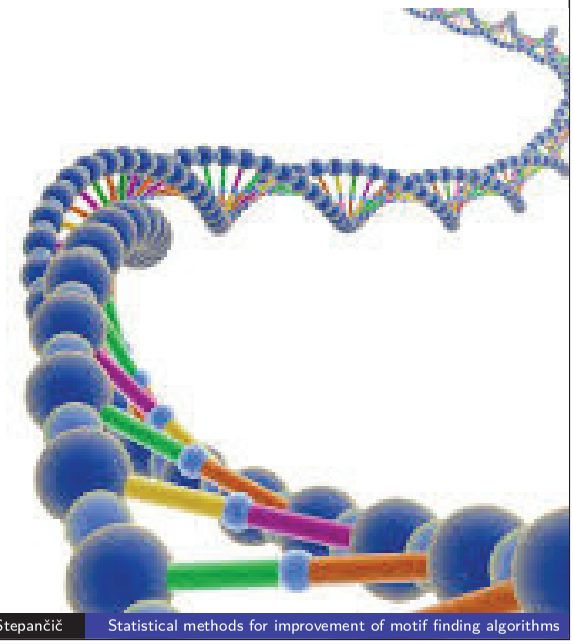
Statistical methods for improvement of motif finding algorithms

Živa Stepančič

Dornbirn, 11.9.2013

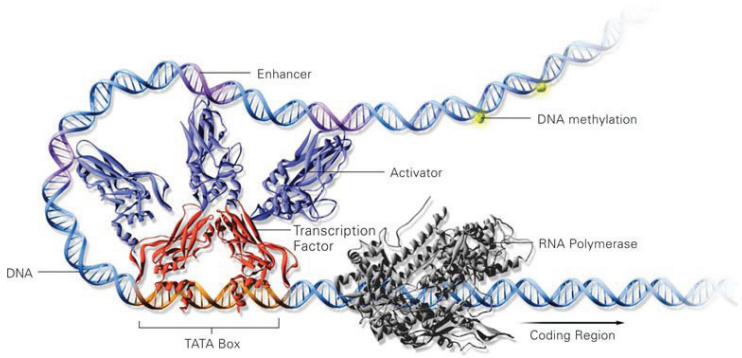
Overview

1. Motivation
2. Different approaches
3. Gibbs Sampling
4. Direction



Motivation

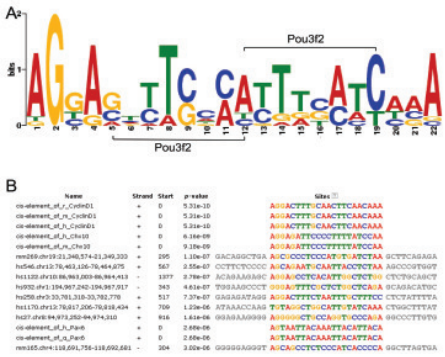
The motif finding problem arrives from biology, where they are trying to identify regulatory elements, more notably the binding sites in DNA for transcription factors.



Source: *www.tolpa.com*

Motivation

Basically this problem can be formulated as the problem of finding short similar motifs in a set of given sequences.



Source: *www.intechnopen.com*

Different approaches

In earlier literature motif finding algorithms were mostly divided into two major groups:

- word-based (string-based) methods
- probabilistic sequence models

- Oligo-Analysis
- algorithm YMF (Yeast Motif Finder)
- MITRA (Mismatch Tree Algorithm)

Different approaches

In earlier literature motif finding algorithms were mostly divided into two major groups:

- word-based (string-based) methods
- probabilistic sequence models

- Consensus
- EM methods (expectation maximazation methods)
- Monte Carlo methods (Gibbs sampling)

Different approaches

In a survey of DNA motif finding algorithms, Modan K. Das and Ho-Kwok Dai proposed three differenting categories:

- algorithms that use promoter sequences of coregulated genes:
  - word-based algorithms
  - probabilistic algorithms
    - Gibbs sampling methods
  - machine learning techniques
- algorithms based on phylogenetic footprinting
- algorithms based on both

Gibbs sampling metods

'Original' Gibbs sampler developed by Lawrance et al. and described in the following article:

Lawrence, C.E. et al., Detecting subtle sequence signals: A Gibbs sampling stategy for multiple alignment, *Science*, **8** (1993), Vol. 262, 208-214.

Gibbs sampling methods

Few variations of the algorithm:

1. AlignACE (Aligns Nucleic Acid Conserved Elements), developed by Roth et al.
   http://arep.med.harvard.edu/mrnadata/mrnasoft.html
2. MotifSampler, designed by Thijs et al.
   http://bioinformatics.psb.ugent.be/webtools/MotifSuite/motifsampler.php
3. BioProspector, developed by Liu et al.
   http://ai.stanford.edu/ xsliu/BioProspector/
4. Gibbs Centroid Sampler (Gibbs Motif Sampler), developed by Thomson et al.
   http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html
5. Info-gibbs, developed by Matthieu Defrance and Jacques van Helden
   http://rsat.ulb.ac.be/info-gibbs_form.cgi

Basic algorithm

We start with a set of N sequences (DNA/protein) $S_1, \ldots, S_N$.

```
        ... TAGGGCGTAATGACACCCACATATAACATAAGTCACAGTGACAGCCAC ...
    ... TCTTTAACATAAGTCACAGTGACAGCCCATTTGGATCATTTCGGACCGCTTCCGTG ...
  ... CTAAAACCAGAAGTTGTGGCCCATACTAACATAAGTCACAGTGACAGCCGTTTACTGGGT ...
    ... CCTCAACATAAGTCACAGTGACAGCCCACACCCCACGATACAAAAACAAGTTACCG ...
          ... GAATCAACATAAGTCACAGTGACAGCGTAACTCCACCACAGAGTA ...
          ... TGGAACTTGAACATAAGTCACAGTGACAGCGAGTCTACGAGGGTTTC ...
  ... CAGTACGATTGACGTTTGCTCACGGTTTGAACATAAGTCACAGTGACAGCTAACGGTGCGG ...
  ... AAAGACTGCACTACTAGTCAGAACATAAGTCACAGTGACAGCTGTATCCATGGTATCTGT ...
    ... TTCCGCTTGCGACCGTTAACATAAGTCACAGTGACAGCTAGATGTTTCAGAACAGG ...
  ... TGAACATAAGTCACAGTGACAGCACCGTATTCCGGGTCCTGTCGGTAGGATTCTAGCCTAC ...
```

Each sequence contains a motif of fixed width W.

AACATAAGTCACAGTGACAGC

Basic algorithm

Two data structures:

1. pattern description

pos	A	T	C	G
1	q_{11}	q_{12}	q_{13}	q_{14}
2	q_{21}	q_{22}	q_{23}	q_{24}
$\vdots$	$\vdots$	$\ddots$		$\vdots$
$\vdots$	$\vdots$		$\ddots$	$\vdots$
W	q_{W1}	q_{W2}	q_{W3}	q_{W4}

   background description

$$[p_A, p_T, p_C, p_G]$$

2. alignment

$$\{a_k \mid k = 1, \ldots, N\}$$

Basic algorithm

The goal of the algorithm is locating the alignment that maximizes the ratio of the corresponding pattern probability to background probability.

Figure 2: Final Alignment for Site Sampler

Source: *Rouchka: A Brief Overview of Gibbs Sampling, 1997*

Basic algorithm

Algorithm framework:

1. Initialization.
2. Begin iteration.
3. 2 steps of Gibbs sampler:
   - Predictive update step
   - Sampling step
4. Evaluating alignment.

Basic algorithm

Predictive update step.

- Chosing and removing one sequence z from the set.
- Calculating pattern and background descriptions from the rest of the sequences:

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B} \qquad \begin{aligned} i &= 1, \ldots, W \\ j &= 1, \ldots, 4 \\ B &= \sum_j b_j \end{aligned}$$

The background frequencies are calculated analogously, with corresponding counts taken over all nonpattern positions.

Basic algorithm

Sampling step.

- Look at every possible segment x in sequence z.
- To each such segment, assign a weight:

$$A_x = \frac{Q_x}{P_x} = \frac{\prod_{i=1}^{W} q_{i,x_i}}{\prod_{i=1}^{W} p_{x_i}}$$

- Randomly select a new possible motif with probability

$$A_x / \sum_s A_s,$$

and its starting position becomes the new value of a_z in the alignment.

Basic algorithm

Evaluating the alignment:

- fscore:

$$F = \sum_{i=1}^{W} \sum_{j=1}^{4} c_{i,j} \log \frac{q_{i,j}}{p_j},$$

- gscore:

$$G = F - \sum_{i=1}^{N} \left(\log L_i + \sum_{j=1}^{L_i} Y_{i,j} \log Y_{i,j} \right)$$

- Information per parameter:

$$I = \frac{G}{3W}.$$

Variants of the basic algorithm

Some of new methods included into the basic algorithm:

- MAP score (maximum a-priori log likelihood), to judge the motif significance;
- incorporation of a higher-order Markov-chain background model;
- using the probability distribution to estimate the number of copies of the motif in a sequence;
- motif score distribution estimated by a Monte Carlo method (to judge motif significance);
- modeling of spaced dyad motifs and motifs with palindromic patterns;
- centroid alignment;

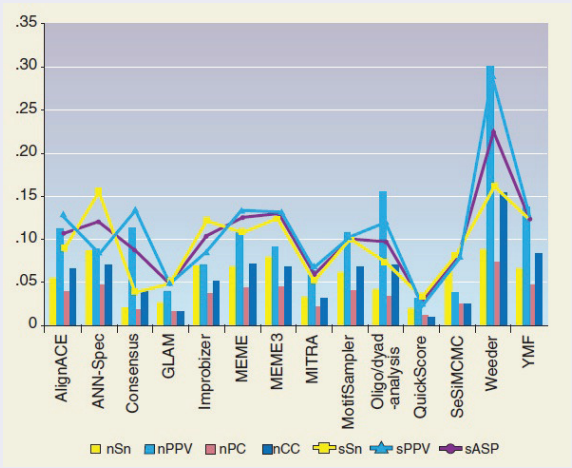
Variants of the basic algorithm

Statistics used to access tool performance quality:

- accuracy:
$$\frac{TP + TN}{TP + TN + FP + FN}$$
- sensitivity:
$$Sn = \frac{TP}{TP + FN}$$
- Positive Predictive Value:
$$PPV = \frac{TP}{TP + FP}$$
- Specificity:
$$SP = \frac{TN}{FP + TN}$$
- Correlation Coefficient:
$$CC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Variants of the basic algorithm

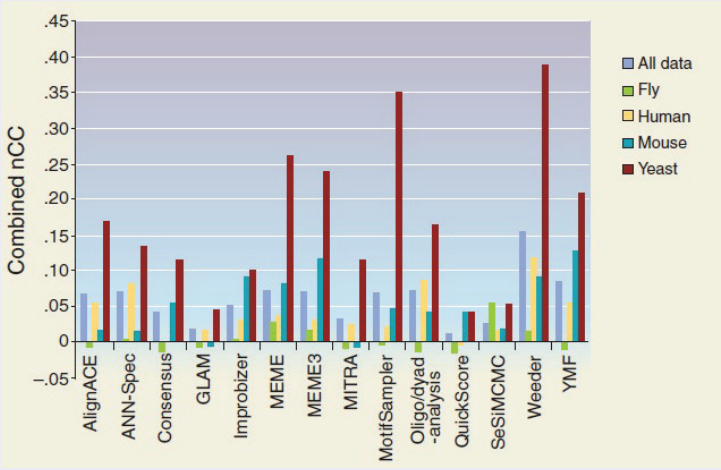
Statistics used to access tool performance quality:



Source: *M. Tompa et al., Assesing computational tools for discovery of transcription factor binding sites*

Variants of the basic algorithm

Statistics used to access tool performance quality:



Source: *M. Tompa et al., Assesing computational tools for discovery of transcription factor binding sites*

- Composing several data sets of different species.
- Evaluation of available tools and their combinations.
- Sampling sheme.

Thank you for your attention!