

- PROSITE fingerprints are described by regular expressions
 - Rules:
 - Each position is separated by a hyphen
 - One character denotes residuum at a given position
 - [...] denoted a set of allowed amino acids
 - (n) denotes repeat of n times
 - (n,m) denoted repeat between n and m inclusive
 - X – any character
- Example [EDQH]-x-K-x-[DN]-G-x-R-[GACV]
Ex. ATP/GTP binding motive [SG]-X(4)-G-K-[DT]
- There is a number of programs that allow to search databases for PROSITE patterns

Finding motifs

- **Method I:** extracted from multiple sequence alignment .
 - EMOTIF
 - PRINTS
- **Method II:** Gibbs sampling – a method that allows to find motifs in the absence of multiple sequence alignment
Reference: Lawrence, C.E. et al (1993) Science 263, 208-214
- **Method III:** Exhaustive or dedicated search

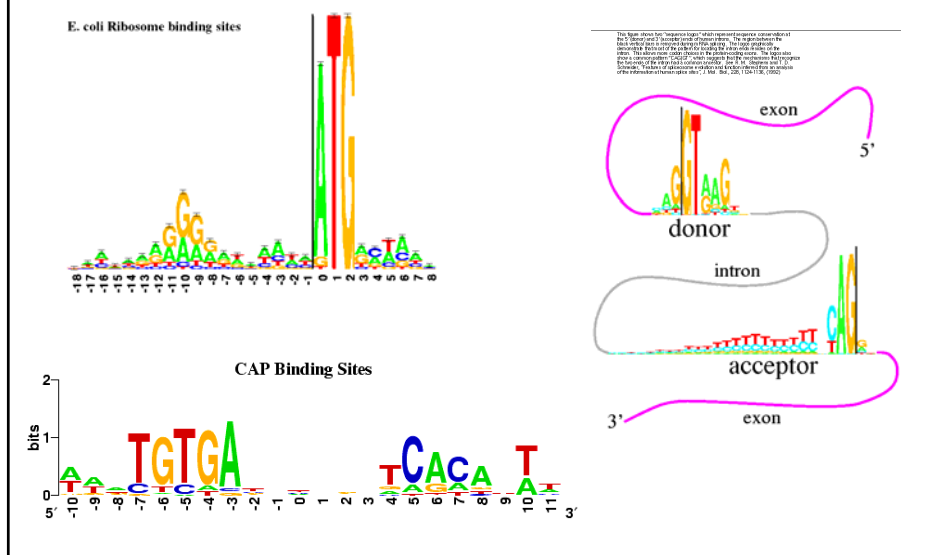
Recognition of transcription factor binding sites

- Transcription Factors = proteins that bind DNA, typically upstream or close to the transcription start site and regulate the expression of the gene by activating or inhibiting the transcription machinery
- Little is known about most of transcription factors it particular what binding sites most of them are.
- Co-regulated genes – genes to which are regulated by the same transcription factor

Typical setting for computational finding of transcription binding sites

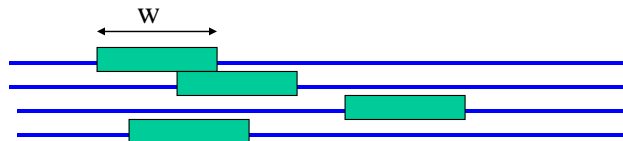
- Give is collection of regulatory regions of genes that are believed to be co-regulated.
- Goal – find sort DNA motifs that are overrepresented.
- So what is the problem?
 - Binding motifs are typically short
 - They have significant variability
- There is a large number of other algorithms: (AlignACE, MEME, Weeder, YMF...)

Examples of binding sites profiles



Finding Motifs with Gibbs Sampling Method

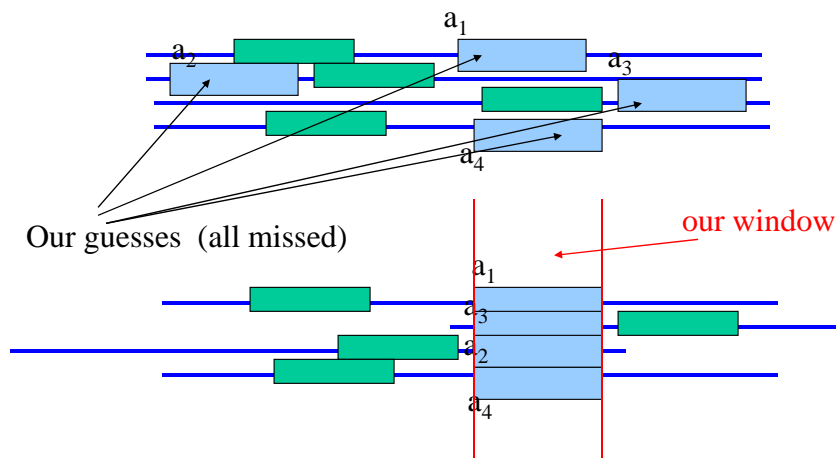
- Assumption: Given is a set of sequences that are believed to share (one) common motif
- Motif is assumed to have length w



Idea: Look for a signal using a Monte Carlo method

Initialization: Make a guess

- Choose randomly at each sequence a candidate position for the motif



The basic algorithm

- Initialization: Choose randomly at each sequence a candidate position for the motif
- Iterate the following two steps until convergence:
 - Predictive update: detecting current signal from the motif identified so far
 - Sampling: improving the signal

Predictive update step – leave one out

- One of N sequences, say z, is removed (randomly or in specific order)
- The pattern frequencies (position specific scoring matrix) and background probabilities are computed with z excluded

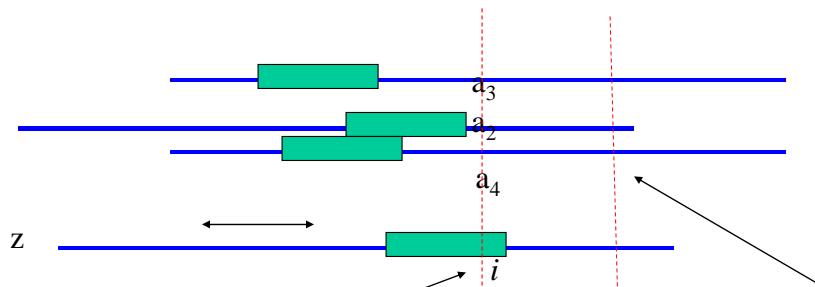
Two evolving data structures maintained by the algorithm

- Pattern description in the form of probabilistic model of residue frequencies
 - q_{i1}, \dots, q_{i20} ; $i = 1, \dots, w$ (q_{ik} is the frequency of amino-acid k on position i of the pattern)
 - p_1, \dots, p_{20} ; background frequency
- Local alignment description
 - a_1, \dots, a_N ; N-number of sequences;
 - a_i – beginning of the pattern in the i^{th} sequence

Sampling Step

- Every possible sequence x of length w is aligned with the profile in the window
- Calculate probability Q_x of generating x according to probability distribution defined by current pattern description (profile) (q_{i1}, \dots, q_{i20} ; $i=1..w$)
 - e.g, probability of ATCA = $q_{1A} q_{2T} q_{3C} q_{4A}$
- Calculate probability P_x of generating x according to background probability distribution p_1, \dots, p_{20}
- Assign weight $A_x = Q_x / P_x$ to the sequence x
- Choose with probability weighted by A_x the sequence x to be aligned with current patten

Sampling Step



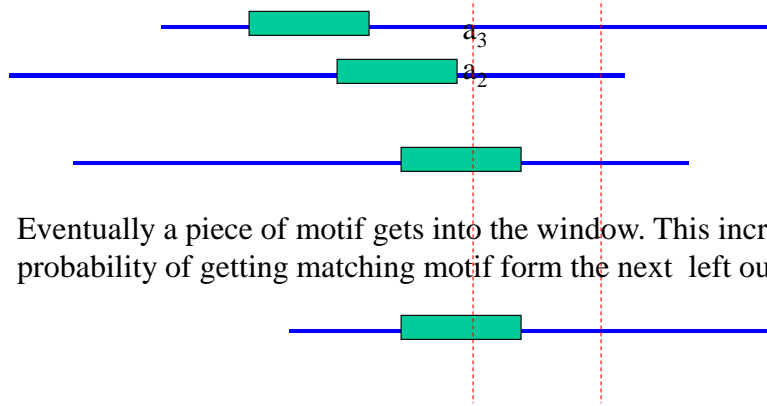
• Try all possible alignments of z to the profile defined by the pattern we found so far.

• Each position i has some probability p_i of being good (if no pattern all position should be equally likely)

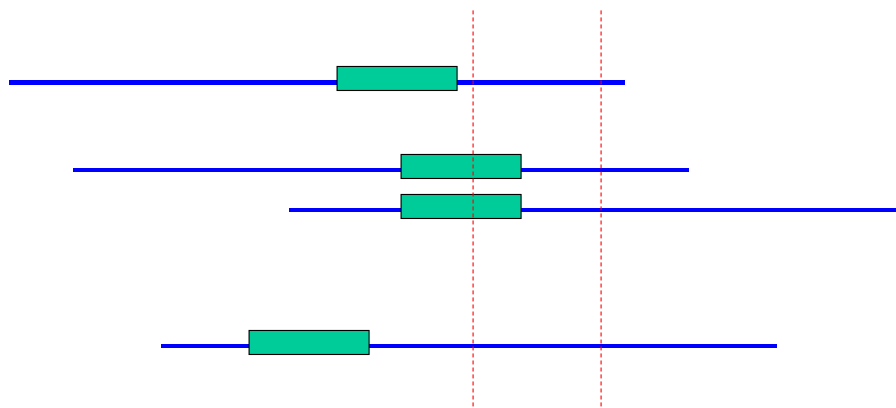
• Chose fragment from i to $i+w$ to be aligned to the window with probability p_i .

We know the Profile of this alignment (still no pattern found so this profile should correspond to random a.a. distribution)

Why it is supposed to work

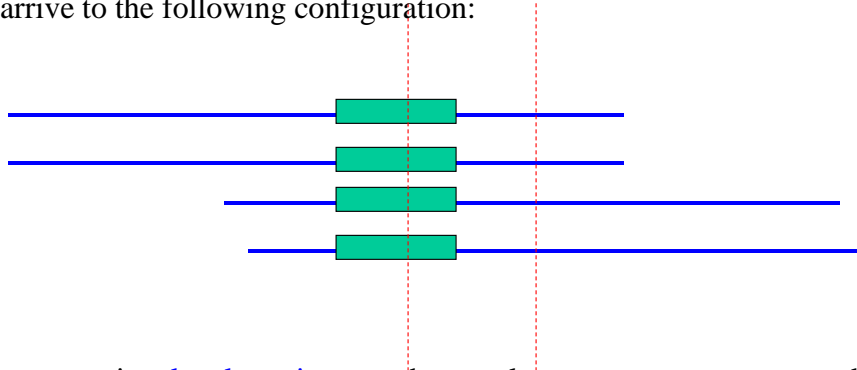


Eventually a piece of motif gets into the window. This increases the probability of getting matching motif from the next left out sequence

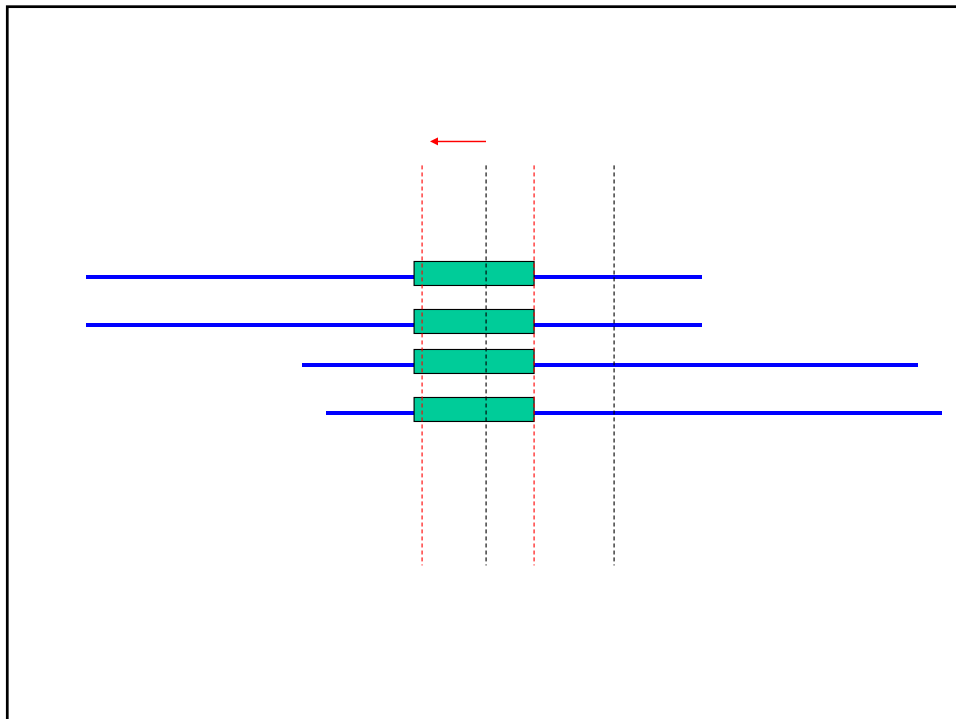


Note that this one has a good chance to fit the pattern in the window

Finally after a number of iteration we have a good chance to arrive to the following configuration:



Now we are in a **local maximum**, when we leave one sequence out and put it back out it has a high probability to realign where it was. When no further improvement is observed we assume that **have a pattern or a part of it** in the window and we try to move the window slightly to the sides to discover the rest of the motif



Uncovering sequence specificity of TF by Chi-Seq technology

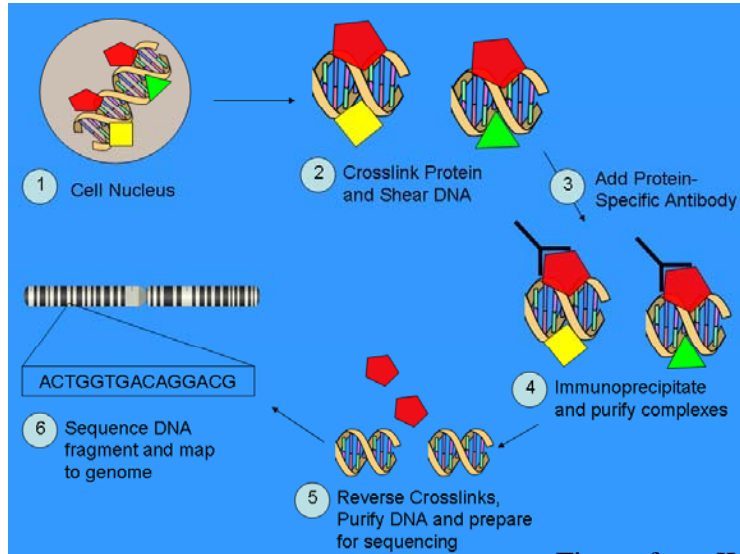
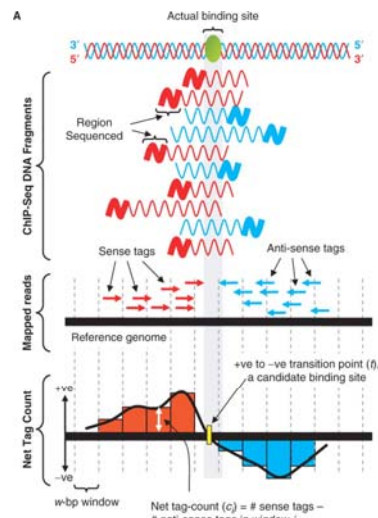


Figure from Wikipedia

Schematic overview of SISR algorithm



Jothi, R. et al. *Nucl. Acids Res.* 2008 0:gkn488v1-488; doi:10.1093/nar/gkn488

Copyright restrictions may apply.

Nucleic Acids Research

Example of an approach to identify binding sites

After identifying where the binding sites are uncovering binding motifs is much simpler