A Fast and Simple Method for Mining Subsequences with Surprising Event Counts

Jefrey Lijffijt

Helsinki Institute for Information Technology HIIT
Department of Information and Computer Science
Aalto University, Finland
jefrey.lijffijt@aalto.fi

Abstract. We consider the problem of mining subsequences with surprising event counts. When mining patterns, we often test a very large number of potentially present patterns, leading to a high likelihood of finding *spurious* results. Typically, this problem grows as the size of the data increases. Existing methods for statistical testing are not usable for mining patterns in *big data*, because they are either computationally too demanding, or fail to take into account the dependency structure between patterns, leading to true findings going unnoticed. We propose a new method to compute the significance of event frequencies in subsequences of a long data sequence. The method is based on analyzing the joint distribution of the patterns, omitting the need for randomization. We argue that computing the p-values exactly is computationally costly, but that an upper bound is easy to compute. We investigate the tightness of the upper bound and compare the power of the test with the alternative of post-hoc correction. We demonstrate the utility of the method on two types of data: text and DNA. We show that the proposed method is easy to implement and can be computed quickly. Moreover, we conclude that the upper bound is sufficiently tight and that meaningful results can be obtained in practice.

Keywords: Big data, pattern mining, multiple hypothesis testing, event sequence, frequency of occurrence

1 Introduction

The amount of collected data is growing rapidly. As a result, the focus in data mining research is more than ever on faster and simpler methods, where fast currently means linear or sublinear in the size of the data. However, *big data* presents more challenges. For example, when mining *patterns*—local structure, as opposed to global structure [15]—the number of patterns potentially present in the data is often exponential in the size of the data. Testing more patterns is nice, because it increases the likelihood of finding interesting results. However, testing more patterns is also dangerous, as it increases the likelihood of finding *spurious* results, i.e., patterns caused by randomness.

Several methods have been developed in the past decade for testing the statistical significance of various types of patterns, and a few studies investigated post-hoc corrections to avoid finding many spurious patterns. Unfortunately, none of the proposed methods is usable for big data, because they rely either on randomization, Bonferroni-style post-hoc correction, or both.

Randomization testing is computationally expensive; a single randomization has a computational cost linear in the size of the data or higher, and thousands or millions of randomizations may be required for sufficient resolution. Bonferroni-style post-hoc correction is also problematic, because the studied patterns (which each correspond to a hypothesis test) are typically dependent, in which case the p-values become conservative, i.e., many true findings will go unnoticed. The problem is worse for large data, as the conservativeness depends on the number of patterns, which may be exponential in the size of the data [6].

We propose a new method for mining subsequences with surprising event counts that does not suffer from these problems. We formulate a statistical test that includes a correction for testing multiple hypotheses, i.e., the p-value for an observation will depend on the observation itself, as well as on the size of the data. This allows us to avoid using a conservative post-hoc correction. Although the method is not directly applicable to other data or pattern types, it may act as a model for methods on other data.

The method provides strong control over the *family-wise error rate* (FWER), that is, the probability that any of the significant results is a false positive. Put less formally, we ask the question "what is the probability that *any of the considered patterns* would have a statistic equal to or higher than the observed statistic?", where the *statistic* can be any interestingness measure: support, lift, WRAcc, etc. We illustrate FWER control in the following example.

Assume that the interestingness measure, and thus the test statistic, is the support of a pattern, and that the data is a transaction database in tabular form. For simplicity assume that all items have equal support. The probability that the statistic of a specific pattern P is significantly high can be assessed by, for example, using swap randomization [5] to generate randomized samples[1] and then computing how often we observe a similar or higher statistic for pattern P in the randomized samples. The obtained p-value corresponds to the question "what is the probability that *this specific pattern* has a test statistic equal to or higher than the observed statistic?".

Now assume that we repeat this procedure for all itemsets of some fixed size. Because we are testing many hypotheses, we are liable to finding many small p-values. To prevent this, we can instead compare the observed statistic with the maximum observed statistic over all itemsets of that size in each randomization. In that case, the p-values correspond to the question "what is the probability that *any of the considered patterns* would have a statistic equal to or higher than the observed statistic?", which is the same as FWER control. Significance

[1] Which randomization method to use depends on the assumptions that one wants to make.

testing with FWER control using randomization for mining frequent itemsets has been studied extensively by Hanhijärvi [7].

As stated earlier, randomization is unpractical for large data, and the method proposed in this paper is based on computing the p-values analytically. This means that we have to analyze the joint distribution of the statistics of all potential patterns. We discuss a specific type of data and patterns. We show that, although exact p-values are computationally costly to obtain, an upper bound can be computed efficiently. We show empirically that the upper bound is sufficiently tight.

The data that we consider are event sequences, and the aim is to find subsequences of a fixed length where a certain event is significantly frequent or infrequent. This is essentially a *subgroup discovery* problem: the target is a specific event, the descriptions or patterns are subsequences, and the aim is to find all descriptions where the target is exceptionally frequent or infrequent. This problem setting has many applications. For example, biologists are interested in detecting *isochores* and *CpG sites* in DNA sequences, which are regions that are especially rich or poor in CG content and rich in the dinucleotide CpG respectively [2], and another example is that in text analysis it is useful to identify text fragments where a certain word is under or overused.

Summary of contributions. We propose a new method to test the significance of event frequencies in subsequences that provides p-values under control of the family-wise error rate. That is, the p-value corresponds to the probability of observing the observed statistic or higher in *any* of the subsequences of a given length in a single long sequence. We show that computing the p-values exactly is computationally costly, but that an upper bound can be computed fast. We investigate the tightness of the upper bound and compare the power of the test against using a generic post-hoc correction. We demonstrate the utility of the method by applying the method to two types of data: text and DNA. We show that the proposed method is easy to implement and can be computed quickly. Moreover, we conclude that the upper bound is sufficiently tight and that meaningful results can be obtained in practice.

Outline. The method is introduced in Section 2. Results from the experiments on the tightness of the upper bound, comparison with the generic post-hoc correction, and the experiments on the two data sets are presented in Section 3. Related work is discussed in Section 4 and conclusions are given in Section 5.

2 Method

2.1 Notation

Given a finite set of *event labels* L, an *event sequence* S is defined as $S = (s_1, \ldots, s_n), \forall i \in \{1, \ldots, n\} : s_i \in L$, where n is the *length* of the sequence. We denote a *subsequence* of S as $S_{i,m} = (s_i, \ldots, s_{i+m-1})$, where m is the *length* of the subsequence. The *count* of an event $a \in L$ in subsequence $S_{i,m}$ is given by $\sigma(S_{i,m}, a) = \sum_{k=i}^{i+m-1} \mathbf{1}_{\{a\}}(s_k)$, where $\mathbf{1}_A(s_k)$ is the indicator function that

equals 1 if $s_k \in A$ and 0 otherwise. The *frequency* of an event $a \in L$ in subsequence $S_{i,m}$ is $\zeta(S_{i,m}, a) = \sigma(S_{i,m}, a)/m$. The count and frequency of an event in a sequence S are defined as $\sigma(S, a) = \sigma(S_{1,n}, a)$ and $\zeta(S, a) = \zeta(S_{1,n}, a)$.

2.2 Background

Our aim is to test the hypothesis that an event is *significantly* frequent or infrequent in a given subsequence. To determine if an observed frequency is significant, we use the notion of *p-values*. Denote Z a random variable that represents the count of an event under the null hypothesis. The p-value for an observed count k is the probability of observing that count or higher, under the null hypothesis:

$$p_H = Pr(Z \geq k)$$

The observed count is *significantly high* if the probability of a observing that count or higher under the null hypothesis is less than or equal to the pre-specified threshold α:

$$p_H \leq \alpha$$

Vice versa, the observed count is *significantly low* if the probability of observing that count or lower is less than or equal to α:

$$p_L = Pr(Z \leq k) \leq \alpha$$

The null hypothesis that we are interested in is that the data has no structure, i.e., that all events in the sequence are i.i.d. samples:

Definition 1 (Null Hypothesis). *The null hypothesis is that the sequence is generated by a sequence of random variables $X_1, \ldots, X_n$, where each random variable X_i is defined by an independent Bernoulli distribution: $X_i \in \{0, 1\}$, and $Pr(X_i = 1) = p$.*

We assume that the parameter p, which represents the expected frequency of an event, is fixed. The parameter p can be, for example, estimated from the sequence S, in which case the method will find regions in the sequence where the event frequency is significantly high (or low) with respect to the rest of the sequence. Alternatively, p can be based on background knowledge, for example an estimate derived from a database of sequences.

Furthermore, we assume that we are going to test subsequences of a fixed length m, which is a parameter defined beforehand by the user, and we assume that the user chooses a priori the significance threshold α.

2.3 Computing P-Values when Testing One Subsequence

Given a sequence of independent random variables $X_1, \ldots, X_n$, each following a Bernoulli distribution with parameter p, define $Z_{i,m}$ as

$$Z_{i,m} = \sum_{j=i}^{i+m-1} X_j.$$

Because $Z_{i,m}$ is the sum of m independent and identically distributed Bernoulli variables, the probability distribution for $Z_{i,m}$ is a binomial distribution:

$$Pr(Z_{i,m} = k) = Bin\,(k; m, p) = \binom{m}{k} p^k (1-p)^{m-k}.$$

We find that, as expected, the distribution is independent of the location i.

We can now define the one-tailed p-value under the null hypothesis for a single subsequence at a random location. For the high frequency direction, the one-tailed p-value is given by

$$\begin{aligned} p_H &= Pr(\sigma(S_{i,m}, a) \geq k) \\ &= Pr(Z_{i,m} \geq k) \\ &= \sum_{j=k}^{m} \binom{m}{j} p^j (1-p)^{m-j}, \end{aligned} \tag{1}$$

while the one-tailed p-value in the low frequency direction is given by

$$\begin{aligned} p_L &= Pr(\sigma(S_{i,m}, a) \leq k) \\ &= Pr(Z_{i,m} \leq k) \\ &= \sum_{j=0}^{m} \binom{m}{j} p^j (1-p)^{m-j}. \end{aligned} \tag{2}$$

As can be seen, the p-values correspond to the cumulative distribution function of the binomial distribution. These tests are also known as the *binomial test*. Many statistical software packages contain a function for computing its value.

2.4 Computing P-Values when Testing All Subsequences

When testing a single subsequence at a random location, the probability of rejecting the null hypothesis while it is actually true—a *false positive* or *type I error*—is exactly α, and thus the result is easy to interpret. However, if we test the significance of the event frequency in multiple subsequences, or in a subsequence at an optimized location, we increase the probability of false positives.

Let us assume that we test the observed counts for all subsequences of a given length, using a sliding window with step size one. In that case, the probability under the null hypothesis of observing a certain count or higher in at least one subsequence of length m is

$$Pr(\bigcup_{i=1,\ldots,n-m+1} Z_{i,m} \geq k). \tag{3}$$

When we test the event frequency in all subsequences, it seems reasonable to use this probability as a p-value. This is also theoretically justified: the probability expressed in Eq. (3) is equal to the probability of obtaining at least one

false positive, thus, using this as the p-value corresponds to strong control of the family-wise error rate [18].

Thus, we redefine the one-tailed p-value, in the high direction, as

$$p_H = Pr(\bigcup_{i=1,\ldots,n-m+1} Z_{i,m} \geq k).$$

The p-value can be decomposed as

$$\begin{aligned} p_H = Pr(Z_{1,m} \geq k) + Pr(Z_{2,m} \geq k \cap \bigcap_{i=1} Z_{i,m} < k) + \ldots \\ + Pr(Z_{n-m+1,m} \geq k \cap \bigcap_{i=1,\ldots,n-m} Z_{i,m} < k), \end{aligned} \tag{4}$$

which highlights that the p-value equals the standard case (Eq. (1)) *plus* a correction term.

This correction term is in general difficult to compute exactly. A straightforward approach would be to define a column vector v with a probability for each possible initial state, and a transition matrix W that specifies the transition probabilities between the states, and use one sink state for all subsequences with at least k ones. Then the exact p-value is given by computing $W^{n-m} \cdot v$. However, the matrix W will have $O(2^{2m})$ entries, so this approach works only when the length of the subsequences, m, is very small.

The main result of this paper is that we can instead obtain an upper bound that is very easy to compute. Let us define the following approximation:

$$\tilde{p}_H = Pr(Z_{1,m} \geq k) + (n-m) \cdot Pr(Z_{2,m} \geq k \cap Z_{1,m} < k).$$

Theorem 1. *$\tilde{p}_H$ is an upper bound on the exact p-value p_H, i.e., $\tilde{p}_H \geq p_H$.*

Proof. Notice that for the correction terms of p_H it holds that

$$\begin{aligned} Pr(Z_{2,m} \geq k \cap \bigcap_{i=1} Z_{i,m} < k) &\geq Pr(Z_{3,m} \geq k \cap \bigcap_{i=1,2} Z_{i,m} < k) \\ &\geq Pr(Z_{4,m} \geq k \cap \bigcap_{i=1,2,3} Z_{i,m} < k) \\ &\geq \ldots \\ &\geq Pr(Z_{n-m+1,m} \geq k \cap \bigcap_{i=1,\ldots,n-m} Z_{i,m} < k). \end{aligned} \tag{5}$$

Combining Eqs. (4) and (5) gives

$$\begin{aligned} p_H &= Pr(Z_{1,m} \geq k) + Pr(Z_{2,m} \geq k \cap \bigcap_{i=1} Z_{i,m} < k) + \ldots \\ &\quad + Pr(Z_{n-m+1,m} \geq k \cap \bigcap_{i=1,\ldots,n-m} Z_{i,m} < k) \\ &\leq Pr(Z_{1,m} \geq k) + (n-m) \cdot Pr(Z_{2,m} \geq k \cap Z_{1,m} < k). \end{aligned}$$

Thus, $\tilde{p}_H$ is an upper bound on the exact p-value p_H. □

Notice that the first term of $\tilde{p}_H$ can be computed using Eq. (1), while the second term can be rewritten as follows:

$$\begin{aligned}
&Pr(Z_{2,m} \geq k \cap Z_{1,m} < k) \\
&= Pr(Z_{1,1} = 0 \cap Z_{2,m-1} = k-1 \cap Z_{m+1,1} = 1) \\
&= Pr(Z_{1,1} = 0) \cdot Pr(Z_{2,m-1} = k-1) \cdot Pr(Z_{m+1,1} = 1) \\
&= (1-p) \cdot Bin\,(k-1; m-1, p) \cdot p.
\end{aligned}$$

Thus, the upper bound $\tilde{p}_H$ is easy to compute.

We propose to use the upper bound $\tilde{p}_H$ as a statistical test. This test may be conservative, but that only means that results may be statistically more significant. As the exact p-value p_H is difficult to compute, we cannot analyze directly how tight the upper bound is. In Section 3.1 we study empirically how tight the approximation is, and in Section 3.2 we compare the power of this test to the alternative of combining the binomial test with a general post-hoc correction.

To complete the story, we obtain an upper bound to the one-tailed p-value in the low direction analogously to the previous case. For brevity we just list the result. Define

$$\tilde{p}_L = Pr(Z_{1,m} \leq k) + (n-m) \cdot Pr(Z_{2,m} \leq k \cap Z_{1,m} > k).$$

Theorem 2. *$\tilde{p}_L$ is an upper bound on the exact p-value p_L, i.e., $\tilde{p}_L \geq p_L$.*

Proof. Analogous to Theorem 1. □

The correction term can be computed using

$$Pr(Z_{2,m} \leq k \cap Z_{1,m} > k) = p \cdot Bin\,(k; m-1, p) \cdot (1-p).$$

2.5 A Generalization for Sliding Windows with Constant Step Size

If we use a sliding window with step size larger than one, we test fewer hypotheses, but the dependency between the consecutive subsequences will also change. The upper bound from Section 2.4 is also an upper bound when using a larger step size, but a tighter bound can be obtained relatively easily.

Let r be the user-defined step size. The p-value in the high direction is

$$p_H = Pr\Big(\bigcup_{i=1,1+r,1+2r,\ldots,1+\lfloor \frac{n-m}{r} \rfloor r} Z_{i,m} \geq k\Big)$$

Since there are $1 + \lfloor \frac{n-m}{r} \rfloor$ subsequences, we define $\tilde{p}_H$ as

$$\tilde{p}_H = Pr(Z_{1,m} \geq k) + \left\lfloor \frac{n-m}{r} \right\rfloor \cdot Pr(Z_{1+r,m} \geq k \cap Z_{1,m} < k).$$

Theorem 3. *$\tilde{p}_H$ is an upper bound on the exact p-value p_H, i.e., $\tilde{p}_H \geq p_H$.*

Proof. p_H can be decomposed as

$$\begin{aligned} p_H = Pr(Z_{1,m} \geq k) + Pr(Z_{1+r,m} \geq k \cap \bigcap_{i=1} Z_{i,m} < k) + \dots \\ + Pr(Z_{1+\lfloor \frac{n-m}{r} \rfloor r,m} \geq k \cap \bigcap_{i=1,1+r,1+2r,\dots,1+(\lfloor \frac{n-m}{r} \rfloor - 1)r} Z_{i,m} < k). \end{aligned} \tag{6}$$

Also, it holds that

$$\begin{aligned} Pr(Z_{1+r,m} \geq k \cap \bigcap_{i=1} Z_{i,m} < k) \geq \\ Pr(Z_{1+2r,m} \geq k \cap \bigcap_{i=1,1+r} Z_{i,m} < k) \geq \\ \dots \end{aligned} \tag{7}$$

Combining Eqs. (6) and (7) gives

$$p_H \leq Pr(Z_{1,m} \geq k) + \left\lfloor \frac{n-m}{r} \right\rfloor \cdot Pr(Z_{1+r,m} \geq k \cap Z_{1,m} < k).$$

Thus, $\tilde{p}_H$ is an upper bound on the exact p-value p_H. □

In this setting, the correction term is more involved. For convenience, we split the correction term into three parts: the overlap between the two subsequences, $Z_{1+r,m-r}$, and the two non-overlapping parts, $Z_{1,r}$ and $Z_{1+m,r}$. We have that

$$Z_{1+r,m} \geq k \Rightarrow Z_{1+r,m-r} + Z_{1+m,r} \geq k, \text{ and}$$
$$Z_{1,m} < k \Rightarrow Z_{1,r} + Z_{1+r,m-r} < k.$$

Both right hand sides are satisfied simultaneously if and only if

$$\begin{aligned} Z_{1+m,r} \geq k - Z_{1+r,m-r}, \quad Z_{1+r,m-r} \geq k - Z_{1+m,r}, \\ Z_{1,r} < k - Z_{1+r,m-r}, \quad Z_{1+r,m-r} < k - Z_{1,r}. \end{aligned} \tag{8}$$

Since $Z_{1+m,r}$ and $Z_{1,r}$ are both by definition between 0 and r, we have that

$$k - r \leq Z_{1+r,m-r} < k. \tag{9}$$

We can rewrite the correction term to an explicit sum using Eqs. (8) and (9):

$$\begin{aligned} & Pr(Z_{1+r,m} \geq k \cap Z_{1,m} < k) \\ &= \sum_{j=max(0,k-r)}^{k-1} Pr(Z_{1+r,m-r} = j \cap Z_{1+m,r} \geq k-j \cap Z_{1,r} < k-j) \\ &= \sum_{j=max(0,k-r)}^{k-1} Pr(Z_{1+r,m-r} = j) \cdot Pr(Z_{1+m,r} \geq k-j) \cdot Pr(Z_{1,r} < k-j) \\ &= \sum_{j=max(0,k-r)}^{k-1} \left(Bin(j; m-r, p) \cdot \sum_{l=k-j}^{r} Bin(l; r, p) \cdot \sum_{l=0}^{k-j-1} Bin(l; r, p) \right). \end{aligned}$$

One may verify that the result for $r = 1$ is the same as in Section 2.4. The binomial pmf and cmf can be computed in constant time [14], thus the computational complexity of the correction term is $O(min(k, r)$ and independent of the size of the full sequence. An upper bound $\tilde{p}_L$ can be derived analogously.

3 Experiments

We studied the power of the test on synthetic data and compared the power of the test with the alternative of post-hoc correction, results of which are discussed in Sections 3.1 and 3.2. We also investigated the practical utility of the test on two types of data: an English novel and a part of the human reference genome. The findings of these experiments are presented in Sections 3.3 and 3.4.

3.1 Tightness of the Upper Bound

Since the proposed test provides strong control over the family-wise error rate, we know that the probability of observing one or more false positives is at most α. Unfortunately, this provides no information on the *power* of the test, i.e., the probability of rejecting a false null hypothesis. Ideally, we would study the probability or rate of false negatives directly. But that is not possible, unless we specify an alternative hypothesis; there is no general false negative rate. Instead, we use the fact that there is a trade-off between the probability false positives and the probability of false negatives.

By definition we have that the probability of false negatives is minimized when the probability of false positives is maximized. Thus, preferably, the probability of observing one or more false positives should be as close to α as possible. To study how close the probability of encountering one or more false positives is in practice, we designed the following experiment.

The tightness of the upper bound may depend both on the length of sliding window, as well as the event probability. Thus, we tried various window lengths ($m \in \{100, 1000, 10000\}$) and event probabilities ($p \in \{0.001, 0.01, 0.1\}$). For each combination, we generated 1,000 sequences of length $n = 9,999 + m$ (such that there are 10,000 p-values per sequence) and computed the p-values $\tilde{p}_H$ for all subsequences using a sliding window with step size 1.

The quantity of interest is the minimal p-value per sequence, because if the minimal p-value in a sequence is below the threshold α, then we have at least one false positive. Ideally, the distribution of minimal p-values over the sequences is uniform, which means that for any value α, the probability of observing one or more p-values below α is exactly α itself. This ensures that the probability of false positives is maximal (while providing FWER control), and that the probability of false negatives is minimal. Note that this holds by definition for the exact p-values under the null hypothesis, but the upper bound that we propose to use instead may have a higher probability of false negatives.

The results of the experiment are presented in Figure 1. We find that the p-values are reasonably close to the optimal distribution and that they are further

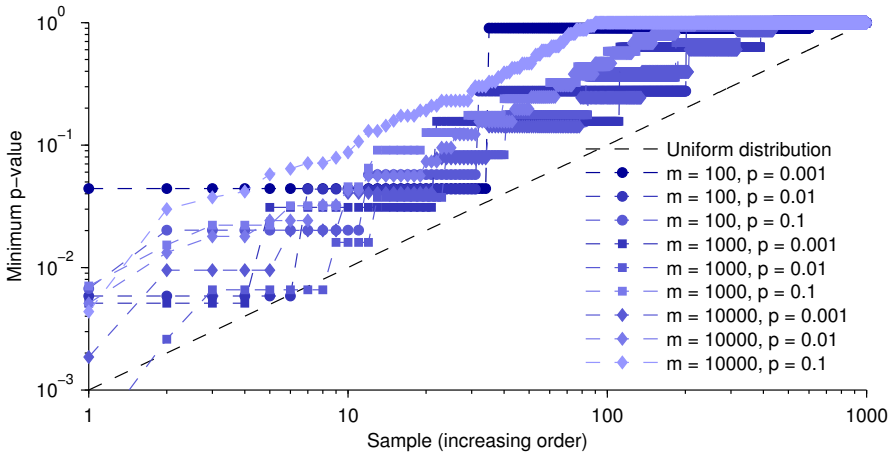

Fig. 1. The distribution of minimal p-values over 1,000 synthetic sequences for the proposed method, using various window lengths m and event probabilities p, compared to the ideal distribution. We find that the p-values are reasonably close to the uniform distribution and that they are further from uniform when the expected number of events ($= m \cdot p$) is higher.

from the optimal distribution when the expected event count ($= m \cdot p$) is larger. The largest observed effect is approximately 1 order of magnitude ($m = 10{,}000$, $p = 0.1$), indicating that the p-values are 1 order of magnitude too high in that case. Note that the results for very low expected counts (e.g., $m = 100$, $p = 0.001$) may appear more conservative, but they are skewed mostly because there are very few distinct p-values: the highest number of events observed in any subsequence is 3 ($\tilde{p}_H = 0.0437$), and for $k \in \{0, 1\}$, we have $\tilde{p}_H = 1$.

We expect that p-value estimates that are conservative by one order of magnitude will not be a problem in most practical settings; much larger differences in the choice of α can be observed in the literature: from $\alpha = 0.1$ to $\alpha = 0.00001$. Also, because the p-values are controlled for family-wise error rate, use of a 'large' α, such as 0.05, still guarantees that obtaining any false-positive results has very low probability.

3.2 Comparison to Hochberg's Step-Up Procedure

An alternative approach to obtaining p-values for the tested hypotheses under strong control of the family-wise error rate is to use the binomial test (Eqs. (1) and (2)) with post-hoc correction. The correction with largest power that we are aware of that provides strong control for the family-wise error rate, and which is applicable in this setting, and that does not require specifying the dependency structure of the p-values, is Hochberg's step-up procedure [8]. Hochberg's procedure is valid for independent and positively dependent p-values [17]. The latter is the case here, as the p-values for overlapping windows have positive correlation.

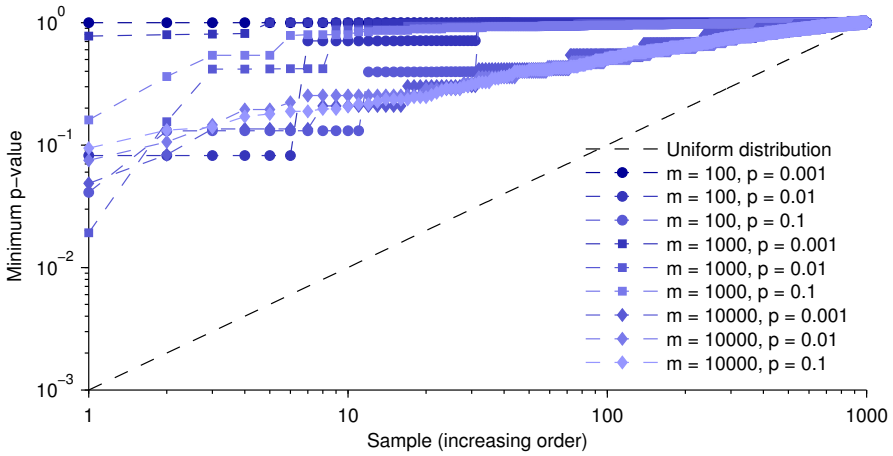

Fig. 2. The distribution of minimal p-values for the binomial test with Hochberg's post-hoc correction, on the data from Figure 1. We find that the p-values are far from the uniform distribution, for any combination of parameters, while the distribution is more uniform when the expected number of events ($= m \cdot p$) is larger.

We computed the p-values for the binomial test for each sequence generated in the previous experiment (Section 3.1), using a sliding window of the same length, and adjusted these using Hochberg's procedure. Thus, the p-values are directly comparable to those in the previous experiment. We computed the minimal p-value per sequence, and compared the results with those from the upper-bound method.

The distribution of minimal p-values is shown in Figure 2. We observe that p-values from the method with post-hoc correction are far from uniform, for any combination of parameters, while the distribution becomes more uniform as the expected number of events per subsequence increases. The proposed method outperforms the post-hoc approach for any combination of parameters, although we cannot be certain that this holds for much larger expected number of events.

3.3 Bursty and Non-Bursty Words in an English Novel

The prime motivation for this work comes from the domain of text analysis. Church and Gale [3] and Katz [9] both studied *burstiness* of words in the context of probabilistic modeling of word counts, and the concept is related to relevance measures in information retrieval, such as inverse document frequency [19]. More recently, using a quantification of burstiness based on the inter-arrival time distributions of words, burstiness of words has been related to semantic categories [1], statistical tests for comparing corpora that take into account burstiness have been proposed [13], and the impact of burstiness on choosing appropriate window lengths for sequence analysis has been studied [12].

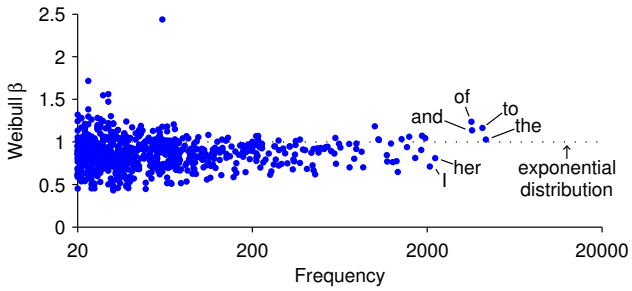

Fig. 3. The relationship between burstiness, measured using the Weibull distribution, and frequency of words. Each dot represents a word in the novel *Pride and Prejudice*.

Table 1. We studied the local behavior of the five least and most bursty words in two frequency bins to investigate the suitability of our method to locate over and underuse of words in text.

Frequency	Low [σ = 40–50]	High [σ = 300-600]
Non-bursty	hardly, help, perfectly, point, scarcely	an, elizabeth, more, there, when
Bursty	marry, pride, read, rosings, william	are, me, their, will, your

For the purpose of text analysis, it is useful to know if there are fragments in a text where a certain word is over or underused and to locate such fragments. We investigated the suitability of the proposed method to this task. As an experiment, we downloaded the book Pride & Prejudice by Jane Austen, which is freely available via Project Gutenberg[2]. We computed the frequency and the maximum-likelihood estimates for the Weibull distribution [1,13] for all words, and then selected the five most and least bursty words in two frequency bins, see Table 1. An overview of the relation between the frequency and burstiness of words is given in Figure 3.

For each of the selected words, we tracked the frequency throughout the book using a sliding window of length 5,000 and step size 1. The book contains n = 121,892 words, thus there are 116,893 windows. We chose a window length of 5,000 to ensure that low event counts could also be significant; for example, for a window length of 2,000 and event probability $p = 1/300$, we have that the p-value for $k = 0$ is $\tilde{p}_L = 0.4833$. Thus, an event count of zero is not significant, even for fairly frequent words. With a window length of 5,000, event counts of 3 and less are significant at $\alpha = 0.05$ ($\tilde{p}_L = 0.0164$).

We computed the significance of the observed frequencies, for both the high and low direction. Because the results are for illustrative purposes, we did not apply any additional correction for testing multiple sets of hypothesis. Figure 4 shows the results for three words. The word *an* is frequent and non-bursty, and no parts of the book show significant under or overuse of the word. For the pronoun *me*, which is frequent and bursty, we observe two areas of overuse, and

[2] `http://www.gutenberg.org/`

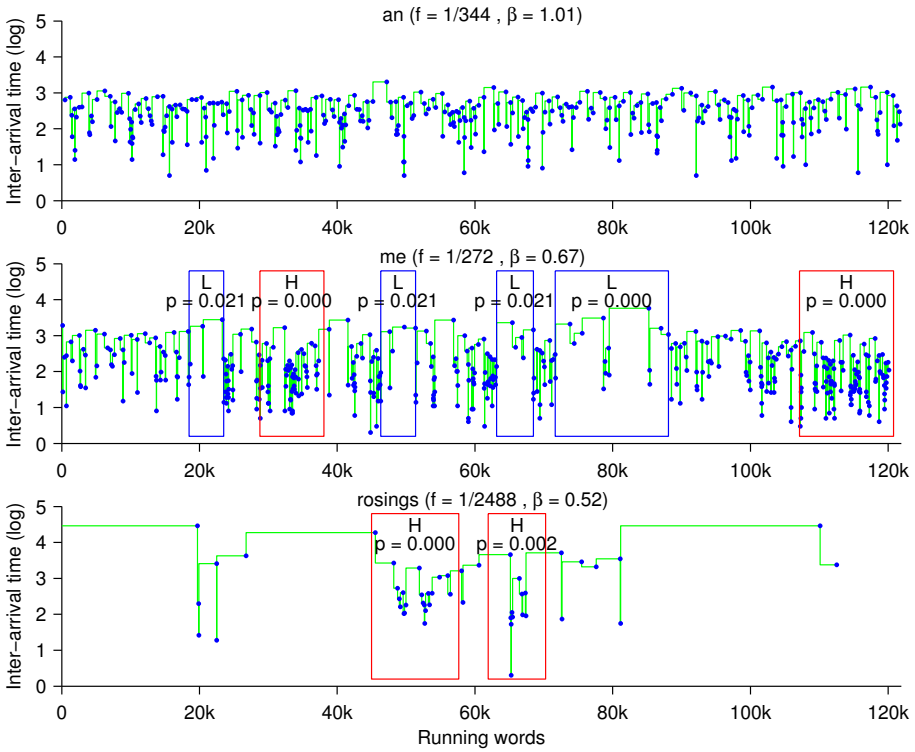

Fig. 4. Significant over and underuse of three words in the novel *Pride and Prejudice*, compared to the average frequency in the book. Each blue dot corresponds to an occurrence of the word in the text. To aid the visualization of the results, all overlapping significant subsequences have been merged together. We observe that for *an*, no parts of the book show significant under or overuse of the word, while for the pronoun *me*, two areas show significant overuse, and four areas show underuse of the word. Finally, the family name *rosings* is used mainly in two parts of the book.

four areas of underuse, compared to the average frequency. Finally, the family name *rosings*, which is infrequent and bursty, is used a lot in two text fragments and occurs a few times in other parts of the book.

A full overview of results is given in Table 2. As expected, we find that each of the bursty words is significantly over or underrepresented in at least one fragment of the book. Surprising is that some frequent words that are non-bursty according to the Weibull distribution estimate are also under or overused in one or more fragments. This indicates that there is local structure that is not captured by the Weibull measure of word burstiness. The results from the proposed method are confirmed by visual inspection of the data and we conclude that the method has a clear potential to find novel and interesting patterns.

Table 2. Number of areas with significant underuse (L) or overuse (H) for each of the twenty words. Each of the bursty words is significantly more or less frequent in some part of the book, and some frequent words that are non-bursty according to the Weibull distribution estimate are also under or overused in one or more book parts.

Non-bursty						Bursty					
Frequent			Infrequent			Frequent			Infrequent		
Word	L	H	Word	L	H	Word	L	H	Word	L	H
an	0	0	hardly	0	0	are	1	0	marry	0	1
elizabeth	2	0	help	0	0	me	4	2	pride	0	1
more	0	0	perfectly	0	0	their	1	0	read	0	2
there	0	1	point	0	0	will	2	3	rosings	0	2
when	0	0	scarcely	0	0	your	2	3	william	0	1

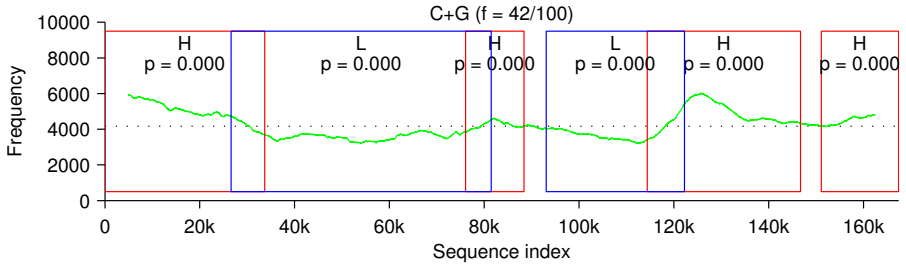

Fig. 5. Analysis of the GC content at the start of Chromosome 1 of the Homo Sapiens reference genome, using a sliding window of length 10,000. All overlapping significant parts have been merged. We observe that the frequency of GC is quite volatile: parts where the content is significantly high overlap with parts where the content is significantly low. We also observe that the test is sufficiently powerful, there are many significant results, even though we are testing a total of 225,270,622 hypotheses.

3.4 Variation in GC and TA Content in DNA

Variation of GC content in DNA sequences is used to define *isochores*, which in turn are used to identify gene structure [2]. We tested if we could find significant variation in GC and TA content in chromosome 1 from the Homo Sapiens reference genome, which we downloaded from the NCBI repository[3]. We computed the frequency of C+G using a sliding window of length 10,000 and step size 1. Chromosome 1 of the reference genome (build 37, patch 9) contains 225,280,621 fixed nucleotides, thus the number of tested hypotheses is in this case very large.

Analysis of the first consecutive fixed part can be found in Figure 5. We observe that the test is sufficiently powerful, because several parts of the sequence are identified as having significantly high or low GC content. We find that the GC content is quite volatile: the parts where the content is significantly low and high overlap each other. We conclude again that the proposed method has potential for finding novel and interesting patterns in the data.

[3] `http://www.ncbi.nlm.nih.gov`

4 Related Work

The popularity of significance testing methods in data mining has increased considerably over the past decade. Gionis et al. [5] introduced swap randomization for mining significant patterns while maintaining row and column margins, while De Bie [4] proposed a maximum-entropy approach that can also take into account other types of constraints. Webb [20] and Hanhijärvi [7] studied the problem of multiple testing for mining patterns. These studies are all restricted to mining itemsets or tiles. A generic approach to mining structure in data using statistical testing has been presented by Lijffijt et al. [11].

There are only a few studies on statistical testing approaches for mining sequential data. Most related is the statistical test proposed by Kifer et al. [10] for detecting change points in streams. However, they rule out the possibility of controlling the family-wise error rate, as they consider only streams of infinite length. Another drawback of that method is that the critical points cannot be computed analytically, but require randomization.

Complementary to this work are the randomization-based statistical tests for comparing event counts between databases of sequences put forward by Lijffijt et al. [13]. Segmentation methods may provide an alternative to modeling frequency variation, although the focus is then on global modeling, while the aim here is to find local structure. Mannila and Salmenkivi [16] study efficient methods for sequence segmentation, while the approach by Lijffijt et al. [11] can be used to assess the significance of such a segmentation.

5 Conclusions

We have introduced a novel statistical test for assessing the significance of event frequencies in subsequences when using a sliding window. The test provides strong control of the family-wise error rate and takes into account the dependency structure of overlapping subsequences. We have shown that, although exact p-values under the null hypothesis are difficult to compute, an easy-to-compute upper bound can be used instead. We have shown empirically that the upper bound is sufficiently tight and that the test offers increased power compared to combining the binomial test with a generic post-hoc correction.

We have also investigated the utility and practicality of the test on linguistic and biological sequences and found several novel and interesting patterns. We have shown that meaningful results can be obtained, and that the method remains sufficiently powerful even when testing a very large number of hypotheses. We conclude that the proposed method is simple, fast and powerful and that it can produce meaningful results on various types of data.

Acknowledgements. The author has received support from the Finnish Doctoral Programme in Computational Sciences (FICS) and the Academy of Finland's Centre of Excellence in Algorithmic Data Analysis (ALGODAN). I thank Heikki Mannila and Petteri Kaski for useful discussion and feedback.

References

1. Altmann, E.G., Pierrehumbert, J.B., Motter, A.E.: Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. PLoS ONE 4(11), e7678 (2009)
2. Bernardi, G.: Isochores and the evolutionary genomics of vertebrates. Gene 241(1), 3–17 (2000)
3. Church, K.W., Gale, W.A.: Poisson mixtures. Nat. Lang. Eng. 1(2), 163–190 (1995)
4. De Bie, T.: Maximum entropy models and subjective interestingness: an application to tiles in binary databases. Data Min. Know. Disc. 23(3), 407–446 (2011)
5. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. ACM TKDD 1(3), 14 (2007)
6. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Min. Know. Disc. 8(1), 53–87 (2004)
7. Hanhijärvi, S.: Multiple hypothesis testing in pattern discovery. In: Elomaa, T., Hollmén, J., Mannila, H. (eds.) Discovery Science, LNCS. vol. 6926, pp. 122–134. Springer, Heidelberg (2011)
8. Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75(4), 800–802 (1988)
9. Katz, S.M.: Distribution of content words and phrases in text and language modelling. Nat. Lang. Eng. 2(1), 15–59 (1996)
10. Kifer, D., Ben-David, S., Gehrke, J.: Detecting change in data streams. In: Nascimento, M.A., Özsu, M.T., Kossmann, D., Miller, R.J., Blakeley, J.A., Schiefer, K.B. (eds.) Proc. of VLDB. pp. 180–191. VLDB Endowment (2004)
11. Lijffijt, J., Papapetrou, P., Puolamäki, K.: A statistical significance testing approach to mining the most informative set of patterns. Data Min. Know. Disc. *in press*
12. Lijffijt, J., Papapetrou, P., Puolamäki, K.: Size matters: Finding the most informative set of window lengths. In: Flach, P.A., de Bie, T., Cristianini, N. (eds.) ECML PKDD, LNCS. vol. 7524, pp. 451–466. Springer, Heidelberg (2012)
13. Lijffijt, J., Papapetrou, P., Puolamäki, K., Mannila, H.: Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD, LNCS. vol. 6912, pp. 341–357. Springer, Heidelberg (2011)
14. Loader, C.: Fast and accurate computation of binomial probabilities (2000), unpublished manuscript
15. Mannila, H.: Local and global methods in data mining: Basic techniques and open problems. In: Widmayer, P., Eidenbenz, S., Triguero, F., Morales, R., Conejo, R., Hennessy, M. (eds.) Automata, Languages and Programming, LNCS. vol. 2380, pp. 57–68. Springer, Heidelberg (2002)
16. Mannila, H., Salmenkivi, M.: Finding simple intensity descriptions from event sequence data. In: Proc. of ACM SIGKDD. pp. 341–346. ACM, New York (2001)
17. Sarkar, S.K., Chang, C.K.: The Simes method for multiple hypothesis testing with positively dependent test statistics. J. Am. Stat. Ass. 92(440), 1601–1608 (1997)
18. Shaffer, J.P.: Multiple hypothesis testing. Ann. Rev. Psych. 46, 561–584 (1995)
19. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. 28(1), 11–21 (1972)
20. Webb, G.I.: Layered critical values: A powerful direct-adjustment approach to discovering significant patterns. Mach. Learn. 71(2–3), 307–323 (2008)