Point Process and Marked Point Process Models of Features on Genomes

International Symposium

Recent Challenges for Statistics in the Biosciences

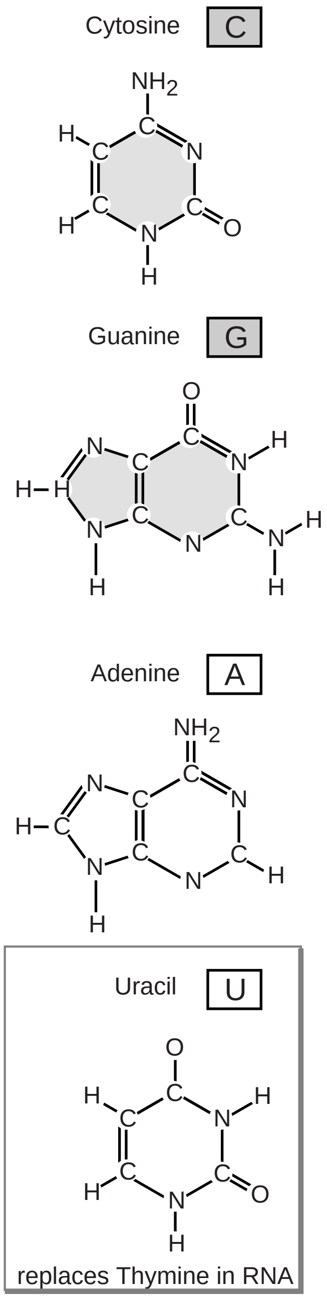
100 Years after Gustav Zeuner

Niels Richard Hansen

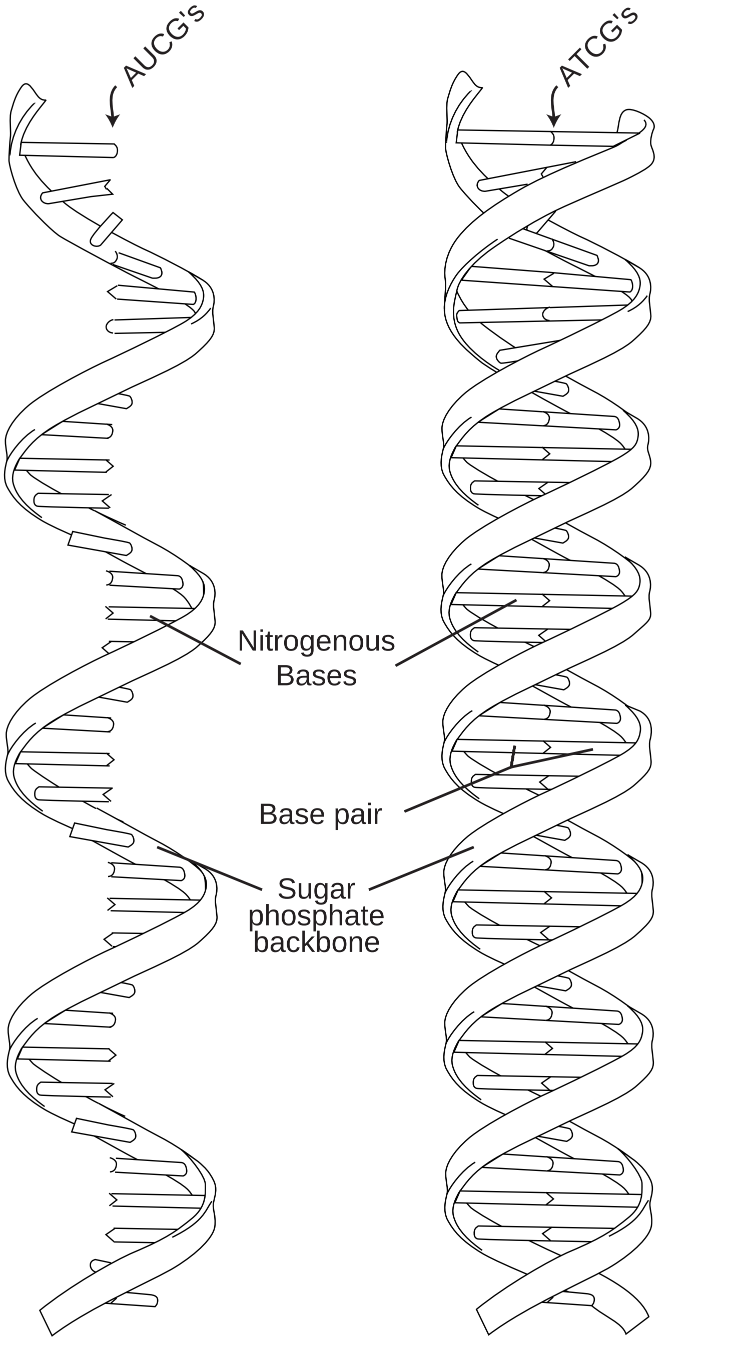
University of Copenhagen

Department of Mathematical Sciences

Ribonucleic acid(RNA)



Nitrogenous Bases

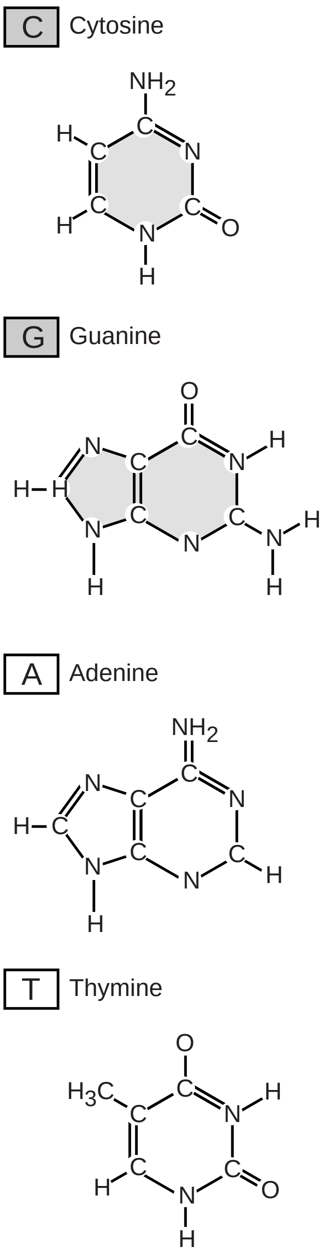


RNA

Ribonucleic acid

DNA

Deoxyribonucleic acid



Nitrogenous Bases

RNA molecular structure

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

UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGAACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA

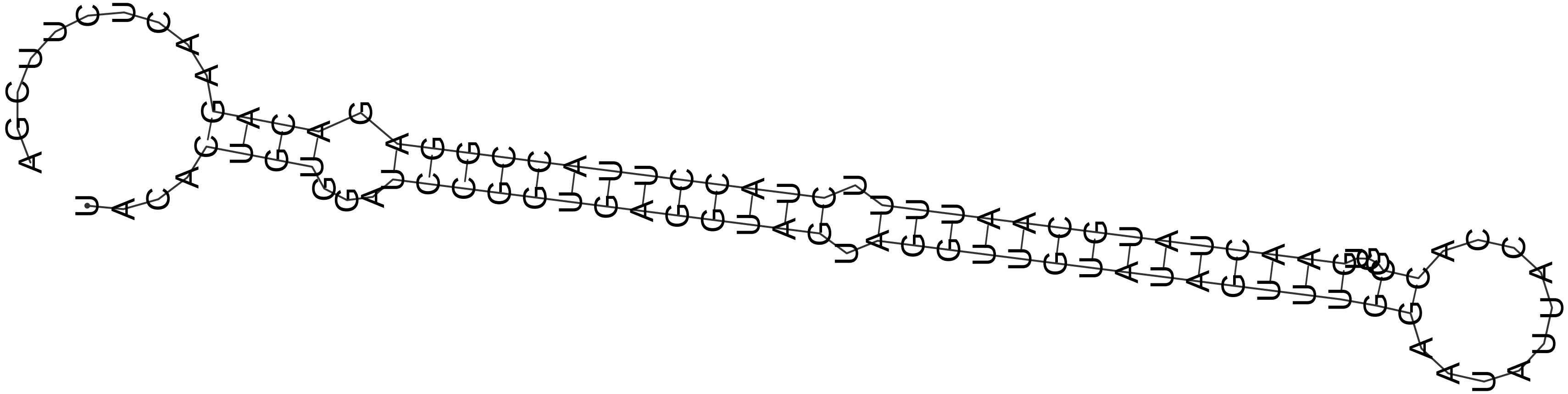
Member of the family of micro RNAs that terminate or inhibit the translation of mRNA to protein.

RNA molecular structure

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

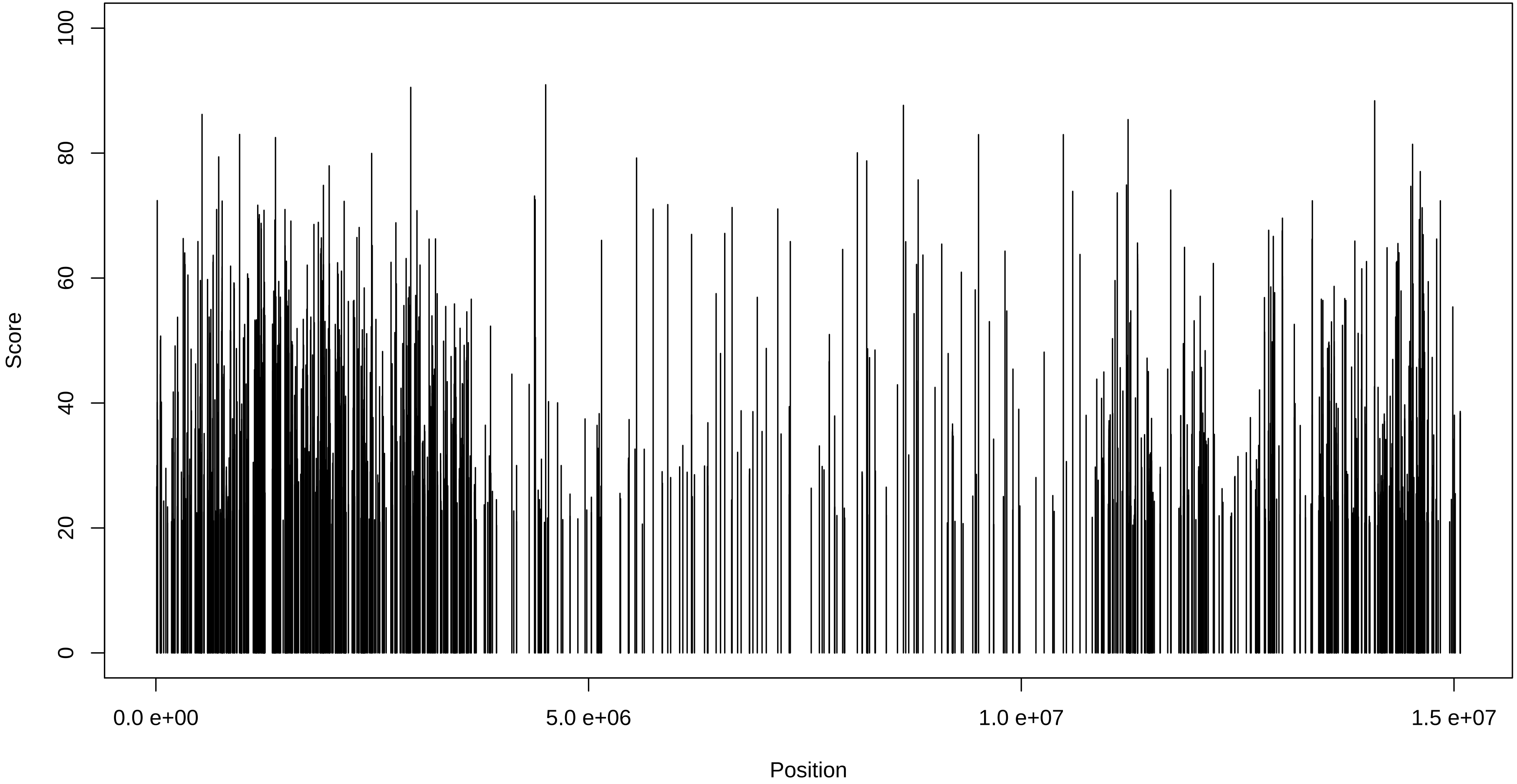
UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGAACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA

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.

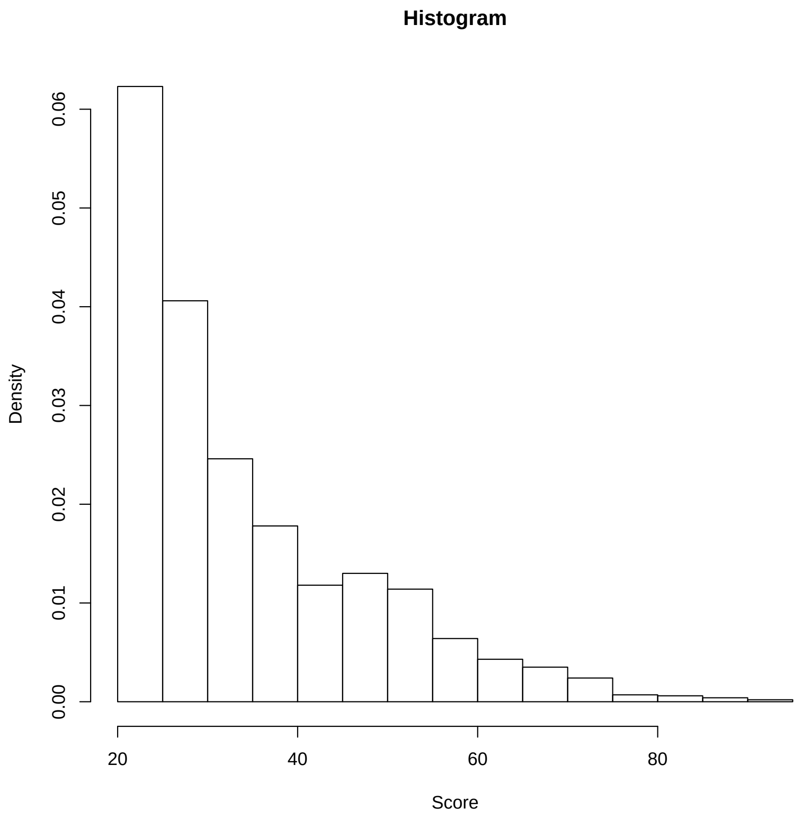


Marked point process view

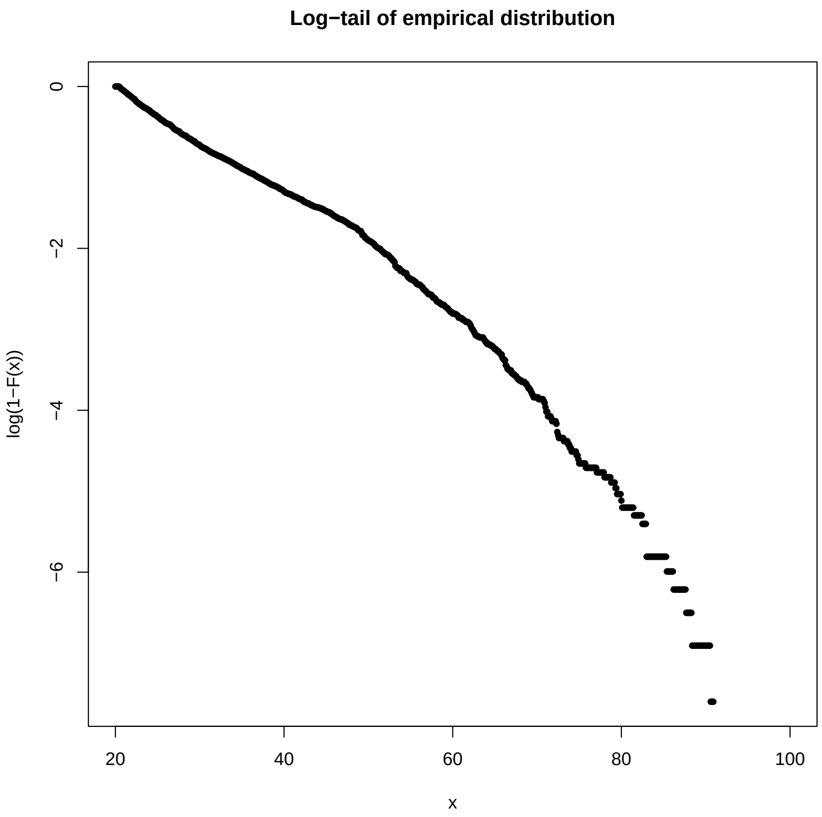
An implementation (StemSearch) gives for *C.Elegans*, chromosome I:



Distribution of overshoots



Histogram



Log−tail of empirical distribution

Objectives

- Want a statistical (null) model of the random occurrences (biologically non-significant) of high-scoring points.

Objectives

- Want a statistical (null) model of the random occurrences (biologically non-significant) of high-scoring points.
- Also want statistical models of biologically significant occurrences.

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.

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., ...

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

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

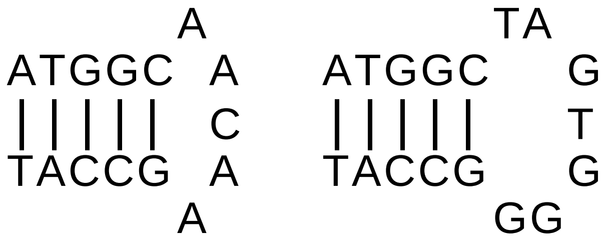
A motif is specified as $\{w | S_w > t\}$.

Stem-loop motifs

The regular expression:

ATGGC.{5,7}GCCAT

corresponds to stem-loop structures



```
                 A                   TA
ATGGC  A         ATGGC     G
|||||  C         |||||     T
TACCG  A         TACCG     G
                 A                   GG
```

with 5-7 letters in the loop.

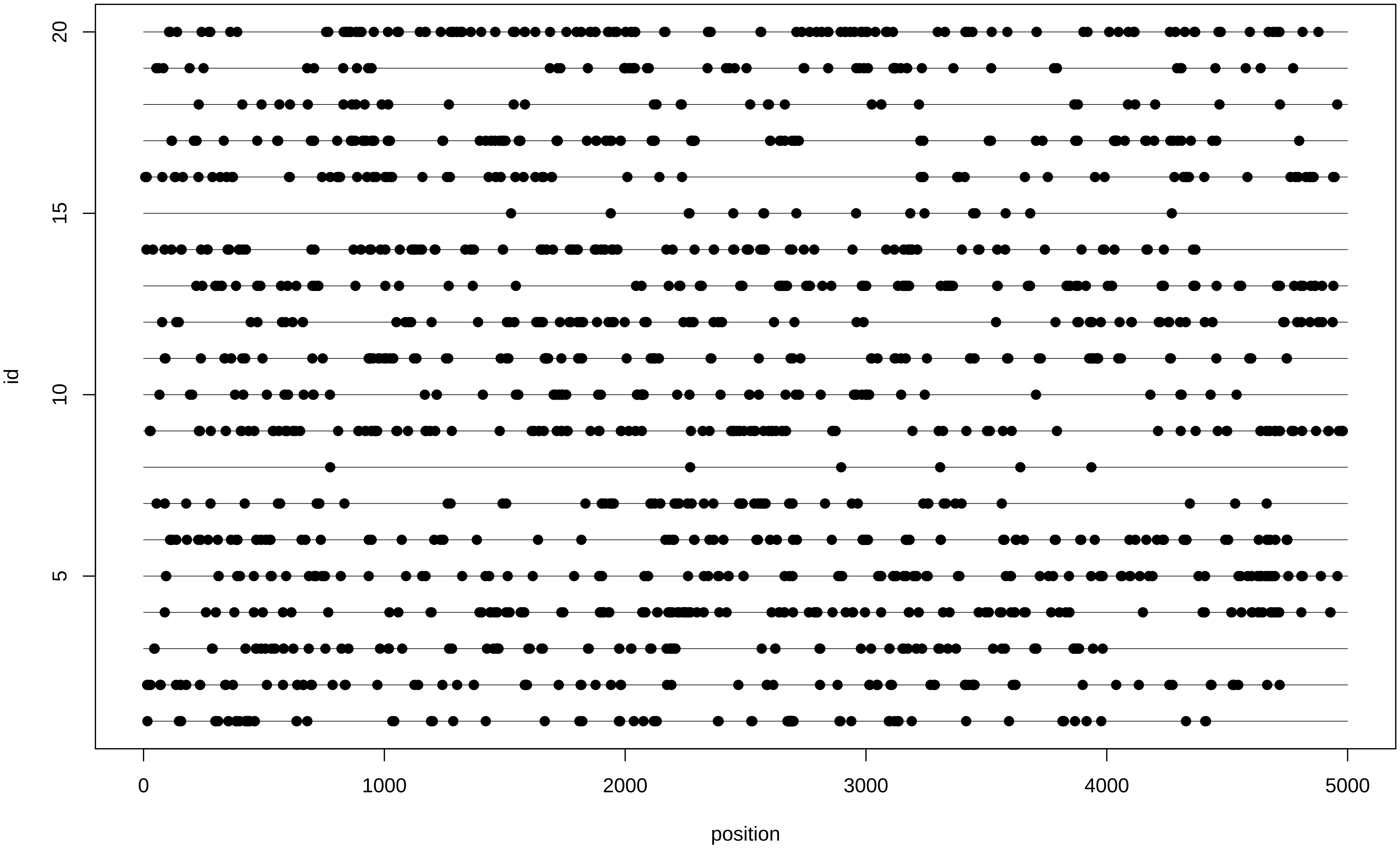
Reinert and Schbath [5] investigate Poisson approximations focusing on exact error bounds for motifs in homogeneous Markov chains – including stem-loop motifs as the above.

MEF2

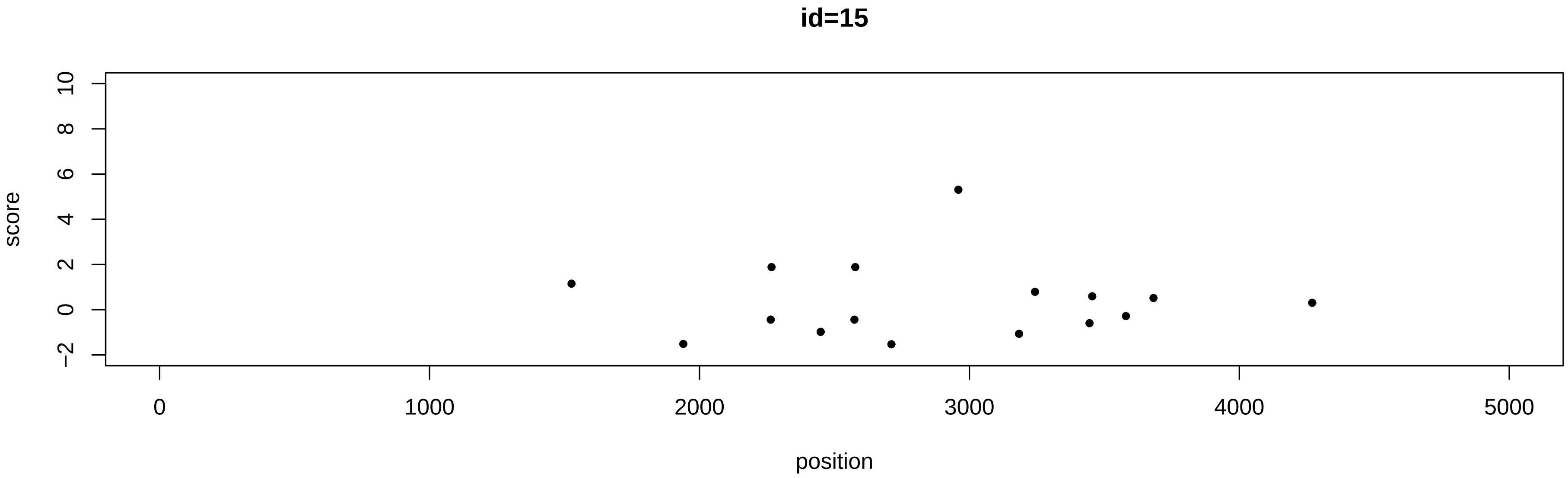
Potential binding sites for the myocyte-specific enhancer factor 2 (MEF2), which is involved in the muscle-specific expression of a number of genes, can be located using a weight matrix:

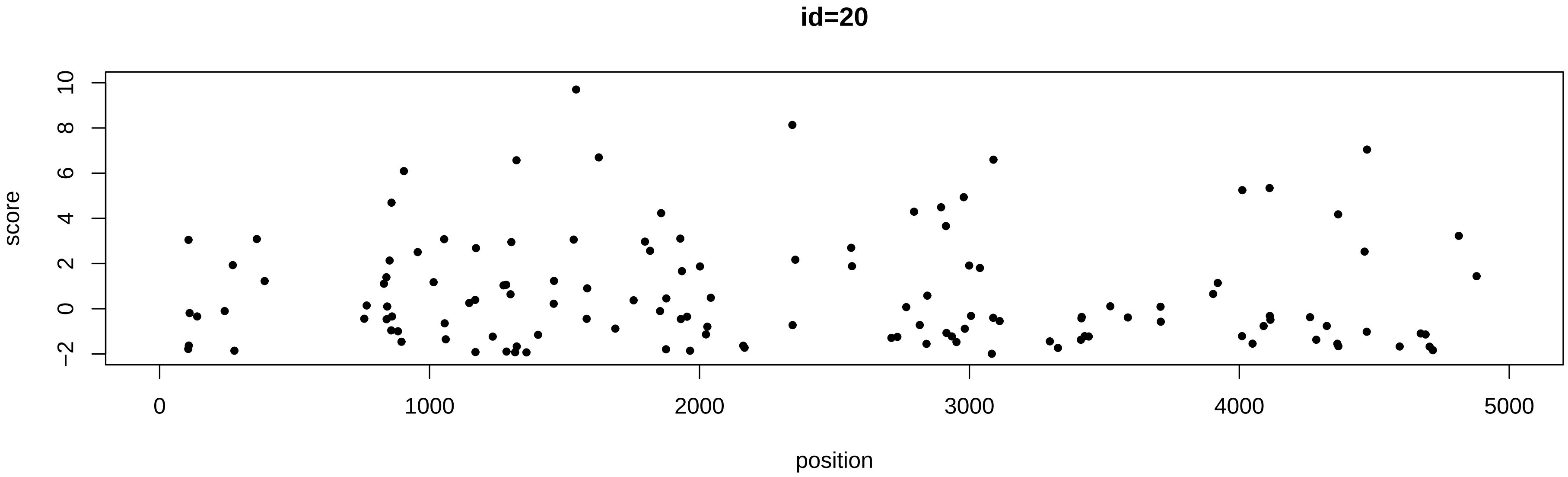
	Position										
	1	2	3	4	5	6	7	8	9	10	11
A	-1.93	-1.93	1.17	0.80	1.25	1.30	1.27	-3.32	1.34	-1.01	0.27
C	1.25	-1.05	-3.25	-3.25	-3.25	-3.25	-3.25	-3.25	-3.25	-3.25	0.67
G	-1.89	-3.28	-3.28	-3.28	-2.58	-3.28	-2.58	-3.28	-3.28	1.28	-0.79
T	-1.04	1.20	-0.51	0.46	-1.15	-1.73	-1.40	1.31	-3.34	-2.65	-1.04

Potential MEF2 binding sites



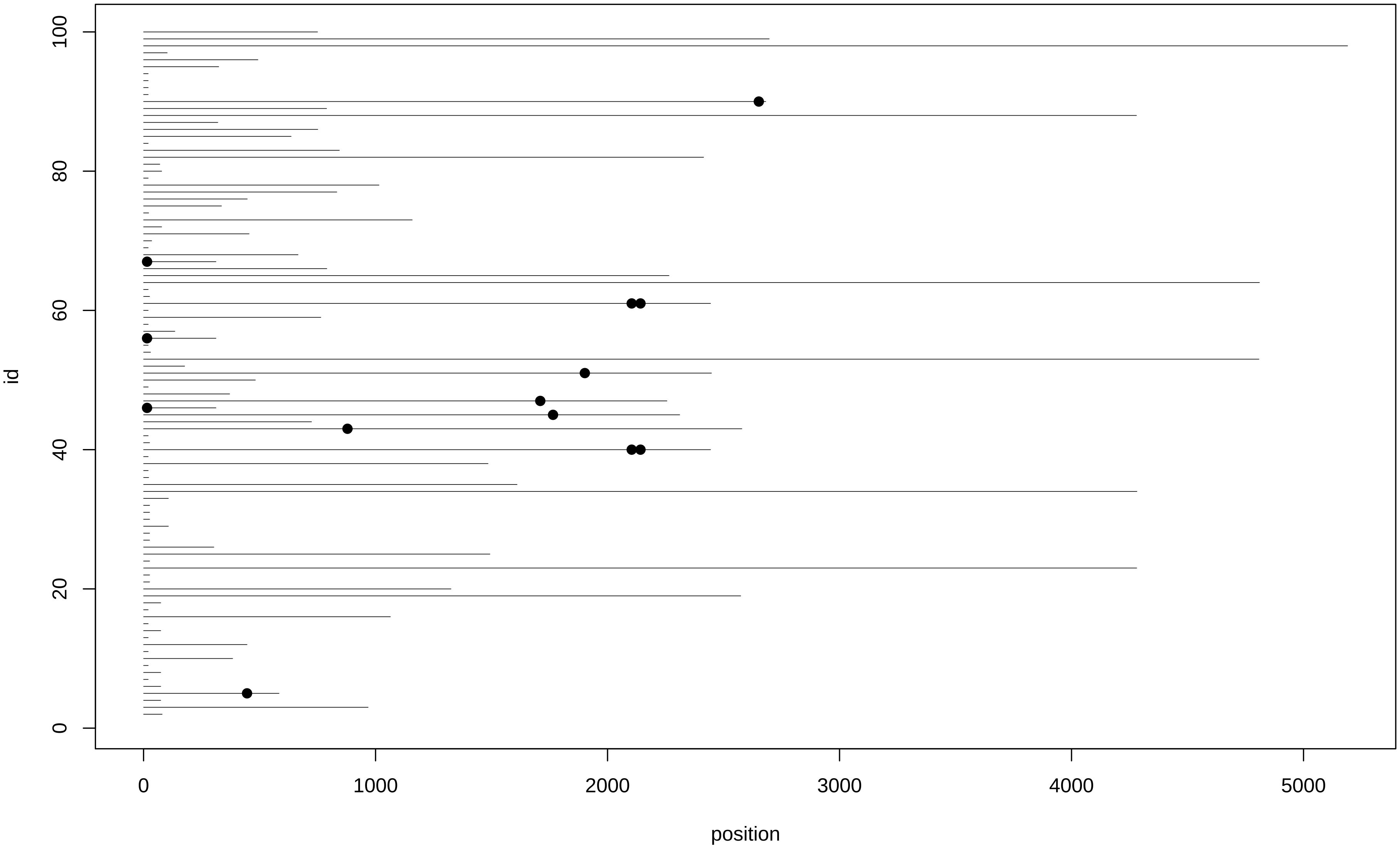
– and with scores as marks





Potential miRNA target sites

Occurrences of AACCTGG in 3’ UTR



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.

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 $n \to \infty$.

Problems

The homogeneous Poisson process model suffers from several problems - even as a null model:

- Non-Markov nature of genomic sequences.

Problems

The homogeneous Poisson process model suffers from several 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.

Problems

The homogeneous Poisson process model suffers from several 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 the simultaneous occurrence of the two words $w_1 =$ AACCTGG and $w_2 =$ ATGCCAT in the sequences $x_1, \ldots, x_m$ ($x_i = x_{i1} \ldots x_{in(i)}$)?

Example

Is there an over-representation of the 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

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

Example

Is there an over-representation of the 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.

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

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

Example - continued

- Marc showed that the model gave a distribution of the test statistic R clearly superior to common distributions based on sequence shuffling.

Example - continued

- Marc showed that the model gave a distribution of the test statistic R clearly superior to common distributions based on sequence shuffling.
- The model is based on “global” sequence covariates and does not attempt to capture heterogeneity in a single sequence.

Example - continued

- Marc showed that the model gave a distribution of the test statistic R clearly superior to common distributions based on sequence shuffling.
- The model is based on “global” sequence covariates and does not attempt to capture heterogeneity in a single sequence.
- It would be desirable to use local covariates also like local (windowed) nucleotide frequencies.

Example - continued

- Marc showed that the model gave a distribution of the test statistic R clearly superior to common distributions based on sequence shuffling.
- The model is based on “global” sequence covariates and does not attempt to capture heterogeneity in a single sequence.
- It would be desirable to use local covariates also like local (windowed) nucleotide frequencies.
- One example is Aalens non-parametric additive hazards model known from survival analysis.

Example - continued

- Marc showed that the model gave a distribution of the test statistic R clearly superior to common distributions based on sequence shuffling.
- The model is based on “global” sequence covariates and does not attempt to capture heterogeneity in a single sequence.
- It would be desirable to use local covariates also like local (windowed) nucleotide frequencies.
- One example is Aalens non-parametric additive hazards model known from survival analysis.
- Another approach include spline-based expansions of position and position-covariate effects.

ENCODE

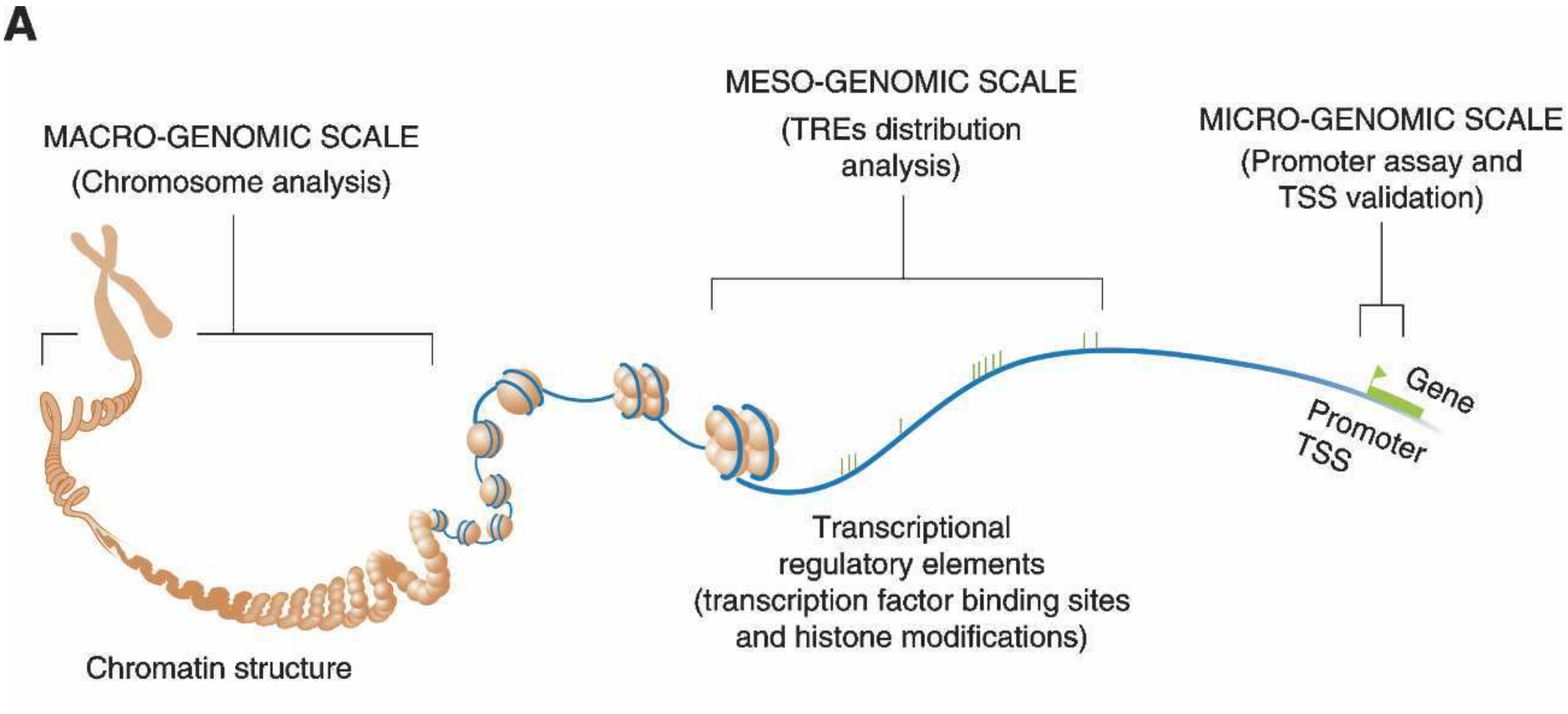


Illustration from [6] – a statistical analysis of regulatory elements in the ENCODE regions.

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

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

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

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

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 [2].

Hawkes processes

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

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

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.

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

- Modeling biologically sequences is a multi-scale problem with sequence motif occurrences being on a meso-genomic scale.

Concluding remarks

- Modeling biologically sequences is a multi-scale problem with sequence motif occurrences being on a meso-genomic scale.
- The micro-genomic scale models based on iid or homogeneous Markov chain models for sequences have a hard time capturing the organization of motifs on the meso-genomic scale.

Concluding remarks

- Modeling biologically sequences is a multi-scale problem with sequence motif occurrences being on a meso-genomic scale.
- The micro-genomic scale models based on iid or homogeneous Markov chain models for sequences have a hard time capturing the organization of motifs on the meso-genomic scale.
- I believe that biologically relevant questions are better addressed with statistical models directly at the meso-genomic scale.

Concluding remarks

- Modeling biologically sequences is a multi-scale problem with sequence motif occurrences being on a meso-genomic scale.
- The micro-genomic scale models based on iid or homogeneous Markov chain models for sequences have a hard time capturing the organization of motifs on the meso-genomic scale.
- I believe that biologically relevant questions are better addressed with statistical models directly at the meso-genomic scale.
- Thanks for your time and for the invitation ... and thanks to Lisbeth and Marc and the 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] GUSTO, G. AND SCHBATH, S. (2005) FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model.

[3] HANSEN, N. R. (2007). Asymptotics for Local Maximal Stack Scores with General Loop Penalty. *Advances in Applied Probability* **39**(3), 776-798.

[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.