Finding Surprisingly Frequent Patterns of Variable Lengths in Sequence Data

Reza Sadoddin* sadoddin@ualberta.ca
Joerg Sander* jsander@ualberta.ca
Davood Rafiei* drafiei@ualberta.ca

Abstract

We address the problem of finding 'surprising' patterns of variable length in sequence data, where a surprising pattern is defined as a subsequence of a longer sequence, whose observed frequency is statistically significant with respect to a given distribution. Finding statistically significant patterns in sequence data is the core task in some interesting applications such as Biological motif discovery and anomaly detection. We show that the presence of few 'true' surprising patterns in the data could cause a large number of highly-correlated patterns to stand statistically significant just because of those few significant patterns. Our approach to solving the 'redundant patterns' problem is based on capturing the dependencies between patterns through an 'explain' relationship where a set of patterns can explain the statistical significance of another pattern. This allows us to address the problem of redundancy by choosing a few 'core' patterns which explain the significance of all other significant patterns. We propose a greedy algorithm for efficiently finding an approximate *core* pattern set of minimum size. Using both synthetic and real-world sequential data, chosen from different domains including Medicine and Bioinformatics, we show that the proposed notion of *core* patterns very closely matches the notion of 'true' surprising patterns in data.

1 Introduction

Finding "surprising" patterns in sequence data is a key problem in many data mining applications in domains as diverse as Bioinformatics, Computer security and medicine. In Bioinformatics, the surprising patterns, often referred to as "motifs", are believed to have some important biological significance and regulate gene expressions [9]. *Motif discovery* in this domain is the problem of finding subsequences in a DNA sequence that are *overrepresented* relative to a background distribution. In computer security, anomalies may correspond to a sequence of commands or system calls executed by an attacker or a malicious program [12]. For time series data, "discords" have been defined as subsequences in a longer time series that are of maximal distance to their nearest neighbour(s) and are shown to capture anomalies in ECG data and space telemetry [19].

Finding surprising patterns without knowing their lengths poses several challenges. The length parameter is not intuitive and difficult to set, in many applications. Without knowing the true length(s), one should run the particular algorithm exhaustively, each time for a specific length. Our study shows that running a length-dependent method with different lengths of patterns will result in a large number of "highly-correlated" patterns, making it difficult to find 'true' surprising patterns. We refer to such presence of highly correlated patterns as the "redundant patterns" problem. The ineffective results produced by length-dependent methods motivate the development of a technique that generates a concise, redundancy-free set of statistically significant patterns. Another drawback of previous methods is the robustness of their results. The definition of surprising patterns and the required parameters have a large impact on the convenience of use and the robustness. For instance, a true surprising pattern might be observed a few times in a sequence. A definition that compares the observed frequency of a pattern with a fixed threshold might or might not capture the pattern based on a selected threshold value. Our experiments show that different parameter settings produce largely deviating results, making these methods less reliable in a real setting where the best parameter settings are unknown.

This paper addresses the aforementioned problems by eliminating the need for non-intuitive input parameters (such as length, frequency, *etc.*) and producing robust results represented by a concise, non-redundant set of relevant patterns. Another important factor in the utility of a surprising pattern technique is the variation allowed in the surprising patterns. In many applications, such as motif discovery, a surprising pattern (*e.g.* motif) is characterized by a group of subsequences which look 'similar' to each other noticeably, with a degree of 'variation'. Most traditional methods (*e.g.* HMM, KNN[7]) which are based on a simple string representation of patterns allow no variations among pattern occurrences.

The contributions of the paper are as follows: first, we provide a domain-independent formulation of the

*Department of Computing Science, University of Alberta

problem of finding surprising patterns based on statistical hypothesis testing. Second, we investigate the problem that a few embedded significant patterns can lead to a large number of "redundant" patterns. Third, we propose a statistical method that captures an "explain" relationship where a set of patterns can explain the statistical significance of another pattern. Fourth, using this "explain" relationship, we address the problem of redundancy by choosing a few 'core' patterns which explain the significance of all other significant patterns. Fifth, we extend the proposed model to capture a degree of variation in patterns, and propose approximation techniques to find the new patterns. Finally, we evaluate our methodology on both synthetic and real-world data, and compare it with anomaly detection and motif discovery methods.

2 Related Work

Chandola *et al.* perform a comparative evaluation of three types of anomaly detection techniques on sequence data [7], including the *Kernel-based*, *Windows-based*, and *Markovian* techniques. The Kernel-based category includes the Nearest Neighbour-based technique, in which a model is trained using normal sequences in a training phase, and each test sequence is compared against the trained model, where an *anomaly score* is computed from the closest 'distance' (or K^{th} nearest distance, in general) with the model. In the Window-based techniques (*e.g.* t-STIDE [36]), a normal profile is created from a dataset of normal sequences by extracting all windows of a fixed length w. In the test phase, all subsequences of length w are extracted from each test sequence and an 'anomaly score' is computed by comparing the frequencies of observed subsequences with those of existing subsequences in the normal profile. The main idea behind Markovian techniques is computing the probability of observing each symbol a in the test sequence conditioned on a limited number of symbols preceding the symbol a. The Markovian techniques used in Chandola's study are chosen from different models including Hidden Markov Model (HMM) [36], Probabilistic Suffix Tree (PST) [32], Finite State Automaton (FSA) [24], and FSAz (a variant of FSA [7]), and often use parameters such as length, probability threshold and K, which may not be easy to set. Moreover, all the methods used in this study consider a definition of a pattern with exact matching, which means that all occurrences of a pattern must exactly match with the subsequence represented by the pattern.

A more *generalized* notion of patterns has been targeted in some research works, in which the patterns approximately describe subsequences. Floratou *et al.* [11] have proposed a *motif model* that depends on four parameters (L, M, s, k), where L denotes the *length* of the pattern, M denotes the *distance matrix* (which is used to compute the similarity between a given string and the reference motif), s denotes the maximum distance threshold, and k denotes the minimum support required for a pattern to qualify as a *motif*. The model proposed in this work depends on parameters that cannot be determined intuitively.

The problem of *motif discovery* in the field of Bioinformatics is also related to our work. Many computational tools have been proposed for finding motifs of a specific length. Examples include AlignACE [18], ANN-Spec [38], Consensus [16], GLAM [13], Improbizer [4], MEME [5],, MotifSampler [34], QuickScore [30], SeSiMCMC [10], Weeder [28], and YMF [31]. While these computational tools have been developed particularly for finding motifs in DNA sequences, our work presented here proposes a general framework for finding *surprising* patterns in sequence data (not limited to Bioinformatics data). Also, the motif discovery methods use biology-motivated heuristics for finding the length of binding sites from a large number of statistically significant patterns. Furthermore, there are some evidences [28] that the patterns with highest scores do not necessarily correspond to true binding sites, leaving the problem open for more sound formulations such as ours.

Detecting anomalies in time series data is also related to our work [20, 19]. Keogh *et al.* introduce the problem of finding 'discords', which are subsequences of a longer time series that are maximally different from all other subsequences of the time series [19]. In this work, every subsequence is compared with its nearest neighbour (or K^{th} nearest neighbour, in general), and the one with the largest distance is returned as the top discord. The work in this category targets time series data (instead of symbolic data), and is also dependent on the input parameters K (*nearest neighbour* rank) and *length* (of subequences which are investigated as potential anomalies). Another line of research related to our work is 'motif' discovery in time series data. Motifs are defined as similar subsequences in time series data that are observed frequently [27], and different variations of motifs are studied by Yankov [8] and Castro [6]. These works are different from our work in the definition of patterns, the type of the data, and also the required input parameters.

Subsequence mining is a close research area to finding statistically significant patterns in sequence data. This topic has been addressed in numerous publications, including the seminal work by Agrawal *et al.* [1] and improvements proposed over Agrawal's work in algorithms such as SPADE [41] and BIDE [35]. The primary focus in this line of research is on mining a sequence of

symbols with arbitrary gaps between them, whereas our work is focused on finding contiguous patterns. Some algorithms such as cSPADE [40], CloSpan [39], PrefixSpan [29], Gap-BIDE [21], and Gap-Connect [21] allow certain constraints on the maximum gap between two consecutive symbols, and as such can be adapted to mine for contiguous subsequences, as defined in our work. However, the applications of these algorithms are limited to finding exactly matching subsequences due to the fact that no notion of noise or approximation is allowed in the pattern definition.

Gwadera *et al.* address the problem of finding significant episodes in an event sequence [14], where the definition of episodes is limited to subsequences occurring in a time window of fixed size. Tatti *et al.* address the problem of summarizing a data sequence with the "best" set of serial episodes based on the MDL principle [33]. There is no notion of deviation from a model in these types of patterns, making them different from statistically significant patterns in our work. Webb *et al.* propose a method for finding statistically significant association rules [37], and Gwadera *et al.* propose a method for evaluating and ranking the significance of sequential patterns [15]. These methods are proposed for transaction-based data, which is different from the continuous sequence data model assumed in our work.

3 Problem Statement & Proposed Method

Intuitively, our goal is to find subsequences in a symbolic data sequence S that occur more frequently than expected — based on some model of how sequences like S would look like under "normal conditions". To formalize this notion, we can conceptually model the sequence S, consisting of symbols of a finite alphabet set Σ, as being generated by some random process Θ, in which, at certain positions, subsequences may occur that are generated by a process different from Θ. We can think of Θ as describing the normal behaviour of the process, and we refer to Θ in the rest of the paper as the "*background*" distribution. We can think of the set of the subsequences of S that are not generated by Θ as *deviations* in the process modelled by Θ, and we refer to these subsequences, which we want to identify, as *deviating patterns* (*e.g.*, anomalies, biological motifs, *etc.*).

It is necessary to distinguish between a subsequence u as a pattern and its occurrences (or instances) in a sequence S. For instance, given the alphabet $\{1, 2, ..., 9\}$, $u =$"123" is a pattern occurring in $S =$"123764123913" at positions 0 and 6.

We assume that the anomalies are generated by processes, which are sufficiently different from the background distribution. Under this condition, *i.e.*, it is expected that the characteristics of a sequence produced by the background distribution will, in general, have changed in a way that allows us to detect those anomalies as deviating patterns. However, we can also expect that, in addition to the embedded deviating patterns, many other subsequences of S, which are partly generated by the background distribution but partly overlap with deviating patterns, may also look deviating. Such patterns are a form of *redundant* deviating patterns. In fact, this phenomenon poses a challenge when trying to identify the truly embedded, deviating patterns.

General Problem Statement: Given a symbolic data sequence S and a model Θ from which sequences can be generated, find a set of patterns O so that:

1. the patterns in O have instances in S;
2. the patterns in O are (with high probability) not generated by Θ;
3. every part of S that does not belong to the instances of O is generated with high probability by Θ;
4. among all sets that have properties 1 to 3, O is a smallest set with these properties.

We assume that we have access to some "training" sequences *like* S, which can be considered to represent "normal behaviour" of the process, and from which we can learn a model $\hat{\Theta}$. To represent an unknown background distribution for a sequence, we use a *Markov chain* model of some order m, represented by Θ_m.

The approach we propose to detect deviating patterns is based on statistical hypothesis testing. Our *Null Hypothesis* H_0 is that a given data sequence S is generated only by a background distribution, described by a stationary Markov chain model Θ_m. We can then estimate a *p-value* for a given subsequence u that occurs in S t times, which is the *probability* that u occurs at least t times in S under the Null Hypothesis. The p-value is then compared with a *significance level* α(*e.g.* 0.01 or even lower); if the p-value is less than α, the null hypothesis is rejected and the results is said to be *statistically significant*. A subsequence that has a lower p-value than a given significance level threshold is simply refereed to as a *significant pattern*.

In this approach, many patterns may pass the significance test only because they overlap with one of the truly embedded, deviating patterns, and determining which of all the significant patterns constitute the true deviating patterns is a challenge. To illustrate the problem, consider a time series sequence of length 10,000 generated by a *random walk* model given by the formula $y(t) = y(t-1)+\lambda$, where $y(1) = 0$, and λ is drawn from a random distribution with mean $\mu = 0$ and standard deviation $\sigma = 2$. Suppose the time series data is

discretized using SAX [22] with input parameters *segment size* and *alphabet size* set to 8 and 6, respectively. The result is a sequence of size 1250 with symbols in the set $\{1, 2, ..., 6\}$. Suppose the Markov model parameters are learned from the discretized data. Now suppose two subsequences of lengths 4 and 6 are inserted (*implanted*) at random locations in the sequence. Let "6545", and "112233" be the implanted subsequences that are inserted at different locations into S with frequencies 5 and 4, respectively. Consider extracting all subsequences in S of lengths 2 to 10 and computing the p-value for each subsequence (details of the p-value calculation are described in supplementary document). Figure 1 shows the number of statistically significant patterns of lengths 2 to 10, at a significance level $\alpha = 0.001$. We can observe that (1) the number of distinct, statistically significant patterns increases when the length of subsequences increases, (2) the bins corresponding to lengths of the implanted patterns (4 and 6) do not stand out in any way from the overall trend, and (3) the number of significant patterns overall is considerably larger than 2, the number of implanted patterns.

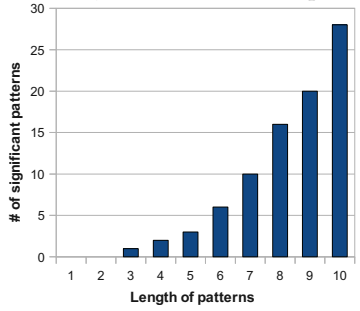


Figure 1: *Number of significant patterns vs. length.*

If we would know the set of embedded patterns E, we could try to "explain" the statistical significance of a pattern u in S by this set E. The main intuition behind this idea of an explain relationship is to devise a statistical test for the frequency of u with respect to a different Null Hypothesis; the new Null Hypothesis assumes that sequences are mostly generated by Θ_m, except with the additional constraint that each sequence must contain the patterns in a set E at the exact same locations as they occur in S.

Definition 3.1. *Given a set of patterns E and their instances in a sequence S, let the total number of the instances be K, and suppose each instance has a start index i_j and a length l_j, $1 \leq j \leq K$. The constraint set $C_{E,S}$ on an arbitrary sequence S' of length $|S|$ is the conjunction of constraints $\bigwedge_{1 \leq j \leq K}(\bigwedge_{0 \leq x \leq l_j}(S'[i_j+x] = S[i_j+x]))$ (**i.e.** the sequence S' has the same symbols as S at locations of instances in E $\big(S'[i_1] = S[i_1] \wedge S'[i_1+1] = S[i_1+1] \wedge \ldots \wedge S'[i_1+l_1] = S[i_1+l_1]\big) \wedge \big(S'[i_2] = S[i_2] \wedge \ldots \wedge S'[i_2+l_2] = S[i_2+l_2]\big) \wedge \ldots \wedge \big(S'[i_K] = S[i_K] \wedge \ldots \wedge S'[i_k+l_k] = S[i_k+l_k]\big)$).*

To illustrate this concept, assume an alphabet set $\{1, \ldots, 9\}$, a model Θ_m, and a data sequence S ="12135443512132351" of length 17. When computing the p-value of a subsequence u with frequency t in S (*e.g.*, u="12", $t = 2$) with respect to Θ_m we consider the probability of having at least 2 occurrences of u in *all* sequences that can be generated by Θ_m and that have the form S' =_,_,_,_,_,_,_,_,_,_,_,_,_,_,_ (*i.e.*, the only constraint is the length of the sequence). In contrast, given a set of patterns $E = \{121, 44\}$ with their instances occurring at positions $\{0, 5, 9\}$, and with the constraint set $C_{E,S}$, we consider all sequences of length 17 of the form S'' =121_,_,44_,_,121_,_,_,_,_, where the symbols at "open" positions are generated by Θ_m. The constraint set $C_{E,S}$ is specified by $(S''[0] = S[0]) \wedge (S''[1] = S[1]) \wedge (S''[2] = S[2]) \wedge (S''[5] = S[5]) \wedge (S''[6] = S[6]) \wedge (S''[9] = S[9]) \wedge (S''[10] = S[10]) \wedge (S''[11] = S[11])$.

The p-value of a pattern u with frequency t under a Null Hypothesis *with constraint set* E, is the probability $P = Prob(frequency(u) \geq t | C_{E,S})$ that the pattern u occurs at least t times in a sequence that satisfies the constraint set $C_{E,S}$, and is otherwise generated by Θ_m. Now, we define an explain relationship $\succ$ as follows:

Definition 3.2. *Given a model Θ_m and a sequence S, we say that a* set E explains *(the significance of) a pattern u with frequency t in S, formally $E \succ u$, iff $Prob(frequency(u) \geq t | C_{E,S})$ is higher than the given significance level α. When E explains all patterns u in a set of patterns U, i.e., if $\forall u \in U : E \succ u$, we also say that E explains U and write $E \succ U$.*

This explain relationship allows us to define a non-redundant set of significant patterns of possibly variable lengths as a subset of significant patterns in S that explains the statistical significance of all significant patterns in S. It is easy to verify that $E \succ E$, and it is clear that there may be several sets of patterns E that can then explain all significant patterns. We are interested in explaining sets E that satisfy the minimality condition.

Definition 3.3. *Given a model Θ_m and a sequence S, a* core pattern set *is a set E that explains all significant patterns in S, and that has the minimum cardinality among all such sets.*

Specific Problem Statement: Given a symbolic data sequence S and a Markov model Θ_m, find a set of subsequence patterns O with the following properties:

1. the patterns in O have instances in S;

2. the patterns in O are statistically significant assuming Θ_m as the generating model;

3. every statistically significant subsequence of S with respect to Θ_m is either in O, or is not statistically significant with respect to (Θ_m, $C_{O,S}$), i.e., when given the occurrences of patterns from O in S as additional constraints on the sequences generated by Θ_m;

4. O is a smallest set that has properties 1 to 4.

The overall method that we propose to solve this problem, after estimating the parameters of a Markov model of order m from a '*training*' sequence is as follows:

1. Extract all subsequences in a desired range of lengths from the data sequence S, using a sliding window, and determine their frequencies in S.

2. For every extracted subsequence u with frequency t, calculate the p-value $Prob(frequency(u) \geq t)$, *i.e.*, the probability that the pattern u occurs in a sequence of length $length(S)$, generated by model Θ_m, at least t times.

3. Add all subsequences whose p-value is smaller than the significance level α (after multiple testing correction) to the set of significant patterns P_{sig}.

4. Find a core pattern set $E \subseteq P_{sig}$.

3.1 Computing a Core Pattern Set Finding all significant patterns based on a significance level α involves multiple statistical tests. The significance level α represents the probability of the null-hypothesis being rejected when in fact it is true (*i.e.* the probability of a "false positive" or *Type I* error). However, the probability of false positives is not equal to α when multiple tests are performed. In statistics, this problem is referred to as *multiple testing* [25] and we adopt the method by Holm [17] for adjusting the significance level.

A naive, exhaustive search for a core pattern set is computationally expensive (exponential in the number of significant patterns). We present an algorithm for constructing an *approximate* solution based on a greedy forward selection strategy. The approximate solution P_{sol} is constructed by adding one pattern at a time from P. At each step, let P_{rest} be the set of patterns that cannot be explained by the current solution P_{sol}. We choose a pattern from P_{rest} that explains the largest number of remaining patterns. In other words, we select a pattern q^* so that $\{q^*\} \cup P_{sol}$ explains at least as many of the remaining patterns P_{rest} than any other choice. In case of a tie, a pattern whose length is not longer than any other candidate is selected (If several patterns satisfy this condition, one pattern is selected randomly) The process continues by removing the newly selected pattern q from P_{rest} and adding it to P_{sol}. Initially, $P_{sol} = \emptyset$ and $P_{rest} = P$; The algorithm terminates when P_{rest} becomes empty. The Pseudocode of the algorithm is presented in the supplementary document.

3.2 Model Extension for Approximate Pattern Matching In many applications, 'surprising' patterns occur in data sequences with *variations*. Well-known examples are *DNA binding sites* of *transcription factors* (motifs). The model we are proposing for capturing *generalized surprising patterns* is a '*mismatch model*', represented by $Q_{s,d}$, where s represents a string of symbols and d is the maximum number of allowed mismatches, w.r.t the string s. In other words, $Q_{s,d}$ represents the set of all subsequences that deviate from s by at most d mismatches. All instances of the mismatch pattern $Q_{s,d}$ are assumed to be of the same length $|s|$. The definitions presented in Section 3 can be adapted to mismatch patterns as well when taking into account that the instances of a *mismatch pattern* $Q_{s,d}$ are allowed to have up to d mismatches w.r.t. string s.

For a mismatch pattern $Q_{s,d}$, the probability of its occurrence at any location is the sum of probabilities of the occurrences of all the strings $s^{'}$ of length $|s|$ which have at most d mismatches w.r.t. string s:

$$\mu = Prob(Q_{s,d}|\theta_m) = \sum_{s' \in \Sigma^{|s|}, dis(s',s) \leq d} Prob(s^{'}|\Theta_m)$$

where Σ is the alphabet (set of symbols), *dist* is the *edit distance* between two strings, and θ_m is the background distribution of the data sequence. Calculating the exact p-values becomes computationally intensive for *mismatch patterns*. The practical alternatives to exact p-value calculation are approximation methods, which will be discussed in the supplementary document.

4 Experimental Evaluation

Our synthetic data is generated using a Markov model of order 1, where we first generate a transition probability matrix that describes the background and other transition probability matrices that model anomalies. The transition probability matrix has most of its transition probability mass in the diagonal, describing the behaviour of a random walk process where with high probability the walk stays in the same state and with small probability it moves to a different state. Formally, we use the following conditions in generating the transition probabilities out of a state s_i: (1) $P(s_i|s_i) > \sum_{i \neq j} P(s_j|s_i)$, i.e. it is more likely to stay in the same sate rather than changing the state, and (2) $\frac{P(s_j|s_i)}{P(s_k|s_i)} =$

$\frac{|j-i|}{|k-i|}$, i.e. the chance of a move from a state s_i to any other state is proportional to its distance from s_i, assuming an order for states based on their indices.

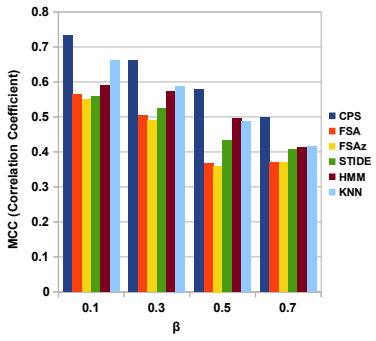
For real data, we use two publicly-available datasets; the first one is an ECG recording [2], showing the electrical impulses of a heart during electrocardiogram tests, and we want to detect anomalies in the form of *arrhythmias*. The second dataset is a motif discovery benchmark [3] that we use it to evaluate our extended model for approximate pattern matching.

On synthetic and ECG datasets, we run our method based on the simple model (with exact matching definition of patterns). In these experiments, the *Matthews Correlation Coefficient* (MCC) [23] is used as a measure of matching between the true positions of anomalies in the data with those positions predicted by our test methods. We chose the five best methods from the comparative study performed by Chandola *et al.* [7], including the KNN, STIDE, FSA, FSAz, and HMM (discussed in Section 2). All of these methods require a length parameter and a probability threshold parameter, which are used for extracting subsequences in a test sequence, estimating their probability of occurrence, and labeling those subsequences as 'anomaly' or 'normal'. The KNN method requires an additional parameter K. For our method, the significance level and Markov model order are set to 0.01, and 3, respectively.

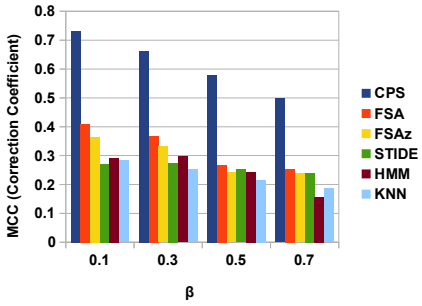
4.1 Results on the Synthetic Data We ran several experiments, keeping the background distribution fixed and varying the distributions of the anomalies, where a parameter β is used to control the amount of deviation from the background distribution. This is achieved by defining the self-transition probability as $P(s_i|s_i) = \frac{1+\beta(n-1)}{n}$, where n is the number of states of the Markov model and $0 \leq \beta \leq 1$. Using values 0.7, 0.5, 0.3 and 0.1 of β gave us four different anomaly models in increasing deviation. For each experiment, we generated a sequence of a fixed length (set at 400) using the background distribution and implanted shorter sequences that were generated by the anomaly models. The implanted sequences in each test sequence varied in length from 4 to 10. We generated 10 datasets under each setting and measured the average performance of methods on these 10 datasets.

The values of the parameters for the comparison methods usually are not known in a real setting and our method doesn't require these parameters. For our comparison, we explore a reasonable part of the parameter space and report their 'best' and 'average' performances. To that end, we ran these methods with all *lengths* in the range $[4 : 10]$, and all possible probability thresholds (which can be limited to the set of scores calculated for all extracted subsequences). For KNN, we considered the range $[1 : 10]$ for K.

Figures 2(a) and 2(b) compare the performance of CPS (our proposed method) with the 'best' and 'average' performance of other methods. respectively. One can clearly see that our method outperforms all the other methods, even when giving them the "unfair" advantage of providing the 'best' parameter setting. Comparing with the expected performance, our method dominates the comparison methods to an even larger extent. The results of the experiments also demonstrate that as the deviation between the background model and anomaly models increases (*i.e.* as the β value decreases), it becomes easier to detect anomalies.



(a) CPS vs. best results of other methods.

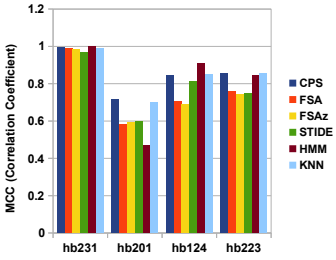


(b) CPS vs. average results of other methods.

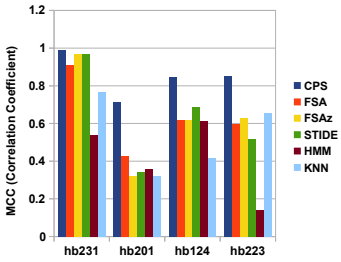
Figure 2: Performance comparison between methods on synthetic data, varying the β of anomaly models.

4.2 Results on the ECG Dataset We randomly selected 4 patient records from MIT-BIH dataset [26], and for each record, we used part of the data that does not include any arrhythmia for training. The test data consists of both normal and arrhythmia intervals. For our competitors, the parameter space was explored similar to what described in the synthetic dataset, to compute the 'best' and 'average' performance of comparison methods. Figure 3(a) compares the results of our method and the 'best' possible performance of our competitors on sample records. The results reveal that

our method outperforms or performs close to the best results achieved by other methods when giving them the advantage of knowing best parameter settings. When comparing to a more realistic expected performance of our competitors in Figure 3(b), our method achieves significantly higher MCCs in all cases, except for the record 231, on which most methods perform well.



(a) CPS vs. best results of other methods.



(b) CPS vs. average results of other methods.

Figure 3: Performance comparison on ECG dataset.

Table 1 shows the patterns found by our method (column 2) and true arrhythmias in the data (column 3). The results show that the notion of a core pattern set in our method closely matches with the definition of arrhythmias. For instance, the pattern *VNNV* in record 201 precisely corresponds to the definition of a "T" (Ventricular trigeminy) arrhythmia, which happens when two normal beats are observed between two V beats. Also, the pattern *VVV* in record 124 precisely specifies the definition of an "IVR" (Idioventricular rhythm) arrhythmia, which happens when three (or more) V beats are observed consecutively.

Table 1: Found patterns vs. true Arrhythmia in MIT-BIH records

Rec	Found patterns	True arrhythmias
201	1)VNNV 2) jNjAj	1) T : VNNVNNV... 2) NOD: jNjAj
223	1) VVV 2) VNV	1) IVR : VVV... 2) B : NVNVNV... 3) T : VNNVNNV...
124	1)VVV 2) NNV	1) IVR : VVV... 2) T : VNNVNNV...
231	1)MRM 2) MNM	1) BII : MRMRMRM... 2) BII : MNMNMN 3) BII : MRMRM...MNMNMN...

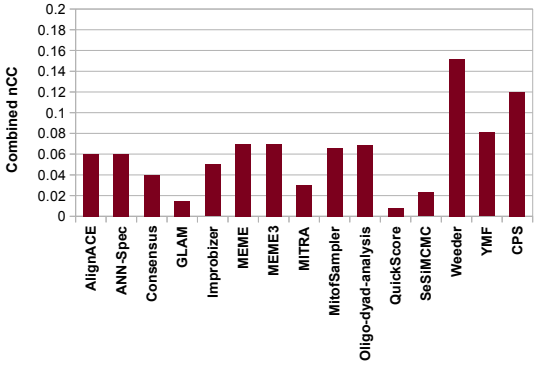


Figure 4: *Combined correlation coefficient (nCC) over all 56 datasets.*

4.3 Results on Motif Discovery benchmark In this section, we evaluate the performance of the extended model on a widely used motif discovery benchmark [3]. The goal of a motif discovery method is to identify regulatory elements, notably the *binding sites* in DNA sequences, for *transcription factors*. The transcription factors in this benchmark are selected from 4 species, including *yeast*, *fly*, *mouse*, and *human*. For each dataset in the benchmark (totally 56), a prediction tool is supposed to select the single *best motif* and report the positions of that motif's binding sites, or to report that the dataset contains no significant motif.

Figure 4 compares the overall performance of *CPS* with all 15 competitors on the given benchmark, based on a combined *correlation coefficient* measure (combined nCC) computed on all the 52 datasets. As it is shown in Figure 4, our proposed method achieves a performance close to Weeder [28], which performs the best among the other competitors overall.

As discussed by providers of the benchmark [3], the low value of combined correlation coefficient should not be taken as indictment of computational methods for motif discovery. One of the reasons is that each method is supposed to return the best motif in each dataset (or none). This makes the selection process a very challenging task. In practice, it might be useful to return a list of top K motifs, which can later be verified by domain experts. This could potentially increase the number of 'true' binding sites which can be found by a motif discovery method. To pursue this scenario, we designed a new experiment using the same benchmark where we evaluated the performance of the 'best' motif among the top 32 motifs returned by a discovery tool. As comparison partner, we selected *Weeder* [28] which achieved the best performance on this benchmark.

The result of the comparison between *Weeder* and *CPS* is shown in Figure 5. As can be observed in this figure, our proposed *CPS* algorithm performs better than Weeder on all species, except on the *human*

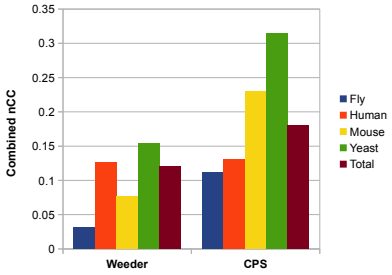


Figure 5: *Combined correlation coefficient (nCC) on top 32 motifs.*

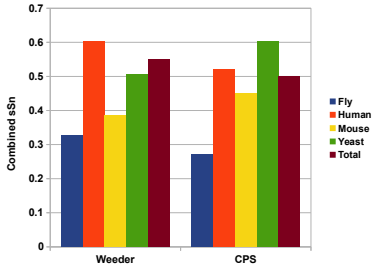


Figure 6: *Combined site-level sensitivity (sSn) on top 32 motifs.*

datasets in which both techniques are comparable. We also observed that the performance of Weeder in the top-K experiment is weaker than in the 'best' motif selection settings used in the original competition.

In previous experiments, we evaluated the performance of methods based on the *nucleotide-level Correlation Coefficient* (nCC) because it provides a comparison at a finer granularity. We can look at other measures to evaluate the performance of a motif discovery tool in terms of its ability to match (even partially) with true binding sites. One of these measures is *Site-level Sensitivity* (sSn) [3], which is defined to be the proportion of true binding sites that overlap with predicted sites to the total number of true binding sites. Figure 6 shows the performance of the top 32 motifs chosen by the CPS and Weeder algorithms based on the *site-level sensitivity*. Despite a very low measure on the nucleotide-level correlation coefficient, these two motif discovery methods perform fairly good in (partial) matching with *true binding sites*. The achieved statistics show that we are able to *partially* match with around half of the *true binding sites* of transcription factors if we just look at the top 32 motifs returned by these two methods.

5 Conclusions

We investigated the problem of finding surprising patterns in sequences. The notion of 'surprise' was formalized by the statistical significance of observed frequency of a pattern compared to its 'expected' frequency. We demonstrated the drawbacks of doing multiple statistical tests on subsequences of all possible lengths, which is an alternative when the lengths of 'true' surprising patterns are unknown. We developed statistical methods to capture the dependencies between patterns through an "explain" relationship, and introduced the new notion of *core pattern set* as a non-redundant set of patterns that can explain the statistical significance of all subsequences in the data.

Our experimental study demonstrates the effectiveness of our method for capturing anomalies in synthetic and real-world data, in which our method clearly dominates the 'expected' performance of our competitors. We also showed that the extended mismatch model provides support for a broader class of applications in which an approximate representation of surprising patterns are required, including the interesting problem of finding transcription factor binding sites in DNA sequences.

References

[1] Agrawal, R., Srikant, R.: Mining sequential patterns. In: IEEE International Conference on Data Engineering, pp. 3–14 (1995)

[2] Goldberger et al., A.L.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), 215–220 (2000)

[3] Tompa et al., M.: Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology **23**(1), 137–44 (2005)

[4] Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., Mango, S.E.: Environmentally induced foregut remodeling by pha-4/foxa and daf-12/nhr. Science **305**(5691), 1743–1746 (2004)

[5] Bailey, T.L., Elkan, C.: The value of prior knowledge in discovering motifs with MEME. In: International Conference on Intelligent Systems for Molecular Biology, pp. 21–29 (1995)

[6] Castro, N., Azevedo, P.: Time Series Motifs Statistical Significance. In: SIAM International Conference on Data Mining, pp. 687–698 (2011)

[7] Chandola, V., Mithal, V., Kumar, V.: Comparative evaluation of anomaly detection techniques for sequence data. In: IEEE Internation Conferenece in Data Mining, pp. 743–748. Washington (2008)

[8] Chiu, B., Keogh, E., Lonardi, S.: Probabilistic discovery of time series motifs. In: SIGMOD international conference on Management of data, pp. 493–498 (2003)

[9] Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press (1999)

[10] Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Ravcheev, D.A., Mironov, A.A., Makeev, V.: A gibbs sampler for identification of symmetrically structured, spaced dna motifs with improved estimation of the signal length. Bioinformatics **21**(10), 2240–2245 (2005)

[11] Floratou, A., Tata, S., Patel, J.M.: Efficient and Accurate Discovery of Patterns in Sequence Data Sets. IEEE Transactions on Knowledge and Data Engineering **23**(8), 1154–1168 (2011)

[12] Forrest, S., Hofmeyr, S.A., Somayaji, A., Longstaff, T.A.: A sense of self for unix processes. In: IEEE Symposium on Security and Privacy, pp. 120–128 (1996)

[13] Frith, M.C., Hansen, U., Spouge, J.L., Weng, Z.: Finding functional sequence elements by local alignment. Nucleic Acids Research **32**(1), 189–200 (2004)

[14] Gwadera, R., Atallah, M.J., Szpankowski, W.: Markov models for identification of significant episodes. In: SIAM International Conference on Data Mining, pp. 404–414 (2005)

[15] Gwadera, R., Crestani, F.: Ranking sequential patterns. In: Pacific-Asia Knowledge Discovery and Data Mining, pp. 286–299 (2010)

[16] Hertz, G.Z., Stormo, G.D.: Identifying dna and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics **15**(7), 563–577 (1999)

[17] Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6**(2), 65–70 (1979)

[18] Hughes, J.D., Estep, P.W., Tavazoie, S., Church, G.M.: Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae1. Molecular Biology **296**(5), 1205–1214 (2000)

[19] Keogh, E., Lin, J., Fu, A.: HOT SAX: efficiently finding the most unusual time series subsequence. In: IEEE International Conference on Data Mining, pp. 226–233 (2005)

[20] Keogh, E., Lonardi, S., Chiu, B.Y.: Finding surprising patterns in a time series database in linear time and space. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 550–556 (2002)

[21] Li, C., Yang, Q., Wang, J., Li, M.: Efficient mining of gap-constrained subsequences and its various applications. ACM Transactins on Knowledge Discovery from Data **6**(1), 1–39 (2012)

[22] Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 2–11 (2003)

[23] Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta **405**(2), 442–451 (1975)

[24] Michael, C.C., Ghosh, A.: Two state-based approaches to program-based anomaly detection. In: IEEE Annual Computer Security Applications, pp. 203–237. Washington, DC (2000)

[25] Miller, R.G.: Simultaneous Statistical Inference. Springer-Verlag, New York (1991)

[26] Moody, G., Mark, R.: The impact of the MIT-BIH arrhythmia database. IEEE Engineering in Medicine and Biology **20**(3), 45–50 (2001)

[27] Patel, P., Keogh, E., Lin, J., Lonardi, S.: Mining motifs in massive time series databases. In: IEEE International Conference on Data Mining, pp. 370–377 (2002)

[28] Pavesi, G., Mereghetti, P., Mauri, G., Pesole, G.: Weeder: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Research **32**(Web-Server-Issue), 199–203 (2004)

[29] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: Mining sequential patterns by Pattern-Growth: The PrefixSpan approach. IEEE Transactions on Knowledge and Data Engineering **16**(11), 1424–1440 (2004)

[30] Régnier, M., Denise, A.: Rare events and conditional events on random strings. Discrete Mathematics and Theoretical Computer Science **6**(2), 191–214 (2004)

[31] Sinha, S., Tompa, M.: YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Research **31**(13), 3586–35,888 (2003)

[32] Sun, P., Chawla, S., Arunasalam, B.: Mining for Outliers in Sequential Databases. In: SIAM International Conference on Data Mining, pp. 94–105 (2006)

[33] Tatti, N., Vreeken, J.: The long and the short of it: summarising event sequences with serial episodes. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 462–470 (2012)

[34] Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., Moreau, Y.: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics **17**(12), 1113–1122 (2001)

[35] Wang, J., Han, J.: BIDE: Efficient mining of frequent closed sequences. In: IEEE International Conference on Data Engineering, pp. 79–90 (2004)

[36] Warrender, C., Forrest, S., Pearlmutter, B.: Detecting intrusions using system calls: Alternative data models. In: Symposium on Security and Privacy, pp. 133–145 (1999)

[37] Webb, G.I.: Discovering significant patterns. Machine Learning **68**(1), 1–33 (2007)

[38] Workman, C.T., Stormo, G.D.: ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pacific Symposium on Biocomputing **5**, 467–478 (2000)

[39] Yan, X., Han, J., Afshar, R.: CloSpan: Mining closed sequential patterns in large databases. In: SIAM International Conference on Data Mining (2003)

[40] Zaki, M.J.: Sequence mining in categorical domains: Incorporating constraints. In: ACM International Conference on Information and Knowledge Management, pp. 422–429 (2000)

[41] Zaki, M.J.: SPADE: An efficient algorithm for mining frequent sequences. Machine Learning **42**(1-2), 31–60 (2001)