Regulatory sequence analysis

Word count statistics

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Background sequences

- The frequencies observed for a *k*-letter word in a reference sequence set (background sequence) can be used to estimate the expected frequencies of the same *k*-letter word in the sequences to be analyzed.
- Typical background models:
  - whole genome
    - But this will bias the estimates towards coding frequencies, especially in microbial organisms, where the majority of the genome is coding.
  - whole set of intergenic sequences
    - More accurate than whole-genome estimates, but still biased because intergenic sequences include both upstream and downstream sequences
  - Whole set of upstream sequences, same sizes as the sequences to be analyzed
    - Requires a calibration for each sequence size
  - Whole set of upstream sequences, fixed size (default on the web site)
    - Reasonably good estimate for microbes, NOT for higher organisms.

Topics

- Word probabilities
  - What is the probability to observe a given word at a given position of the sequence ?
- Expected word frequencies
  - Given the word probability and the sequence length, what is the expected word frequency ?
- Significance
  - We observe x occurrences of a word. Is this word significantly
    - Over-represented ?
    - Under-represented ?
  - Choice of a scoring scheme
    - Which theoretical distribution should we use to score this significance ?
  - Choice of a threshold
    - How to select a threshold to consider that the score is significant ?

Estimating a sequence probability with a Markov model

$$P(S) = P(S_{1,m}) \prod_{i=m+1}^{L} P\left(r_i \mid S_{i-m,i-1}\right)$$

In a Markov chain model, the probability to find a letter at position *i* depends on the residues found at the *m* preceding residues.

Example:

Markov model order 1

$P_{CACG} = P_{CA} \times P_{(A|C)} \times P_{(C|G)}$

$= F_{CA} \times F_{AC} / F_A \times F_{CG} / F_G$

where $P_{(A|C)}$ *is the probability for a A to be followed by a C*

F_{CA} *is the frequency of CA in the sequence set*

Markov model order 2

P(CACG)= P(CA) x P(C|CA) x P(G|AC)

= P(CA) x P(CAC)/P(CA) x P(ACG)/P(AC)

= P(CAC) x P(ACG) / P(AC)

P(CACGTG) = P(CA) x P(C|CA) x P(G|AC) x P(T|CG) x P(G|GT)

= P(CA) x P(CAC)/P(CA) x P(ACG)/P(AC) x P(CGT)/P(CG) x P(GTG)/P(GT)

= [P(CAC) x P(ACG) x P(CGT) x P(GTG)]/ [P(AC) x P(CG) x P(GT)]

Markov chain models

$$P(S) = P(S_{1,m}) \prod_{i=m+1}^{L} P\left(r_i \mid S_{i-m,i-1}\right)$$

The expected frequency of a sequence (e.g. : a k-letter word) is estimated on basis of sub-word frequencies.

The probability to find a letter at position *i* depends on the residues found at the *m* preceding residues.

Example:

Markov model order 1

$P_{CACG} = P_{CA} \times P_{(A|C)} \times P_{(C|G)}$

$= F_{CA} \times F_{AC} / F_A \times F_{CG} / F_G$

where $P_{(A|C)}$ *is the probability for a A to be followed by a C*

F_{CA} *is the frequency of CA in the sequence set*

Markov model order 2

$P(CACG) = P(CA) \times P(C|CA) \times P(G|AC)$

$= P(CA) \times P(CAC)/P(CA) \times P(ACG)/P(AC)$

$= P(CAC) \times P(ACG) / P(AC)$

$P(CACGTG) = P(CA) \times P(C|CA) \times P(G|AC) \times P(T|CG) \times P(G|GT)$

$= P(CA) \times P(CAC)/P(CA) \times P(ACG)/P(AC) \times P(CGT)/P(CG) \times P(GTG)/P(GT)$

$= [P(CAC) \times P(ACG) \times P(CGT) \times P(GTG)] / [P(AC) \times P(CG) \times P(GT)]$

Regulatory sequence analysis

Expected word occurrences

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Expected occurrences

- How many occurrences do we expect in the sequence ?

$$E_W = P_W \; x \; T$$

where

E_w *is the expected number of occurrences*

P_W *is the probability to observe W at each position*

T *is the number of possible positions*

- The number of possible positions depends on
  - The sequence length
  - The word length
  - The counting mode
    (accept or prevent overlapping matches for periodical words)

Number of possible positions

- The number of possible positions (T) depends on
  - The sequence length
  - The word length
  - The counting method

Overlap accepted

$$T = \sum_{i=1}^{s} \left(L_i - k + 1 \right)$$

L_i *length of the* i^{th} *sequence*

T *Number of possible positions for the word*

Overlapping matches discarded : each match forbids the k-1 next positions

$$T = S_j (L_j - k + 1) - (k-1)C_w \qquad \text{(1 strand)}$$

C_w *number of occurrences of the word W*

Expected number of matching sequences

In some cases, one is interested by the first occurrence of each sequence only.

In how many sequences do we expect to find at least one match for the word W ?

$$Em_W = S_j \, Pm_{W_j}$$

Em_w *expected number of sequences containing at least one W*

Pm_{wj} *probability for each sequence to include at least one match*

j *sequence index (from 1 to S)*

$$Pm_{W_j} = 1 - P_{!m_{W_j}} = 1 - (1 - P_W)^{t_j}$$

$P_{!mw_j}$ *probability for a given sequence to have* ***no*** *match with W*

t_j *number of possible positions in the j*th *sequence of the set*

Regulatory sequence analysis

Scoring the significance of word occurrences

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Scoring schemes

- Several statistics can be used to score the significance of the observed number of occurrences
  - Ratio
  - Log likelihood
  - Binomial distribution
  - Poisson distribution
  - Normal distribution
  - Chi-square
  - Compound Poisson
  - Non-parametric methods
- How to choose among all these possibilities ?
  - Some statistics are biased and should be avoided
  - Among the other ones, the choice depends on the parameters of the analysis, which will determine whether the conditions of applicability are satisfied.

Scoring scheme - Representation ratio

- $r = C_W / E_W$
- Advantages
  - easy to calculate (note: this is a poor advantage, giving the weakness !)
- Weaknesses
  - overestimates the importance of words with weak expected frequencies
  - no correction for self-overlapping patterns
- Recommendation
  - Never use the observed/expected ratio to estimate over/under representation !

Scoring scheme - log likelihood

- $K = F_W \ ln(F_W / P_W)$
- Advantages
  - correction for the words with weak expected frequency
  - easy to calculate
- Weaknesses
  - no correction for self-overlapping patterns
  - no estimation of the P-value

Scoring scheme - Binomial P-value

- Advantages
  - Rigorous probability
  - No bias for the words with weak expected frequency
  - Can be computed efficiently (see following slides)
- Weaknesses
  - No correction for the bias observed with self-overlapping words
- The binomial distribution indicates the probability to observe exactly C_w matches

$$P(X = C_w) = \frac{T!}{C_w!(T - C_w)!} p^{C_w} (1 - p)^{T - C_w} = C_T^{C_w} p^{C_w} (1 - p)^{T - C_w}$$

Over- and under-representation

- One usually wants to measure the degree of over- or under-representation of a word, in which cases one measures the probability associated to the left or right tail of the distribution.
  - The degree of **over-representation** is reflected by the probability to observe **at least** C_w
    (the lower is the probability, the higher the over-representation)
  - The degree of **under-representation** is reflected by the probability to observe **at most** C_w
    (the lower is the probability, the higher the over-representation)

$$\boxed{P(X \geq C_w) = \sum_{i=C_w}^{T} \frac{T!}{i!(T-i)!} p^i (1-p)^{T-i}}$$

$$\boxed{P(X \leq C_w) = \sum_{i=0}^{C_w} \frac{T!}{i!(T-i)!} p^i (1-p)^{T-i}}$$

Binomial : efficient computation

- The binomial probability can be computed efficiently by using a recursive formula.
- This dramatically reduces the computation time.

$$
\boxed{\begin{aligned}
&P(X=0)=\left(1-p\right)^{T}\\
&P(X=s+1)=P\left(X=s\right)\frac{p\left(T-s\right)}{\left(1-p\right)\left(s+1\right)}
\end{aligned}}
$$

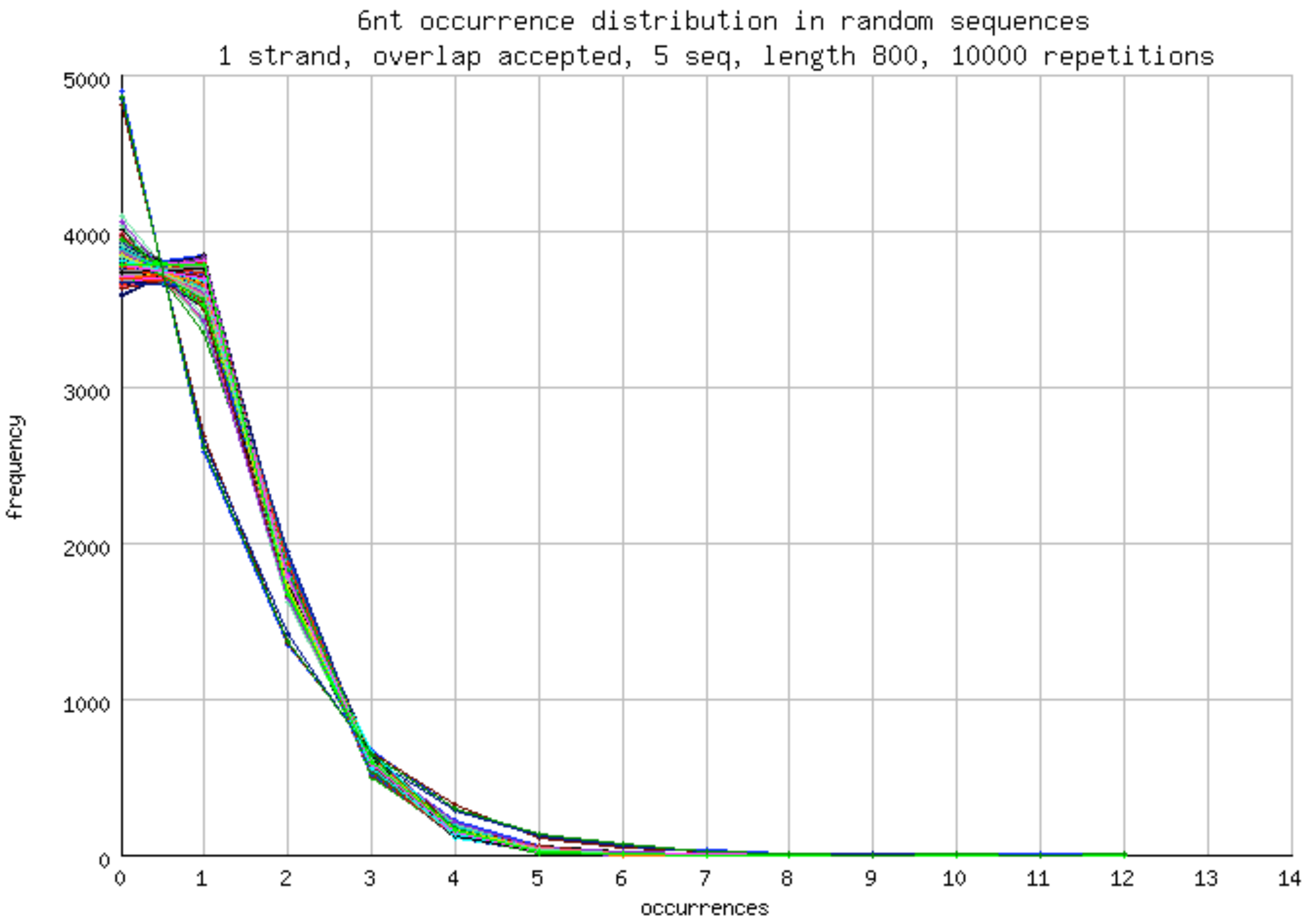
Validation of the statistical model : analysis of random sequences

- Allows to validate the choice of a statistical model
- In many publications, used in replacement of a statistical model
- The random sequence itself can be generated according to different models (independently distributed nucleotides, Markov chains, background frequencies)
- Warning: be sure to generate enough random sequences to have a chance to detect discrepancies

Binomial : Conditions of Applicability

- The binomial distribution relies on an assumption of independence of the successive trials.
- It is important to notice that this condition is never fulfilled in biological sequences !
    - non periodic words: each match prevents the k-1 following positions to include a match
    - periodic words: each match increases dramatically the probability to observe a match at the next period
- How sensitive is this effect ?

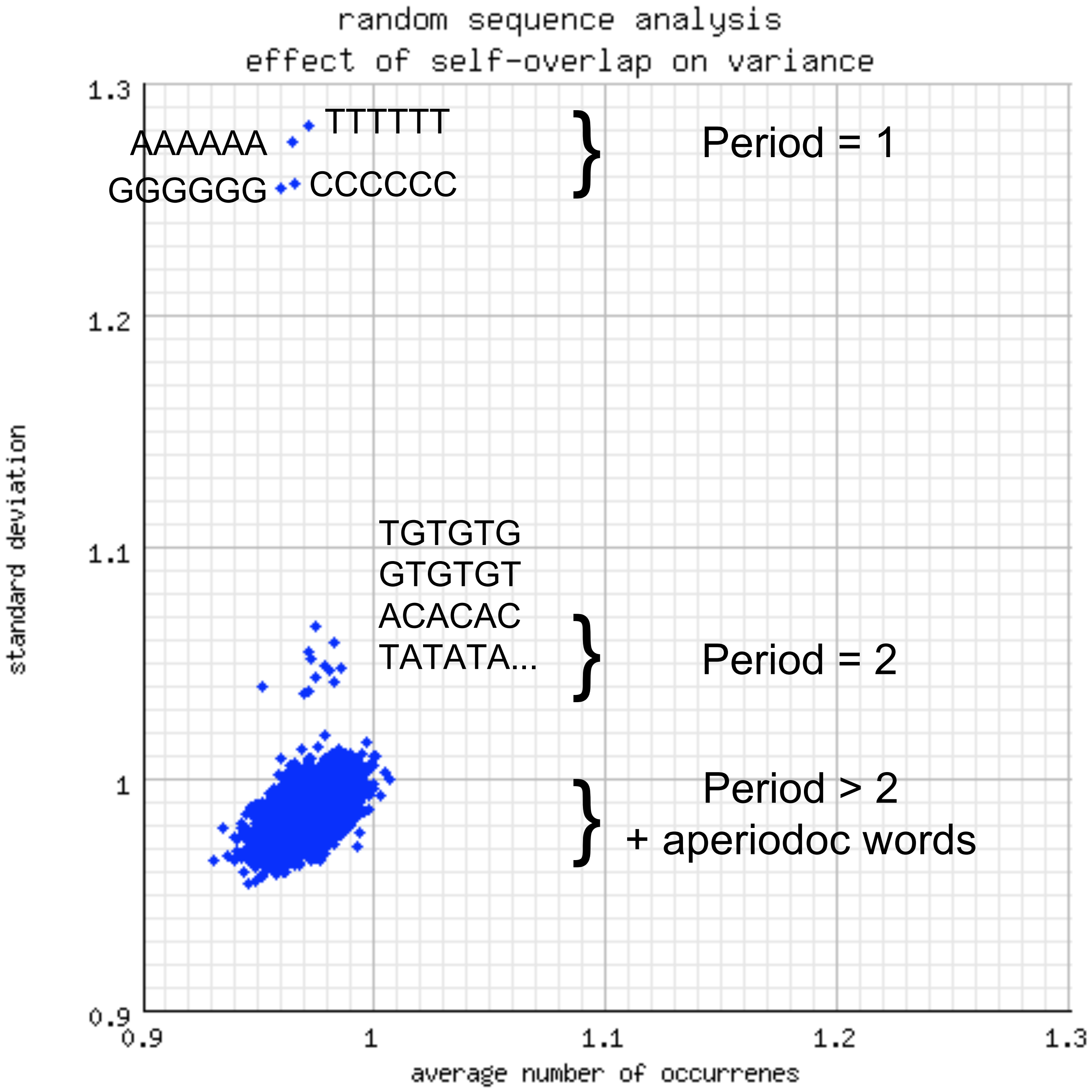
Deviation from binomial distribution for self-overlapping words



Random expectation

- The “outlier” curves seem “flatter” than the other ones, they have a larger variance than expected.
- We can measure, for each word separately, the mean and the variance of the occurrences observed in the different trials.
- Random expectation
  - According to the Poisson law: $\sigma^2=Tp= E_w$
  - According to the binomial law:
    - $\sigma^2=Tpq = Tp(1-p)=E_w(1-p)$
  - For large words (e.g. hexanucleotides: k=6) the Poisson is a valid approximation of the binomial, and the variances tend to the same value.
    - $p \sim 4^{-k} << 1 \rightarrow \sigma^2=Tp(1-p) \sim Tp$

10,000 trials

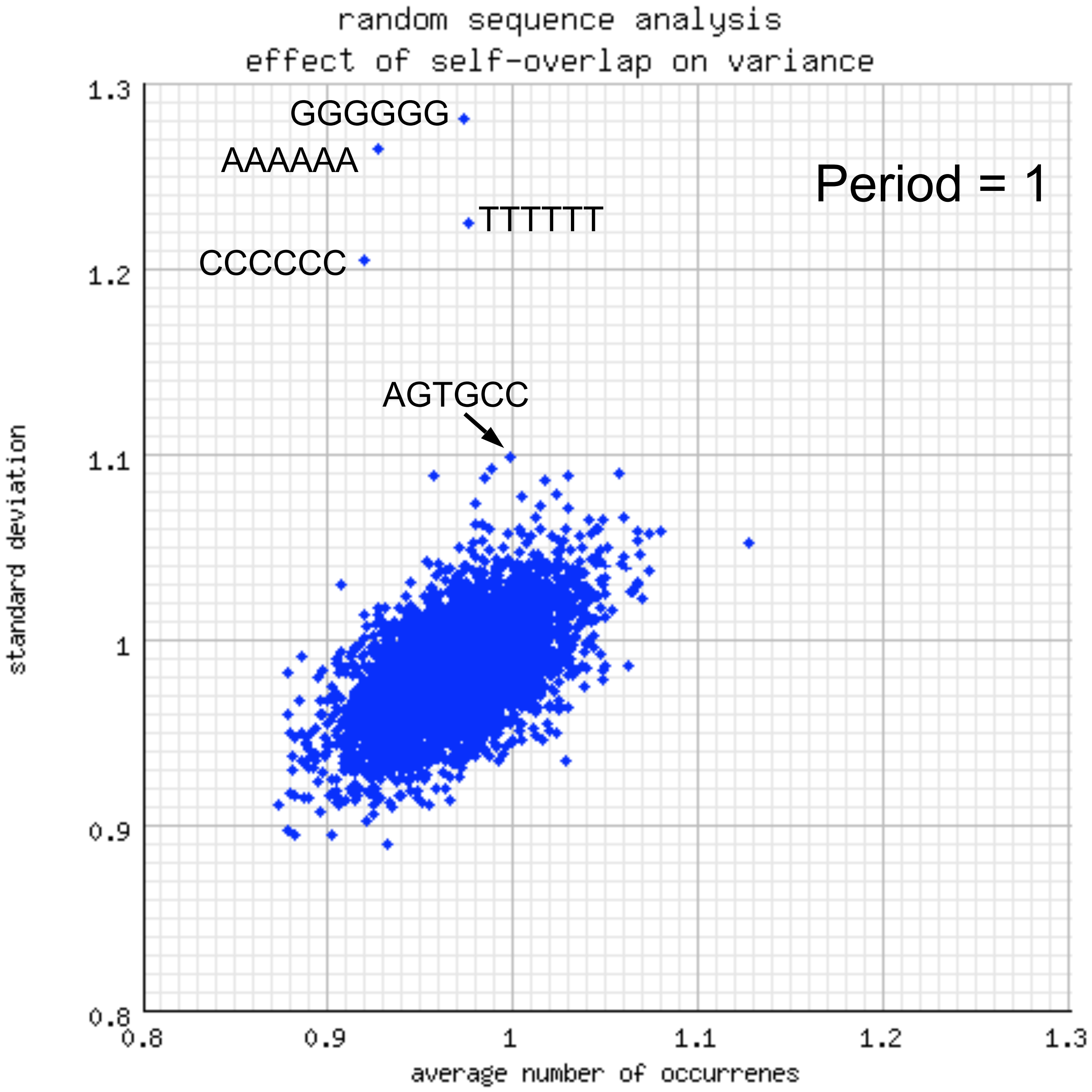


} Period = 1

} Period = 2

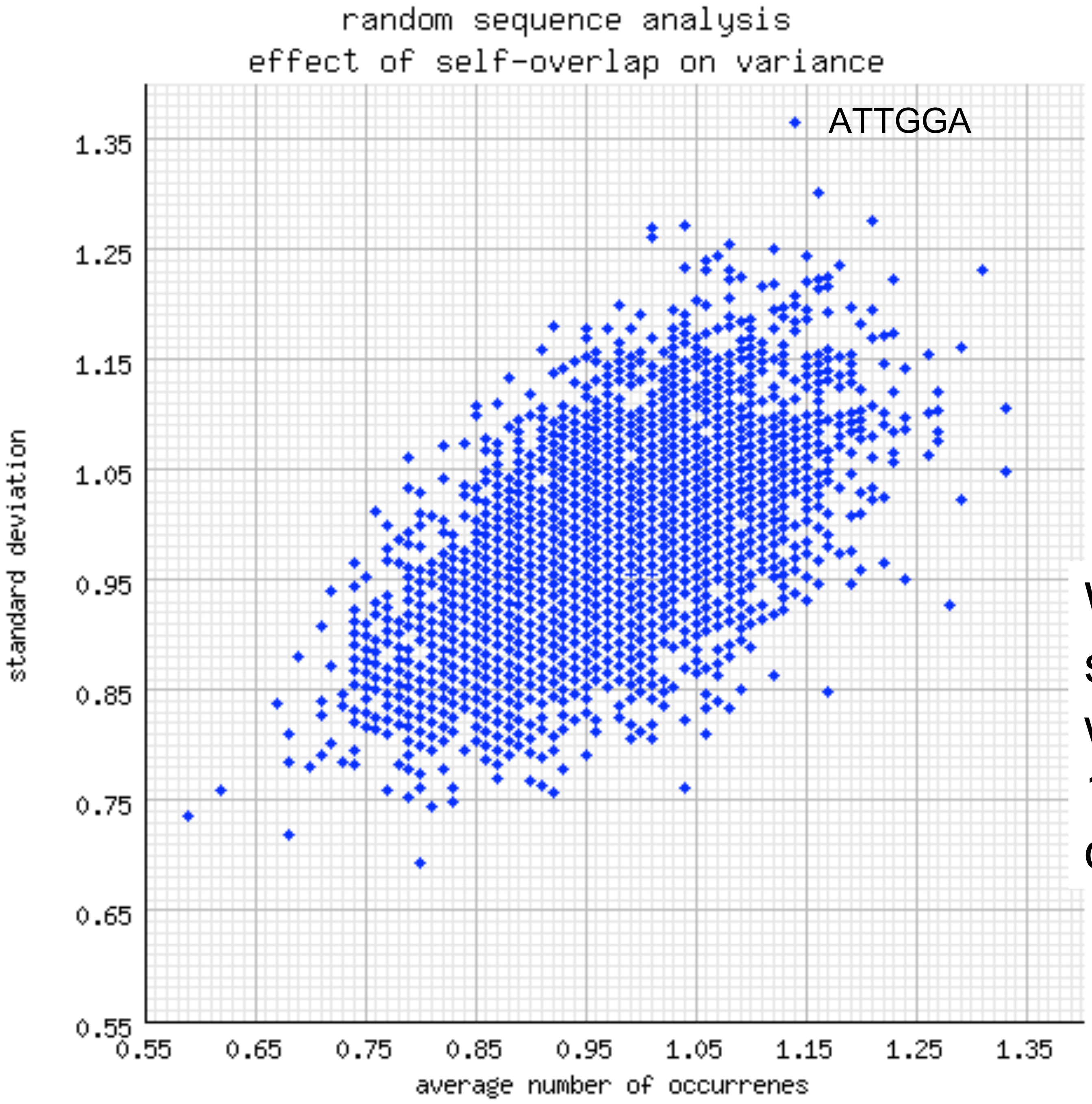
} Period > 2
+ aperiodoc words

With 10,000 random sequences, the effect of periods 1 and 2 are visible, but not of higher periods (3,4,5)



1,000 trials

With 1,000 random sequences, the effect of period 2 is overlooked



100 trials

With 100 random sequences, even words with period 1 are not discriminated

Scoring scheme - Poisson

$$P(X = C_w) = \frac{e^{-E_w} E_w^{\,C_w}}{C_w!}$$

- Advantages
  - A good approximation of the binomial
- Weaknesses
  - no correction for self-overlapping patterns

Poisson - efficient computation

- The Poisson probability can be calculated efficiently with a recursive formula

$$P(X = 0) = e^{-E_W}$$

$$P(X = C_w + 1) = P(X = C_w)\frac{E_w}{(z+1)}$$

Scoring scheme
Poisson : conditions of applicability

- The Poisson approximates the binomial when the following conditions are fulfilled:

$P_W \rightarrow 0$; $T \rightarrow \infty$; $E_w = P_w$ x $T \rightarrow m$ *finite*

E_w *small (generally* $E_w < 5$*)*

Scoring scheme - Z-score

- $Z = (C_w - E_w)/s_w$
- The P-value can then be obtained from the normal distribution
- Advantages
  - correction for self-overlapping patterns can be introduced by including a self-overlap coefficient in the estimation of the variance
- Weaknesses
  - Only applies to large sequence sets ($T \rightarrow \infty$)
  - Even for large sequence sets, the Gaussian distribution is a good estimation for the binomial, but only in the centre of the distribution
  - Particularly bad estimation for under-represented words (Mathias Vandenbogaert, PhD thesis).

Z-score : estimation of the variance

- For **non-periodic** words, the variance is estimated according to the Poisson distribution.
  - Poisson distribution:the variance equals the mean.
- For **periodic words**, one can incorporate a self-overlap coefficient.

$$\boxed{\hat{\sigma}^2 = m}$$

Self-overlap coefficient

- Pevzner et al.(1989). J. Biomol. Struct & Dynamics 5:1013-1026 proposed to correct the variance estimate by an overlap coefficient (K_{ov}).
- Where
  - k the word length
  - i the position index, comprised between 0 and k
  - v_i takes value 1 if there is an overlap starting at position i of the word, 0 otherwise.
- The variance is then estimated as

$$\boxed{K_{ov} = \sum_{i=1}^{k} v_i 4^{-i}}$$

$$\boxed{\hat{\sigma}^2 = C_W\left[2K_{ov} - 1 - (2k+1)P(W)\right]}$$

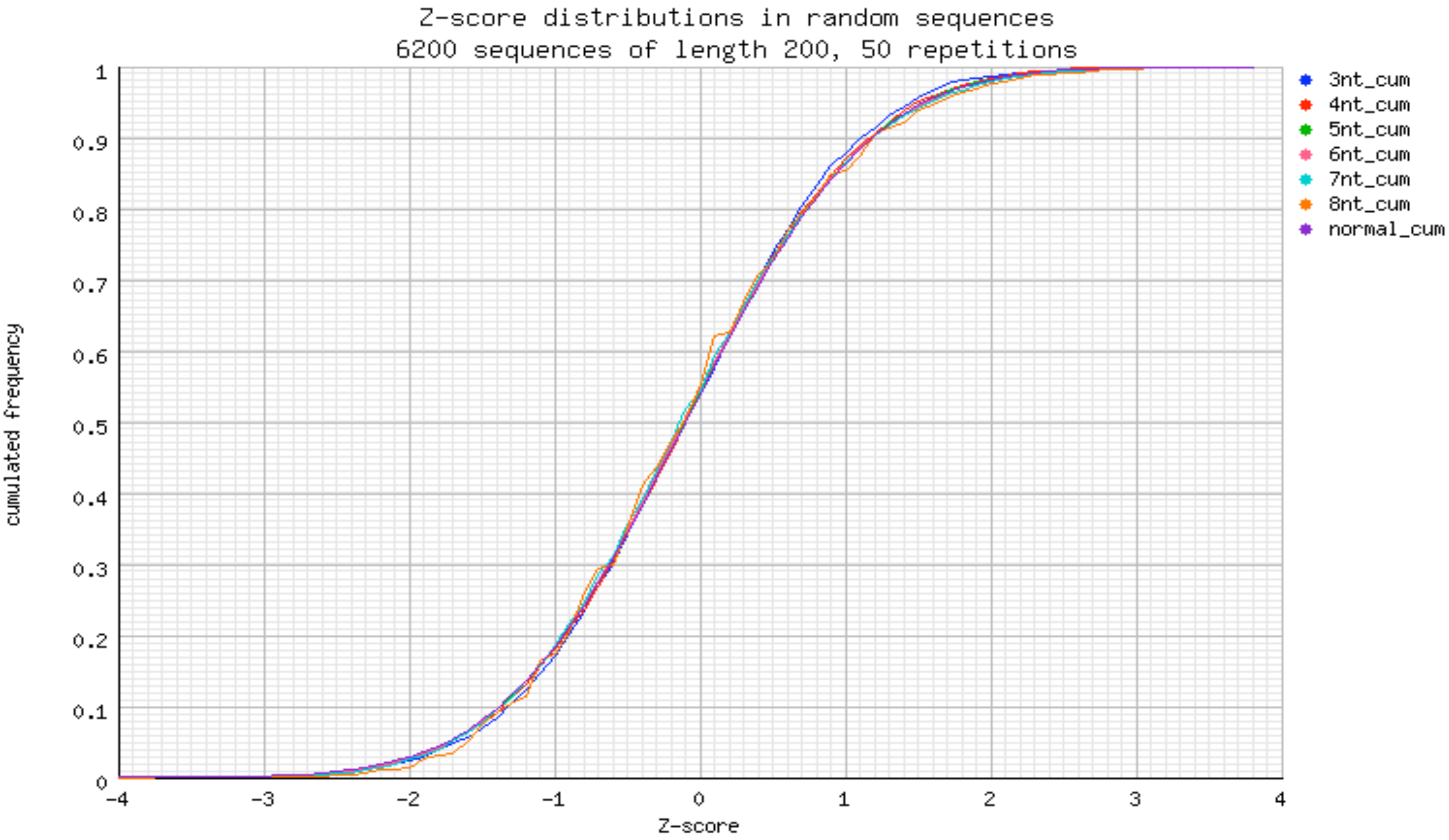
Scoring scheme
Z-score: conditions of applicability

$$T \rightarrow \infty$$

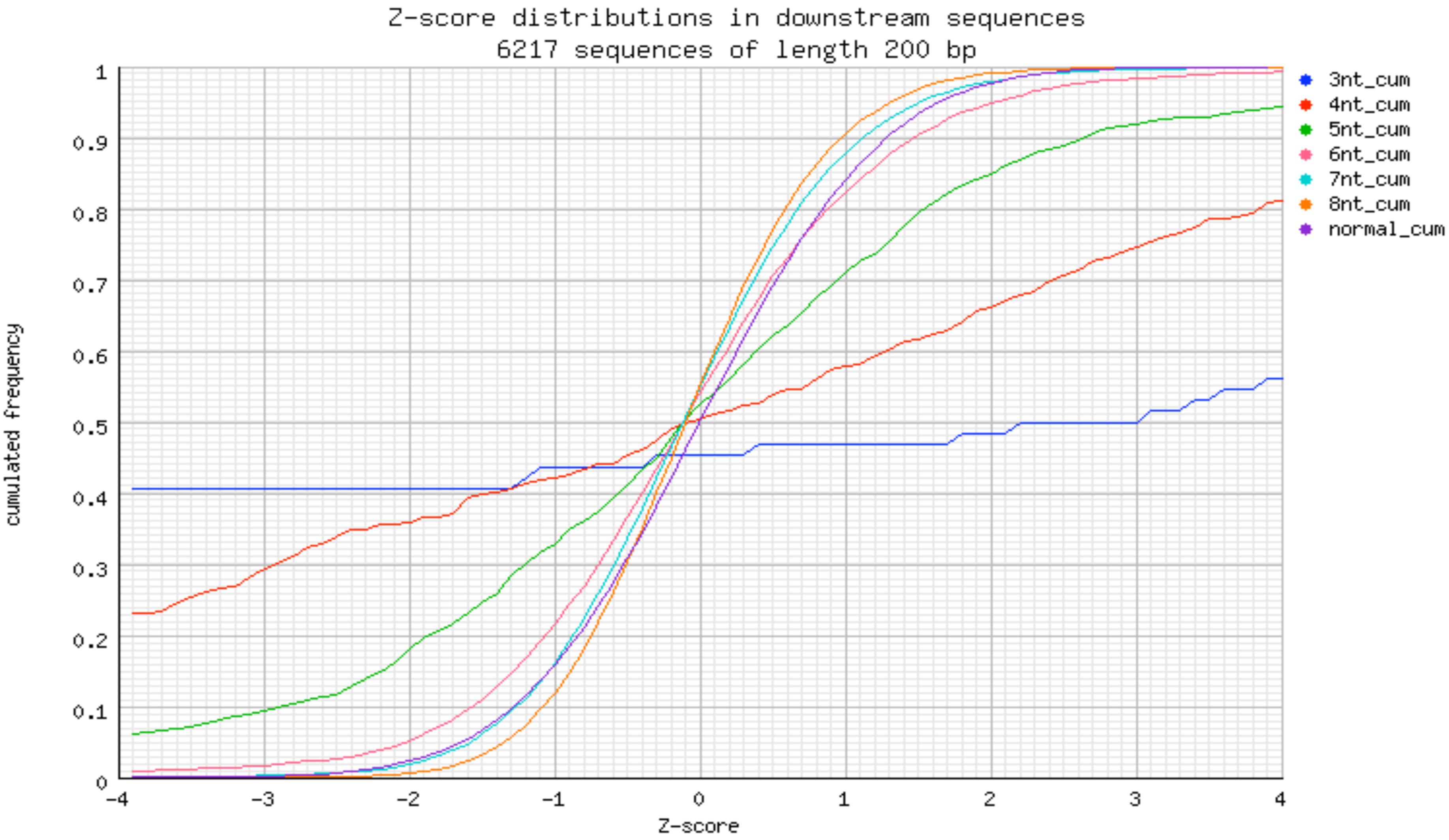
$$E_w >> 1$$

- The significance of a z-score can be calculated with the normal distribution.
- However, this makes strong assumptions on the model:
  - The expected number of occurrences should be sufficiently large (>> 1).
    - This condition is satisfied for very large sequences (e.g. genome-wise analyses), but **not** for small gene families (like the PHO example). Indeed, the expected number of occurrences is smaller than 1 for several words!
  - The normal distribution is symmetric. However, we are generally interested by rare events, and their distribution is bound by 0 (one cannot observe less than 0 occurrences).
- The normal distribution is thus **not** appropriate
  - for small sequence sets
    (example: regulons, groups of co-expressed genes)
  - to calculate the significance of under-representation.

Normality of word population in synthetic sequences random sequences generated with a Bernoulli process



Normality of word population in biological sequences
All yeast downstream sequences



Regulatory sequence analysis

Choice of a significance threshold

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Choice of the threshold

- The scoring scheme allows you to sort the words, then how many of them are significant ?
- You need a p-value (binomial, Z-score), then

  G = 1/M

  where G = threshold, M = number of possible words
- M depends on word length, alphabet, counting method

  Example:

  k=6, 1 strand $M = 4^6 = 4096$ $G = 0.00024$

  k=6, 2 strands $M = (4^6 + 4^3)/2 = 2080$ $G = 0.00048$

Significance index

$$E\text{-}value = P(C_W >= z)\ T$$

$$sig = -\log_{10}(E\text{-}value)$$

Where

T is the number of tested words

- Takes into consideration the dependency of the threshold on word length
- Provides an intuitive perception of the level of over-representation

sig > 0	1 such word at random in each sequence set
sig > 1	1 such word expected every 10 sequence sets
sig > 2	1 such word expected every 100 sequence sets
...	
sig > s	1 such word expected every 10s sequence sets

Choice of the threshold

- Enables your program to return a negative answer llike "there are no over-represented patterns in your sequence set".
- Applicable ONLY if the population behaves normally !!! (not the case for very small words).

Choice of the word length

- How to choose the length of the words to be analyzed ? The answer of course depends on the biological signals to be detected.
- The choice of the appropriate length can be crucial
  - Too small: discrepancy from the normal distribution
  - Too large: loss of significance
- Example
  - In yeast, a size of 6 is generally satisfactory for detecting the core of most transcriptional regulation sites in DNA. This is related to the structure of the protein domains that bind DNA (HLH, zinc fingers, homeodomains, …)

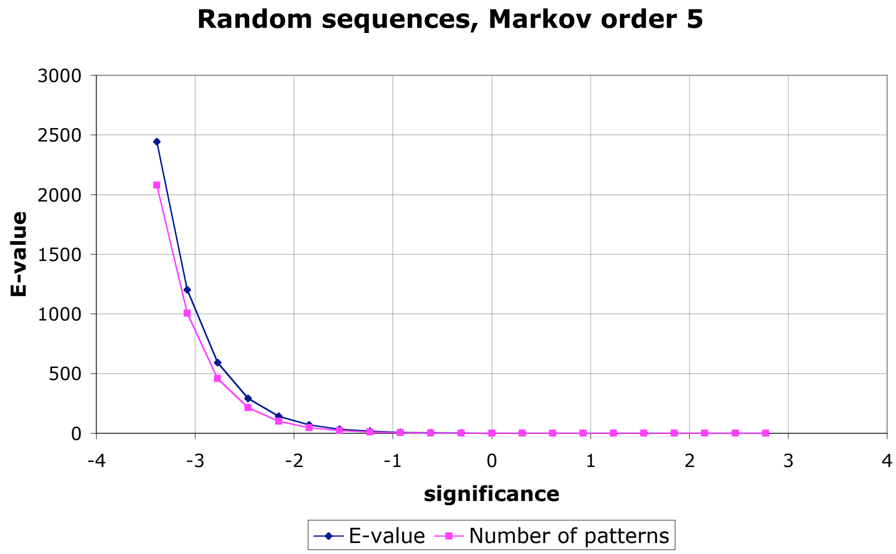
Regulatory Sequence Analysis

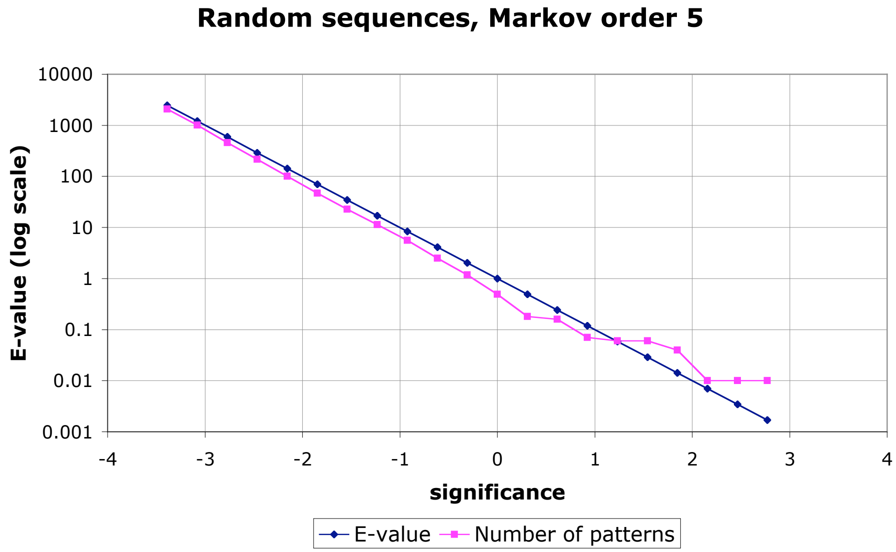
Checking background models and scoring statistics with random sequences

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

A trivial negative control: rate of false positives in random sequences

- A first control is to measure the number of significant patterns in a set of random sequences.
  - Random sequences can be generated according to different background models (Bernoulli or Markov chains of various orders).
- Test
  - Measure the number of patterns above each score value (empirical E-value).
  - Compare it with the theoretical E-value (calculated with the binomial)
- Weakness
  - This test mainly checks the appropriateness of the scoring statistics for the chosen background model.
  - It does not tell much about the behaviour of the program with real biological sequences.


Random sequences, Markov order 5


Random sequences, Markov order 5