Significant Motifs in Time Series

Nuno C. Castro* and Paulo J. Azevedo

HASLab / INESC TEC, University of Minho, Braga, Portugal

Received 16 June 2011; revised 15 November 2011; accepted 30 November 2011
DOI:10.1002/sam.11134
Published online in Wiley Online Library (wileyonlinelibrary.com).

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 the data mining communities to evaluate 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. © 2012 Wiley Periodicals, Inc. Statistical Analysis and Data Mining, 2012

Keywords: time series; motif discovery; statistical significance tests; significant patterns

1. INTRODUCTION

Extracting 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 [1]. 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 three different motifs (displayed in blue, green, and red), as typically outputted by existing motif discovery algorithms.

Since the problem formulation in ref. 2, many proposals on how to extract motifs from a time series database have been introduced [1,3–12]. 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. Restraining the most frequent motifs to expert analysis 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 randomly generated database of length 65 536 from the work of Keogh and Folias [13], for example, 65 motifs are discovered. The top motif reaches four 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 shown 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 [14]; in networks analysis, to find significant subgraphs [15]; in association rules mining to discard redundant rules [16]. 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

* *Correspondence to:* Nuno C. Castro (castro@di.uminho.pt)

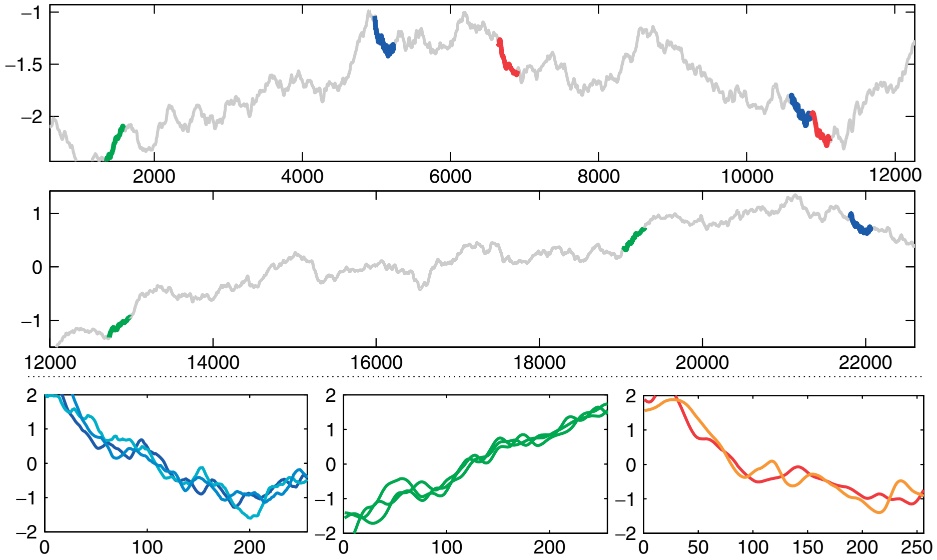


Fig. 1 Example of a time series with several motifs. Above: in its original context; below: detail of each motif. Blue and green: three instances; red: two instances. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

as to 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 [14]. 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: one is to provide an approach to assess the statistical significance of time series motifs and the other 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 properly evaluate 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 7 we derive conclusions.

2. RELATED WORK

Since the introduction of the time series motif discovery problem [2], many approaches have been proposed [1,4–12]. 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 Ferreira and Azevedo [17], motif mining evaluation measures can be classified in the following categories: class-based, theoretical 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. Theoretical 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 theoretical 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*. [18] 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 [20] 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 ref. 21, 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*. [22] 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 following a Binomial distribution. The p-values that measure over-representation are then calculated. Hollunder *et al*. [23] 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 [24] 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 ref. 25, the drawbacks of the Gaussian approximation are analyzed. Schbath [26] studies the statistical distributions of word counts in Markov chains. Formulae are derived for the estimated expected counts under these distributions. In ref. 14, 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 modeled 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 ref. 27, a binomial test is used to evaluate the statistical significance of frequent subgraphs in a database of chemical compounds' graphs. Milo *et al*. [15] 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 carried out in complex networks from biochemistry, neurobiology, ecology, and engineering. In ref. 28 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 modeling data. Ribeca and Raineri [29] 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*. [30] 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 ref. 31, the authors consolidate a decade of research in biosequence 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 [1] 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. Keogh *et al*. [3] 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*. [32] 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 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 [33].

We are typically interested in mining a collection of time series with arbitrary lengths.

DEFINITION 2: A *time series database* D is a set of $|D|$ unordered time series [9].

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

DEFINITION 3: Given a time series T of length n, a *time series subsequence* $S = s_i, \ldots, s_{i+m-1}$ is a sampling of $m \leq n$ contiguous positions of T, such that $1 \leq i \leq n - m + 1$ (definition from ref. 3).

For simplicity, we treat each subsequence of database as a different time series in D. In practice, sliding 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 [3].

Since our work is inspired by the biosequences' motif mining, we are interested in the symbolic representation of time series and their subsequences.

DEFINITION 4: A *word* $w = w_1 w_2 \ldots w_l$ 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 ref. 34. The symbolic aggregate approximation (iSAX) [35] representation takes a time series as input and transforms it into a sequence of symbols, as shown in Fig. 2.

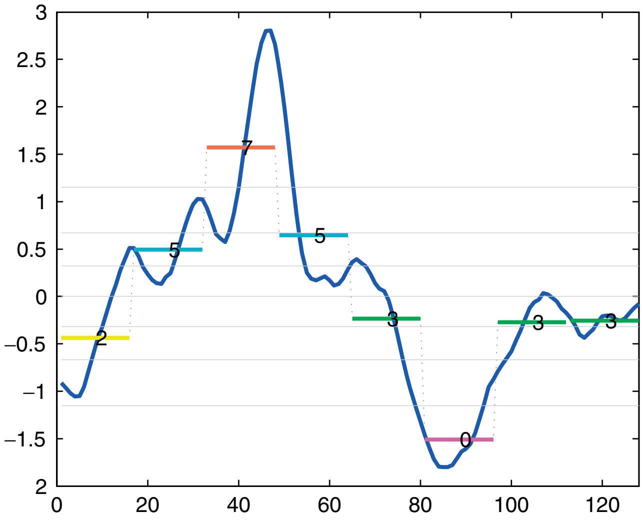


Fig. 2 Conversion of a time series into its iSAX representation, generating word {2, 5, 7, 5, 3, 0, 3, 3}. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

As shown in Fig. 3, slightly different subsequences can originate the same word in the representation space. These subsequences are called instances of the given word.

DEFINITION 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:

DEFINITION 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 three instances is shown in Fig. 3.

DEFINITION 7: The word w is a *Motif* in database D if the count of all instances of w in D is greater than 1.

DEFINITION 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.

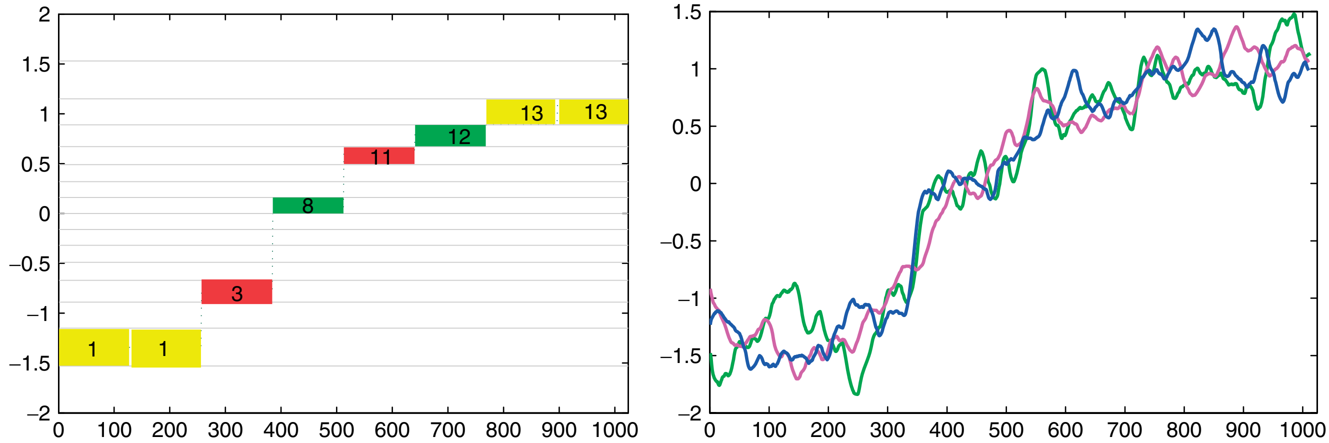


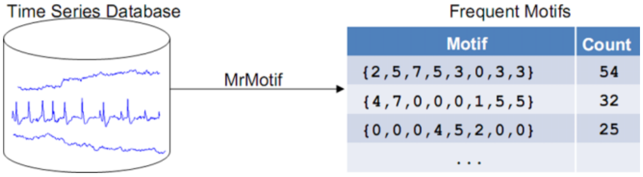
Fig. 3 Motif {1, 1, 3, 8, 11, 12, 13, 13} (left) and its three instances in the database (right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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 [9] 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 [3,12]. 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 [12] 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 Fig. 4.



Motif	Count
{2,5,7,5,3,0,3,3}	54
{4,7,0,0,0,1,5,5}	32
{0,0,0,4,5,2,0,0}	25
...	

Fig. 4 Extraction of frequent motifs from the time series database. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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 refs. 18,19,28.

4.2. Reference Model

On its own, motif support is not a good interestingness measure. Similarly, to support in item sets mining, frequency does not guarantee that motifs are significant. A trivial example highlighting this problem is shown in Section 1 (random time series data are used). This example shows that patterns can be found to be frequent even in random data. However, those frequent patterns are meaningless because they were randomly generated. A better approach is to consider the difference between the observed motif count and the motif expected count. The expected count is the number of motifs one should expect in random sequences similar to our database (under some similarity definition). This knowledge is obtained considering a reference model that reflects the background distribution of the motifs.

Random sequences are typically modeled using Bernoulli trials or Markovian sequences [26]. The former assume that words are i.i.d., although word symbols are possibly neither independent nor identically distributed in real data [36]. The latter take the composition of the words into account. That is, they consider the time dependency characteristic of time series data. Also, there are analytical probability calculations available which prevents expensive simulations [36]. A Markov chain is a mathematical system to model random processes. It is composed by states and the transitions between them in a chainlike fashion. That makes them suitable to model sequential symbolic data, where each state is a symbol in the sequence. They have been widely used [15,20,26,27,29,30]. A simple Markov chain is shown in Fig. 5.

The time dependency is considered by assuming that the distribution of each symbol depends on its previous symbols. In particular, each symbol depends on the symbols that immediately precede it, rather than its entire past (Markov property) [37]. Given a random sequence of symbols from an alphabet $\mathscr{A} = \{0, 1, 2, 3\}$:

$$S = (X_1, X_2, \ldots, X_i, \ldots)$$

the probability of each symbol depends only on the recent past of the sequence, i.e. it depends only on the symbols $(X_{i-m}, \ldots, X_{i-1})$. The integer m is the order of the Markov chain model. A Markov chain model of order m is denoted by Mm. For example, in model M1, each symbol depends only on the previous one. This dependence is expressed by the probability to transit from one symbol to the next. This probability is the transition probability $\pi(x, y)$, where $x, y \in \mathscr{A}$. Formally, the probability that X_i is at state y, given X_{i-1} is at state x is:

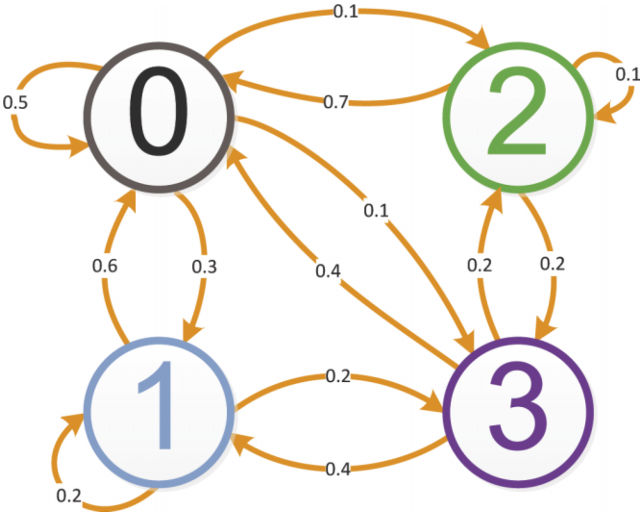


Fig. 5 Example of a Markov chain. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$\pi(x, y) = \mathbb{P}\{X_i = y | X_{i-1} = x\}$$

The set of all transition probabilities form a transition matrix [37]. For example, the transition matrix for $\mathscr{A}$ is:

$$\Pi = \begin{pmatrix} \pi(0,0) & \pi(0,1) & \pi(0,2) & \pi(0,3) \\ \pi(1,0) & \pi(1,1) & \pi(1,2) & \pi(1,3) \\ \pi(2,0) & \pi(2,1) & \pi(2,2) & \pi(2,3) \\ \pi(3,0) & \pi(3,1) & \pi(3,2) & \pi(3,3) \end{pmatrix}$$

Considering the Markov chain shown in Fig. 5, we have:

$$\Pi = \begin{pmatrix} 0.5 & 0.3 & 0.1 & 0.1 \\ 0.6 & 0.2 & 0 & 0.2 \\ 0.7 & 0 & 0.1 & 0.2 \\ 0.4 & 0.4 & 0.2 & 0 \end{pmatrix}$$

The general model Mm, for arbitrary order m, is defined by the following transition probabilities:

$$\pi(x_1 x_2, \ldots x_m, y) =$$
$$\mathbb{P}\{X_i = y \mid X_{i-m} = x_1, X_{i-m+1} = x_2, \ldots, X_{i-1} = x_m\}$$

In our approach, we have previously extracted frequent motifs from the time series database. Fig. 6 shows a motif that has been extracted from the database (above) and the transition probabilities for that motif (below), for the several orders from M1 up to M6. We follow the approach described in refs. 14 and 37, where Markov chains are used 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. As they are *a priori* unknown, they are estimated according to the observed sequence. In fact, they are estimated according to the corresponding observed counts $N(xy)$, i.e. their support. For example, under M1:

$$\pi(x, y) = \frac{N(xy)}{N(x+)}$$

where $N(xy)$ is the number of times y follows x in the database and $N(x+)$ the number of times x occurs in the database, followed by any other symbol. In practice, we

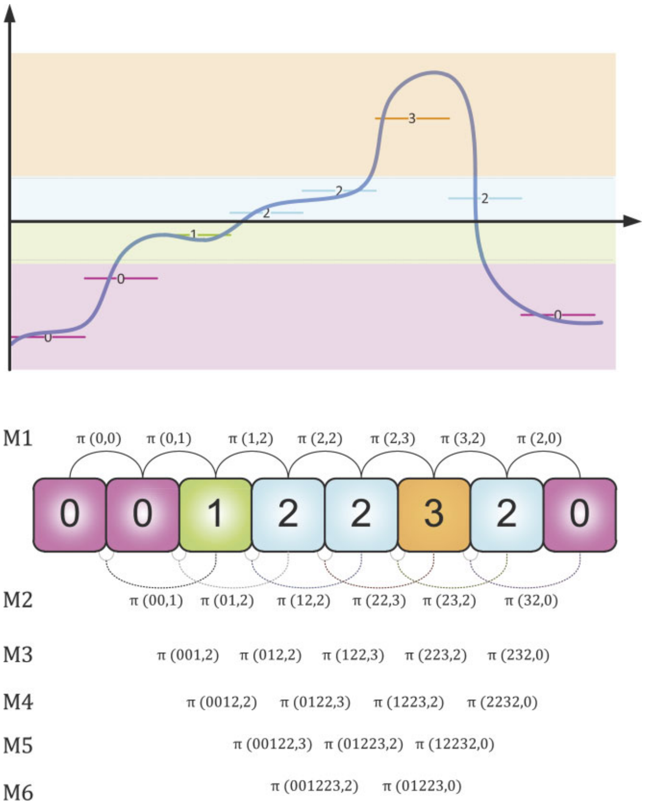

0
0
1
2
2
3
2
0
M1
π (0,0) π (0,1) π (1,2) π (2,2) π (2,3) π (3,2) π (2,0)
0 0 1 2 2 3 2 0
M2
π (00,1) π (01,2) π (12,2) π (22,3) π (23,2) π (32,0)
M3
π (001,2) π (012,2) π (122,3) π (223,2) π (232,0)
M4
π (0012,2) π (0122,3) π (1223,2) π (2232,0)
M5
π (00122,3) π (01223,2) π (12232,0)
M6
π (001223,2) π (01223,0)

Fig. 6 The transition probabilities for orders M1–M6 (below) of the {00122320} motif (above). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

replace $N(x+)$ with the close approximate $N(x)$. For the general model Mm, we have:

$$\pi(x_1x_2,\dots x_m, y) = \frac{N(x_1x_2,\dots x_m y)}{N(x_1x_2,\dots x_m+)}$$

The observed counts contain all the information necessary to estimate the transition probabilities. The probability μ of a word $w = w_1w_2,\dots w_l$ can be estimated by the product of the transition probabilities. For M1, we have:

$$\mu(w) = \mu(w_1) \times \pi(w_1, w_2)\times \cdots \times \pi(w_{l-1}, w_l)$$

Generalizing to Mm:

$$\mu(w) = \mu(w_1,\dots w_m) \times \pi(w_1,\dots w_m, w_{m+1}) \\ \times \cdots \times \pi(w_{l-m},\dots w_{l-1}, w_l)$$

As we can estimate transition probabilities using the observed counts for each subsequence, the probability of a word is calculated using the observed counts of its subwords of length m and $m+1$. For example, the probabilities μ for the motif {00122320} shown in Fig. 6 are calculated as follows:

M0	$\mu = \dfrac{N(0)\,N(0)\,N(1)\,N(2)\,N(2)\,N(3)\,N(2)\,N(0)}{n_s^8}$
M1	$\mu = \dfrac{N(00)\,N(01)\,N(12)\,N(22)\,N(23)N(32)\,N(20)}{n_s\,N(0)\,N(1)\,N(2)\,N(2)\,N(3)\,N(2)}$
M2	$\mu = \dfrac{N(001)\,N(012)\,N(122)\,N(223)\,N(232)N(320)}{7n\,N(01)\,N(12)\,N(22)\,N(23)\,N(32)}$
M3	$\mu = \dfrac{N(0012)\,N(0122)\,N(1223)\,N(2232)\,N(2320)}{6n\,N(012)\,N(122)\,N(223)\,N(232)}$
M4	$\mu = \dfrac{N(00122)\,N(01223)\,N(12232)\,N(22320)}{5n\,N(0122)\,N(1223)\,N(2232)}$
M5	$\mu = \dfrac{N(001223)\,N(012232)\,N(122320)}{4n\,N(01223)\,N(12232)}$
M6	$\mu = \dfrac{N(0012232)\,N(0122320)}{3n\,N(012232)}$

where $N(x)$ is the count of motif x in the total (symbolic) length of the time series database n_s. We generalize the formulae for calculating μ in M0, M1 and the general order M$(l-2)$. For a motif of length l, the maximal order is $l-2$.

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$(l-2)$	$\mu = \dfrac{N(w_1,\dots w_{l-1})\,N(w_2,\dots w_l)}{n\,(l-m+1)\,N(w_2,\dots w_{l-1})}$

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:

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

Model influence: The order of the model (m) is selected according to the length of the subwords composition we are interested, as it is known that Mm depends on its subwords of length $m+1$ and m. For example, in M2 we consider random sequences having on average the same composition as the observed sequence (regarding their subsequences of length 3 and 2). Fig. 7 shows an example of subsequences of length 3 and 2 that compose M2 and possible instances

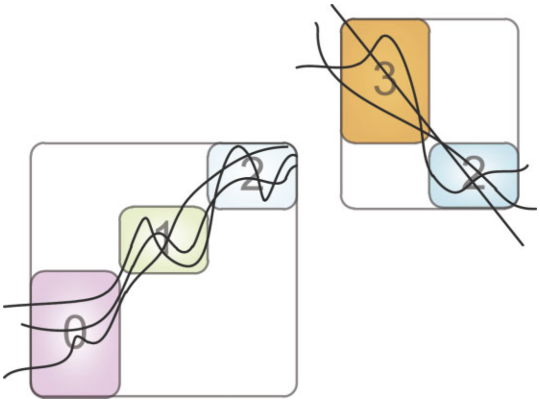


Fig. 7 Example of subsequences of length 3 and 2 that compose M2. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

that match submotif {012} (left) and {32} (right). Using a resolution (alphabet size) higher than 4 narrows the size of each block and allows to match more similar motifs [12].

The larger the order of the model, the larger the subsequences that are selected to compose the model. For that reason, larger models capture more information and mimic better the occurrences of a given motif. However, they may miss important information contained in the smaller subsequences (e.g. of length 2 and 3). A more suitable solution could be to consider all the orders and study how a motif's significance varies along the increasing order of the model [37]. For instance, the high expected frequency presented by a motif can be caused by a very frequent subsequence of length 1. However, in this work we focus on a single model at a time. The model is essential to determine whether a motif is exceptional or not [37].

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. We declare this motif statistically significant. After the hypothesis definition, it is necessary to define a test statistic and characterize its distribution. Our subject of interest is the motif count. Motif counts' distribution in the observed time series can be characterized as follows. Let the observed motif 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 cannot guarantee that this assumption holds, due to the internal dynamics of the process that generated the time series at hand. Statistical significance of motif 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:

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

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 tend to increase the false discovery rate [38]. We use the Holm adjusted significance level (α') to control the number of false discoveries in the entire time series. This topic is 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 [14,15,19,24,25,31], as they can reduce the computation time by one order or magnitude (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 [14]. Assuming this approximation, the motif count has mean and variance λ, i.e.

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

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

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

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:

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

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

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

The quality of the described approximations is experimentally analyzed in Section 5.

4.5. Controlling the Risk of False Discoveries

In classical hypothesis testing, the p-value is compared to a fixed α significance level, such as 0.05 or 0.01. 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 100 000 simultaneous tests to motifs that follow the null hypothesis, one would expect to find 5000 significant motifs by chance alone [39]. 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 (MHTP) and occurs when multiple statistic hypothesis tests are performed simultaneously [16,39]. This will cause some motifs to be erroneously declared significant, i.e. false discoveries derivation. Traditionally, to control the number of false discoveries the significance level is set to values stricter than 0.05 or 0.01, to avoid an abundance of false positives [40].

The Bonferroni adjustment [39] is the classical and most simple approach to tackle the problem. The approach adjusts α to $\alpha' = \alpha/d$, where d is the number of hypothesis tests performed—one test for each distinct motif N_{d}. However, this value tends to be extremely strict [16,39]. An alternative method is the Holm procedure [38]. This method provides a more reasonable α' level, while still maintaining the experiment-wise 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_d. For all $1 \leq j \leq d, \alpha'$ is set to the maximum p-value p_j that rejects $p_j \leq \alpha/(d - j + 1)$ [16].

The aim of both the Bonferroni and Holm procedures is to control the probability of at least one false positive, i.e. controlling the risk of committing even one type I error across the entire family of hypothesis tests—family-wise error rate (FWER) [41]. This type of strong control is necessary when even a single false positive would be disastrous. For example, the hypothesis that several drugs are safe to use [41], or the prediction of an earthquake leading to a city's evacuation. However, most of the time the FWER is too conservative [40], as guarding against any single false positive occurring will lead to missed findings (false negatives—type II error). That is to say, the strong control is achieved at the expense of true motifs that are missed. In some applications, to enforce not having even one false motif is not necessary and is too strict [42]. In particular, in problems where the number of hypothesis tested is very large, the probability to make any discovery at all becomes very small [43]. However, it may prove more suitable for applications where the number of significant motifs is extremely large. In some other applications, where it may be more appropriate to identify hypothesis for further study (exploratory analysis), it can end up being too strict.

Table 1. Outcomes of the motif statistical significance approach.

	Declared non-significant	Declared significant	Total
Spurious motifs	U	V	d_0
True motifs	T	S	$d - d_0$
	$d - R$	R	d

We are in the presence of two extremes: the abundance of false positives caused by ignoring the multiple hypothesis problem and the too strict control of FWER approaches leading to potential important motifs being missed. To bridge these two extremes, the false discovery rate (FDR) was introduced [42]. It is the expectation of the proportion of rejected true null hypothesis, among the rejected hypothesis [43]. That is, the expected proportion of falsely rejected hypothesis. It aims to describe the number of erroneous rejections, rather than whether any single error was committed [42]. Formally, we simultaneously test d hypothesis (motifs), of which d_0 are true. Table 1 shows the summary of the approach outcomes:
where R is an observable count and U, V, S, and T are unknown. The FDR can be defined by dividing the number of false positives by the number of significant motifs output by the approach: the random variable $Q = V/(V + S) = V/R$ — the proportion of rejected null hypothesis erroneously rejected [42]. The FDR is different from the false-positive rate. Although the former is the rate that spurious motifs will be declared significant $V/(V + U)$, the latter is the rate that motifs declared significant are spurious. A false-positive rate of 5% means that, on average, 5% of the spurious motifs in the database will be called significant. An FDR of 5% means that, on average, 5% of the motifs called significant are spurious [40]. For that reason, it provides a sensible balance between true and false positives (S and V).

The FDR can be controlled at level q using the Linear step up procedure [42,43]. Similarly to the Holm procedure, we sort all p-values increasingly from the smallest p_1 until p_d. Let

$$k = \max\left\{i : P_i \leq \frac{iq}{d}\right\}$$

then k motifs associated with $P_1, \ldots, P_k$ are significant. That is, we compare sequentially the ordered p-values to the constants linearly interpolated between q and q/m [43], and keep rejecting the null hypothesis (declaring motifs as significant) until one fails to reject $P_i \leq iq/d$. This procedure controls the FDR at level q (typically set to 0.05) and aims to maximize the number of significant motifs for this level. That is, it assumes one should expect a given number of false discoveries and focuses on estimating what the FDR actually is [44]. Thus, more significant motifs will be derived by this approach. However, it allows the user to specify what percentage of the discovered motifs are spurious. In this work we compare the Holm and FDR approaches.

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. Second, the datasets and their sources are described. Then, our approach is applied to datasets from various application domains and the results are shown. The scalability of the approach is next analyzed. Finally, 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 [38] and FDR [42] approaches. The Java implementation provided by MrMotif [12] 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 [45]. 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 4 GB 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 statistically significant motifs is analyzed, considering the comparison between the Holm and FDR approaches to control the MHTP. A p-value is derived for each motif, assisting in the ranking of the different motifs. The scalability of the proposed approach is studied next. 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 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. The use of a model combining several orders remains as open research. Motifs of possible different sizes are accounted by treating each time series' subsequence as a different time series (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 ref. [6], projectile shapes (*arrowhead*), brain activity (*eeg*), and motion-capture (*mocap*) data. Electrooculogram (*eog*) data are from ref. 10. Sensor networks monitoring (*sensorsnetwork*), telecommunication traffic (*telecom*), and protein data (*sasa*) are from ref. 12. Random walk data (*10*) are from ref. 9. Data from chlorine concentration measurements (*cl2*), Electrocardiogram (*koskiecg*), star light curves (*lightcurves*), graphical passwords (*pen*), exchange rate (*tickwise*) are from ref. 35. From ref. 46 we choose respiration (*nprs*), power demand (*powerdata*), and space shuttle data (*TEK*). Finally, datasets from a variety of sources are aggregated in ref. 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 section the proposed approach is applied to the 52 different datasets generating more than 110 000 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, to compare the Holm and FDR approaches in our data and ultimately, to show the need for statistical tests in time series motifs' mining. We first analyze the relation between sequence length (n_s), number of discovered motifs (N_d), number (NSM) and percentage (%) of significant motifs, and the cutoff level (α and α'). We show results for several datasets and false discoveries controlling approaches − the standard 0.05, Holm and FDR − in Table 2.

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.

In particular, when the MHTP is not considered ($\alpha = 0.05$), the number of significant motifs is large. Using the Holm procedure, however, there is no clear relation between dataset size and significant motifs in terms of percentage. The Holm 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. Using the FDR to control false discoveries increases the number of discovered motifs, as expected, when compared to the Holm approach. The increase is large for some of the datasets. In some of the data, most of the frequent motifs clearly exceed their expectation (e.g. rateeg, ERP, cl2). In these cases, the FDR control is too permissive; it tends to converge to the standard 0.05 control. In those applications, to further reduce the FDR rate or using any FWER approach may prove a more suitable solution. Despite it appears that higher percentages of accepted motifs are present in larger datasets, some large datasets present a small portion of accepted motifs. The relation between FDR accepted motifs and dataset size is not linear. Nevertheless, the results suggest larger datasets present a larger percentage of FDR controlled significant motifs.

The main conclusion to draw from these results is that the control of false discoveries is clearly required. Using the standard significance level at 0.05, without considering the MHTP, generates large percentages of significant motifs. It seems clear that most of these motifs are meaningless. The results suggest that the decision on which technique to control false discoveries is application dependent. In this work, we cover the Holm and FDR approaches. The user should select the approach taking several aspects into consideration. First, if the data will be further analyzed by the domain expert, we suggest the FDR approach. It outputs more statistically significant motifs, giving the opportunity to have more powerful interpretability. In the case of a prohibitively large number of significant motifs, the FDR rate should be made stricter (i.e. smaller than 0.05). Considering only the higher ranked motifs (in terms of significance) is also desirable. In case the motif mining

Table 2. Motif results for all datasets.

			0.05			Holm			FDR		
Dataset	n_s	N_d	NSM	α	%	↓ NSM	α'	%	NSM	Cutoff	%
ERP	47 616	2 620	2 612	0.05	99.69	757	2.684E-05	28.89	2 612	4.987E-02	99.69
rateeg	576 694	100 438	70 009	0.05	61.20	281	4.381E-07	0.25	45 030	1.968E-02	39.36
lightcurves	5 327	376	205	0.05	54.52	70	1.634E-04	18.62	154	2.061E-02	40.96
cl2	4 310	54	44	0.05	81.48	36	2.778E-03	66.67	43	4.074E-02	79.63
koskiecg	2 394	344	178	0.05	51.74	23	1.558E-04	6.69	106	1.555E-02	30.81
mallat	803	30	24	0.05	80.00	18	4.167E-03	60.00	22	3.833E-02	73.33
motor	420	60	25	0.05	41.67	8	9.615E-04	13.33	16	1.417E-02	26.67
stocks	18 000	1 402	532	0.05	37.95	7	3.584E-05	0.50	25	9.272E-04	1.78
arrowheads	1 231	161	51	0.05	31.68	5	3.205E-04	3.11	13	4.348E-03	8.07
pen	510	46	17	0.05	36.96	4	1.190E-03	8.70	4	5.435E-03	8.70
sasa	81 280	8 146	5 498	0.05	67.49	4	6.141E-06	0.05	403	2.480E-03	4.95
eog	67 493	3 101	1 558	0.05	50.24	3	1.614E-05	0.10	34	5.643E-04	1.10
10	10 000	754	259	0.05	34.35	2	6.649E-05	0.27	5	3.979E-04	0.66
powerdata	1 838	291	128	0.05	43.99	2	1.730E-04	0.69	52	9.107E-03	17.87
shapemixed	160	14	6	0.05	42.86	2	4.167E-03	14.29	2	1.071E-02	14.29
TEK	180	50	19	0.05	38.00	2	1.042E-03	4.00	3	4.000E-03	6.00
burstin	1 310	224	84	0.05	37.50	1	2.242E-04	0.45	18	4.241E-03	8.04
eegheartrate	373	85	38	0.05	44.71	1	5.952E-04	1.18	1	1.176E-03	1.18
insect	1 471	73	20	0.05	27.40	1	6.944E-04	1.37	2	2.055E-03	2.74
leaf	442	72	29	0.05	40.28	1	7.042E-04	1.39	1	1.389E-03	1.39
network	1 121	42	12	0.05	28.57	1	1.220E-03	2.38	1	2.381E-03	2.38
sensorsnetwork	555	9	1	0.05	11.11	1	6.250E-03	11.11	1	1.111E-02	11.11
attas	96	5	0	0.05	0	0	1.000E-02	0	0	1.000E-02	0
burst	488	4	0	0.05	0	0	1.250E-02	0	0	1.250E-02	0
chaotic	109	1	0	0.05	0	0	5.000E-02	0	0	5.000E-02	0
darwin	171	12	1	0.05	8.33	0	4.167E-03	0	0	4.167E-03	0
earthquake	209	6	0	0.05	0	0	8.333E-03	0	0	8.333E-03	0
ecg	56	8	2	0.05	25.00	0	6.250E-03	0	0	6.250E-03	0
eeg	62 700	2 880	1 107	0.05	38.44	0	1.736E-05	0	0	1.736E-05	0
fluid	1 662	17	1	0.05	5.88	0	2.941E-03	0	0	2.941E-03	0
fortune	500	9	0	0.05	0	0	5.556E-03	0	0	5.556E-03	0
infrasound	425	4	0	0.05	0	0	1.250E-02	0	0	1.250E-02	0
laser	194	29	11	0.05	37.93	0	1.724E-03	0	0	1.724E-03	0
leleccum	107	9	0	0.05	0	0	5.556E-03	0	0	5.556E-03	0
logistic	2 000	181	64	0.05	35.36	0	2.762E-04	0	0	2.762E-04	0
lsf5	1 496	7	1	0.05	14.29	0	7.143E-03	0	0	7.143E-03	0
memory	260	10	1	0.05	10.00	0	5.000E-03	0	0	5.000E-03	0
mocap	1 370	70	18	0.05	25.71	0	7.143E-04	0	0	7.143E-04	0
muscle	188	27	7	0.05	25.93	0	1.852E-03	0	0	1.852E-03	0
nprs	1 095	15	1	0.05	6.67	0	3.333E-03	0	0	3.333E-03	0
ocean	124	8	0	0.05	0	0	6.250E-03	0	0	6.250E-03	0
oceanshear	124	8	0	0.05	0	0	6.250E-03	0	0	6.250E-03	0
packet	2 332	195	66	0.05	33.85	0	2.564E-04	0	0	2.564E-04	0
powerplant	154	5	1	0.05	20.00	0	1.000E-02	0	0	1.000E-02	0
random	1 718	61	8	0.05	13.11	0	8.197E-04	0	0	8.197E-04	0
sp	921	28	3	0.05	10.71	0	1.786E-03	0	0	1.786E-03	0
speech	47	2	0	0.05	0	0	2.500E-02	0	0	2.500E-02	0
sunspot	146	3	0	0.05	0	0	1.667E-02	0	0	1.667E-02	0
synthetic	600	59	7	0.05	11.86	0	8.475E-04	0	0	8.475E-04	0
telecom	71	3	1	0.05	33.33	0	1.667E-02	0	0	1.667E-02	0
tick	903	16	1	0.05	6.25	0	3.125E-03	0	0	3.125E-03	0
tide	2 906	9	0	0.05	0	0	5.556E-03	0	0	5.556E-03	0
TS	30	9	0	0.05	0	0	5.556E-03	0	0	5.556E-03	0

approach output is not further inspected by experts, the Holm procedure seems the more suitable approach. It controls the probability of any single false positive. Second, the user should consider the weight of false positives versus false negatives in the application at hand. If missing a significant motif is more problematic than erroneously declaring a motif significant, the FDR control should be selected.

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. We consider only the FWER or FDR levels to control false discoveries, and we set these to the standard values in the literature (0.05). The adjusted cutoff value is then automatically derived by our approach. In practice, 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. Ranks organize the motifs in order of their evidence against the null hypothesis. Declaring a motif significant means that any other motif with more evidence against the null hypothesis should also be declared significant [40]. 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 motifs for significance in the domain of study. In Table 3 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 ref. 47.

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 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 [7]. 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 [7]. By introducing statistical tests in time series motif discovery, we intend to shed light on the motifs that are considered to present 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 expert's analysis. For example, the five 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 five motifs rather than 5000. 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. Significant motifs are patterns whose actual frequency clearly exceeds their expected frequency. In this sense we can expect significant motifs to be frequent. However, there are also significant motifs whose frequency is low. For example, if the expected frequency of a motif is 0.45 and the actual frequency is 2, it can be marked as significant. We also observe that some of higher ranked motifs are very similar to each other, for example in the cl2 dataset. This may suggest a real bias in the data worthy of further investigation by the domain expert. It remains as open research to analyze the potential merge of these very similar motifs.

The MrMotif parameter setting does affect the extracted motifs. For example, by slightly increasing the resolution and length of the SAX discretization, the number of possible motifs (*resolution*^{l}) increases significantly. This obviously impacts the significance computation. By having a much larger number of possible motifs, the probability of each individual motif becomes very low. This can lead to very low expected counts for motifs and inconsistent results for p-values. We follow the recommendation by SAX authors that 8 is a reasonable value for both SAX parameters and maintain these constant for all motifs, resulting in a consistent behavior. From our experience, this value is a good choice, as it provides a sensible balance. It guarantees a sufficiently large number of distinct motifs, but not too large to cause results degeneration.

The approach allows different length time series. However, all subsequences of a particular time series are limited to a fixed length. For different time series in the database the length can be different. As we are interested only in the shape of the series, if a series of length 1000 has a similar shape to a series of length 100 then they will match the same motif. In practice, this is achieved by seamlessly converting all series in the database to a SAX word of

Table 3. 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	gddddddbg	40	0.002734	6.544641	2.37E-12
	dddddbfh	34	0.00299	7.157086	2.88E-12
	hedddddb	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

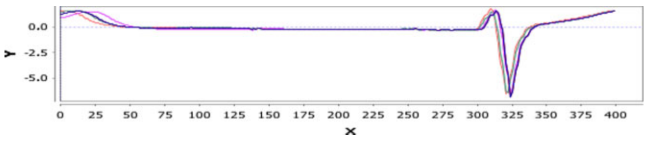


Fig. 8 Motif with highest statistical significance in dataset *koskiecg*. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

length 8. This enables different length series to match if they generate the same SAX word.

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. 8. This motif is a well-known pattern in ECG data—the K-complex [9].

5.4. Scalability Experiments

In this section we study the execution time for the proposed approach. We use ten different sets of increasing size, ranging from 10 000 to 100 000 time series of length 1024, from ref. 9. We use these data for two reasons: they have been used before and results on similar datasets are encouraged in order to walk toward data mining benchmarks [12]; in addition, the size (in the gigabytes)

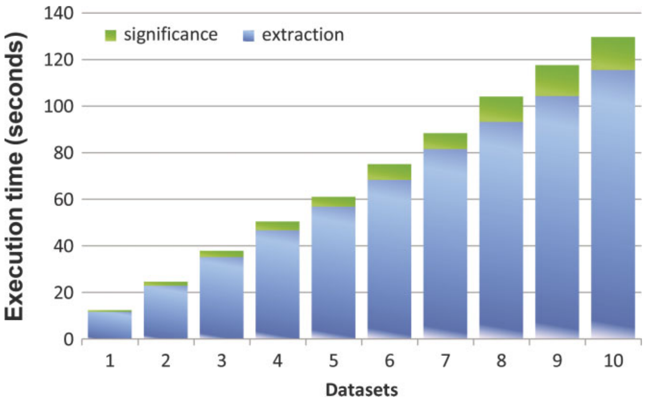


Fig. 9 Execution time of the approach in ten increasingly sized datasets. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

makes them attractive to test any approach. We apply our approach to each of the datasets. For each dataset, we record the motif extraction (MrMotif execution) and the statistically significant steps' duration. We show the results in Fig. 9.

We observe that the approach increases linearly with the size of the time series dataset. The motif extraction step takes about 90% and the motif statistical significance analysis 10% of the total execution time, for all datasets. In

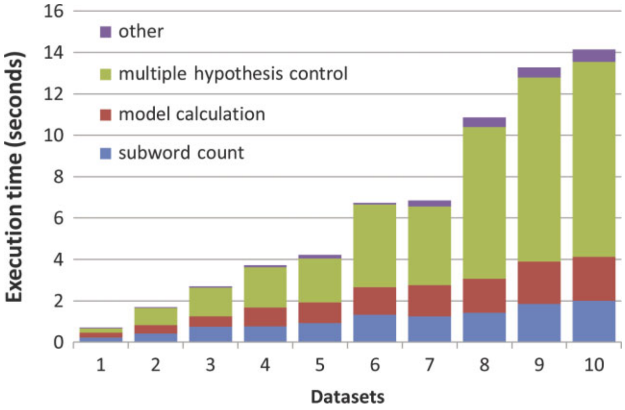


Fig. 10 Execution time of the statistical significant part. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Fig. 10 we zoom in the execution time of the statistical significance part of our approach.

The larger portion of the approach is taken by the multiple hypothesis control substep (Holm and FDR). This substep sorts the p-values in order to find the appropriate adjusted cutoff value. The $O(n \log n)$ complexity of the sorting operation can be perceived. The other substeps of our approach seem to be negligible, considering we are only looking at 10% of the total execution time. Nevertheless, they increase linearly with the number of time series analyzed.

To execute the motif extraction step (MrMotif) with $K = \infty$ does not decrease the performance of the approach. It always extracts and counts all motifs. The complexity of MrMotif is linear, as it performs only one sequential disk scan. Each motif is stored in a constant access time structure (hash table) and updated as the scan evolves. At the end of the scan we have the top-K motifs in main memory. In practice, to execute the algorithm with $K = 10$ or $K = \infty$ is equivalent.

5.5. Measuring the Poisson and Gaussian Approximations' Quality

The exact binomial p-value calculation is computationally expensive for extremely large time series and motif counts. For example, with $n = 100\,000$ 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, the Poisson and the Gaussian approximated p-values. It is calculated as follows:

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

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:

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

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 4 the average and standard deviation of the executed calculations are shown.

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 ref. 25, 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 four groups of 13 datasets each and sorted according to their length. Results for the group RMSE average are shown in Table 5.

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

Table 4. RMSE and d_{TV} average and standard deviation.

	RMSE(B.P)	d_{TV}(B.P)	RMSE(B.G)	d_{TV}(B.G)
Average	0.000193	0.002103	0.124324	24.6976
Std. Dev.	0.000251	0.00228	0.032015	105.1292

Table 5. 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
803–1 838	4.93E-05	0.121433
2 000–576 694	2E-05	0.094968

whether p-values are under- or over-estimated by each representation. To answer this question we have plotted all motifs for nine of the datasets and their location in the chart with respect to the identity function ($f(x) = x$). Fig. 11 compares the binomial and Poisson, and Fig. 12 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 under-estimates statistical significance.

6. FUTURE WORK

Several directions to extend this work are possible. Studying a Markov chain model using all the orders combined is currently underway. It may be interesting to analyze the impact of selecting another type of models, e.g. fractal models [48]. To adapt our approach to the streaming time series data is another interesting possibility. It remains to be investigated whether this adaptation is trivial. Our approach enables further research on time series motifs extraction algorithms, reference models, motif evaluation measures, and false discoveries controlling procedures. Any approach to tackle these tasks can be effectively plugged into our framework.

7. CONCLUSION

We have proposed an approach to evaluate the significance of time series motifs using statistical significance

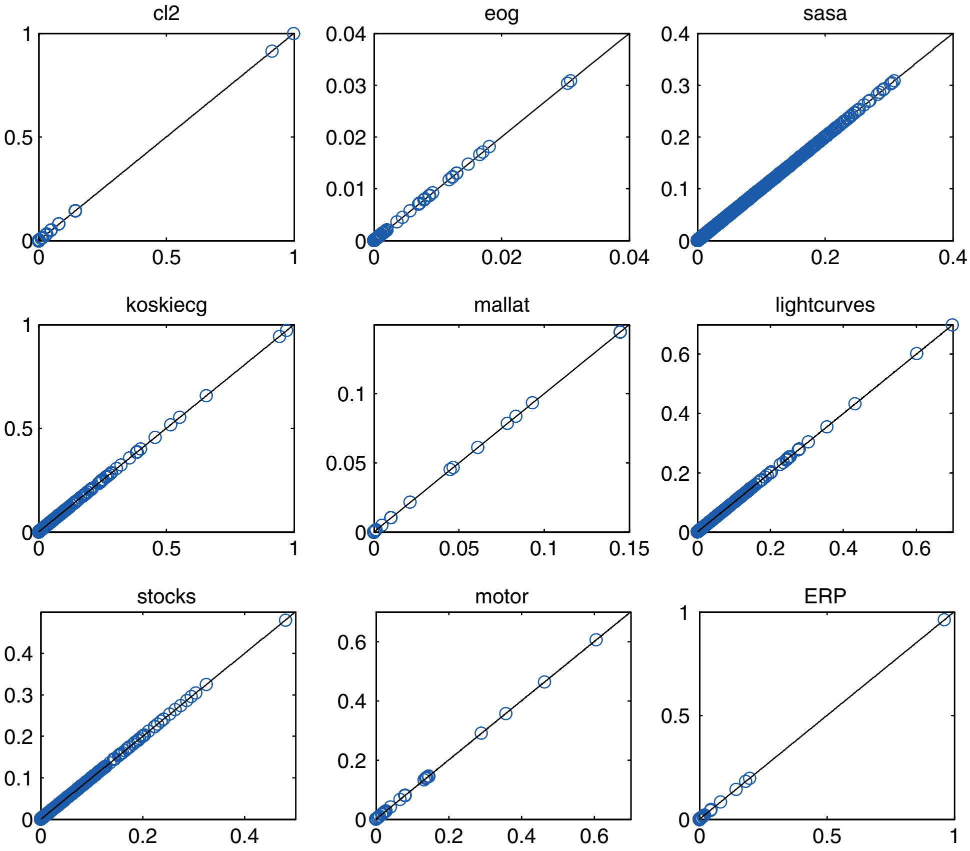


Fig. 11 p-Values of the binomial (X-axis) versus p-values of the Poisson approximation (Y-axis). The diagonal line is the graphical representation of the identity function. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

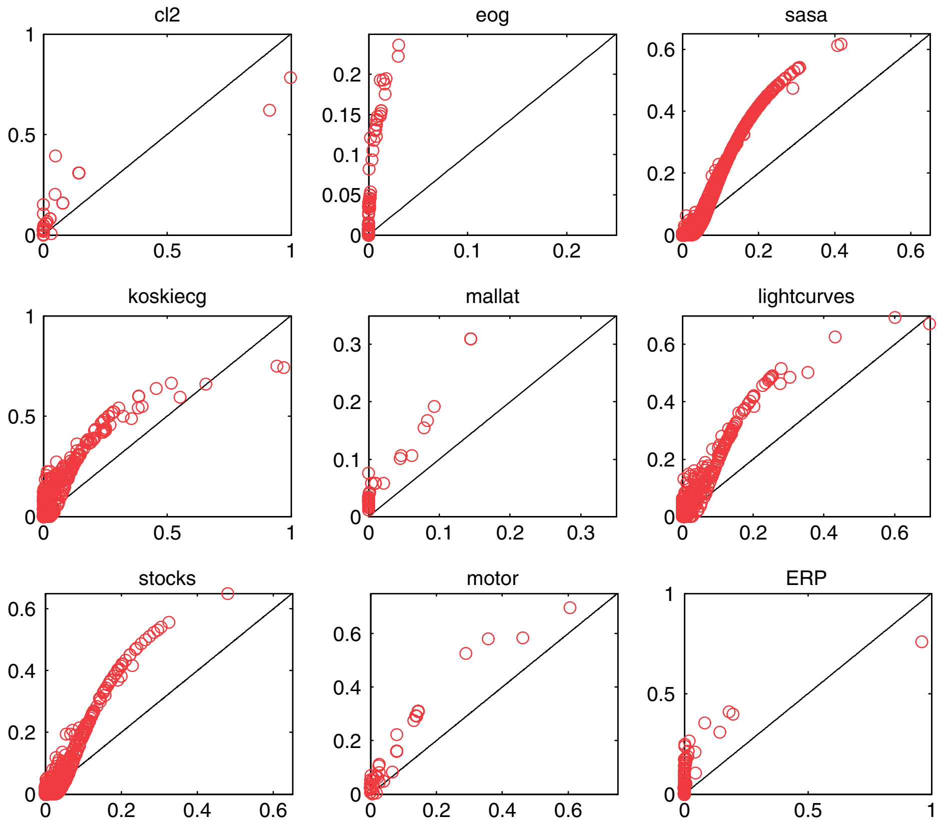


Fig. 12 p-Values of the binomial (X-axis) versus p-values of the Gaussian approximation (Y-axis). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

tests. Our approach innovates by computing, for the first time in the literature, each time series motif p-value and declares a motif significant if its p-value is smaller than an automatically derived significance level. 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 a useful task in practice.

ACKNOWLEDGMENTS

The authors are supported by *Fundação para a Ciência e a Tecnologia*. Nuno Castro is supported by grant SFRH/BD/33303/2008. Paulo J. Azevedo is supported by Project ProtUnf, FEDER, and Programa de Financiamento Plurianual de Unidades de I&D. We would also like to thank the reviewers who helped to significantly improve this paper with their invaluable feedback.

All experiments, data and source code used in this paper are available online [47].

REFERENCES

[1] P. Ferreira, P. Azevedo, C. Silva, and R. Brito, Mining approximate motifs in time series, in Discovery Science, Secaucus, New Jersey, Springer, 2006, 89–101.

[2] 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, 53–68.

[3] 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, 498.

[4] Y. Tanaka, K. Iwamoto, and K. Uehara, Discovery of time-series motif from multi-dimensional data based on mdl principle, Mach Learn 58(2) (2005), 269–300.

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

[6] 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, 844–853.

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

[8] F. Mörchen and A. Ultsch, Efficient mining of understandable patterns from multivariate interval time series, Data Min Knowl Discov 15(2) (2007), 181–215.

[9] 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, 473–484.

[10] 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, 1089–1098.

[11] 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, 367–376.

[12] N. Castro and P. Azevedo, Multiresolution motif discovery in time series, In Proceedings of the Tenth SIAM International Conference on Data Mining, 2010, 665–676.

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

[14] S. Robin, S. Schbath, and V. Vandewalle, Statistical tests to compare motif count exceptionalities, BMC Bioinformatics 8(1) 2007, 84.

[15] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network motifs: simple building blocks of complex networks, Science 298(5594) (2002), 824.

[16] G. Webb, Discovering significant patterns, Mach Learn 68(1) (2007), 1–33.

[17] 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, 171–178.

[18] J. Zhang, B. Jiang, M. Li, J. Tromp, X. Zhang, and M. Zhang, Computing exact P-values for DNA motifs, Bioinformatics 23(5) (2007), 531.

[19] T. Marschall and S. Rahmann, Efficient exact motif discovery, Bioinformatics 25(12) 2009, i356.

[20] G. Nuel, Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and to pattern statistics, Algorithms Mol Biol 1(1) 2006, 5.

[21] 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 Mol Biol 2(1) (2007), 13.

[22] C. Low Kam, A. Mas, and M. Teisseire, Mining for unexpected sequential patterns given a Markov model, 2008. http://www.math.univ-montp2.fr/∼mas/lmt_siam09.pdf.

[23] 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(1) (2007), 77.

[24] S. Robin and S. Schbath, Numerical comparison of several approximations of the word count distribution in random sequences, J Comput Biol 8(4) (2001), 349–359.

[25] M. Régnier and M. Vandenbogaert, Comparison of statistical significance criteria, J Bioinformatics Comput Biol 4(2) (2006), 537–552.

[26] S. Schbath, An overview on the distribution of word counts in Markov chains, J Comput Biol 7(1-2) (2000), 193–201.

[27] 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, 885–890.

[28] S. Jacquemont, F. Jacquenet, and M. Sebban, Mining probabilistic automata: a statistical view of sequential pattern mining, Mach Learn 75(1) (2009), 91–127.

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

[30] C. Matias, S. Schbath, E. Birmelé, J. Daudin, and S. Robin, Network motifs: mean and variance for the count, REVSTAT Stat J 4(1) (2006), 31–51.

[31] F. Picard, J. Daudin, M. Koskas, S. Schbath, and S. Robin, Assessing the exceptionality of network motifs, J Comput Biol 15(1) (2008), 1–20.

[32] 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, 550–556.

[33] E. Keogh and S. Kasetty, On the need for time series data mining benchmarks: a survey and empirical demonstration, Data Min Knowl Discov 7(4) (2003), 349–371.

[34] 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, Proc VLDB Endowment 1(2) (2008), 1542–1552.

[35] 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, 623–631.

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

[37] S. Robin, F. Rodolphe, and S. Schbath, DNA, Words and Models, New York, Cambridge Univ. Press, 2005.

[38] S. Holm, A simple sequentially rejective multiple test procedure, Scand J Stat 6(2) (1979), 65–70.

[39] S. Hanhijärvi, K. Puolamäki, and G. Garriga, Multiple hypothesis testing in pattern discovery, STAT 1050 (2009), 29.

[40] J. Storey and R. Tibshirani, Statistical significance for genomewide studies, Proc Natl Acad Sci USA 100(16) (2003), 9440.

[41] S. Hanhijärvi, K. Puolamäki, and G. Garriga, Multiple hypothesis testing in pattern discovery, Arxiv preprint arXiv:0906.5263, 2009.

[42] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J R Stat Soc B 57(1) (1995), 289–300.

[43] Y. Benjamini and M. Leshno, Statistical methods for data mining, Data Mining and Knowledge Discovery Handbook, Tel-Aviv, Israel, Springer, 2005, 565–587.

[44] H. Zhang, B. Padmanabhan, and A. Tuzhilin, On the discovery of significant statistical quantitative rules, In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, 374–383.

[45] 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, Comput Stat Data Anal 53(7) (2009), 2446–2452.

[46] 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, 233.

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

[48] Y. Gao, J. He, J. Zou, R. Zeng, and X. Liang, Fractal simulation of soil breakdown under lightning current, J Electrostat 61(3-4) (2004), 197–207.