Point processes in biological sequence analysis

Statistical modeling

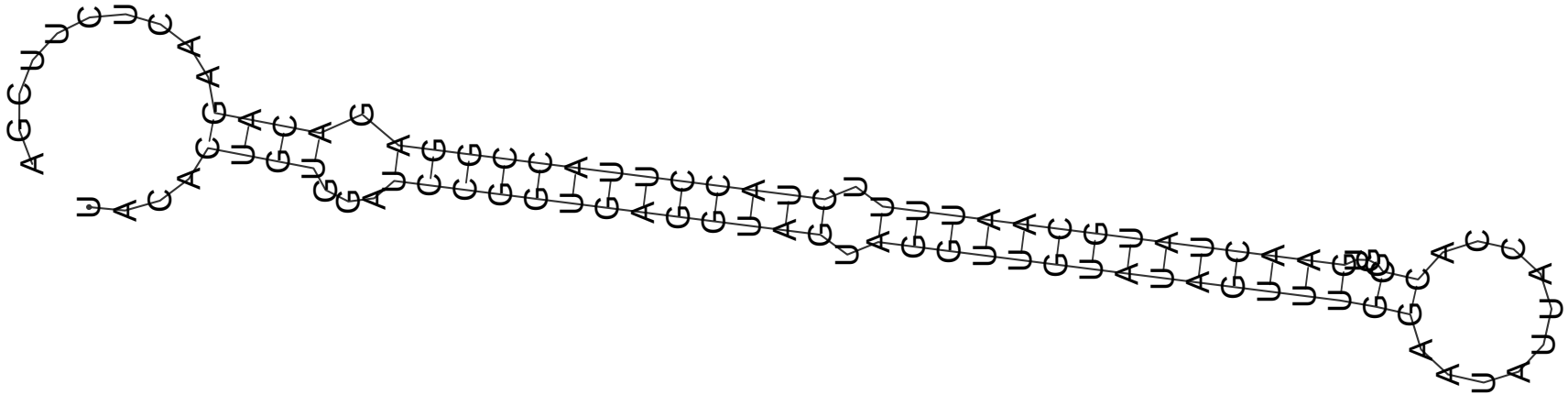
Niels Richard Hansen
University of Copenhagen
Department of Mathematical Sciences

RNA molecular structure

Let-7 (pre-cursor) from C. Elegans.

UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGAACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA

Member of the family of micro RNAs that terminate or inhibit the translation of mRNA to protein. The pre-cursor is embedded as a gene in the DNA – we want to find genes with similar structure.



StemSearch

StemSearch is an implementation of a search algorithm for general stem-loop motifs.

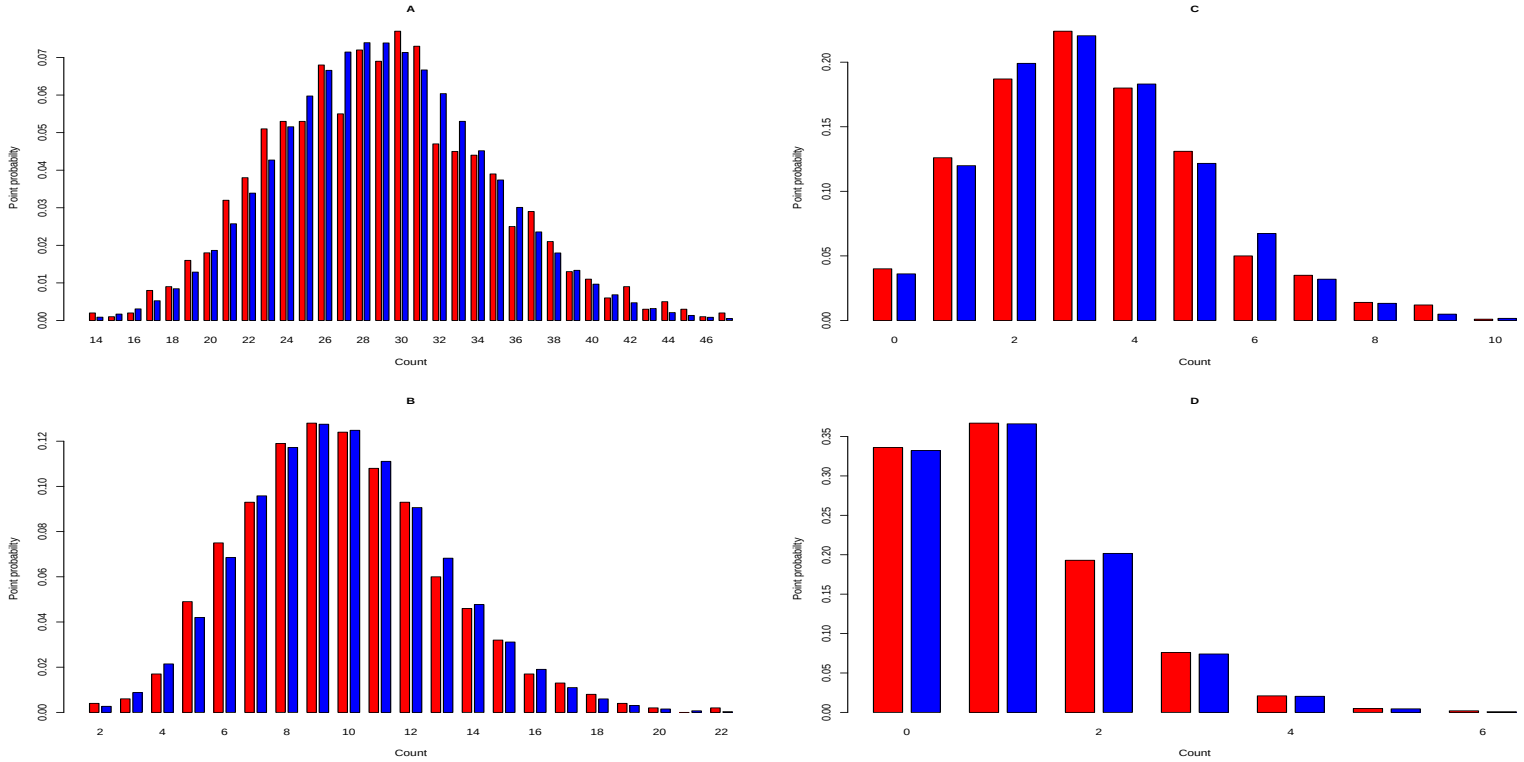
For a fixed threshold t and a sequence of length n we are given the N_t (declumped) findings with a score exceeding t. The statistical null model states the N_t is Poisson distributed with

$$\mathbb{E}(N_t) = nK \exp(-\lambda t)$$

and the excesses are iid exponentially distributed with parameter λ.

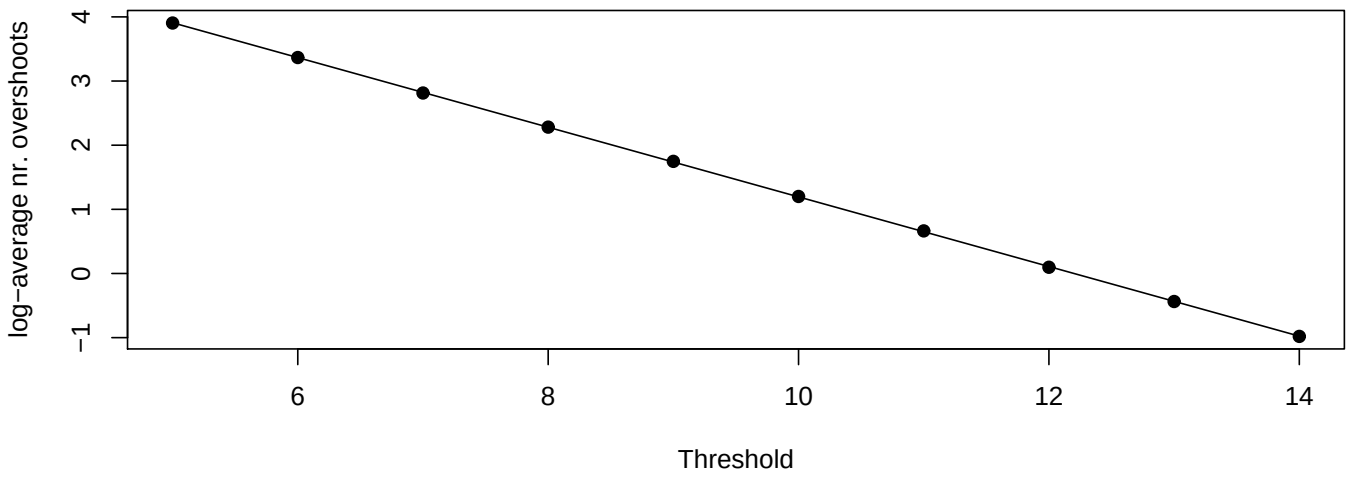
The model is only valid for t sufficiently large. [4]

StemSearch



Empirical (red) and theoretical Poisson point probabilities (blue) using *StemSearch* with $n = 5000$, $b = 200$, $v = (-4, -1, -1)$ and *C.Elegans* genome first order Markov transition probabilities. Here 1000 sequences with thresholds $t = 6$ (A), $t = 8$ (B), $t = 10$ (C) and $t = 12$ (D). The variance-to-mean ratio for the empirical counts are 1.117 (A), 1.034 (B), 1.052 (C) and 1.037 (D).

StemSearch



The log-average number of overshoots as a function of the threshold for a simulation study using *StemSearch* with $n = 5000$, $b = 200$ and $v = (-4, -1, -1)$ on sequences generated by a first order Markov chain with *C.Elegans* genome transition probabilities. The line is the least squares fit to the points with slope -0.54 and intercept 6.62.

StemSearch

We use the standard Hill estimator from extreme value statistics;

$$\hat{\lambda} = \left(\frac{1}{N} \sum_{i}^{N} S_{(m-i+1):m} - S_{(m-N):m} \right)^{-1}$$

of λ, where

$$S_{1:m} < \ldots < S_{m:m}$$

denote the ordered m overshoots of a (suitable) threshold t.

$$\hat{K} = \exp(\hat{\lambda} S_{(m-N):m}) \frac{N}{n}.$$

Poisson process limits

With $(X_k)_{k\geq 1}$ a sequence of random variables we can often associate a random measure

$$\mu_n = \sum_i \delta_{(t_i, m_i)} \in \mathcal{M}([0,1] \times E)$$

which places motif m_i at position t_i.
With restrictions of the following type:

- Stationarity or asymptotic stationarity of $(X_k)_{k\geq 1}$.
- Rare motifs – $\mathbb{E}(\mu_n([0,1] \times E)) \simeq \lambda$ for large n and rare motifs.
- Motifs are declumped.
- Weak – or moderate – dependence in $(X_k)_{k\geq 1}$.

Then $\mu_n(\cdot \times E)$ converges weakly to an homogeneous Poisson random measure (Poisson process) on $[0,1]$ for $n \to \infty$.

Motifs in genomes

The DNA-alphabet is $E = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.

- Words are finite strings; ACGTTA, GTAACA, AGA, ...
- A collection of words is a Motif.
- Regular expressions; A.*[CG]TT., G.[AG][AG]C., ...
- Weight matrices; $W = \{W_{x,i}\}_{x \in E, i=1,\ldots,k}$.
  A word $w = x_1 \ldots x_k$ receives the score

$$S_w = \sum_{i=1}^{k} W_{x_i,i}.$$

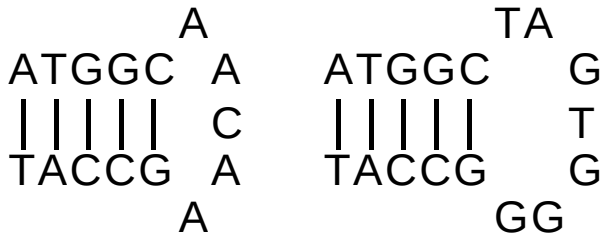
A motif is specified as $\{w | S_w > t\}$.
See [2] for a probabilistic treatment.

Stem-loop motifs

The regular expression:

$$\texttt{ATGGC.}\{5,7\}\texttt{GCCAT}$$

corresponds to stem-loop structures



with 5-7 letters in the loop.

Reinert and Schbath [5] investigate Poisson approximations focusing on exact error bounds for motifs in homogeneous Markov chains – including stem-loop motifs as the above.

Problems

The homogeneous Poisson process model suffer from some problems - even as a null model:

- Non-Markov nature of genomic sequences.
- Heterogeneity of genomic sequences:
  - Heterogeneous nucleotide frequencies.
  - Low-complexity and repeat patterns (fixed by repeat masker?).
  - Heterogeneous distribution of larger motifs.
- Dependence structures of biologically relevant motifs.

How to get beyond the null model?

Example

Is there an over-representation of simultaneous occurrence of the two words $w_1 =$ AACCTGG and $w_2 =$ ATGCCAT in the sequences $x_1, \ldots, x_m$ ($x_i = x_{i1} \ldots x_{in(i)}$)?

Null model: The words occur as independent Poisson processes in each sequence (intensities λ_1^i and λ_2^i), and the sequences are independent.

$$R = \sum_{i=1}^{m} 1(w_1 \in x_i, w_2 \in x_2) \overset{\text{approx}}{\sim} \text{Poi}(\xi)$$

with

$$\xi = \sum_{i=1}^{m} (1 - e^{-\lambda_1^i})(1 - e^{-\lambda_2^i}).$$

A theoretical foundation is given in Reinert and Schbath [5].

Example - continued

In a concrete application, Marc Riemer Friedländer investigated in his Master's Thesis the co-occurrence of miRNA target sites (7 letter words) in the 3'UTR of mRNA taking

$$\log(\lambda_w^i) = \beta_w + \beta_w(0)\log n(i) + \beta_w(\mathsf{A})\log f_{\mathsf{A}}(i) + \ldots \beta_w(\mathsf{T})\log f_{\mathsf{T}}(i)$$

with $f_{\mathsf{A}}(i), \ldots, f_{\mathsf{T}}(i)$ the relative frequency of nucleotides in sequence i.

Parameters were estimated using Poisson regression with a much better model fit than the iid sequence model where $\beta_w(0) = 1$, $\beta_w = 0$ and

$$\beta_w(\alpha) = \text{number of times } \alpha \text{ occurs in word } w$$

ENCODE

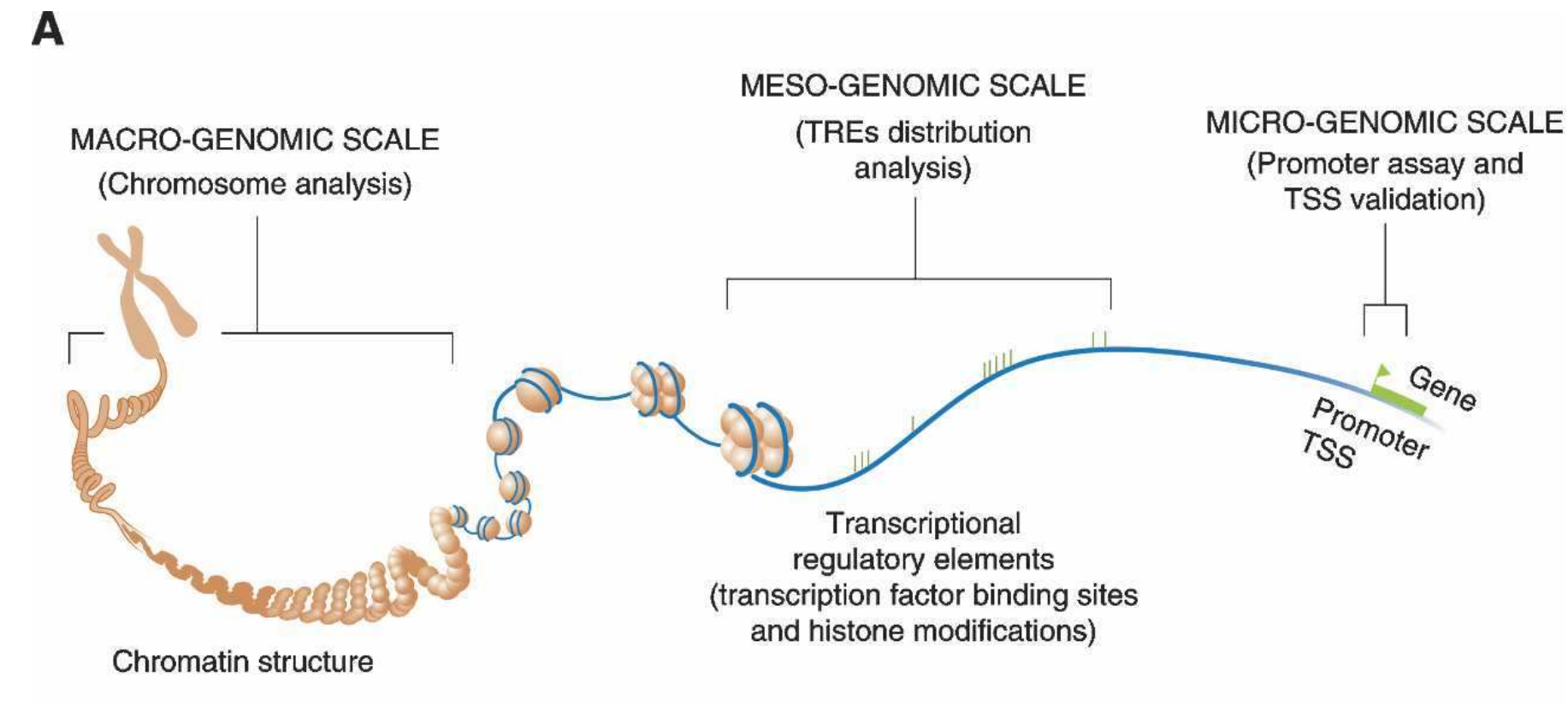


Illustration from [6] – a statistical analysis of regulatory elements in the ENCODE regions.

Transcription Factor Binding Sites

- Protein binding sites on DNA serve a central role in the regulatory mechanisms for gene transcription.
- Typically hard to locate computationally – computational predictions are noisy.
- Better experimental data is becomming available (ChIP-chip), which provides actual binding sites of proteins (e.g. ENCODE data).

Intensity based modeling

For a multivariate point-process $(N_1(t), \ldots, N_k(t))$ with filtration $(\mathcal{F}_t)_{t \geq 0}$ and adapted intensity process $\lambda(t) = (\lambda_1(t), \ldots, \lambda_k(t))$ we have

$$\mathbb{P}(N_i(t+\epsilon) - N_i(t) > 0 | \mathcal{F}_t) \simeq \lambda_i(t)\epsilon$$

We also have the log-likelihood process

$$\sum_{i=1}^{k} \left[\int_0^t \log \lambda_i(t) N_i(\mathrm{d}t) - \int_0^t \lambda_i(t) \mathrm{d}t \right].$$

A statistical modeling approach using Hawkes processes was first attempted by Gaëlle Gusto and Sophie Schbath in [3].

Hawkes processes

- Multivariate point-process $(N_1, \ldots, N_k)$ with intensity

$$\lambda_i(t) = \phi\left(\sum_{j=1}^{k} \int_0^t h_{ij}(t-s) N_j(\mathrm{d}s)\right).$$

- Lisbeth Carstensen (Ph.D.-student, Copenhagen) has an implementation fitting two-dimensional Hawkes processes with spline-based expansions of h_{ij} – including additional local sequence covariates.
- Ongoing projects: More then two dimensions, inclusion of a Cox-process component, superpositions, model selection and test-statistics for $h_{ij} = 0$.

Concluding remarks

- Actually proving an asymptotic Poisson result can be an arbitrary hard mathematical challenge.
- The iid or homogeneous Markov chain models for sequences attempt modeling on a microscopic scale.
- I believe that biologically relevant questions are better addressed with statistical models directly at the mesoscopic scale.
- Thanks for your time, thanks to Richard Gill for inviting me, and thanks to Lisbeth and Mark and the entire Bioinformatics Centre in Copenhagen.

References

[1] Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.* **17**, 9–25.

[2] Goldstein, L and Waterman, M. S. (1994) Approximations to Profile Score Distributions *Journal of Computational Biology* **1**(2), 93-104.

[3] Gusto, G. and Schbath, S. (2005) FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Statistical Applications in Genetics and Molecular Biology* **4**(1), Article 24.

[4] Hansen, N. R. (2007). Statistical models of local RNA stem-loop scores *Submitted to Bioinformatics*.

[5] Reinert, G. and Schbath, S. (1998). Compound poisson and poisson process approximations for occurrences of multiple words in markov chains. *Journal of Computational Biology* **5**, 223–253.

[6] Zhengdong D. Zhang, Alberto Paccanaro, Yutao Fu, Sherman Weissman, Zhiping Weng, Joseph Chang, Michael Snyder and Mark B. Gerstein (2007). Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res.* **17**, 787-797.