41. Poisson approximation of the distribution of the number of repeats in biological sequences modelled by Markov chains

Touyar Narjiss [1] **Dominique Cellier** [2] **Sophie Schbath** [3] **Helene Dauchel** [1]

Abstract: The aim of this work is to approximate the distribution of the number of repeats in biological sequences modelled by Markov chains. Because of the inaccessibility of this distribution, we approximate it thanks to the Chen-Stein method using the Poisson distribution.
The goal of the application is to find the statistical significance of the repeats in the genome of a biological species.
Keywords: repeated sequences, Chen-Stein method, Poisson approximation.

1 Introduction

Genomes are dynamic and redundant structures: over a person's lifetime or over generations, genomes are regularly subject to mutations, deletion, duplications and inversions. The phenomenon of sequence repeats can concern the genes, or the extragenic portions of the genome. It is therefore necessary to study the repeats in order to understand better genome structures and how they evolve. All the algorithmic methods dedicated to the detection of repeats must be accompanied by a suitable mathematical analysis providing a statistical interpretation of the detected repeats (count, length and distribution). This analysis must make it possible to distinguish the significant repeats from those simply imputed randomly and so direct the biological analyses.
The statistical study presented here is based on the modelling of the sequences by a succession of random variables in a Markov chain which belong to a finite alphabet.

2 Model and notations

We model the sequence $S = X_1 \cdots X_n$ of length n by a stationary Markov chain of order one on the alphabet $\mathcal{A} = \{A, C, G, T\}$ with a transition matrix $\Pi = (\pi(a,b))_{a,b\in\mathcal{A}}$.
We define $\rho = \max_{a,b} \pi(a,b)$.
Definitions
The sequence has a repeat of length t at positions i and j if and only if the word of t letters starting at i is identical to the one starting at j, that is $X_i \cdots X_{i+t-1} = X_j \cdots X_{j+t-1}$.
We will say that a repeat starts at (i,j) if there is a repeat at (i,j) but not at $(i-1, j-1)$. That is $X_{i-1} \neq X_{j-1}$ and $X_i \cdots X_{i+t-1} = X_j \cdots X_{j+t-1}$.
We will be only interested here in the case of non-self-overlapping repeats, that means those relating to two disjoined occurrences in the sequence S.

[1] Atelier Biologie Informatique Statistiques et Sociolinguistique (ABISS), UMR CNRS 6037, 76821 Mont Saint Aignan Cedex France. E-mail: {Narjiss.Touyar,Helene.Dauchel}@univ-rouen.fr
[2] Laboratoire de Mathématiques Raphaël Salem (LMRS), UMR CNRS 6085, Université de Rouen, France. E-mail: Dominique.Cellier@univ-rouen.fr
[3] INRA, Unité Mathématique Informatique et Génome (MIG), 78352 Jouy-en-Josas, France. E-mail: Sophie.Schbath@jouy.inra.fr

By denoting I the index set of the possible positions $\alpha = (i, j)$, we have

$$I = \{\alpha = (i,j) \ / \ 1 \leq i < i + t - 1 < j \leq n - t + 1\}.$$

We will neglect the edge effect, e.g. when $i = 1$.
For $\alpha = (i, j) \in I$, we consider the random indicator function that a repeat starts at α

$$Y_\alpha \equiv Y_{(i,j)} = \mathbf{1}\{X_{i-1} \neq X_{j-1}, X_i \cdots X_{i+t-1} = X_j \cdots X_{j+t-1}\}.$$

Finally, we define the random variable W counting the number of repeats of length t by $W = \sum_{\alpha \in I} Y_\alpha$ and let λ be its expectation : $\lambda = \mathbb{E}(W)$.
The study of the statistical significance of the repeats relies on the possibility of evaluating the distribution of the random variable W in the Markov chain model M_1. Unfortunately, this probability distribution is not available and approximation techniques are then used.

3 Result

Proposition *If $n \to +\infty$ and if $t = O(\log_{\frac{1}{\rho}} n)$ then the distribution of W can be approximated by a Poisson distribution with mean λ. Consequently the p-value associated to the existence of at least one repeat of lenght t is given by :* $p = \mathbb{P}(W \geq 1) \approx 1 - e^{-\lambda}$.
This proposition is obtained by getting a bound of the total variation distance between the distribution of W and the Poisson distribution with mean λ.(the Chen-Stein method)
The calculation of this bound relies on both judicious choice of the neighborhoods system B_α and calculation of upper bounds of b_1, b_2 and b_3 that tend to zero on the condition that: $t = O(log_{\frac{1}{\rho}} n)$.
Choice of the neighborhood : We define B_α by the following relation
$\beta \in B_\alpha \Leftrightarrow min(|i - i'|, |j - i'|, |i - j'|, |j - j'|) < 2t$.
Calculation of the bounds on the error : Concerning b_1, we show that : $b_1 = O(\lambda^2 \frac{t}{n})$.
On the condition that $t = O(\log_{\frac{1}{\rho}} n)$, the bound $b_1 \to 0$ when : $n \to +\infty$.
For the calculation of b_2, for instance, we need to distinguish several cases depending on the overlaps between the repeats starting at α and $\beta \in B_\alpha$.
The calculation of b_3 is based on the Markov property.
Limiting Poisson distribution : The parameter $\lambda = \sum_{\alpha \in I} p_\alpha$ is not easily computable. We have however the following upper bound $\lambda^* = \frac{3}{2}(n - 2t - 1)(n - 2t + 2)\rho^t$.
Consequently $p \leq 1 - e^{-\lambda^*}$.

4 Application

The results obtained are now used for the statistical study of the repeats detected over an entire genome, in order to establish the common or distinctive characteristics between the chromosomes of the same organism (intra-genomic comparison) or between the chromosomes of two different organisms (inter-genomic comparison). This statistical study will be implemented and associated to the programs ForRepeat and EvoRep developed within the ABISS laboratory.

References

[1] Arratia, R., Goldstein, L. and Gordon, L. 1998. Two moments suffice for Poisson approximation: the Chen-Stein method, *Ann.Prob.*17, pp. 9–25.

[2] Arratia, R., Martin, D., Reinert, G. and Waterman, M. 1996. Poisson process approximation for sequence repeat and sequencing by hybridization, *J.Comp.biol.* 3(3), pp. 425–463.