Statistically Significant Patterns in DNA Sequences – Part II

part of “Genomik der Genregulation”

Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, SS 2011

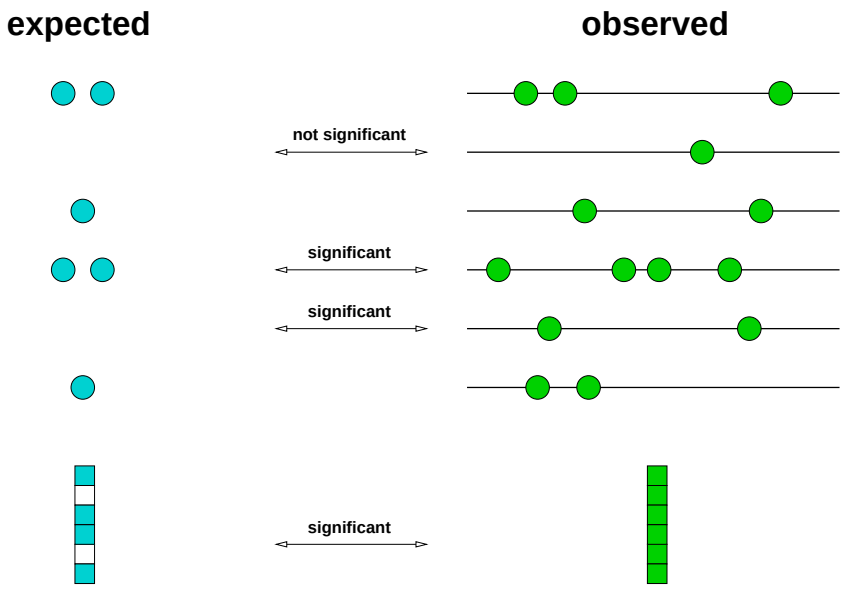
Finding Sequence Motifs Enriched/Depleted in a Set of Sequences

Given:

- a set of N functionally related but not evolutionarily homologous sequences

Find:

- all motifs of length w which are significantly shared (or avoided) in the N sequences



“Was bisher geschah”

for a single sequence $i \in N$ of lenght L

a word k of length w and characters u_1 to u_w is expected to occure $E_i(k)$ times.

$$E_i(k) = \frac{f_i(u_1, u_2) \times f_i(u_2, u_3) \times \ldots \times f_i(u_{w-1}, u_w)}{f_i(u_2) \times f_i(u_3) \times \ldots \times f_i(u_{w-1})} \times (L - w + 1)$$

Comparing Occurences in Single Sequences

Given we are interested in the probability $\pi_i(k)$ that the motif k is found in sequence i at least once and w-mers are poisson distributed:

expected

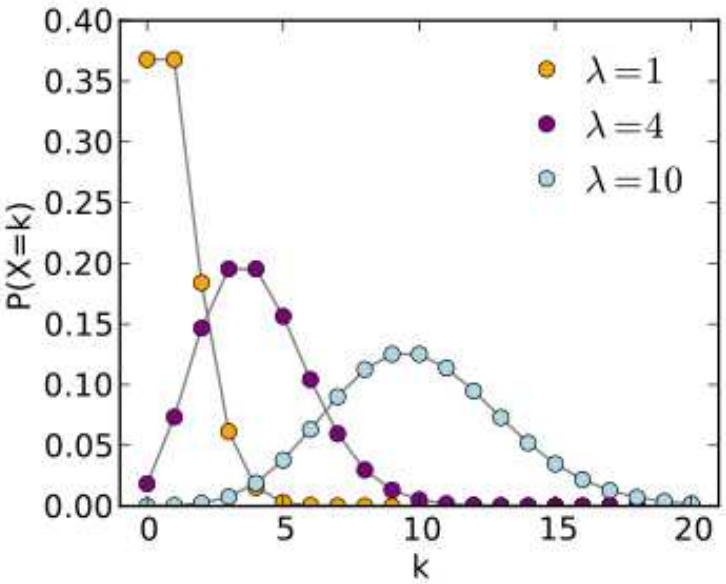
$$\pi_i(k) = 1 - e^{-E_i(k)}$$

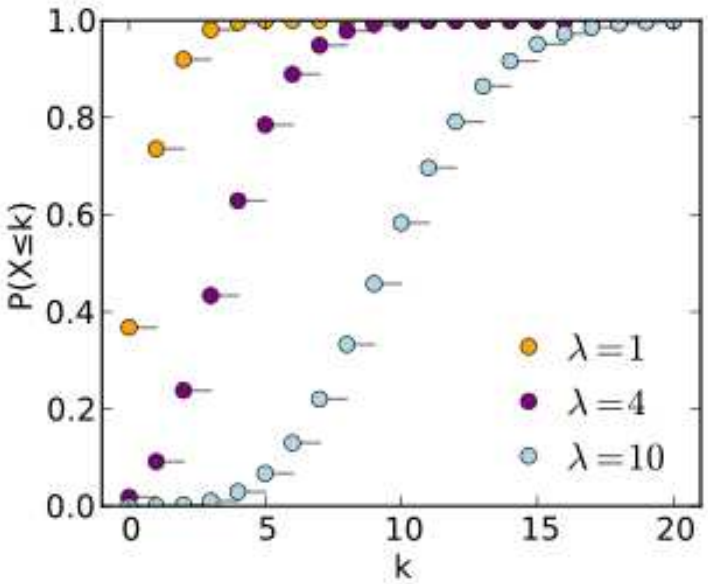
observed

$p_i(k)$ is one if motif k is found in sequence i or zero otherwise

We can now compare the expected and observed value for motif k and sequence i. Is there a significant difference?

Poisson Distribution

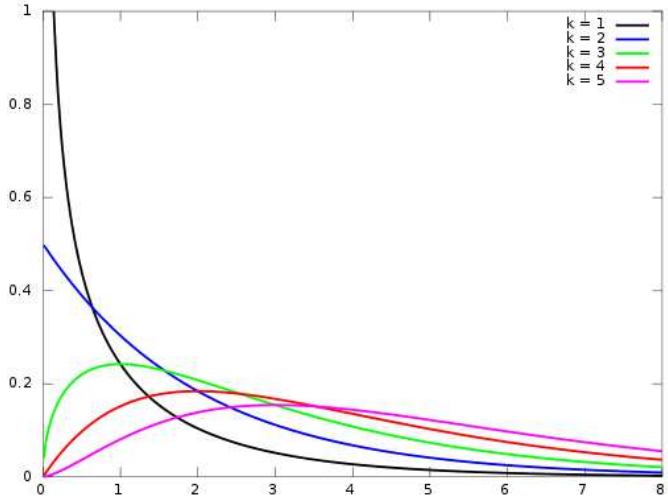


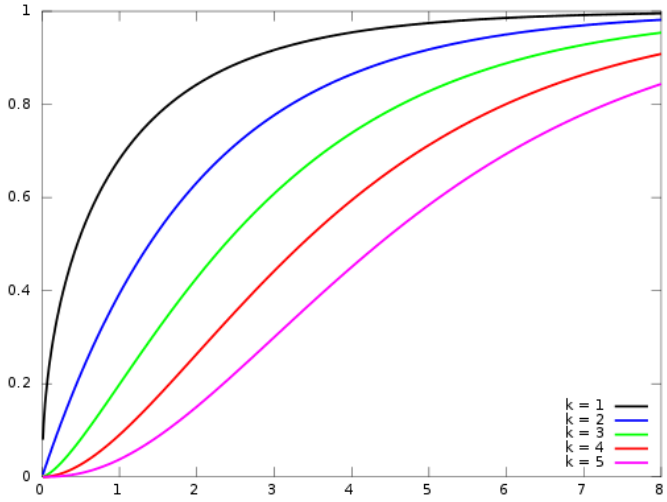


- k is the number of occurrences
- λ is the expected number of occurrences
- $P(X = k)$ is the probability that the observable occures exactly k times
- $P(X \leq k)$ is the probability that the observable occures fewer or exactly k times
- mean: λ, variance: λ

χ^2-Test

In the chi-squared test, the chi-squared distribution is used to tests for goodness of fit of an observed distribution to a theoretical one.





$$\chi^2 = \sum_{i=1}^{N} \frac{(p_i(k) - \pi_i(k))^2}{\pi_i(k)} \tag{1}$$

In the table of χ^2-quantiles one can look up the significance threshold given the degrees of freedom ($N - 1$) and the desired level of significance. E.g. for $N = 20$ and a level of 1% significance, χ^2 has to acceed 36.19 to reject the hypothesis that the distributions are “the same”.

Comparing Occurences in Sets of Sequences

Here, the (expected or observed) occurence per sequence are summed over all sequences N.

expected

$$\Pi(k) = \sum_{i=1}^{N} \pi_i(k)$$

observed

$$P(k) = \sum_{i=1}^{N} p_i(k)$$

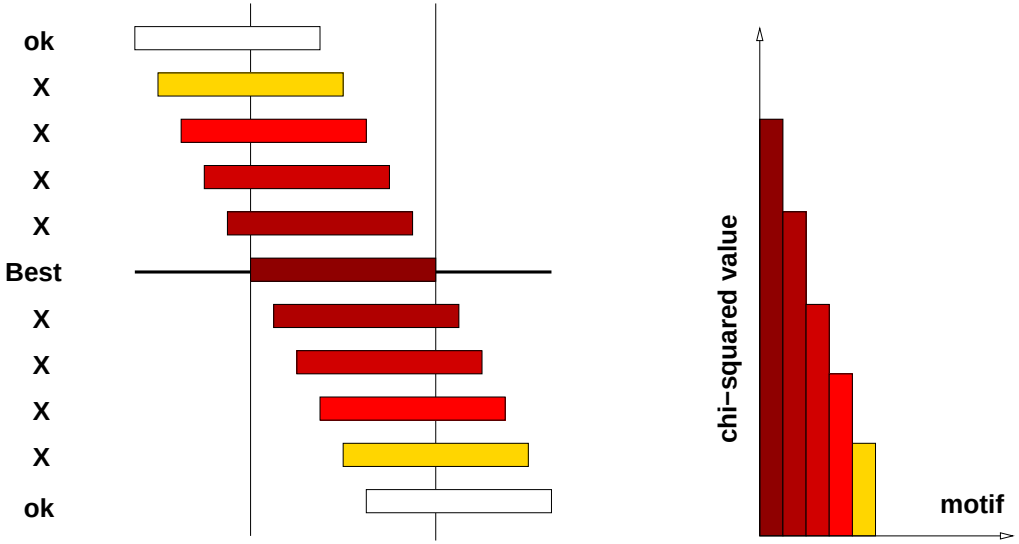
To compare the motifs among each other, a score is calculated. The higher the score, the more difference between the expected and observed motif occurences.

$$\chi_k^2 = \frac{(P(k) - \Pi(k))^2}{\Pi(k)} \tag{2}$$

Scoring Motifs

- for all 4^w possible motifs
  - calculate the expected occurences for all N sequences
  - count the number of occurances for all N sequences
  - calculate χ_k^2
- rank the motifs by χ_k^2
- remove unsignificant motifs resulting in M significant motifs

Problem! Motifs might overlap!



Solving this Problem

1. take the best scoring motif out of the set M
2. for all $M - 1$ other motifs
   - remove the motif if it overlaps by at least $w/2$ consecutive motifs with the best scoring motif of this round
3. goto step 1 with the remaining ist of motifs being the new M

Result: a ranked list of significant motifs that do not overlap more than by half

Literature

[Pesole, 1992] Pesole G., Prunella N., Liuni S., Attimonelli M. and Saccone C. *WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences* Nucl. Acids Res. 1992; 20(11):2871-2875