SGN-6156, Lecture 6
Biological sequence analysis

Harri Lähdesmäki, harri.lahdesmaki@tut.fi

(part of the material by Juha Kesseli)

Department of Signal Processing,

Tampere University of Technology

16.04.2008

DNA sequence motifs

- DNA sequence motifs are short sequences of DNA the are found throughout the genome and that are presumed to have a biological function
- A particularly important type of sequence motif is so called binding site of a transcription factor (TF). We will mainly focus on that in the following
- We can consider DNA sequence motifs more generally, e.g., short subsequences of DNA which are recognized (and bound by) some DNA-binding molecule

Transcriptional regulation

- Recall the central dogma: DNA $\rightarrow$ RNA $\rightarrow$ protein
- Transcriptional regulation generally involves DNA-binding proteins, transcription factors (TF), that control gene expression by recognizing and binding to short regulatory sequence motifs in gene promoters.
- DNA-binding specificities of TFs are encoded in their DNA-binding domains that specialize them to recognize and bind specific types of binding sites.
- This mechanism is the basis of control in complex transcriptional regulatory networks.
- Revealing these regulatory mechanisms is one of the key problems in understanding genome-wide transcriptional regulation and transcription level control in general

Sequence motif identification

- A sequence motif can be found by empirically studying the binding sites of a specific molecule
- TF binding sites are reported in databases (TRANSFAC and JASPAR)
- Binding site identification can also be done in a high-throughput fashion using the ChIP-chip (Chromatin immunoprecipitation on chip) technology
- Alternatively/In addition, algorithmic prediction can be used to find putative candidate motifs
  - The prediction can be done based on sequence similarity. Other measurements, e.g. similar expression profiles, can be used to select the regions for the sequence similarity search
  - Note that predicting a motif does not directly tell us the transcription factor binding to the motif

Consensus sequences and position frequency matrices

- Some molecules have very specific binding sequences so that they can be described by a simple consensus sequence, e.g. GAATTC
- Note that since the consensus sequence is short it will occur randomly, on average, once every $4^6 = 4096$bp
- Many other enzymes bind to a degenerate consensus sequence that can have one of several nucleotides in at least one position. There exist standard IUPAC-IUB codes for these cases, e.g. Y = C or T
- A position frequency matrix (a position-specific frequency matrix) giving the frequency $f_{b,i}$ of nucleotide b at position i is a more useful way of describing motifs in general

Sequence logos

- Sequence logos are used to show the information content (conservation) at each position of the motif along with frequency information (see example).

- If $f_{b,i}$ denotes the frequency of base b at position i and q_b contains background frequencies of bases in the genome, the height of the sequence logo at position i in bits is typically computed as the Kullback-Leibler distance (relative entropy)

$$I_{\mathrm{seq}} = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{q_b}.$$

Alignment matrix and a sequence logo

```
# example alignment matrix
# RFX1 binding site alignment, S. cerevisiae
PO   1   2   3   4   5   6   7   8   9  10  11  12  13  14
A    0   0   1   3   0   0  17   0   5   4   0  15  17   1
C    0   0   1   1  13  13   0   1   0   1  13   0   0  16
G   16   0   0  13   1   0   0   0  12  12   1   1   0   0
T    1  17  15   0   3   4   0  16   0   0   3   1   0   0
```

Figure 1: An example sequence logo created with enoLOGOS.

Position weight matrix

- Position weight matrix (PWM) or position specific scoring matrix (PSSM) can be obtained from the frequencies as

$$W(b,i) = \log_2 \frac{f_{b,i}}{p_b}$$

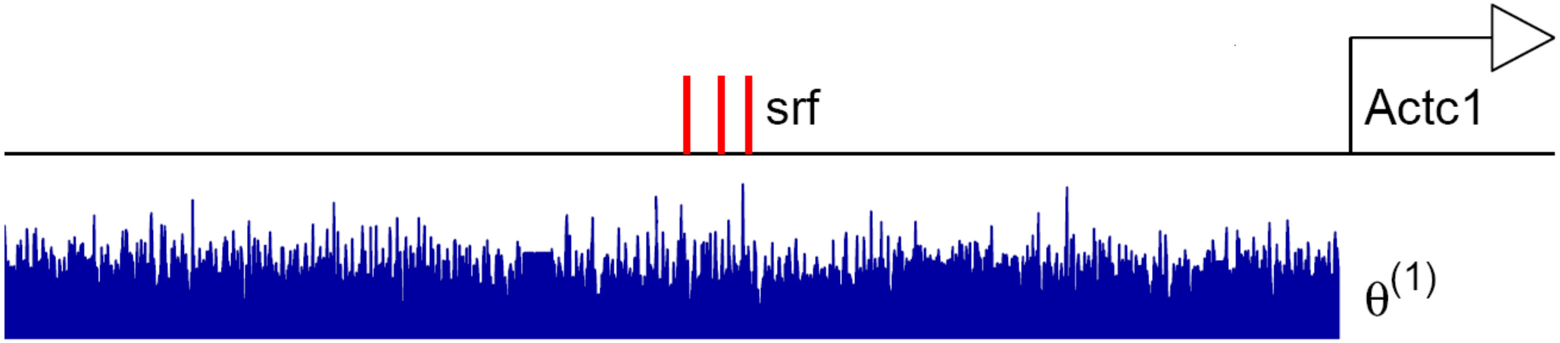
  where $W(b,i)$ defines the score for seeing nucleotide b in the ith position of the binding site and the width of the motif is ℓ

- Each subsequence $x_k, \ldots, x_{k+\ell-1}$ of the DNA can be scored by

$$S_k = \sum_{i=k}^{k+\ell-1} W(x_i, i).$$

- In order to identify putative binding sites, the whole DNA can be scanned this way. This is the most common approach

- Typically, PSSM is defined using a higher Markovian order background model, i.e., $P(x_k = a_k | x_{k-1} = a_{k-1}, \ldots, x_{k-d} = a_{k-d})$
- The search for binding sites can be formulated as a hypothesis testing problem:
  1. H_0: kth location is not a TF binding site
  2. Use test statistic is S_k
  3. A null distribution of S_k, $\hat{F}$, can be obtained e.g. by computing the test statistic on a set of intergenic sequences (or on all promoter sequences)
  4. Compute S_k on the given subsequence
  5. Choose a significance level α
  6. Compute the significance value with respect to $\hat{F}$ (and possibly correct for multiple testing)



Example of binding site prediction

Sequence motif discovery

- Let us consider a simple evaluation method for sequence motifs
- Assume one is evaluating the significance of a motif/consensus sequence w and that there are
  - N genes
  - n genes out of all N genes contain w
  - a known set of m functionally related genes
  - k genes out of m genes contain w
- Assuming that our motif is completely independent of the known set of m genes, we want to know the probability that the motif exists in at least k of the m genes (just by chance)

- The probability of having overlap of exactly k genes, by chance, can be computed from the hypergeometric distribution

$$P(\text{overlap} = k) = \frac{\binom{N-m}{n-k}\binom{m}{k}}{\binom{N}{n}}$$

- The overlap of at least k is

$$P(\text{overlap} \geq k) = \sum_{l=k}^{\min\{n,m\}} \frac{\binom{N-m}{n-k}\binom{m}{k}}{\binom{N}{n}}$$

- If $P(\text{overlap} \geq k)$ is smaller than a chosen significance level then motif w can be considered as significant
- Sequence motifs can also be searched using probabilistic methods
  - Several advanced methods
- Probabilistic multiple alignment methods can also be used

- It is recommended that several motif discovery tools are used to obtain more reliable results.

References