Journal of Discrete Algorithms 24 (2014) 2–11

Contents lists available at ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Large deviation properties for patterns

Jérémie Bourdon [a,b,*], Mireille Régnier [c]

[a] LINA, CNRS UMR 6241, Université de Nantes, France
[b] DYLISS-Inria team, Inria Rennes-Bretagne-Atlantique, France
[c] AMIB-Inria team, LIX-Ecole Polytechnique, 91128 Palaiseau, France

ARTICLE INFO

Article history:
Available online 28 September 2013

Keywords:
Pattern matching
Statistics
Large deviation

ABSTRACT

Deciding whether a given pattern is over- or under-represented according to a given background model is a key question in computational biology. Such a decision is usually made by computing some p-values reflecting the "exceptionality" of a pattern in a given sequence or set of sequences. In the simplest cases (short and simple patterns, simple background model, small number of sequences), an exact p-value can be computed with a tractable complexity. The realistic cases are in general too complicated to get such an exact p-value. Approximations are thus proposed (Gaussian, Poisson, Large deviation approximations). These approximations are applicable under some conditions: Gaussian approximations are valid in the central domain while Poisson and Large deviation approximations are valid for rare events. In the present paper, we prove a large deviation approximation to the double strands counting problem that refers to a counting of a given pattern in a set of sequences that arise from both strands of the genome. In that case, dependencies between a sequence and its reverse complement cannot be neglected. They are captured here for a Bernoulli model from general combinatorial properties of the pattern. A large deviation result is also provided for a set of small sequences.

1. Introduction

Counting random words [17] and calculating probabilities is an old and extensively studied problem in theoretical computer science for various applications in bioinformatics including finding motifs [18,7] and calculating p-values [19]. Rare or over-represented words are commonly assumed to be related to biological functions in some genomes.

An exact derivation of pattern occurrences is theoretically solved [17,8]. Nevertheless, despite improvements from Markov chains embedding and automata approaches [5,10], its computation is expensive, not to mention accuracy and numerical issues. Classical approximations of the distribution (Gaussian, Poisson) are known to be inapplicable [6]. Indeed, distribution convergence is valid in the central domain only, while rare events should be studied in the *tail* domain. These validity limits for asymptotic results are pointed out in [6] by numerous simulations on biological data and randomly generated data. A theoretical interpretation of their observations is extensively discussed in [16].

Rare events of a random process are described by the tail of the associated distribution. One typical result in this case is a mathematical proof of a so-called *Large Deviation principle* that consists of providing a big-O bound on the tail distribution. In this paper, we prove that *combinatorial* properties of words can be taken further by providing an explicit and tractable formula for the tail distribution. A survey can be found in [17] in the context of word occurrences. Still, very few results are known.

* Corresponding author.
E-mail addresses: Jeremie.Bourdon@univ-nantes.fr (J. Bourdon), mireille.regnier@inria.fr (M. Régnier).

1570-8667/$ – see front matter
http://dx.doi.org/10.1016/j.jda.2013.09.004

One may also cite [12] and [9]. The latter compares several methods for proving the exceptionality of a single pattern including Gaussian approximation, compound Poisson approximation [15] and large deviation approximation.

Recently, new sequencing methods seem to require pushing away our knowledge on pattern occurrence statistics in order to obtain more precise results while keeping the complexity of the computations under control. Our goal is to obtain some probabilistic results adapted to datasets containing a huge number of sequences (typically reads from a high throughput sequencing) possibly coming from a pair-end sequencing. In this case, patterns and their complements counterparts have to be considered at the same time (in the sequel, this case is referred as the "double strands" counting problem). The main problem here is properly taking into account the dependencies between a DNA sequence and its complement.

We consider here two cases. Large sequences are addressed in Section 3 when one counts occurrences of words from a finite set $\mathcal{H}$, under a Bernoulli model. Large deviation results were obtained for a single word in [12]. The case where $\mathcal{H}$ admits two overlap classes [14] is solved here. This case is fundamental, as it allows to address *double strand* counting. Short sequences are addressed in Section 4. Large deviation results have been known for long for random independent trials in the case of *identical* distributions [2] and [21]. Non-identical distributions are considered here. In both cases, the formula are computable with a low space and time complexity and their tightness is guaranteed.

2. Preliminaries

In this section, we present the basic definitions and framework necessary to understand our work. We make use of an equivalence relation on $\mathcal{H}$, defined in [14], that is *stable* for prefix and suffix relations.

Definition 1. Given a set $\mathcal{H}$, the *overlap set* is the set of words that are prefix and suffix of two (possibly equal) words in $\mathcal{H}$. It is denoted $\mathcal{OV}(\mathcal{H})$.

Two words f and g are said to be *overlap equivalent* iff

$$\max_{w\in\mathcal{OV}(\mathcal{H})}\{w \text{ is a prefix of } f\} = \max_{w\in\mathcal{OV}(\mathcal{H})}\{w \text{ is a prefix of } g\}, \tag{1}$$

$$\max_{w\in\mathcal{OV}(\mathcal{H})}\{w \text{ is a suffix of } f\} = \max_{w\in\mathcal{OV}(\mathcal{H})}\{w \text{ is a suffix of } g\}. \tag{2}$$

Definition 2. A set $\mathcal{H}$ is called a q-set if the quotient set admits exactly q classes.

The overlap set is computable with linear space and time complexity. First, a trie is built with Aho–Corasick algorithm. Second, this trie is traversed from the leaves to the root, following suffix links. The words of $\mathcal{OV}(\mathcal{H})$ are the paths of visited nodes.

A prefix (resp. a suffix) of a member g of a class G is a prefix (resp. a suffix) of any member of that class.

Definition 3. Let w and g be two words such that w is a prefix of g. Let $e(w,g)$ denote the word that satisfies

$$g = w\cdot e(w,g).$$

Let G be an overlap class. Given a word w in $\mathcal{OV}(\mathcal{H})$ that is a prefix of a member of G, let $e(w,\mathrm{G})$ denote the set of words defined as

$$e(w,\mathrm{G}) = \bigcup_{g\in\mathrm{G}}\{e(w,g)\}.$$

Given two overlap classes F and G, one denotes $OV(\mathrm{F},\mathrm{G})$ the set of words that are the suffix of any word in F and the prefix of any word in G.

A probability model on words steadily extends to overlap classes by a summation of the probabilities of its members. This allows an extension to overlap classes of classical generating functions for words [17]. This extension is given below for 2-sets, assuming a Bernoulli distribution.

Definition 4. Given an overlap class F, one denotes $H_{\mathrm{F}}(z)$ being the *probability generating series*

$$H_{\mathrm{F}}(z) = \sum_{f\in\mathrm{F}} Prob(f)z^{|f|}, \tag{3}$$

where $Prob(f)$ is the probability of the word f.

Given two overlap classes F and G, the *probability matrix* $\mathbb{H}(z)$ is defined as

$$\mathbb{H}(z) = \begin{pmatrix} H_{\mathrm{F}}(z) & H_{\mathrm{G}}(z) \\ H_{\mathrm{F}}(z) & H_{\mathrm{G}}(z) \end{pmatrix}. \tag{4}$$