Hidden Markov Models for Surprising Pattern Detection in Discrete Symbol Sequence Data

Ken McGarry[1 [0000−0002−9329−9835]]

School of Computer Science, Faculty of Technology, University of Sunderland, UK
{ken.mcgarry@sunderland.ac.uk}

Abstract. Detecting unusual or interesting patterns in discrete symbol sequences is of great importance. Many domains consist of discrete sequential time-series such as internet traffic, online transactions, cyber-attacks, financial transactions, biological transcription, intensive care data and social sciences data such as career trajectories or residential history. The sequences usually consist of discrete symbols that may form regular patterns or motifs. We use regular expressions to construct the longest repeating sequences and subsequences that compose them, we then define these as motifs (which may or may not represent novel patterns). The sequences are now composed of simpler motifs which are used to build Hidden Markov Models models which can capture complex relationships based on location, frequency of occurrence and position. New data that deviates from established motifs either in location of appearance, frequency of appearance, or motif composition may represent patterns that may be different in some way and hence interesting to the user.

Keywords: Motif · Regex · Sequence · Hidden Markov Model.

1 Introduction

In this work we develop a novel anomaly detection method to uncover patterns or trends in discrete sequence data that differ from normal expectations. However, anomalies cannot always be expected to follow previously known patterns or trends. Anomalies are patterns in data that may be incorrect, noisy, novel or unusual in some respect. This can be with regard to magnitude of values, unusual timing of appearance, changing relationships with other patterns or some other factor that makes them different to the normal values encountered. The problem is that nearly every data mining problem is different and indeed patterns can change over time even within the same problem or area.

The ability to be surprised is fundamental to many human cognitive and intellectual endeavors, it is an essential trait for learning and discovering new knowledge [1,2]. Cognitive scientists generally agree that surprise is an emotion that arises when differences occur between our expectations and the actual results [6]. This mismatch can be accounted for in a principled way using Bayes theory, which is perfectly suited to update beliefs in the light of new information. We implement a modification of Bayesian surprise discussed by Itti which corresponds to subjective beliefs that are revised as new information appears [11]. Bayes theorem allows the conversion of our

prior beliefs into posterior beliefs. Therefore, Bayesian surprise is a criterion to judge discrepancies between a systems prior and posterior beliefs on any given matter. The bigger the discrepancy then the more surprised we should be [12]. The use of surprise and novelty is also finding applications in goal directed learning when developing automated systems [8] and agent based systems [22]. Furthermore, product design has benefited from this approach whereby *surprise* is used as a creative metric to predict if customers will find new features and styles exciting and attractive [3]. In the past, the so-called interestingness measures were used to assess the novelty or unusualness of patterns in a data set, many such measures have been developed [19].

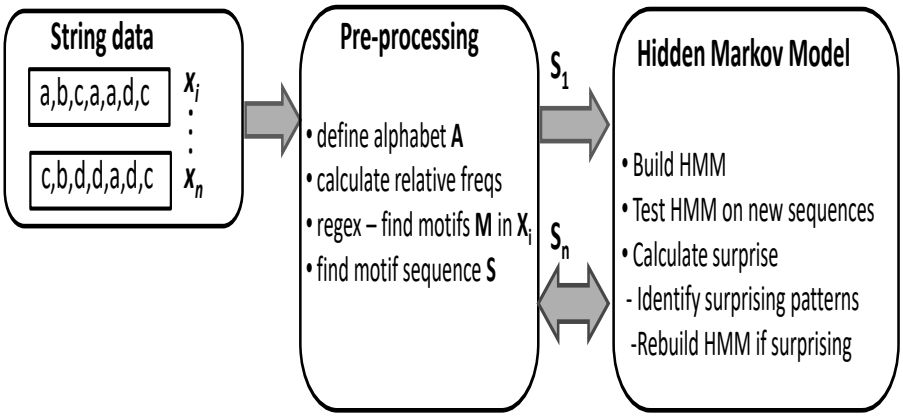


Fig. 1: Motif generation and analysis

In this work we develop novel methods to search for discrete sequences of symbols that may form motifs. Motifs are collections of discrete symbols (effectively a string) from an alphabet of variable but finite size. In Figure 1 we show that motifs are repeated patterns found within discrete symbol sequences and time series data. The sequence of symbols are then pre-processed and will form recurring patterns or motifs that imply important changes or activities in the time-series [24].

2 Related work

The SEQUITUR system of Nevill-Manning builds a structure from sequences of symbols by removing common phrases using a grammatical rule that models the phrase, this process continues until all sequences are read in. SEQUITUR operates by modelling repeated subsequences into if..then type rules. The whole process is recursive and the result is a hierarchical rule based system [20]. Previous work by Lin and Keogh in detecting motifs in time-series led to the development the Piecewise Aggregate Approximation (PAA) algorithm and the Symbolic Aggregate approXimation

(SAX) algorithm [17]. PAA is a very popular algorithm used to convert the time series into discrete levels or symbols. The sequences/sub-sequences of symbols may form patterns or motifs that represent particular activity in the time series data [13]. Further improvements on PAA and SAX symbolic representation method for discretization includes preserving the information contained in the angle of the signals characteristics of the time series. Other Computer Science areas such as temporal association rules [27] have similar issues such as modelling the temporal aspects by integrating interval-based relationships to occurrences of items in the database.

Natural Language Processing and speech recognition provide many insights into modelling sequence information, especially when dealing with strings, letter and word occurrences [23,26]. For example part-of-speech-tagging (POS) allows the sequence of words to be represented to resolve ambiguity. Hidden Markov Models (HMM) are able to label text data in POS applications, we describe HMM in detail later. DNA searching algorithms have elements in common with motif detection, they all search for recurring or interesting patterns in sequential, discrete data [14]. In fact the discovery of motifs has received a great deal of interest in DNA analysis, often suffix trees are common data structures used hold sequential data. Huang *et al* used suffix trees to contain temporal data for regularly occurring patterns of different sorts such as full, inner, and tail patterns [10]. Work by Cohen used segmentation by information theory to decompose strings [5]. Experiments conducted by Reick on the suitability of various hierarchical structures such as tries, suffix trees and sorted arrays for sequence analysis provided valuable insights into their usage [23].

Deep learning has been successfully used in many applications of discrete sequence data. For example, recurrent neural networks (RNN) can deal with variable-length sequences while maintaining sequence order and tracking long-term dependencies [25]. However, in order to model longer term dependencies than is currently possible with RNN, a variation called Long Short-Term Memory (LSTM) should be used. These LSTM networks have generally been applied to speech and handwritting recognition, machine translation, parsing, image captioning. This is made possible through LSTM models ability to keep information over many time steps by using a separate cell state from what of outputted and gates controlling data flow [15]. Other models useful for anomaly or unusual pattern detection in sequences include the range of auto-encoder networks.

Our contribution uses Hidden Markov Models which are simpler methods in comparison to the deep learning models and their architecture is more convenient to provide access to intermediate states using the (i) transition matrix , (ii) prior probability and (ii) emission probability matrices. It is variations in expected and actual emissions that are used by the Bayesian Surprise mechanism to determine the newness/surprisingness of a given pattern.

3 Hidden Markov Models

Hidden Markov Models (HMM) are often used to capture the dynamics of sequence data [21]. Furthermore, a Markov chain is a model we can employ to inform us about the probabilities of discrete variable sequences, states or events. These discrete

variables normally take on predetermined values from an alphabet. The alphabet can be natural language words or symbols that represent changes or events occurring in the problem domain. A Markov chain makes a very strong assumption that if we want to predict the future in the sequence, all that matters is the current state [4]. All the states before the current state have no impact on the future except via the current state. The HMM is considered to be the model of choice for sequential problems, [16] and is used in sequence anomaly detection [7].

The main characteristic of Markov Chains is their *memory-less* property, that is to say each new state will only depend on the last or previous state. This enables the transition probabilities to be calculated for the next event or state, based on the current event or state:

$$P(x_i|x_{i-1}, ..., x_1) = P(x_i|x_{i-1})$$

where: A is the alphabet of symbols and S is the generative model (built on the alphabet) representing the probability A^ℓ. While $x \in A^\ell$ is the sequence presented to the HMM (S). These probabilities are used by the next stage to determine if a pattern or sequence is surprising or interesting.

Therefore, the columns of A^ℓ have to sum up to 1. Breaking down the given sequence, the probability for each symbol can be defined by:

$$P(x) = P(x_L, x_{L-1}, ..., x_1) = P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})...P(x_2|x_1)P(x_1) \quad (1)$$

$P(x_1)$ is obtained from the transition probability matrix, multiplying the initial event probabilities at time $t = 0$ using the transition data, the probabilities of every event at time $t = 1$ can be determined and therefore we also have them for time $t = n$.

$$\forall \ell \in \{0, 1, 2, ...\} : \sum_{x \in A^\ell} P^S(x) = 1. \quad (2)$$

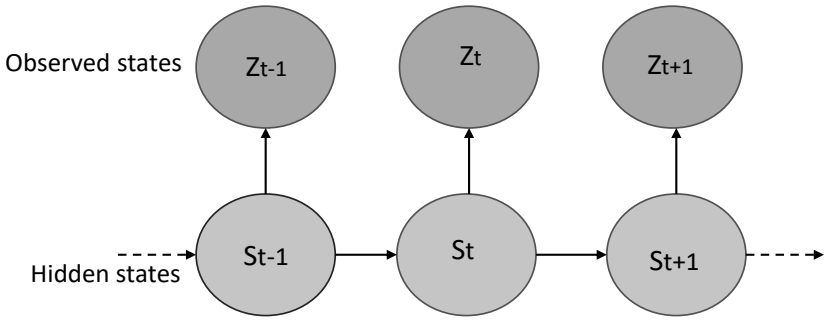


Fig. 2: Example HMM, hidden states and observed states.

As shown in Figure 2 we have a series of observed states and hidden states, by which the HMM will generate transition probabilities and emission probabilities for the hidden states, these are the model's main parameters. They can be setup prior to training or simply determined by the training algorithm. Observing the transition matrix the highest transition probabilities are generally found on the diagonal, this implies that particular state is unlikely to change. Interpreting the matrix we can say the probabilities of making a transition to another state depend on the data types and frequency of occurrence, e.g a transition probability of 0.9 indicates that nine times out of ten, the hidden state remains the same [9]. The emission states indicate how long each *observed state* is likely to remain in a certain *hidden state*. Analysis of these two matrices allows us to estimate periods of stability and change in a dataset.

4 Methods

We now define the notations for representing the data which comprise the alphabets, strings, letters and motifs. We consider strings and sequences to be similar and use the terms interchangeably. Consider the alphabet A comprised of elements or letters, $A = \{a, b, c, d\}$. The string x "abcdabcabcddda" consists of the letters "a", "b", "c" and "d" from A. Four unique letters in a string of length $|x|$ 14 giving the relative frequencies in table 1. A motif M is defined as a recurring pattern or substring in the string x that appears at least twice. This cut-off point β is arbitrary and may be modified. The motifs can be of differing lengths, we use regular expressions to find the largest repeating substring that occurs at least twice in x. All occurrences of the newly discovered motif are removed from x and the subsequent largest motif is discovered, the process repeats until there are no further valid motifs to be found. The algorithm will return two lists; the first is a list of unique motif names M_n, the second is the ordered list of motif occurrences M_o from the original string x.

Table 1: Relative frequencies of letters in example string *abcdabcabcddda*

A	*B*	*C*	*D*
4	3	3	4
0.28	0.21	0.21	0.28

The relative frequency information is used to assist in the generation of motifs. The initial set of motifs is based on the assumption that the largest reoccurring sequences are important in the data mining domain i.e. they are typical patterns as shown in algorithm 1. We set a minimum cut-off point β of at least two occurrences for a set of repeating substrings of length ψ of two letters to be considered as a viable motif. A single letter cannot be a motif since a motif should be a combination of symbols or events.

When building the HMMs, the first stage is to identify best number of hidden states for the model. Examining the smallest AIC value for different hidden states.

Algorithm 1 Motif generation

Input: set of strings: $x(alphabet)$, : A
Output: Motif list of names $\mathbf{M}_n$; Motif ordered sequence $\mathbf{M_o}$

```
Mo; Mn ← 0
Initialize β =0, ψ=0.
repeat
    RegEx find largest subtring in x = ∀x.
    Calculate length of ψ and β
    M_n ← add substring to motif list if ψ ≥ 2 and β ≥ 2
    Remove M_n from x
until M_n ∉ x
RegEx populate Mo with ordered occurrences of M_n ∈ x
Return [Mo; Mn]
```

We use the R, Hidden Markov Model (seqHMM) by Helske [9]. We make our source code (written in the statistical language R) and data available from: https://github.com/kenmcgarry/AI2022.

The Bayesian surprise measure provides a natural and useful method for defining and representing novel and surprising patterns [2]. Equation 3 calculates the distribution over all hypothesis $h \in \mathcal{H}$. The surprise is given as the two-fold difference between $P(h|D)$ and $P(h)$.

$$\begin{aligned} S(D, \mathcal{H}) &= distance[P(h), P(h|D)] \\ &= \sum_{\mathcal{H}} P(M|D) log \frac{P(M|D}{P|M} \end{aligned} \tag{3}$$

However, without the ability to recognize previous interesting patterns, each presentation of data would result in similar outcomes of interesting scores being assigned. We use a decay function over time that will dampen surprise such as that proposed by Baldi [1]. This is used as post-processing stage.

$$\frac{1}{a_n} + log(1 - \frac{1}{a_N + b_N}) \approx \frac{1-p}{p_N} \tag{4}$$

Where: N is the number of data samples, a and b refer to the respectively to the prior and posterior probability values.

5 Data

We now define the notations for representing the data which comprise the alphabets, strings, letters and motifs. The data sets consist of a series of categorical data sequences of fixed length and also variable length vectors. The sequences represent a series of actions or situations that have occurred. The datasets also contain variables such as age, gender, biological measurements using a variety of data types. However, we do not use these and concentrate on the discrete event sequence data.

5.1 Sepsis data set

This data set consists of events of sepsis cases from a Swiss hospital. Sepsis is a life threatening condition usually caused by an infection. One case represents the pathway through the hospital from admission, drug treatments and finally discharge. The events were recorded by the ERP (Enterprise Resource Planning) system of the hospital. There are 1,000 individual patient records with a total 15,000 events that were recorded for 16 different activities. A further, 39 data attributes are recorded, such as the biological values from several tests conducted by the Doctors and the particular lab where the tests were made, along with any medication given. The main feature of interest of this dataset is the variable length of the discrete sequence records.

5.2 Self-rated Health data set

The self-rated health (SRH) data set contains sequences for 2,612 respondents of a survey conducted by the Swiss Household Panel (SHP). The individuals were aged between 20 and 80 years at the start of the survey. The data is organized into 11 variables of 2,612 records (one reading for each person over the 11 years of the survey 1999-2009). The survey has missing data. Respondents' self rated health is collected at each yearly wave of the SHP with the following question: "How do you feel right now?" Possible answers are: very well; well; so, so (average), not very well and not well at all.

- G1 (very well) ; G2 (well); M (so, so (average));
- B2 (not very well) ; B1 (not well at all) ; * (missing)

6 Results

For each data set motifs were derived using equation 1, as the motif detection algorithm analyses the data it will discover over several passes, decreasingly smaller motifs (repeating patterns of characters). For all data, the discovered motifs were split 75/25 for training and test data. In reality, the test data represents new sequences to be passed to the HMM. We wish to determine for all data sets the amount of *surprise* generated when new data becomes available. No cross-fold validation or other methods were used used to train the HMM. The test data was passed through the HMM and the output sequences were assessed by the Bayesian Surprise criteria to judge novelty and the "surprise" value for each sequence passed through the HMM.

6.1 Sepsis data analysis

Prior to motif discovery, the text descriptions of the discrete Sepsis sequences were converted into single letter characters, shown in table 2 - this was to avoid computational issues with large strings. The motif finding results for the Sepsis data set are displayed in table 3. Approximately, 67 motifs were discovered, only the largest are

Table 2: Sepsis string to character conversion

String	Character
ERRegistration	A
Leucocytes	B
CRP	C
LacticAcid	D
ERTriage	E
ERSepsisTriage	F
IVLiquid	G
AdmissionNC	H
IVAntibiotics	I
AdmissionIC	X
ReleaseA	1
ReleaseB	2
ReleaseC	3
ReleaseD	4
ReleaseE	5
ReturnER	Z

Table 3: Sepsis Motifs

ID	Motif
1	AEFBDCGIHCB
2	AEFCDBGIAEF
3	B1ZAEFGICDB
4	1AEFCDBGIH
5	AEFGIBCDHC
6	AEFBCDGIH
7	AEFDBCGIH
8	AEFGCDBIH
9	1ZAEFAEF
10	AECDBFGI

shown. Prior to motif discovery, the text descriptions of the discrete sequences were converted into single letter symbols - this was to avoid computational issues.

The Sepsis HMM model was built using the motifs, the data divided into train and test sets. We used five hidden states, although empirically the most optimum number can be discovered by building several models and comparing the Bayesian Information Criteria (BIC) and the Akike Information Criteria (AIC), the highest value usually implies the best fit. These are shown in fig 3, the diagram on the left indicates the transition probabilities from one state to the next after training. The diagram on the right indicates the changes to the model as a result of the test data passed through the HMM and then refitting the model. The differences between the model's is used by the Bayes surprise algorithm to determine how interesting and unusual the data actually is. The decay function is then applied to reduce the effect of previously seen patterns and thus reduce *surprise*.

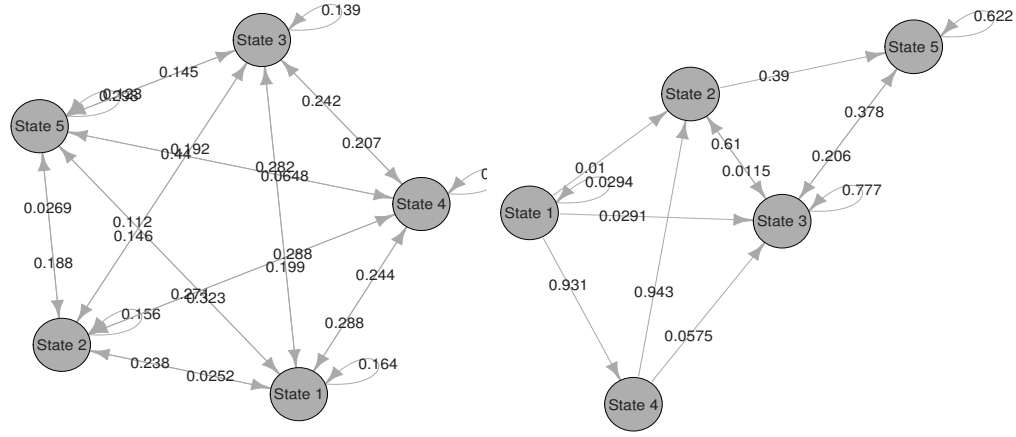
In table 4 the statistically significant events from the Sepsis data sub-sequences that discriminate between interesting and not-interesting sub-sequences are presented, using the tagged sequences with the surprising label. The Surprising patterns have values for the key variables between 0.6 and 0.9, while the not-surprising sequences events have Pearson coefficients much lower between 0.05 and 0.5 - only the Lecoucyte -> CRP sequence has a significantly higher value.

In fig 3 the HMM models are plotted, arbitrarily five hidden states were chosen to fit the data for both train and test conditions. Test data was passed through the trained model in fig 3a to produce a second model shown in fig 3b. The differences between the two models in terms of transition and emission probabilities are noted and used by the Bayesian Surprise algorithm.

Table 5 gives the transition probabilities for the trained model and how it changes from one state to the next. The initial starting probabilities are State 1 (1.0), State 2(0), State 3(0), State 4(0), State 5(0). Interpreting table 5 we see that the sequences go from State 1 through to State 5. In fig 4 we break down the *surprising* sequences

Table 4: Top sub-sequences discriminate between surprising and not-surprising sequences for Sepsis data

ID	subseq	Sup	pvalue	statistic	surp	not-surp
1	(Leucocytes>CRP)-(CRP>Leucocytes)	0.39	0.00	75.48	0.61	0.06
2	(Leucocytes>CRP)-(Leucocytes>CRP)	0.35	0.00	67.87	0.56	0.05
3	(ERRegistration>ERTriage)-(Leucocytes>CRP)-(CRP>Leucocytes)	0.37	0.00	67.86	0.58	0.06
4	(ERRegistration)-(Leucocytes>CRP)-(CRP>Leucocytes)	0.36	0.00	66.39	0.57	0.06
5	(ERRegistration>ERTriage)-(Leucocytes>CRP)-(Leucocytes>CRP)	0.33	0.00	63.75	0.53	0.04
6	(ERRegistration)-(ERRegistration>ERTriage)-(Leucocytes>CRP)	0.35	0.00	62.11	0.55	0.06
7	(ERRegistration)-(Leucocytes>CRP)-(Leucocytes>CRP)	0.33	0.00	60.82	0.53	0.05
8	(ERRegistration)-(ERRegistration>ERTriage)-(Leucocytes>CRP)	0.32	0.00	59.67	0.51	0.04
9	(Leucocytes>CRP)	0.73	0.00	58.26	0.91	0.47
10	(CRP>Leucocytes)-(Leucocytes>CRP)	0.37	0.00	58.07	0.57	0.09



(a) The trained sepsis HMM (b) The Sepsis HMM model with test data

Fig. 3: The sepsis HMM

against the *not-surprising* sequences, the surprising patterns have Bayes values of 0.6 to 0.8, while the less interesting sequences have Bayes values of around 0.05.

The sepsis data set is rich in terms of the number of records and the large number of discrete symbols. The HMM was able to process these and to determine when new data was passed through it could identify sequences that were sufficiently different to what it had been trained on.

6.2 Self-reported Health (SRH) data analysis

The motifs for the SRH data are shown in table 6, the motifs are composed of only six symbols and therefore more challenging to detect interesting patterns over time. However, it has found motifs that represent those individuals who have consistently good health over the 11 years.

The Self-reported Health data between surprising and not-surprising test data subsequences. In figure 6 the Swiss Health event sub-sequences are shown greater detail,

Table 5: State transition probabilities for sepsis HMM

	State 1	State 2	State 3	State 4	State 5
State 1	0.03	0.00	0.00	0.01	0.96
State 2	0.00	0.83	0.15	0.01	0.01
State 3	0.00	0.00	1.00	0.00	0.00
State 4	0.00	1.00	0.00	0.00	0.00
State 5	0.00	0.00	0.00	1.00	0.00

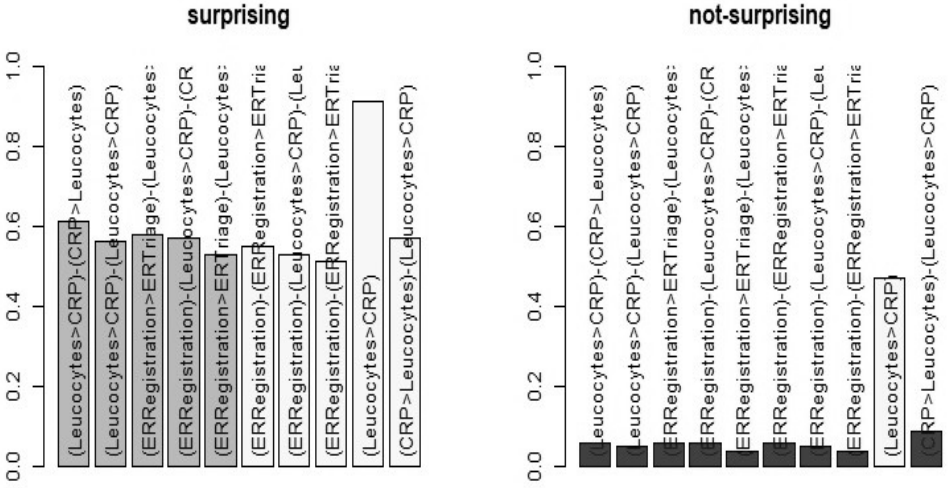


Fig. 4: The 10 most discriminating sub-sequences for Sepsis test data

we discover the most discriminating sequences between surprising and not-surprising sequences. The values of the event sequences are generally lower for the surprising patterns with the exception of the "well" event sequence. The not-surprising event sequences generally vary between 0.3 and 0.5 for the Pearson correlation coefficient. The reason for the not-surprising sequences having greater coefficients is probably because there is little difference between the two but the not-surprising patterns are more numerous. For the other Sepsis data the differences are more striking and noticeable, due to the larger number of symbols which provide more discrimination.

In fig 5 the HMM models are plotted, again arbitrarily five hidden states were chosen to fit the data for both train and test conditions. Test data was passed through the trained model in fig 5a to produce a second model shown in fig 5b. The differences between the two models in terms of transition and emission probabilities are noted and used by the Bayesian Surprise algorithm. The top 10 sub-sequences are displayed in table 7, the surprising patterns generally do not have higher Bayesian Surprise scores.

The transition probabilities for the self-reported health is shown in table 8. The initial starting conditions are State 1 (0.4), State 2 (0.04), State 3(0.19), State 4(0.3), State 5 (0.04), thus State 1 is the most probable.

Table 6: Motifs for self-reported health data

id	symbol sequence	motif number	motif
1	G2G2G2G2G2G2G2G2G2G2	11	MMG2G2G
2	G2G2G1G1G1G2G2G1G2G	12	2G2MG2
3	G1G1G2G1G2G2G	13	G1G1G1
4	G2G1G2G1G1G2G	14	2G1MG
5	G2G2G1G2G2G2G	15	2G2MM
6	G2G1MG1G2G2G	16	1G2M
7	G2G2G1G2G1G2	17	2MG2
8	G2G1G1G1G	18	21M
9	G2G1G2G1	19	22G
10	G1G2G2G	20	G2G

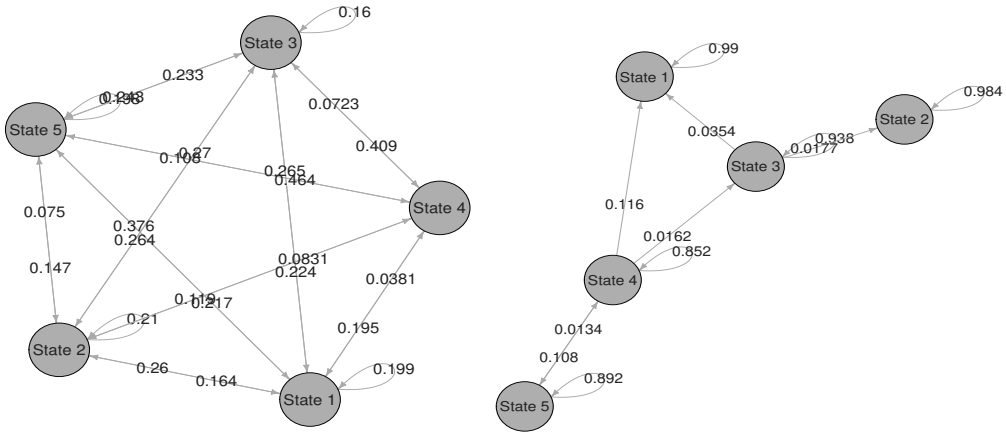
Table 7: Top sub-sequences discriminate between surprising and not-surprising sequences for Self-reported Health data

ID	subseq)	Sup	pvalue	statistic	surprising	not-surprising
1	(very well>well)-(well>very well)	0.40	0.00	50.70	0.27	0.55
2	(very well>well)-(very well>well)	0.34	0.00	40.42	0.22	0.46
3	(very well>well)-(well>very well)-(very well>well)	0.33	0.00	38.79	0.22	0.46
4	(well)	0.54	0.00	37.12	0.65	0.41
5	(very well)-(well>very well)	0.22	0.00	31.96	0.13	0.32
6	(very well)-(very well>well)-(well>very well)	0.22	0.00	31.11	0.13	0.32
7	(so)-(so (average)>well,so)-(so (average))	0.16	0.00	28.54	0.08	0.24
8	(so)-(so (average)>well)-(so (average))	0.16	0.00	28.54	0.08	0.24
9	(very well)-(well>very well)-(very well>well)	0.20	0.00	28.52	0.12	0.29
10	(very well>well)-(very well>well)-(well>very well)	0.19	0.00	28.34	0.11	0.27

7 Discussion

This work only uses the sequences of discrete symbols to detect deviations and novel patterns. It does not use domain knowledge from experts, nor does it use the other variables provided with each data set. Data sets consisting of variable lengths have historically posed a problem for most probabilistic methods, however the Hidden Markov Model (HMM) and its variable length Markov chain property are ideal for this type of data. In fact variable length sequences such as the Sepsis data are the most interesting in terms of the results. They enable a richer variety of patterns to be encountered. In addition, having many different symbols in the data sets also enables a richer variety of patterns. We find that data sets with small variety of symbols and fixed length sequences such as the Self-reported health tend not to generate that many interesting patterns and their surprise measure rapidly falls away when the decay parameter is used.

Once trained, the HMM outputs a set of probability values for each new input (test) sequence. These new values (one for each symbol in the sequence) are then used as posteriors to be compared with the priors for each symbol based on the trained HMM estimates for those symbols. The Bayesian Surprise method will evaluate the differences and determine how anomalous or interesting these patterns are. The decay parameter is essential, otherwise there would be no "memory" of previously observed interesting patterns and the system would continuously view all patterns higher than the Bayesian Surprise cutoff point as interesting. This is in keeping with the cognitive reasoning of a human conducting the analysis.



(a) The trained SRH HMM (b) The HMM SRH model with test data

Fig. 5: The SRH HMM

Table 8: State transition probabilities for self-reported health HMM

	State 1	State 2	State 3	State 4	State 5
State 1	0.29	0.16	0.40	0.14	0.02
State 2	0.31	0.40	0.26	0.03	0.00
State 3	0.41	0.05	0.15	0.15	0.25
State 4	0.20	0.26	0.27	0.13	0.13
State 5	0.23	0.16	0.16	0.05	0.40

8 Conclusions

In this work we have demonstrated that an unsupervised approach to seeking motifs in sequential, discrete data can provide a reasonable solution to capturing the features of the data. We have shown how the motifs can be modelled by Hidden Markov Models and surprising motifs can be identified. There are certain limitations of this work, mainly the computational burden of increasing the number of symbols in the alphabet. The regular expressions create motifs that currently require an exhaustive search. Another issue is the limitation of data such as the Self Reported Health that has a small number symbols present in the alphabet (6 symbols) and regularity of symbols forming the motifs. The Sepsis data has a larger alphabet (16 symbols) and has a more varied set of motifs. Thus it is easier to differentiate between surprising and not-surprising patterns. Future work will address the computational issues of motif generation and will explore the use of Generative Adversarial Networks (GAN), Probabilistic Suffix Trees (PST), Recurrent Neural Networks (RNN) and Long Term-Short Term memory (LSTM).

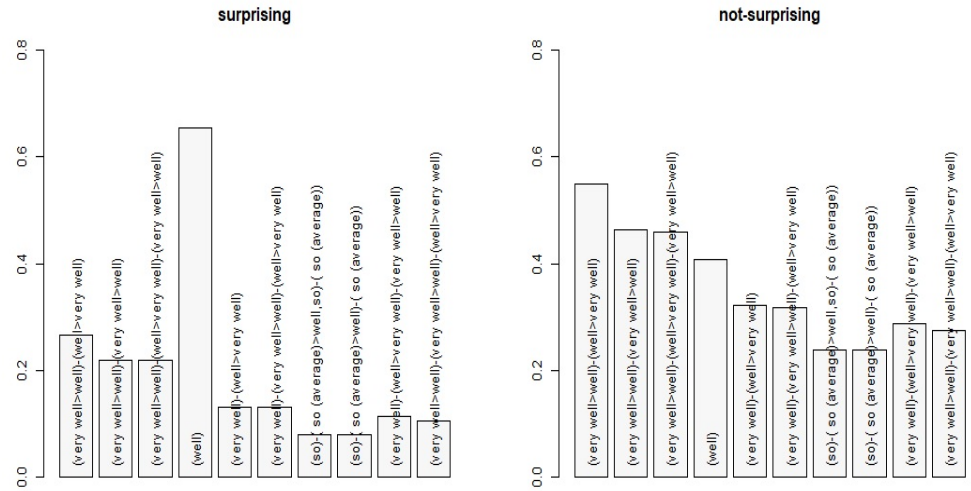


Fig. 6: The 10 most discriminating sub-sequences for Self-reported Health test data

Acknowledgements We would like to thank Satu Helske for useful advice on the seqHMM package and the two reviewers for their advice to improve the quality of the paper.

References

1. Baldi, P., Itti, L.: Of bits and wows: A bayesian theory of surprise with applications to attention. Neural Networks **23**, 649–666 (2010). https://doi.org/10.1016/j.neunet.2009.12.007
2. Barto, A., Mirolli, M., Baldassarre, G.: Novelty or surprise? Frontiers in Psychology **4**, 907 (2013). https://doi.org/10.3389/fpsyg.2013.00907
3. Becattini, N., Borgianni, Y., Cascini, G., Rotini, F.: Surprise and design creativity: investigating the drivers of unexpectedness. International Journal of Design Creativity and Innovation **5**(1-2), 29–47 (2017). https://doi.org/10.1080/21650349.2015.1090913
4. Boldt, M., Borg, A., Ickin, S., Gustafsson, J.: Anomaly detection of event sequences using multiple temporal resolutions and markov chains. Knowledge and Information Systems **62**, 669–686 (2019). https://doi.org/10.1007/s10115-019-01365-y
5. Cohen, P., Heeringa, B., Adams, N.: Unsupervised segmentation of categorical time series into episodes. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings. pp. 99–106 (2002). https://doi.org/10.1109/ICDM.2002.1183891
6. Ekman, P., Davidson, R.: The nature of emotion: Fundamental questions. McGraw-Hill (1960)
7. Florez-Larrahondo, G., Bridges, S., Vaughn, R.: Efficient modeling of discrete events for anomaly detection using hidden markov models. Information Security **3650**, 506–514 (2005)
8. Gottlieb, J., Oudeyer, P., Lopes, M., Baranes, A.: Information-seeking, curiosity, and attention: computational and neural mechanisms. Trends Cognitive Science **17**, 585–593 (2013). https://doi.org/doi:10.1016/j.tics.2013.09.001.

9. Helske, S., Helske, J.: Mixture hidden Markov models for sequence data: The seqHMM package in R. Journal of Statistical Software **88**(3), 1–32 (2019). https://doi.org/10.18637/jss.v088.i03
10. Huang, J., Jaysawal, B., Wang, C.: Mining full, inner and tail periodic patterns with perfect, imperfect and asynchronous periodicity simultaneously. Data Mining and Knowledge Discovery **35**, 1225–1257 (2021)
11. Itti, L., Baldi, P.: A principled approach to detecting surprising events in video. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 631–637 vol. 1 (2005). https://doi.org/10.1109/CVPR.2005.40
12. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research **49**(10), 1295 – 1306 (2009). https://doi.org/https://doi.org/10.1016/j.visres.2008.09.007
13. Keogh, E., Lonardi, S., Chiu, B.: 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. p. 550–556. Association for Computing Machinery, New York, NY, USA (2002)
14. Li, H., Homer, N.: A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics **11**(5), 473–483 (2010)
15. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computing **9**(8), 1735–1780 (1997)
16. Liao, T., Fasang, A.: Comparing groups of life-course sequences using the bayesian information criterion and the likelihood-ratio test. Sociological Methodology **51**, 44–85 (2021)
17. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing sax: A novel symbolic representation of time series. Data Mining Knowledge Discovery. **15**(2), 107–144 (2007)
18. Maguire, P., Moser, P., Maguire, R., Keane, M.: Seeing patterns in randomness: A computational model of surprise. Topics in Cognitive Science **11**(1), 103–118 (2019)
19. McGarry, K.: A survey of interestingness measures. Knowledge Engineering Review **20**(1), 39–61 (2005)
20. Nevill-Manning, C., Witten, I.: Identifying hierarchical structure in sequences: A linear-time algorithm. J. Artif. Int. Res. **7**(1), 67–82 (1997)
21. Rabiner, L., Juang, B.: An introduction to hidden markov models. IEEE ASSP Magazine **3**(1), 4–16 (1986). https://doi.org/10.1109/MASSP.1986.1165342
22. Rhienberger, C., Hammitt, J.: Dinner with bayes: On the revision of risk beliefs. Journal of Risk and Uncertainty **57**(3), 253–280 (2018)
23. Rieck, K., Laskov, P.: Linear-time computation of similarity measures for sequential data. Journal Machine Learning Research **9**, 23–48 (Jun 2008)
24. Ritschard, G.: Measuring the nature of individual sequences. Sociological Methods & Research **0**(0) (2021). https://doi.org/10.1177/00491241211036156
25. Shen, Z.: Bao, W., Huang, D.S.:Recurrent Neural Network for Predicting Transcription Factor Binding Sites. Scientific Reports **8**(15270). https://doi.org/10.1038/s41598-018-33321-1 (2018)
26. Wilson, W., Birkin, P., Aickelin, U.: The motif tracking algorithm. International Journal of Automation and Computing **5**(1), 32–44 (2007). https://doi.org/10.1007/s10453-004-5872-7
27. Yang, P., Chen, K., Ching-Chi, H.: Subjective Association Rule Mining: From Point-based Ranking Sequence to Interval-based Temporal Sequence. In: Proceedings of the 10TH International Conference on Machine Learning and Computing (ICMLC 2018). pp. 167–171. Assoc Computing Machinery, 1515 Broadway, New York, NY 10036-9998 USA (2018)