Time Series Motifs Statistical Significance

Nuno Castro* Paulo J. Azevedo†

Computer Science and Technology Center
Department of Informatics
University of Minho, Portugal
{castro, pja}@di.uminho.pt

Abstract

Time series motif discovery is the task of extracting previously unknown recurrent patterns from time series data. It is an important problem within applications that range from finance to health. Many algorithms have been proposed for the task of efficiently finding motifs. Surprisingly, most of these proposals do not focus on how to evaluate the discovered motifs. They are typically evaluated by human experts. This is unfeasible even for moderately sized datasets, since the number of discovered motifs tends to be prohibitively large. Statistical significance tests are widely used in bioinformatics and association rules mining communities to evaluate the extracted patterns. In this work we present an approach to calculate time series motifs statistical significance. Our proposal leverages work from the bioinformatics community by using a symbolic definition of time series motifs to derive each motif's p-value. We estimate the expected frequency of a motif by using Markov Chain models. The p-value is then assessed by comparing the actual frequency to the estimated one using statistical hypothesis tests. Our contribution gives means to the application of a powerful technique - statistical tests - to a time series setting. This provides researchers and practitioners with an important tool to evaluate automatically the degree of relevance of each extracted motif.

Keywords

Time Series; Motif Discovery; Statistical Significance tests; Significant Patterns.

1 Introduction

To extract previously unknown recurrent patterns (motifs) from time series databases is an important data mining problem. Motifs are relevant because they can summarize the time series database and provide useful insight to the domain expert [6]. A large number of applications exist from a broad variety of areas such as health and finance. Fig. 1 shows an example of a time series with 3 different motifs (displayed in blue, green and red), as typically outputted by existing motif discovery algorithms.

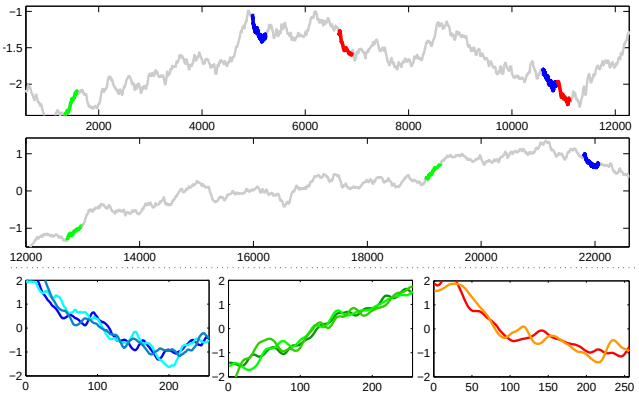

Figure 1: Example of a time series with several motifs. Above: in its original context; below: detail of each motif. Blue, Green: 3 instances; Red: 2 instances.

Since the problem formulation in [17], many proposals on how to extract motifs from a time series database have been introduced [3, 4, 6, 22–26, 28, 38, 40]. Surprisingly, most of these proposals do not focus on how to evaluate the extracted motifs. Returned motifs tend to be subjectively evaluated by humans because they are application dependent and not previously labeled - motif discovery is an unsupervised task. In practice, this is unfeasible. Datasets are often large and motif mining algorithms typically return a prohibitively large number of patterns. To restrain to expert analysis the most frequent motifs is not an interesting approach, as frequent patterns are not necessarily the most interesting ones. Many frequent patterns are spurious, trivial or simply expected: they are not meaningful to the user. In a

*Nuno Castro is supported by *Fundação para a Ciência e a Tecnologia* grant SFRH/BD/33303/2008.

†Paulo J. Azevedo is supported by *Fundação para a Ciência e Tecnologia*, Project ProtUnf, FEDER and Programa de Financiamento Plurianual de Unidades de I&D.

randomly generated database of length 65536 from [13], for example, 65 motifs are discovered. The top motif reaches 4 repetitions, and the average motif count is 2.17. Since a random process generated the database, all discovered motifs are meaningless. In fact, this example is depicted in Fig. 1. It highlights the need for automatic time series motifs evaluation.

Statistical tests have been successfully applied to other pattern mining problems. For example, in bioinformatics they have been used to detect DNA segments with significantly unexpected frequency [33]; in networks analysis, to find significant subgraphs [21]; in association rules mining to discard redundant rules [39]. In all these examples the common question to be addressed is: "Can this pattern be observed so many times just by chance?". These approaches consider the observed count (frequency) of a pattern which is typically compared to its expected count. This difference is then statistically analyzed. However, this method cannot be directly applied to time series data since it is not clear how to calculate the expected frequency of a given section of the series.

To overcome this limitation and take advantage of the wealth of available algorithms for symbolic data (DNA sequences, text, etc.), we use a symbolic definition of time series motifs. Our approach is based on work from bioinformatics [33]. We estimate the probability of occurrence of a word (motif) using Markov Chain Models. In these models, the probability of a motif is estimated according to its subword count. Given a motif, we compare the difference between its observed count and estimated expected count in terms of statistical significance. Namely, we calculate the *p-value* of this difference, aiming to answer whether we can observe such a count solely by chance.

Our contributions are twofold: to provide an approach to assess the statistical significance of time series motifs, and to compare the performance of several simple statistical hypothesis tests on motifs extracted from real datasets. The novelty of our work is that it enables the calculation of time series motifs p-values. To the best of our knowledge, this has not been attempted in the literature. It has been shown to be an important problem in DNA, protein, and network motifs (discrete motifs). We provide the link between the well studied discrete motif significance problem and time series motif evaluation. This allows time series data mining practitioners to evaluate better the motifs extracted from their data. It also provides researchers with a method to evaluate properly the output of motif discovery algorithms using statistical significance.

The remainder of the paper is organized as follows: section 2 describes the state of the art in motif statistical significance; background and notation used throughout the paper are described in section 3; in section 4 an approach for assessing time series motifs significance is proposed; the experimental analysis is described in section 5; finally, in section 6 we derive conclusions.

2 Related Work

Since the introduction of the time series motif discovery problem [17], many approaches have been proposed [3, 6, 22–26, 28, 38, 40]. Most of these works tackle the algorithmic details of the motif extraction process. Surprisingly, the critical aspect of evaluating the extracted motifs has not received much attention by researchers. The results are typically interpreted by experts on the domain at hand. This approach is untenable for large real-world datasets that can reach terabytes of data. Automatic motif evaluation procedures are required.

According to [7], motif mining evaluation measures can be classified in the following categories: class-based, theoretic-information, mixed measures and statistical significance tests. Class-based measures (accuracy related) are calculated by comparing the motif occurrences with the ground truth using a confusion matrix. Examples are precision, recall and specificity. Theoretic-information measures are calculated using probabilistic or information criteria contained in the motif itself. Examples are the Information Gain and the Minimum Description Length. Measures such as Mutual Information and J-measure are mixed, because they use both class-based and theoretic information criteria. From this set of measures, we are particularly interested in statistical significance tests. These tests are very popular in science in general and data mining in particular. They tend to be accepted as the *de facto* standard to evaluate significance or help in the decision making process.

Statistical significance tests are widely used in bioinformatics. Without claiming to be exhaustive we mention a few of these works. Zhang et al. [41] define the problem of evaluating statistical significance of DNA motifs as the ranking of such motifs according to an underlying model, defined using Markov chains. A dynamic programming algorithm (MotifRank) is proposed to compute motif exact p-values. Marschall and Rahmann [19] propose a methodology to calculate p-values with respect to independent and identically-distributed (i.i.d.) and Markov models. A compound Poisson approximation is used for the number of motif occurrences (null distribution). These techniques are integrated in an efficient motif discovery algorithm by exploiting the monotonicity property of the compound Poisson approximation. The algorithm is applied to IUPAC strings (chemical compounds representation) and

Mycobacterium tuberculosis data. Nuel [27] provides recursive algorithms to compute Cumulative Distribution Functions (CDF) using Finite Markov Chain Imbedding (FMCI). The algorithms are applied to discover exact p-values of patterns aiming to find hydrophobic segments in protein data. In [1], the authors introduce an algorithm to calculate the probability of finding multiple occurrences of motif in a random text. This probability is calculated using both the Bernoulli and order one Markov chain models. The approach is applied to find the statistical significance of binding sites frequency in regulatory modules of eukaryotic genes. Mas et al. [18] propose an algorithm to mine unexpected frequent sequential patterns in DNA and protein sequences. Sequential patterns are defined according to a Markov model and patterns support follow a Binomial distribution. The p-values that measure overrepresentation are then calculated. Hollunder et al. [10] introduce the DASS algorithm to estimate the statistical significance of patterns in protein data. Several techniques for determining the expected value of each pattern such as data permutations, shuffling, and the binomial distribution are used. Robin and Schbath [32] perform an experimental comparison of several distributions of word counts in random sequences, regarding accuracy and computational cost. The exact distribution is compared to the Gaussian and compound Poisson approximations in the extraction of exceptional words of the phage *Lambda* genome. In [30], the drawbacks of the Gaussian approximation are analyzed. Schbath [35] studies the statistical distributions of word counts in Markov chains. Formulae are derived for the estimated expected counts under these distributions. In [33], statistical tests are used to compare motif count exceptionalities in two (or more) sequences. The exact binomial and the asymptotic likelihood ratio test are used. The motif count is modelled using Poisson processes. The motifs in the backbone and loops of the *Escherichia coli* K-12 bacterium are compared.

In the networks (graph) mining community, the issue of statistical significance in motif discovery has also received much attention. In [9], a Binomial test is used to evaluate the statistical significance of frequent subgraphs in a chemical compounds graphs database. Milo et al. [21] define network motifs as patterns of interconnections with a significantly higher frequency than those in randomized networks, according to their Z-score. A comprehensive experimental analysis is done in complex networks from biochemistry, neurobiology, ecology and engineering. In [12] the authors convert sequential data to probabilistic automata and then integrate statistical constraints to reduce the search space of the exploratory process. The approach is applied to car flow modelling data. Ribeca and Raineri [31] derive a fast motif Z-scores exact calculation method using discrete finite-state automata (DFA), assuming the sequence is generated by a Markov model of arbitrary order. The authors experimentally test their approach in large scale human genome and yeast binding factors data. Matias et al. [20] provide exact formulas for the expectation and variance of a motif's number of occurrences. This approach also introduces a simple and efficient probabilistic model for the motif distribution in networks, which is much more efficient than the traditional comparison to randomized (simulated) networks. In [29], the authors consolidate a decade of research in biosequences motifs exceptionality and apply it to the network motifs scenario. Several motif distributions approximations are compared such as the compound Poisson distribution and the Gaussian approximation. Approximate p-values are calculated to assess the exceptionality of observed motif counts. The method is applied to protein-protein interaction networks.

There is a handful number of time series motif mining proposals that consider the significance evaluation aspect of extracting motifs. Ferreira and Azevedo [6] use the Information Gain and Log-Odds measures to assess the statistical significance of motifs. However, the order dependency (time) that characterizes time series data is not taken into account. In [4], Keogh et al. use a statistical test as a criterion to stop their iterative motif discovery algorithm, i.e. the algorithm ends the execution when the observed motif count significantly exceeds the expected by chance. In this work, we aim to go one step further and calculate each motif's p-value according to their statistical significance. In the context of time series anomaly detection, Keogh et al. [16] propose an approach to find surprising patterns in time series data. Markov Chain Models are used to predict the expected frequency of patterns, given a collection of previously observed normal data. However, the motif discovery problem is unsupervised. It is not possible to know beforehand which patterns are significant. Moreover, we are not interested in finding anomalous patterns. Rather, we aim at statistically stating which frequent patterns are also significant by calculating each pattern's p-value.

3 Background and Notation

In this section we introduce some notations and useful definitions. First we define our object of study.

DEFINITION 3.1. A *time series* T of length n is an ordered succession of a variable's observations $(t_i, \ldots, t_n)$ over time, with $t_i \in \mathbf{R}$.

For the scope of this work, all time series are normalized in order to remove offset and scaling effects. It has been shown that comparing time series that are not normalized is meaningless [14].
We are typically interested in mining a collection of time series with arbitrary lengths.

DEFINITION 3.2. A *time series database* D is a set of $|D|$ unordered time series [26].

Time series data mining algorithms often use subsections, or subsequences, of the original time series in their calculations.

DEFINITION 3.3. Given a time series T of length n, a *time series subsequence* $S = s_i, ..., s_{i+m-1}$ is a sampling of $m \leq n$ contiguous positions of T, such that $1 \leq i \leq n - m + 1$ (definition from [4]).

For simplicity, we treat each subsequence of database as a different time series in D. In practice, to slide a window through a long time series for the purpose of extracting subsequences is similar to handling each subsequence as a different time series. Possible motif overlaps are handled by taking into account trivial matches [4].
Since our work is inspired by the biosequences motif mining, we are interested in the symbolic representation of time series and their subsequences.

DEFINITION 3.4. A *word* $w = w_1 w_2 \dots w_l$ is the symbolic representation of a subsequence S, with $w_i \in \Sigma$. The Σ is the representation alphabet and its size is named the representation *resolution*.

The symbolization of S by a generic times series representation technique R is denoted by $R(S) = w$. In this work we use the best representation technique available in the literature for time series data, as experimentally shown in [5]. The Symbolic Aggregate Approximation (iSAX) [37] representation takes as input a time series and transforms it into a sequence of symbols, as highlighted in Fig. 2.
As depicted in figure 3, slightly different subsequences can originate the same word in the representation space. These subsequences are called instances of the given word.

DEFINITION 3.5. A subsequence S is an *instance* of a word w if $R(S) = w$, where $R(S)$ is a symbolic representation of S.

Matching between two or more instances of word w is defined as follows:

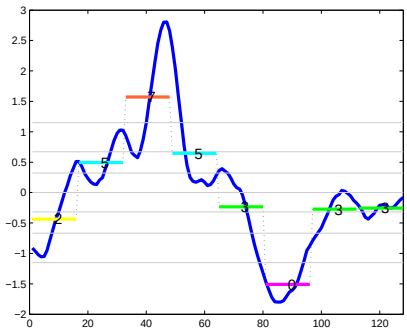

Figure 2: Conversion of a time series into its iSAX representation, generating word {2,5,7,5,3,0,3,3}.

DEFINITION 3.6. Subsequences S_1 and S_2 *match* if their symbolic representations are the same, i.e. $R(S_1) = R(S_2)$.

At this point, we are ready to formalize the notion of time series motif. An example of a motif with 3 instances is shown in figure 3.

DEFINITION 3.7. The word w is a *Motif* in database D if the count of all instances of w in D is greater than 1.

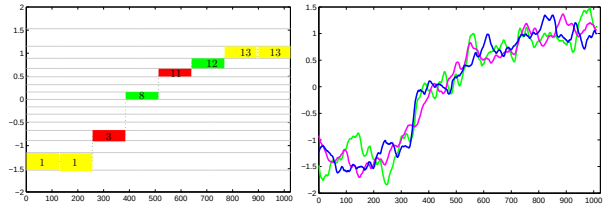

Figure 3: Motif {1,1,3,8,11,12,13,13} (left) and its 3 instances in the database (right).

DEFINITION 3.8. The *motif count* (or frequency) of a motif M is the total number of instances M has in database D.

Time series motifs are typically sorted according to their motif count.

4 Time Series Motifs Statistical Significance

In this section we introduce an approach to assess the statistical significance of time series motifs. To the best of our knowledge, there is no approach available in the literature to calculate the p-value of time series motifs. Our approach methodology is described next. First, motifs are extracted from the database. Second, the

probability of each motif is calculated using Markov Chain models. Statistical hypothesis tests are then applied according to several distributions for the motif counts (Binomial, Poisson and Gaussian distributions) to calculate each motif's p-value. In this section the false discovery rate problem is also considered.

4.1 Extracting Motifs

The first step of the motif significance evaluation is the actual extraction of frequent motifs. There is a plethora of time series motif discovery algorithms in the literature (see section 2). Among those, exact algorithms [26] have been shown to be a sound contribution to the time series motif discovery problem. Despite being less accurate than their exact counterparts, approximate algorithms present a relatively good trade-off between accuracy and efficiency. They are also typically robust to noise [3, 4]. In this work, to leverage the existing work in bioinformatics motif discovery, we are interested in symbolic motifs, i.e. discretized representations of the discovered motifs. Therefore, we select an approximate algorithm, that internally uses a symbolic representation and outputs discrete motifs. It is noteworthy that any motif discovery algorithm can be used, since its output is symbolic, or discretized using iSAX. The recently introduced MrMotif [3] is an excellent candidate to represent the symbolic motifs approach. It uses a symbolic definition of time series motifs, a necessary property to take advantage of the wealth of existing work in the bioinformatics. It also outputs the most frequent motifs in a straightforward manner (a list of words) and it is efficient (linear complexity). MrMotif takes as input a time series database D and a parameter K and derives the top-K motifs in D and their count. This step is shown in figure 4.

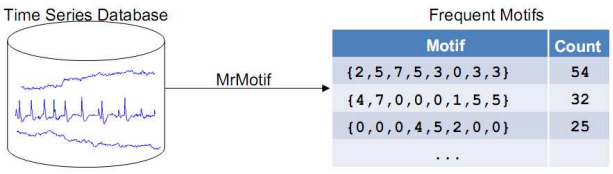

Figure 4: Extraction of frequent motifs from the time series database.

For simplicity, we choose to evaluate motif statistical significance as post processing task. This process can also be integrated in the motif search itself as demonstrated in [12, 19, 41].

4.2 Reference Model

The motif count by its own is not a good interestingness measure. Frequency does not guarantee that motifs are significant, similarly to support in itemsets mining. A trivial example highlighting this problem is shown in section 1, by using random time series data. A better approach is to consider the difference between the observed motif count and the motif expected count, given some knowledge on the time series. This knowledge is obtained regarding a reference model that reflects the background distribution of the motifs. The expected count is the number of motifs one should expect in random sequences that are similar, under some similarity definition, to our database. Random sequences are typically Bernoulli trials or Markovian sequences [35]. The former assume that words are i.i.d., although word symbols are possibly not i.i.d. in real data [36]. Markov Chain models take the composition of the words into account. That is, they consider the time dependency characteristic of time series data. They have been widely used in bioinformatics [9, 20, 21, 27, 31, 35] because they take the time dependency into account [18] and assist in correctly fitting the composition of words of length 1 up to $(m+1)$ (where m is the selected order for the model). Also, there are analytical probability calculations available which prevents the need to refer to expensive simulations [36].

We follow the approach described in [33] to obtain expected counts of DNA motifs. Namely, we use Markov Chain models as the reference model to calculate the (estimated) expected probability μ of a motif to occur in the database. The probability is calculated with respect to transition probabilities, which are estimated according to the observed sequences (see formulae below). The order m of the model ranges from 0 (Bernoulli) up to $l-2$. In Markov model of order m (denoted Mm), the composition of a word $w = w_1 w_2 \dots w_l$ is calculated using the observed counts of its subwords of length m and $m+1$. Hereby we show the expressions for M0 (Bernoulli), M1, and the maximal model M$(l-2)$:

M0	$\mu = \dfrac{\prod_{i=1}^{l} N(w_i)}{n_s^l}$
M1	$\mu = \dfrac{\prod_{i=1}^{l-1} N(w_i\, w_{i+1})}{n_s \prod_{j=2}^{l-1} N(w_i)}$
M$(l-2)$	$\mu = \dfrac{N(w_1 \dots w_{l-1})\, N(w_2 \dots w_l)}{n\,(l-m+1)\, N(w_2 \dots w_{l-1})}$

where $N(x)$ is the count of motif x in the sequence of (symbolic) length n_s. Under this scenario, the expected

count of a motif is the product between the total number of words in the database (n) and the probability of the motif in the database:

$$\hat{N}_m(w) = n\,\mu$$

For example, for the symbolic word *baccdfah* the probabilities are calculated as follows:

M0	$\mu = \frac{N(b)\,N(a)\,N(c)\,N(c)\,N(d)\,N(f)\,N(a)\,N(h)}{n_s^8}$
M1	$\mu = \frac{N(ba)\,N(ac)\,N(cc)\,N(cd)\,N(df)N(fa)\,N(ah)}{n_s\,N(a)\,N(c)\,N(c)\,N(d)\,N(f)\,N(a)}$
M6	$\mu = \frac{N(baccdfa)\,N(accdfah)}{3n\,N(accdfa)}$

The model order m is selected according to the length of the subwords composition we are interested since we know Mm depends on its subwords of length m and $m+1$.

4.3 Assessing Statistical Significance

The expected counts have been estimated by a probabilistic model (Markov chains). However, expected counts by themselves do not provide enough information regarding the significance of motifs. Statistical hypothesis tests are widely used to help in decision making. In this setting, a null hypothesis is defined and then it is tested whether there is enough evidence in the data to reject that hypothesis. In motif discovery, the null hypothesis means that the given pattern is spurious or uninteresting, i.e. the actual motif count is similar to the expected one. It means that if the motif count happens to be greater than expected, given that motif composition, it is so solely by chance. The null hypothesis is rejected in favor of the alternative hypothesis. In our case, that the motif has a frequency significantly greater than the expected count. After the hypothesis definition, it is necessary to define a test statistic and characterize its distribution. Our subject of interest is the motif count. Motifs counts distribution in the observed time series can be characterized as follows. Let the motif observed count w be:

$$N(w) = \sum_{i=1}^{n} Y_i$$

where Y_i is the Bernoulli random variable:

$$Y_i = \begin{cases} 1 & \text{if } w \text{ occurs in position } i \text{ in database } D \\ 0 & \text{otherwise} \end{cases}$$

with probability $p(Y_i) = \mu$. The motif count $N(w)$ is a sum of Bernoulli random variables. Therefore it follows a Binomial distribution:

$$N(w) \sim \mathcal{B}(n, \mu)$$

Note that the possible dependence between the different motifs is not an issue in our approach. Each motif count is treated independently of the others. However, we assume each instance (occurrence) of a motif is independent of one another. We can not guarantee that this assumption holds, due to the internal dynamics of the process that generated the time series at hand. Motif statistical significance is assessed by means of the *p-value*: the probability of the test statistic to present the observed value or a more extreme one, if the null hypothesis is true. That is to say, given the distribution for test statistic (the motif count), the p-value is the probability of the motif count to be at least as large as the observed motif count, just by chance. It can be calculated by the probability of the $\mathcal{B}(n, \mu)$ random variable to be at least as large as $N(w)$. It is calculated by the complement of the Binomial cumulative density function, as follows:

$$\mathbb{P}(\mathcal{B}(n,\mu) \geq N^{obs}(w)) = 1 - \sum_{k=0}^{N(w)-1} \binom{n}{k} \mu^k (1-\mu)^{n-k}$$

The p-value is then compared to a predefined critical value (α). If it is no greater than α, the null hypothesis is rejected and the pattern is accepted as significant. In the literature, the critical value is typically set to 0.05. However, not considering the multiple hypothesis problem and fixing a value as the significance level tends to increase the false discovery rate [11]. We use the Holm adjusted significance level (α') to control the number of false discoveries in the entire time series. This topic will be discussed in detail in section 4.5.

Besides the use of p-values to accept motifs that are statistically significant, they can also be used to sort the motifs of a given time series. This permits to achieve a rank of motifs according to their significance. If a p-value is very small, the motif is significantly frequent (over-represented).

4.4 Approximating p-values

To calculate p-values using the exact Binomial cumulative density function can be a computationally expensive operation, if n and k are large. This is the case in massive real-world data. Further, one should consider that the test must be executed for all extracted motifs. Approximate or asymptotic distributions are widely used in the literature [19, 21, 29, 30, 32, 33], as

they can reduce the computation time by one order or magnitude (see section 5). This difference stretches out along the size of the Binomial parameters. Typically, it is better to compute a computationally lighter analytic expression. They theoretically converge to the correct value as the sample size tends to infinity.

The Poisson approximation has been shown to fit correctly observed counts of words [33]. Assuming this approximation, the motif count has mean and variance λ, i.e.

$$N(w) \sim \mathcal{P}(\lambda), \text{ with } \lambda = n\mu$$

The p-value is approximated by the tail distribution of the Poisson distribution:

$$\mathbb{P}(\mathcal{P}(\lambda) \geq N^{obs}(w)) = 1 - e^{-\lambda} \sum_{i=0}^{N(w)-1} \frac{\lambda^i}{i!}$$

The Gaussian approximation has also been used to approximate motif counts in bioinformatics. In this distribution, the motif count has mean $n\mu$ and variance $n\mu(1-\mu)$. That is,

$$N(w) \sim \mathcal{N}(n\mu, n\mu(1-\mu))$$

The p-value can be approximated by the following expression:

$$\mathbb{P}(\mathcal{N}(\mu, \sigma^2) \geq N^{obs}(w)) = 1 - \frac{1}{2}\left[1 + \frac{erf\Big(N(w) - 1 - \mu\Big)}{\sqrt{2\sigma^2}}\right]$$

where $erf(x)$ is the Gauss error function and is calculated as follows:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

The quality of the described approximations is experimentally analysed in section 5.

4.5 Controlling the risk of false discoveries

In classical hypothesis testing, the p-value is compared to the defined α significance level. In mining for statistical significant motifs we apply a test for each discovered motif, i.e. the number of tests applied is the number of distinct motifs (N_d). If α is set to 0.05 and we apply 100000 simultaneous tests to motifs that follow the null hypothesis, one would expect to find 5000 significant motifs by chance alone [8]. The larger the number of executed tests, the higher the chance to find at least one that incorrectly rejects the null hypothesis. This issue is known as the multiple hypothesis testing problem and occurs when multiple statistic hypothesis tests are performed simultaneously [8, 39]. This will cause some patterns to be discovered in error i.e. false discoveries derivation. To control the false discovery rate one can apply the Bonferroni adjustment [8], the classical and most simple approach. The approach adjusts α to $\alpha' = \alpha/n$, where n is the number of hypothesis tests performed. However this value tends to be extremely strict [8, 39]. An alternative method is the Holm procedure [11]. This method provides a more reasonable α' level, while still maintaining the experimentwise significance level to α. The adjusted significance level is calculated as follows: all p-values are sorted increasingly from the smallest p_1 until p_n. For all $1 \leq j \leq n$, α' is set to the maximum p-value p_j that rejects $p_j \leq \alpha/(n-j+1)$ [39]. We use the Holm adjusted α' for all tests, as shown in the experimental analysis section (5).

5 Experimental Analysis

In this section we describe the experiments performed using the proposed approach to analyze the statistical significance of time series motifs. First, the experimental methodology is outlined. Then, the datasets and their sources are described. Finally, our approach is applied to datasets from various application domains and results are shown. The quality of the Poisson and Gaussian approximations is evaluated according to existing measures.

5.1 Methodology

Motifs are extracted from the data using the MrMotif algorithm, with $K = \infty$, i.e. all patterns are extracted. See section 4.1 for the algorithm selection discussion. The iSAX *length* and *resolution* parameters are both set to 8, resulting in a $\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7\}$. The significance level (α) of the tests is automatically adjusted to cope with multiple testing. Instead of setting α to a typical value such as 0.05, we automatically derive the adjusted threshold using the Holm procedure [11]. The Java implementation provided by MrMotif [3] authors is used. The Colt Library for High Performance Scientific and Technical Computing (v1.2.0) in Java is used for computing the Binomial, Poisson and Gaussian p-values. This library has been shown to provide accurate (long tail region) small p-values [34]. The approach was implemented in the Java language and compiled using JDK 6. All experiments were executed in a machine with a Intel® Core^{TM} i5-530 processor with 4GB of RAM.

Our experimental methodology proceeded as follows. First, we extract frequent motifs from each of the presented datasets and calculate their statistical significance using the proposed approach. The number of

statistical significant motifs (according to a significant threshold) is analyzed. A p-value is derived for each motif, assisting in the ranking of the different motifs. Then, the quality of the Poisson and Gaussian p-value approximations is compared, using several measures, to the Binomial Exact value. The aim of this work is not to provide proof of correctness for the statistical tests. Their theoretical properties are well established. Rather, we aim at showing their applicability and impact in the time series motif evaluation setting.

For clarity, we choose to use only one order for the Markov model from which we derive the motif expected probabilities. The chosen order is the maximal order (M6). We believe that this maximal order is the most representative of the significance we are interested in. However, calculations using smaller orders are also valid and should be used when the application at hand justifies it. Motifs of possible different sizes are accounted by treating each time series subsequence as a different time series (see section 3).

5.2 Datasets

We aim to test our approach on data from a wide range of applications and sizes. A set of 52 time series datasets available in the literature are selected from several sources. From [40], projectile shapes (*arrowhead*), brain activity (*eeg*) and motion-capture (*mocap*) data. Electrooculogram (*eog*) data from [24]. Sensor networks monitoring (*sensorsnetwork*), telecommunication traffic (*telecom*) and protein data (*sasa*) from [3]. Random walk data (*10*) from [26]. Data from chlorine concentration measurements (*cl2*), Electrocardiogram (*koskiecg*), star light curves (*lightcurves*), graphical passwords (*pen*), exchange rate (*tickwise*) from [37]. From [15] we choose respiration (*nprs*), power demand (*powerdata*) and space shuttle data (*TEK*). Finally, datasets from a variety of sources are aggregated in [13]: airplane sensor data (*attas*), elastic burst (*burst*, *burstin*), chaotic time series (*chaotic*), sea level pressure (*darwin*), earthquake (*earthquake*), ECG (*ecg*), EEG heart rate (*eegheartrate*), brain imaging (*ERP*), fluid dynamics (*fluid*), Fortune 500 data (*fortune*), explosion sound (*infrasound*), laser measurements (*laser*), leaf images (*leaf*), electric signal (*leleccum*), logistic surrogate noisy data (*logistic*), fault detection (*mallat*), memory (*memory*), muscle activation (*muscle*), network (*network*), ocean depth (*ocean*, *oceanshear*), network packet delay (*packet*), power plant (*powerplant*), random walk (*random*), EEG (*rateeg*), image shape (*shapemixed*), standard and poor index (*sp*), speech recording (*speech*), stocks (*stock*), sunspots (*sunspot*), synthetic control charts (*synthetic*), and water level observations (*tide*) data.

5.3 Motif Statistical Significance Results

In this subsection the proposed approach is applied to the 52 different datasets generating more than 110000 distinct motifs. The statistically significant motifs returned by the approach are shown. The goals of the experimental analysis are: to show the pruning power of our approach, to highlight that it allows to avoid the use of unintuitive support of Top-K parameters as a pruning mechanism, to discuss whether p-value based motif ranking is an interesting approach and ultimately, to show the need for statistical tests in time series motifs mining. We first analyze the relation between sequence length (n), number of discovered motifs (N_d), number (NSM) and percentage (%) of significant motifs, and the adjusted cutoff value (α) for several datasets. In table 1 we show these outcomes. Results for all datasets are omitted for brevity and can be consulted in [2].

Table 1: Motif results for several datasets

Dataset	n	N_d	NSM	α'	%
ERP	47616	2628	95	1.97E-05	3.61
eog	67493	5882	95	8.64E-06	1.62
rateeg	576694	100438	95	4.98E-07	0.09
lightcurves	5327	376	70	0.000163	18.62
cl2	4310	54	36	0.002632	66.67
sasa	81280	754	29	6.89E-05	3.85
koskiecg	2394	360	24	0.000148	6.67
mallat	803	30	18	0.003846	60.00
motor	420	60	7	0.000926	11.67
stocks	18000	1394	7	3.6E-05	0.50
arrowheads	1231	161	5	0.000318	3.11
pen	510	46	4	0.001163	8.70
burstin	1310	221	4	0.000229	1.81
powerdata	1838	295	4	0.000171	1.36
shapemixed	160	14	2	0.003846	14.29
10000	10000	754	2	6.64E-05	0.27
TEK	180	51	1	0.00098	1.96
eegheartrate	373	85	1	0.000588	1.18
leaf	442	72	1	0.000694	1.39
network	1121	36	1	0.001389	2.78
insect	1471	77	1	0.000649	1.30
chaotic	109	4	0	0.0125	0
random	1718	65	0	0.000769	0
fortune	500	9	0	0.005556	0
logistic	2000	181	0	0.000276	0
packet	2332	187	0	0.000267	0
tide	2906	6	0	0.008333	0
eeg	62700	2767	0	1.81E-05	0

We can observe that larger datasets generate a larger number of frequent motifs. This is expected, since frequent motifs can be found even in random data. We can also see that a larger number of significant motifs are also extracted from larger datasets. Nevertheless, in terms of percentage, there is no clear relation between dataset size and significant motifs. This is a result of considering the motif count in the adjusted cutoff value calculation (Holm procedure). Our approach

prunes most of the false discoveries, since most of the motifs are not statistically significant. The percentage of accepted motifs is small for most of the datasets. For some datasets, all frequent motifs were discarded. Despite some of these data are large, no frequent motif could reject the null hypothesis. This indicates that using statistical tests in time series motif discovery can act as a filter, pruning meaningless motifs. This seems to support the need for statistical tests in time series motif discovery. Pruning the prohibitively large output of pattern discovery algorithms is typically done by support or (Top) K parameters. These parameters are unintuitive and are typically optimized by experimentation. However, this is untenable in practice since the data are massive and it becomes very difficult to re-run the algorithms with a new parameter setting. Assessing motifs p-values avoids the use of unintuitive parameters. Since the adjusted cutoff value is automatically derived by our approach, no threshold setting is necessary to find the most statistically significant patterns in the dataset.

An interesting byproduct of our approach is that the motifs can be ranked according to their statistical significance, i.e. their p-value. To be able to rank motifs is important, since a ranking yields a smooth way to select the patterns in the database that are most representative and relevant. The domain expert can further investigate those patterns for significance in the domain at hand. In table 2 the highest ranked motifs for five of the datasets are presented. For simplicity, the numeric symbols are converted to alphabetic ones (respecting the alphabet index, i.e. $a = 0$ up to $h = 7$). Results for all datasets and full ranks (up to the least ranked motif) can be accessed in [2].

It can be observed that motifs with the smallest p-value, i.e. highest ranking, present a large difference between their expected count and actual number of occurrences. The ranking produced by the approach is calculated by using statistical tests, which are well established in the literature. Therefore, they reflect the degree of difference between expected and observed motif counts, which is the aim of the motif's p-value based ranking. Typically, the ground-truth motifs are not available in time series data, as the motif discovery process is unsupervised. To obtain a ground-truth about time series motifs can only be achieved by a domain expert, motif utility in a specific task (e.g. symbolic language) or interpretability [22]. Even in the presence of a domain expert, some of the errors a motif discovery algorithm can incur are justified by real patterns that are simply unexpected [22]. By introducing statistical tests in time series motif discovery, we intend to shed light on the motifs that are considered to present

Table 2: Most statistically significant motifs for several datasets

Datasets	Motif	$N(w)$	μ	Expected	p-value
sasa	gggfcbbb	17	3.9E-05	3.172479	4.77E-08
	hggdcbbb	8	8.79E-06	0.7143	8.93E-07
	bbbbgggg	14	3.37E-05	2.735099	1.19E-06
	bbbcggfg	10	1.67E-05	1.354194	1.68E-06
	abbdgggg	7	7.16E-06	0.58183	2.7E-06
eog	aacefggg	31	8.79E-05	5.932245	3.69E-13
	caacfggh	11	6.36E-06	0.429089	1.54E-12
	babbeggh	12	8.78E-06	0.592607	2.27E-12
	dbdgggfa	11	7.38E-06	0.497955	7.41E-12
	gabdeggd	12	1.03E-05	0.695669	1.2E-11
cl2	heddddbe	74	0.00193	8.319006	3.98E-13
	hecdccdf	37	0.001998	8.613394	7.54E-13
	hedcdccd	645	0.049903	215.0832	9.33E-13
	hedddcce	80	0.006069	26.1573	1.06E-12
	hedddccd	64	0.004855	20.92584	1.23E-12
koskiecg	gdddddbg	40	0.002734	6.544641	2.37E-12
	dddddbfh	34	0.00299	7.157086	2.88E-12
	heddddb	43	0.006027	14.42812	7.89E-10
	dddddbgh	22	0.001817	4.350855	1.49E-09
	dbggdddd	45	0.00719	17.21198	1.55E-08
mallat	dgbcdche	90	0.03608	28.97219	6E-13
	cgbcdche	97	0.041707	33.49079	6.16E-13
	dgbbdche	92	0.038283	30.74089	6.57E-13
	dgbcdcge	59	0.024542	19.70757	7.29E-13
	dhbcdcge	137	0.056988	45.76165	7.92E-13

the highest statistical significance. As widely mentioned in the literature, statistical significance does not imply significance in a specific domain. However, to use the highest ranked motifs can provide a good starting point for the experts analysis. For example, the 5 highest ranked statistical significant motifs in protein unfolding data can provide the user a starting point to analyze the database for interesting motifs in that specific application. It is important that the expert considers only 5 motifs rather than 754. In some cases, when the number of returned motifs makes the manual analysis very difficult, the use of p-value based rankings may become a requirement. We can also observe that motifs with the highest p-value also exhibit a large frequency. That is expected, since significant motifs are those whose frequency exceed their estimated frequency. There is no clear relation between motif count ranking and p-value ranking. However, some of the motifs with high frequencies are in the top p-value rankings, and vice-versa.

We show another practical example to highlight the relevance of the ranks generated by our approach. The most significant motif (showing the smallest p-value) from the *koskiecg* is displayed in Fig. 5. This motif is a well-known pattern in ECG data - the K-complex [26].

5.4 Measuring the Poisson and Gaussian Approximations Quality

The exact Binomial p-value calculation is computationally expensive for extremely large time series and motif

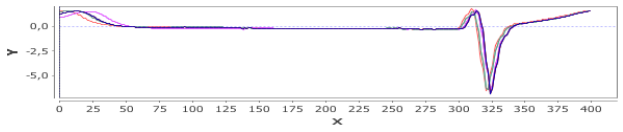

Figure 5: Motif with highest statistical significance in dataset *koskiecg*.

counts. For example, with n=100000 and k=5000, the approximated p-value can be calculated about one order of magnitude faster than the exact one. It is therefore important to evaluate the quality of the p-value derived by approximated approaches. In this work, two measures are used to quantify the agreement among the p-values produced by the different tests. The root mean square error (RMSE) is widely used to measure the difference between estimated values and actual values in prediction algorithms, for example. Hereby it is used to quantify the difference between the Binomial and the Poisson and Gaussian approximated p-values. It is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(E_i - O_i)^2}$$

where E_i is the Binomial test p-value for motif i, and O_i the approximation test's (Poisson or Gaussian) p-value.

The Total Variation Distance (d_{TV}) between the exact distribution p and its approximation $\hat{p}$, measures the greatest error one can make, in terms of probability, when using $\hat{p}$ instead of p:

$$d_{TV}(p,\hat{p}) = \sup_{A\subset\mathbb{N}} \mid p(A) - \hat{p}(A) \mid = \frac{1}{2}\sum_{n\geq 0} \mid p(n) - \hat{p}(n) \mid$$

In this subsection we calculate the RMSE and d_{TV} of the Binomial exact (B) and the Poisson approximation (P), for all datasets. Then, the same measures are applied to the Binomial and Gaussian approximation (G). The results for each measure are averaged for all datasets. In table 3 the average and standard deviation of the executed calculations are shown.

Table 3: RMSE and d_{TV} average and standard deviation

	RMSE(B.P)	dTV(B.P)	RMSE(B.G)	dTV(B.G)
Average	0.000193	0.002103	0.124324	24.6976
Std. Dev.	0.000251	0.00228	0.032015	105.1292

We can observe that the Poisson approximation is highly accurate, as both RMSE and d_{TV} present a very small average (and standard deviation), for all datasets. Therefore, it can be used as a replacement for the Binomial distribution. The Gaussian approximation however, presents relatively large RMSE (average of about 12%) and d_{TV} values. These results support the experiments presented in [30], which has concluded that the Gaussian approximation is not suited to motifs.

To explore a possible relation between approximation quality and dataset size, the datasets are grouped in 4 groups of 13 datasets each and sorted according to their length. Results for the group RMSE average are shown in table 4.

Table 4: RMSE averages for each increasingly sized dataset interval.

N	Average RMSE(B.P)	Average RMSE(B.G)
1-180	0.000519	0.147843
188-600	0.000184	0.13305
803-1838	4.93E-05	0.121433
2000-576694	2E-05	0.094968

It can be observed that the RMSE and d_{TV} decrease as dataset increase in size, i.e. N grows larger, for both approximations. These results suggest that the approximation quality improves with dataset length. This result is somehow expected, since both Binomial and Gaussian approximations are asymptotic and are assumed to converge to the correct result as N grows to infinity.

We have studied the difference between exact and approximated p-values. It is also important to study whether p-values are under or over-estimated by each representation. To answer this question we have plotted all motifs for 9 of the datasets and their location in the chart with respect to the identity function ($f(x) = x$). Fig. 6 compares the Binomial and Poisson, and Fig. 7 the Binomial and Gaussian approximated p-values.

It can be observed that the Poisson and Binomial p-values are mostly situated on the identity function line. This is expected as results show that these two distributions yield very similar p-values (RMSE and d_{TV} comparison). The larger difference is between the Gaussian and Binomial results. It can be observed that most of the points in the scatterplot are above the identity function line. This means that the Gaussian approximation over-estimates p-values and by consequence underestimates statistical significance.

6 Conclusion

We have proposed an approach to evaluate the significance of time series motifs using statistical significance tests. Our approach innovates by computing, for the first time in the literature, each time series motif p-value and accepts a motif as significant if its value is smaller than an automatically derived significance level.

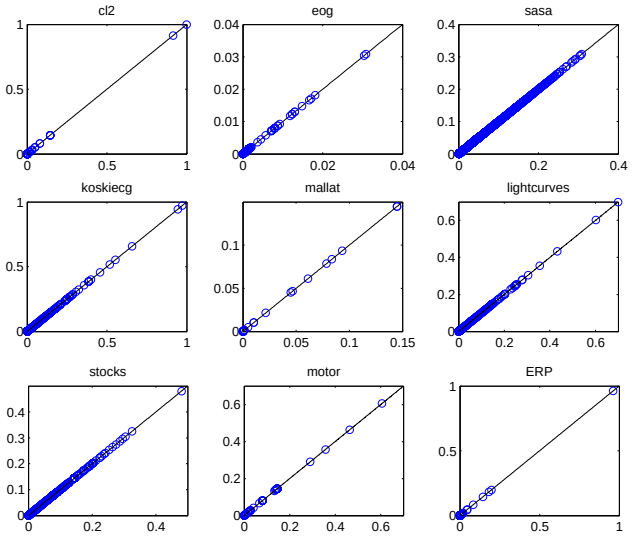

Figure 6: p-values of the Binomial (X axis) vs. p-values of the Poisson approximation (Y axis). The diagonal line is the graphical representation of the identity function.

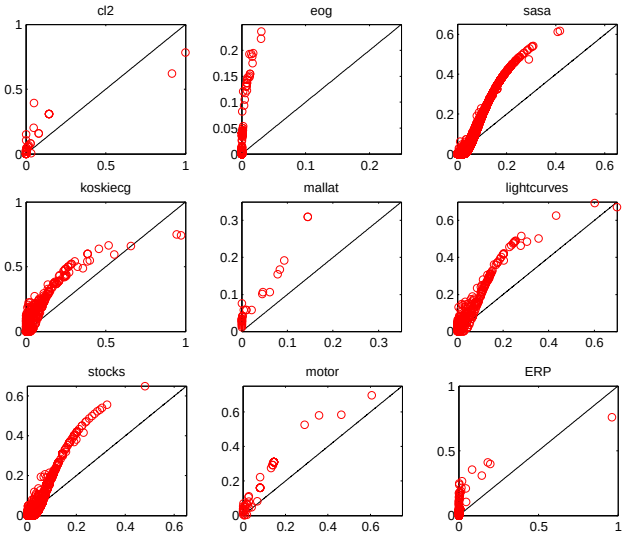

Figure 7: p-values of the Binomial (X axis) vs. p-values of the Gaussian approximation (Y axis).

This circumvents the need to define unintuitive parameters like support or top-K in motif discovery algorithms. Further, it significantly reduces the number of returned patterns. An interesting byproduct is the ranking of motifs obtained by considering their statistical significance. We believe our approach provides researchers and practitioners with an important technique to evaluate the degree of relevance of each extracted motif. We also aim to highlight the importance of evaluating motifs since it is crucial to make motif mining an useful task in practice. Future work includes expanding our proposal to other types of evaluation measures, and to study the power of the used statistical tests.

Reproducibility Note

All experiments, data and source code used in this paper are available online at [2].

References

[1] V. Boeva, J. Clément, M. Régnier, M. Roytberg, and V. Makeev, *Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules*, Algorithms for molecular biology, 2 (2007), p. 13.

[2] N. Castro, *Time series motifs statistical significance website*. `http://www.di.uminho.pt/~castro/stat`.

[3] N. Castro and P. Azevedo, *Multiresolution Motif Discovery in Time Series*, in Proceedings of the Tenth SIAM International Conference on Data Mining, 2010, pp. 665–676.

[4] B. Chiu, E. Keogh, and S. Lonardi, *Probabilistic discovery of time series motifs*, in Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, p. 498.

[5] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, *Querying and mining of time series data: experimental comparison of representations and distance measures*, Proceedings of the VLDB Endowment, 1 (2008), pp. 1542–1552.

[6] P. Ferreira, P. Azevedo, C. Silva, and R. Brito, *Mining approximate motifs in time series*, in Discovery Science, Springer, 2006, pp. 89–101.

[7] P. G. Ferreira and P. J. Azevedo, *Evaluating protein motif significance measures: A case study on prosite patterns*, in IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007), 2007, pp. 171–178.

[8] S. Hanhijärvi, K. Puolamäki, and G. Garriga, *Multiple Hypothesis Testing in Pattern Discovery*, stat, 1050 (2009), p. 29.

[9] H. He and A. Singh, *Graphrank: Statistical modeling and mining of significant subgraphs in the feature space*, in Sixth International Conference on Data Mining (ICDM'06)., 2006, pp. 885–890.

[10] J. Hollunder, M. Friedel, A. Beyer, C. Workman, and T. Wilhelm, *DASS: efficient discovery and p-value calculation of substructures in unordered data*, Bioinformatics, 23 (2007), p. 77.

[11] S. Holm, *A simple sequentially rejective multiple test*

procedure, Scandinavian Journal of Statistics, 6 (1979), pp. 65–70.

[12] S. Jacquemont, F. Jacquenet, and M. Sebban, *Mining probabilistic automata: a statistical view of sequential pattern mining*, Machine Learning, 75 (2009), pp. 91–127.

[13] E. Keogh and T. Folias, *The UCR Time Series Data Mining Archive. Riverside CA*, University of California-Computer Science & Engineering Department, (2002).

[14] E. Keogh and S. Kasetty, *On the need for time series data mining benchmarks: A survey and empirical demonstration*, Data Mining and Knowledge Discovery, 7 (2003), pp. 349–371.

[15] E. Keogh, J. Lin, and A. Fu, *HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence*, in Proceedings of the Fifth IEEE International Conference on Data Mining, IEEE Computer Society, 2005, p. 233.

[16] E. Keogh, S. Lonardi, and B. Chiu, *Finding surprising patterns in a time series database in linear time and space*, in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 550–556.

[17] J. Lin, E. Keogh, S. Lonardi, and P. Patel, *Finding motifs in time series*, in Proc. of the 2nd Workshop on Temporal Data Mining, Citeseer, 2002, pp. 53–68.

[18] C. Low Kam, A. Mas, and M. Teisseire, *Mining for unexpected sequential patterns given a Markov model*. Unpublished, 2000.

[19] T. Marschall and S. Rahmann, *Efficient exact motif discovery*, Bioinformatics, 25 (2009), p. i356.

[20] C. Matias, S. Schbath, E. Birmelé, J. Daudin, and S. Robin, *Network motifs: mean and variance for the count*, REVSTAT–Statistical Journal, 4 (2006), pp. 31–51.

[21] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network motifs: simple building blocks of complex networks*, Science, 298 (2002), p. 824.

[22] D. Minnen, T. Starner, I. Essa, and C. Isbell, *Improving activity discovery with automatic neighborhood estimation*, in Proceedings of the 20th international joint conference on Artifical intelligence, Morgan Kaufmann Publishers Inc., 2007, pp. 2814–2819.

[23] F. Mörchen and A. Ultsch, *Efficient mining of understandable patterns from multivariate interval time series*, Data Mining and Knowledge Discovery, 15 (2007), pp. 181–215.

[24] A. Mueen and E. Keogh, *Online Discovery and Maintenance of Time Series Motifs*, in Proceedings of the sixteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 1089–1098.

[25] A. Mueen, E. Keogh, and N. Bigdely-Shamlo, *Finding Time Series Motifs in Disk-Resident Data*, in 2009 Ninth IEEE International Conference on Data Mining, 2009, pp. 367–376.

[26] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover, *Exact discovery of time series motifs*, in Proceedings of the Ninth SIAM International Conference on Data Mining (SDM), 2009, pp. 473–484.

[27] G. Nuel, *Effective p-value computations using Finite Markov Chain Imbedding(FMCI): application to local score and to pattern statistics*, Algorithms for Molecular Biology, 1 (2006), p. 5.

[28] T. Oates, *PERUSE: An unsupervised algorithm for finding recurring patterns in time series*, IEEE ICDM, Japan, 2 (2002), p. 5.

[29] F. Picard, J. Daudin, M. Koskas, S. Schbath, and S. Robin, *Assessing the exceptionality of network motifs*, Journal of Computational Biology, 15 (2008), pp. 1–20.

[30] M. Régnier and M. Vandenbogaert, *Comparison of statistical significance criteria*, Journal of Bioinformatics and Computational Biology, 4 (2006), pp. 537–552.

[31] P. Ribeca and E. Raineri, *Faster exact Markovian probability functions for motif occurrences: a DFA-only approach*, Bioinformatics, 24 (2008), p. 2839.

[32] S. Robin and S. Schbath, *Numerical comparison of several approximations of the word count distribution in random sequences*, Journal of Computational Biology, 8 (2001), pp. 349–359.

[33] S. Robin, S. Schbath, and V. Vandewalle, *Statistical tests to compare motif count exceptionalities*, BMC bioinformatics, 8 (2007), p. 84.

[34] S. Santosh Bangalore, J. Wang, and D. Allison, *How accurate are the extremely small p-values used in genomic research: An evaluation of numerical libraries*, Computational statistics & data analysis, 53 (2009), pp. 2446–2452.

[35] S. Schbath, *An overview on the distribution of word counts in Markov chains*, Journal of Computational Biology, 7 (2000), pp. 193–201.

[36] S. Schbath, *Statistics of motifs*, Atelier de formation, 1502 (2006).

[37] J. Shieh and E. Keogh, *iSAX: indexing and mining terabyte sized time series*, in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 623–631.

[38] Y. Tanaka, K. Iwamoto, and K. Uehara, *Discovery of time-series motif from multi-dimensional data based on mdl principle*, Machine Learning, 58 (2005), pp. 269–300.

[39] G. Webb, *Discovering significant patterns*, Machine Learning, 68 (2007), pp. 1–33.

[40] D. Yankov, E. Keogh, J. Medina, B. Chiu, and V. Zordan, *Detecting time series motifs under uniform scaling*, in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 844–853.

[41] J. Zhang, B. Jiang, M. Li, J. Tromp, X. Zhang, and M. Zhang, *Computing exact P-values for DNA motifs*, Bioinformatics, 23 (2007), p. 531.