Flexible Exploration and Visualization of Motifs in Biomedical Sensor Data

Arvind Balasubramanian
The University of Texas at Dallas
Richardson, TX 75080
arvind@utdallas.edu

Balakrishnan Prabhakaran
The University of Texas at Dallas
Richardson, TX 75080
bprabhakaran@utdallas.edu

ABSTRACT

Motif discovery in time series is an important data mining task that involves the identification of frequently repeating subsequences, and has been used in biomedical sensor data analysis for pattern detection. However, time series data generated by different biomedical sensors can have noise and variations due to various factors such as calibration issues, drifts while sensing, and environmental conditions. The presence of such noise may induce distortion or variation in some instances of a pattern because of which they might not be identified. A recent work discussed the merits of combining the SAX time series representation with the Sequitur string compression algorithm in finding variable length motifs. Using this motif discovery technique as an example, we propose a more flexible treatment of time series subsequences based on partial similarity matching and an increased margin of tolerance in the interest of finding noisy or distorted motif instances, or possible variations of a particular motif. To complement the proposed approach, we also identify some fundamental requirements in the visualization and exploration of discovered motifs, and present a new motif visualization tool to address them.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data Mining.* H.3.3 [**Information Systems**]: Information Search and Retrieval.

General Terms

Algorithms, Performance, Design.

Keywords

Data Mining, Pattern Discovery, Time Series, Motifs, Noisy Data, Rule Mining, Visualization.

1. INTRODUCTION

Finding approximately repeated subsequences in time series data are referred to as *motifs*. In recent years, motif discovery in time series has been one of the most important data mining tasks [1,6,7,8,9,10], irrespective of domain and nature. The objective of motif discovery techniques is to identify time series subsequences that are similar to each other, or find repetitive patterns in a time series. Data generated by different biomedical sensors can have noise due to various factors such as calibration issues, drifts while sensing, and environmental conditions. The presence of noise may induce distortion in some instances of a pattern because of which they might not be identified. It might also be desirable to detect possible non-trivial variations of a motif. This poses a veritable challenge to motif discovery techniques. While smoothing can be used for cleaning originally noisy signals, general smoothing of the time series is not a solution for overcoming local distortions or variations, since optimality of the smoothing filter is disputable and clarity of potential motifs may be lost. Thus, the solution to capturing variations of motifs is not trivial. There have been many studies [3,8,10,14] that address the issue of motif discovery despite factors such as noise, scaling, translation etc.

In biomedical sensor datasets, it is rarely enough to identify motifs; the interpretation of the detected motif and its relevance to the analysis is highly imperative. The detected motifs need to be associated with specific inferences about the nature of the data or the condition of the subject being monitored. Therefore, it is important to focus on the descriptive aspects of a motif such as the frequency, time duration, amplitude scale etc. that would aid in proper identification and annotation. Efficient and flexible visualization is key to descriptive analysis and inference in time series motif discovery.

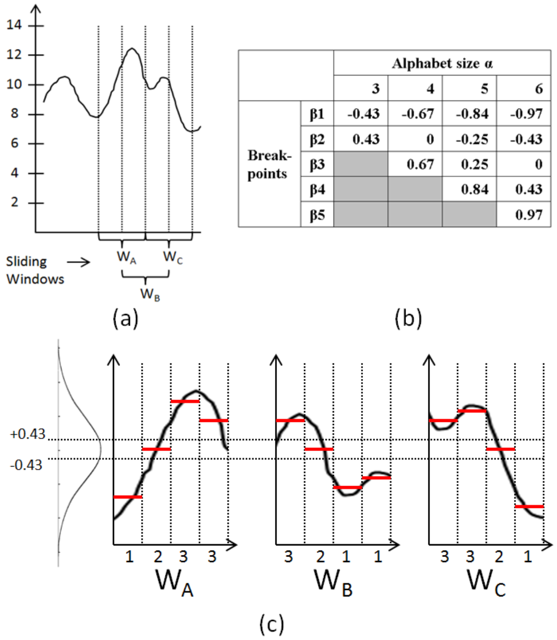
In Section 2, we study the challenges presented to motif discovery by inadequate flexibility by reviewing a recently proposed combination [4] of the popular time series representation technique SAX [2] and a string compression algorithm Sequitur [5] for finding variable length motifs in time series data. We examine the various aspects of these techniques that are affected by noise distortions in data, and propose an approach in Section 3 that provides for a more flexible treatment of time series sequences and lends a margin of tolerance for overlooking localized distortion or variations in the interest of discovering valid patterns instances. Section 4 presents a new motif visualization and exploration tool that addresses some fundamental requirements of motif analysis.

2. BACKGROUND

2.1 SAX and Discretization

The SAX (Symbolic Aggregate approximation) representation [2] is a widely popular symbolic representation for time series data. The two major merits of SAX are (i) efficient dimensionality reduction while retaining essential features; and (ii) lower bounding of the distance measure. Its basic principle is that a time series C of length n can be represented in a w-dimensional space by a vector $\overline{C} = \overline{c}_1, \overline{c}_2 \dots \overline{c}_w$, where the i^{th} element of $\overline{C}$ is given by

KDD-DMH'13, August 11, 2013, Chicago, Illinois, USA.



		Alphabet size α			
		3	4	5	6
Break-points	β1	-0.43	-0.67	-0.84	-0.97
	β2	0.43	0	-0.25	-0.43
	β3		0.67	0.25	0
	β4			0.84	0.43
	β5				0.97

Figure 1. (a) Segmentation of original time series into subsequences using overlapping sliding windows; (b) Chart showing sample of breakpoints that divide the distribution space into equiprobable regions for each alphabet size α; (c) Derivation of SAX representation for each subsequence.

$$\overline{c_i} = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \qquad (1)$$

In SAX, each time subsequence is z-normalized (mean = 0 and SD = 1), and split into *w* equal segments. For each segment, the mean is calculated and a symbol is assigned based on a set of breakpoints that divide the distribution space into α equiprobable regions, where α is the alphabet size. Thus, each time subsequence is converted into a string or word of length *w*, formed by symbols from an alphabet of size α. Both the word length *w* and the alphabet size α are pre-specified. Theoretically, an optimal combination of the two parameters – *w* and α – should be able to efficiently represent the variation in the sequences of any given time series data. Figure 1 shows the usage of SAX to represent time series subsequences as strings of symbols, with *w* = 4 and α = 3{1,2,3}. For the definition of breakpoints, please refer to [2]. In the interest of reducing redundancy while representing motif instances having variable lengths, an approach similar to run-length encoding called *numerosity reduction* has been employed to record consecutive occurrences of identical SAX words.

In their formalization of SAX [2], the authors have defined a distance metric for computing the similarity between two time series subsequences $Q = q_1, q_2 \dots q_w$ and $C = c_1, c_2 \dots c_w$ as

$$MINDIST(Q, C) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} \left(dist(q_i, c_i)\right)^2} \qquad (2)$$

The $dist()$ function is calculated as follows

$$cell_{r,c} = \begin{cases} 0, if\ |r - c| \leq 1 \\ \beta_{\max(r,c)-1} - \beta_{\min(r,c)}, otherwise \end{cases} \qquad (3)$$

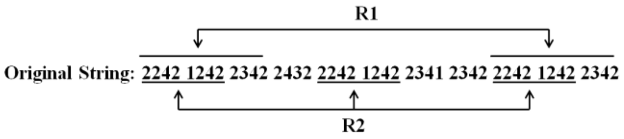
This distance function has the highly desirable quality of lower bounding i.e. the distance between any two time series subsequences in their SAX representation will not be larger than the Euclidean distance between the original subsequences.

Local distortions or variations can cause similar data sequences to be represented using different symbols. While the MINDIST distance measure in combination with a probabilistic approach presented in [3] (discussed in a later section) is robust enough to handle this, motif discovery techniques that rely on *exact string matching* metrics would fail to capture the actual extent of similarity between the two subsequences. This point is of considerable importance to the premise of our study.

2.2 Sequitur and Motif Discovery

While initial motif discovery techniques required the length of the motifs to be pre-specified, recently a grammar induction approach to finding motifs having variable lengths has been proposed in [4]. This approach is based on the Sequitur string compression algorithm [5] that uses a context free grammar based rule building approach to index repeating bigrams in a string of symbols. Every bigram (two consecutive symbols) in the string of symbols is recorded. Whenever a bigram is repeated, all occurrences of the bigram are replaced by a non-terminal symbol. If a rule for such a substitution does not already exist, a new rule is created and added to the existing rule base. Also, rules that are not used more than once are discarded, emphasizing the meaningfulness of the rules that are retained.

Since Sequitur processes strings one symbol at a time from beginning to end, it has been shown to be a useful online technique for finding repetitive patterns. A pattern or motif corresponds to any of the different rules in the rule base, since every rule corresponds to a sequence that occurs at least twice in the symbol string. Sequitur is therefore used to mine rules and identify motifs from a given string of SAX words, where each SAX word is a symbol. The approach provides scope for identifying variable length motifs, as is obvious from the possibility of a non terminal substitution for a bigram of non-terminals. Also, the aforementioned numerosity reduction feature also allows subsequences within each motif to have different lengths.



Original String: 2242 1242 2342 2432 2242 1242 2341 2342 2242 1242 2342

Rules: R0 → R1 2432 R2 2341 2342 R1
R1 → R2 2342
R2 → 2242 1242

Figure 2. Rule mining from a SAX symbol string using Sequitur

The greatest merits of using Sequitur for motif discovery are its efficiency in finding naturally occurring repetitive patterns, and its identification of a hierarchy among the motifs found. Knowledge of the components of a specific motif is highly advantageous for semantic interpretation of the motif and its association with other motifs in the data. However, Sequitur employs exact matching, and does not make use of the distance metric in Eqn(2) to compare two bigrams. Consequently, a match between two bigrams is successful if and only if the bigrams are identical i.e. the edit distance between them is zero.

2.3 Motivation for proposed approach

While the combination of SAX and Sequitur provides for efficient motif discovery and identifying hierarchical associations among motifs, the approach compromises on flexibility in more than once aspect of the motif discovery and analysis.

A distortion in the original sequence may very well be reflected in its SAX representation as well, hindering the identification of distorted instances of a motif. Although it might be possible to capture the variations of the sequences using an optimal selection of the parameters *w* and α, this would necessitate re-segmentation of the original data and thus possible only in an offline environment. Chiu et al [3] addressed the issue of motif distortion in the presence of noise, and proposed a probabilistic approach based on random projection, to find similar SAX words even when they differed by one or more symbols. However, this approach only provides for noise tolerance at the SAX word level through partial matching of SAX words. In other words, this approach can identify distorted instances of a motif, but only as long as the noise is within the range of a SAX segmentation window. Therefore, the approach should be made more flexible to identify motif instances when the distortion occurs over a larger duration than the initial assumption.

As discussed earlier, instead of employing the distance metric in Eqn. 2, Sequitur uses an exact match criterion to match bigrams and build rule hierarchy. Even under the assumption that noise distortion occurs only within a SAX word (resulting in at least a partial dissimilarity in symbols), Sequitur would lose the partial similarity information, and would never match a bigram with its partially distorted counterpart. Thus, even if a rule is formed by matching the 'clean' instances of a potential motif and the motif is identified by the rule, any distorted instances would be lost, no matter how small the distortion.

Since noise might introduce sufficient distortion in an instance of a repeating motif to hinder its identification, one can always measure the extent of similarity between the noisy instance and other instances to determine if they belong to the same motif. An argument can be made against partial similarity that it might attempt to correlate two different patterns as being variations of the same pattern. Such false positives might hinder motif discovery. However, the concept behind partial match is to have a flexible perspective towards motifs, since naturally occurring motifs are seldom exactly similar across their repetitions. Having a margin of tolerance for dissimilarity enables us to look beyond the tolerance levels of SAX and might help in capturing similarities lost due to the strict approach of the SAX representation. At any rate, it could be useful to know that there is a similarity overlap between two distinct time sequences. This would just provide more information and the decision of whether or not the overlap is enough for motif identification can be dealt with manually or through heuristic computation.

3. PROPOSED APPROACH

3.1 Implementation

The approach proposed in this work attempts to induce flexibility in motif discovery while retaining the benefits provided by the original approach. In the original Sequitur algorithm, bigrams are matched exactly and rules formed only upon encountering exact duplicates of a previously recorded bigram. Thus, each rule maps a non-terminal to a single bigram. In contrast, the proposed approach allows *multiple bigrams* to be substituted by the *same rule*. Each rule in the rule base would correspond to not a single bigram, but instead a set of bigrams, all of which are evaluated to be similar to each other than others. To accommodate this situation, a record structure is implemented, where each record consists of a list of bigrams (Figure 3). For each record, a *representative* bigram is calculated from the member bigrams of the record. The representative bigram is not one of the actual

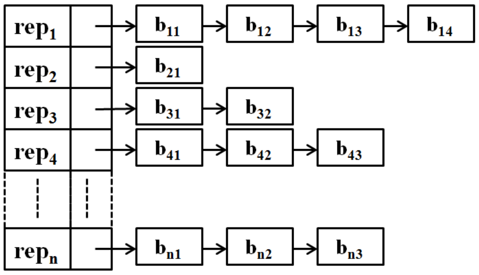


Figure 3. Each record consists of a list of bigrams that are more similar to each other than to others. Each record also stores a representative bigram which is the approximate average of all bigrams in the list at any given point in time.

Table 1. Algorithm for flexible motif discovery

```
Algorithm: Proposed Approach
Repeat till end of symbol stream:
  Read symbol
  Form cand.bigram using previous symbol
  and current symbol
  Search lists for exact match
  If exact match is found
    RuleCheck()
  Else
    Compare to rep.bigrams in record for
    partial match
    If partial match is not found
      Create new list
      Add cand.bigram
      Make bigram the rep.bigram
    Else If partial match is found
      Compare to list members for partial
      match
      If p-match found
        Enter bigram into list
        Update rep.bigram
        RuleCheck()
      Else
        Create new list
        Add cand.bigram
        Make bigram the rep.bigram

RuleCheck()
If Rule exists for record
  Non-terminal substitution using Rule
Else
  Create Rule for bigram substitution
  Replace All instances of bigram set
  (two instances including the current
  one) with Non-terminal
Recursively execute bigram routine for
any new bigrams formed
```

Table 2. Comparing resultant rules from exact match and partial match criteria of Sequitur

Original String:	1234 1134 1234 1244 1243 1343 1342 1432 1423 2423 2313 2314 2324 2224 2124 1124 1134 1144
Exact Match:	R0 -> [1234 1134 1234 1244 1243 1343 1342 1432 1423 2423 2313 2314 2324 2224 2124 1124 1134 1144]

Partial Match:	R0 -> [R1 R2 R2 R3 R3 R1 1134 1144];
	R1 -> [R4 R4];
	R2 -> [1243 1343]; [1342 1432];
	R3 -> [1423 2423]; [2313 2314];
	R4 -> [1234 1134]; [1234 1244];
	[2324 2224]; [2124 1124];

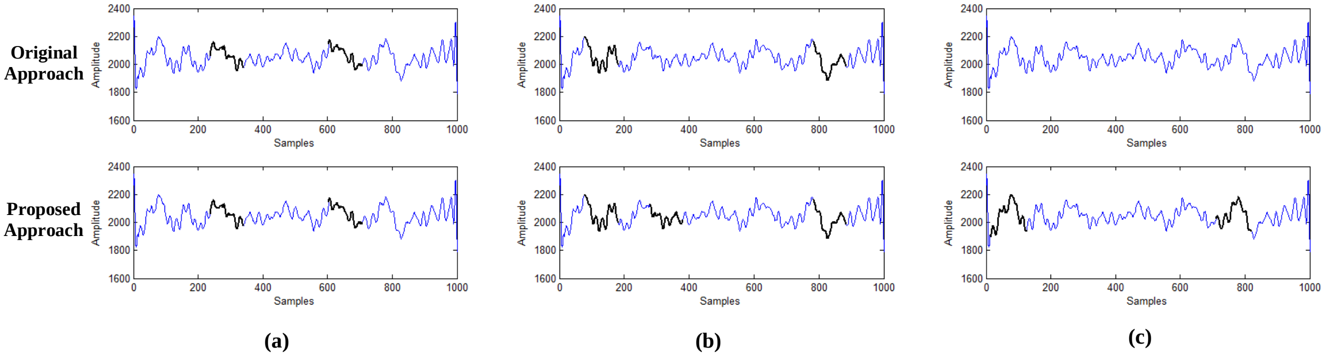


Figure 4. Comparison of Motifs detected by original approach and the proposed approach in a smoothed EEG signal. The proposed approach (a) captured every motif identified by the original approach, (b), identified more instances for certain motifs, (c) identified previously undiscovered motifs that were not captured by the original approach.

member bigrams but the best approximation of the current members of the record. This notion of a representative is similar to that of a cluster mean in cluster analysis. The representative bigram of a list is updated whenever a new bigram is added to it.

Whenever a new bigram (referred to as *candidate* bigram) needs to be processed for entry in the record, an exact match routine is executed to find any bigram in the records identical to the candidate. If an exact match is found, it means it already exists in the list and does not need to be re-entered. The routine can skip the entry and representative update, and proceed with rule creation and/or non-terminal substitution, depending on whether or not a rule exists for that list. If an exact match is not found, then the candidate bigram needs to be either entered into the record corresponding to the list of bigrams that are sufficiently similar to the candidate, or entered into a new record on its own. This is resolved in the following manner.

The candidate bigram is first matched with the representative bigrams for each record for partial similarity. If an admissible partial match with a representative is not found, then a new record is created, with the candidate bigram as the first member of the new record. Thus, it becomes both the representative as well as the first member of its list. However, if an admissible partial match is found with the representative bigram of a certain record, the candidate bigram is then entered into the bigram list corresponding to that record, if and only if there is at least one actual member of that list that has an admissible partial match with the candidate. This ensures that the candidate is indeed being added to a list where it belongs in terms of similarity. If such an actual member does not exist in the list, the candidate is not entered in that list, and a new record is created for the candidate.

Also, if the candidate is entered into a pre-existing bigram list that does not have a rule associated with it, it means the candidate is the second bigram to enter the list. In this case, a new rule is created for the list, with a new non-terminal substituting both bigrams belonging to the list in the string. If a rule already exists for the bigram list, the candidate bigram being entered into the list is substituted in the original symbol string by the non-terminal corresponding to the rule.

Table 1 shows the pseudocode for the proposed approach using the partial match criteria. It is important to note that the modification made to the implementation of Sequitur to accommodate partial similarity does not affect the algorithm's original performance. If no margin is allowed for a partial match (match threshold = 0), then the proposed approach performs identical to the original Sequitur algorithm. The utility of our approach can be demonstrated by the example presented in Table 2, which compares results obtained by using the original exact match criterion with the results obtained by the use of partial match using the MINDIST measure with a match threshold of 0.1. Since the original string does not have any repeating bigrams, no rule is put forward. However, the partial match approach puts forward a set of rules that illustrate the similarities between the various bigrams.

3.2 Experiments and Discussion

In this section, we present some examples that demonstrate the proposed approach and compare its performance with the original approach. The MINDIST measure was used for calculating the similarity between bigrams composed of terminal symbols for evaluating matches. The threshold for match criterion is accepted as input from the user. Two bigrams were considered to be similar only if the similarity value satisfied the threshold. Since a match threshold of zero would enforce the exact match criterion, the original approach can now be considered a strict version of the proposed approach.

The proposed approach was implemented and used for motif discovery on a number of biomedical time series datasets such as EEG, ECG etc. [12]. In our implementation, segmentation of data is done by SAX using an overlapping window with numerosity reduction. The window sizes used were typically 50 to 100 samples long, and the values for the word length and alphabet size ranged from 4 to 5 and 4 to 6, respectively. In a SAX word discretization, two adjacent SAX words that form a bigram can very well be the discretized form of two overlapping windows. However, since, numerosity reduction does not guarantee that, and two adjacent words are more often than not non-overlapping, the two adjacent words were considered to have a sequential association. This assumption holds to all similarity calculations and rule mining tasks.

Figure 4 offers a juxtaposition of motifs identified by the original and proposed approaches. The data used is EEG data, smoothed by a moving average of 50 samples (to ignore local variations and capture higher level motifs). The sliding window size, the SAX word length, the SAX alphabet size and match threshold for partial match were 100 samples, 5, 6 and 0.1. While the old approach retrieved 9 motifs, the proposed approach retrieved 42 motifs. As expected, several of the motifs returned by the partial match approach were spurious and redundant. However, the proposed approach was able to not only able to capture every motif identified by the old approach (Figure 4(a)), it identified more instances for certain motifs (Figure 4(b)) as well as totally new motifs that were not identified by the old approach (Figure 4(c)).

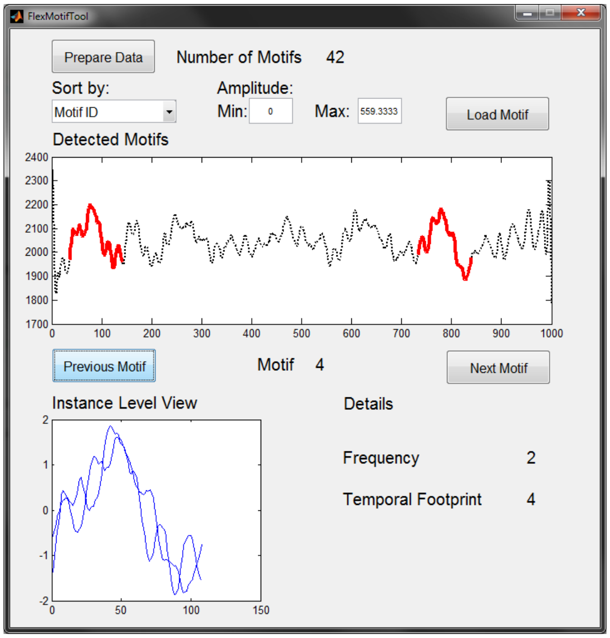
It is a challenge to compare the performance of the exact and partial match approaches in an unsupervised motif discovery task due to number of variables in the process. The quality of motifs returned by both approaches depends heavily on the effectiveness of the initial segmentation using SAX. Just as there is no optimal set of values for the word length and alphabet size parameters of SAX that can be generalized across all datasets, the optimal match threshold for the distance measure used in the partial match criteria also has to be derived empirically. While one might expect that the number of rules returned by a partial match approach would be lesser, this cannot be predicted since the number of rules depends not only on the number of repeating bigrams, but also on the repeated associations of the rules. As noted in [4], the number of rules generated by Sequitur depends largely on the complexity of the dataset as well as the choice of SAX parameters, and these rules represent distinct motifs. While the new flexible approach might result in the merging of some of these into a single motif group by mapping multiple bigrams to each rule, one cannot assume that it returns lesser rules than Sequitur. Due to the inherent flexibility in the rule-building approach, the rule hierarchy may have more levels than before with two or more rules constituting higher level rules, i.e, motifs with larger temporal footprints.

4. Motif Visualization Tool

While inducing greater flexibility in motif discovery enhances the scope of identifying possible variations of a motif, the benefits of this flexible approach can be further exploited through the provision of an efficient visualization and exploration functionality. A number of studies have focused on the visualization and searching of time series motifs [15, 16]. In this section, we present a new motif visualization tool to complement the motif discovery approach presented in the previous sections.

There are certain requirements in the exploration and visualization of the detected motifs that need to be addressed. The merits of finding the most significant motif have been established in previous studies. The 'significance' of a motif can be evaluated based on its frequency of occurrence (number of non-trivial instances), time period or duration of the motif etc., and the choice of such a parameter would be dictated by the purpose of the analysis. It would be beneficial to provide the option of selecting a parameter, and present detected motifs ranked by the selected parameter. Secondly, SAX amplitude-normalizes time series information in a piece-wise manner to result in a representation that is constituted by equiprobable symbols. Due to this normalization step, SAX generates the same (or similar) representations for sections of the time series that may differ in scale of amplitude, but have the same underlying pattern. Therefore, a detected motif might have instances that are not uniform in amplitude scale. In the interest of proper interpretation of the motif, it is important to provide the functionality of specifying a range for amplitude variation in the motif instances. Finally, the integral aspects of visualizing a motif are identifying the location of its occurrences in the original data as well as illustrating the consistency of its different instances.

To exploit the flexibility of the proposed approach as well as to enhance control over motif exploration, we present a simple interactive tool that offers the functionality of visualizing the motifs and analyzing the motifs detected by approach. This program enables the user to (i) view the total count of detected motifs, as well as basic statistics such as the frequency and duration for each motif, (ii) visualize each distinct motif detected by the proposed motif discovery approach, (iii) locate and plot the instances of each motif in the original time series data, as well as



Number of Motifs 42

Sort by: Motif ID

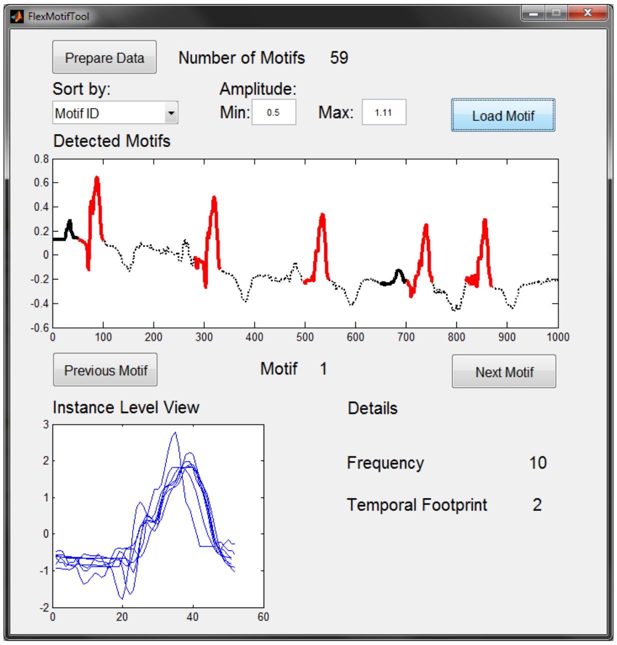
Amplitude: Min: 0 Max: 559.3333

Motif 4

Frequency 2

Temporal Footprint 4

Figure 5. Snapshot of the Motif Visualization Tool. The original curve is plotted as a dotted line and the detected motif instances are shown in red. The instance level view displays a superimposed plot of the detected instances.



Number of Motifs 59

Sort by: Motif ID

Amplitude: Min: 0.5 Max: 1.11

Motif 1

Frequency 10

Temporal Footprint 2

Figure 6. Utility of Filtering Option on Amplitude Variation of Motif; motif instances satisfying specified amplitude range are in red, rest in black.

illustrate their consistency using an instance level superimposed plot, (iv) rank detected motifs based on different parameters such as frequency, temporal footprint (duration) etc., and (v) set ranges for the amplitude scale to filter and highlight motif instances that conform to the scale constraints. The fundamental merit of this visualization tool is its simplicity of operation and efficient controls for navigation through the set of detected motifs. Forward and backward navigation is made possible using navigation controls. The ranking functionality accepts the parameter from the user through a pop-up menu and sorts the detected motifs, such that the navigation controls show motifs in the order of their rank. With respect to filtering instances by amplitude, the tool first reports the maximum and minimum values in the time series data, and accepts any valid amplitude range of interest for filtering. It also provides the option of either showing only the desired instances or showing all instances with the desired instances highlighted. Figure 5 shows a snapshot of the tool displaying motif instances, and Figure 6 demonstrates amplitude filtering. The time duration or temporal footprint of a motif can be viewed in terms of number of samples or the number of SAX words used in the discretized form of the motif. The functionality of filtering and ranking the detected motifs using a parameter of choice would prove highly useful in the analysis of biomedical data streams. We are currently adding more modules to the visualization tool such as enable search by query, advanced statistics report etc. that would prove valuable to clinicians and medical data researchers and analysts.

5. CONCLUSION

The proposed approach presented in this paper offers a flexible approach to motif discovery in biomedical sensor data. The partial match criterion allows consideration of sequences that are not exactly similar, but are sufficiently similar to be considered instances of a motif. The approach facilitates motif discovery with a tolerance margin for distortion or variation of time subsequences in the presence of noise, amplitude scaling etc. The interactive motif visualization and exploration tool complements the flexible motif discovery approach, enabling easy navigation, filtering and ranking of detected motifs. Our future studies would be directed at identifying relationships between the variability in the data and the importance attributed to factors included in the match criteria. The challenges for extending the proposed solution for motif discovery in multi-dimensional multi-sensor biomedical sensor data also need to be addressed.

NOTE: All code, datasets and supplementary examples for the proposed motif discovery approach and the motif visualization tool presented in this paper are available at the support webpage: http://www.utdallas.edu/~arvind/support/exploreMotif.html.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1012975. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Dr. Eamonn Keogh for his inputs on a previous version of this work.

6. REFERENCES

[1] Lin, J., Keogh, E., Lonardi, S. & Patel, P. (2002). Finding Motifs in Time Series. In proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, July 23- 26. pp. 53-68.

[2] Lin, J., Keogh, E., Lonardi, S. & Chiu, B. (2003) A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA. June 13.

[3] Chiu, B. Keogh, E., & Lonardi, S. (2003). Probabilistic Discovery of Time Series Motifs. In the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 24 - 27, 2003. Washington, DC, USA. pp 493-498.

[4] Yuan Li, Jessica Lin, and Tim Oates. 2012. Visualizing variable-length time series motifs. In Proceedings of the 2012 SIAM International Conference on Data Mining. Anaheim, CA, April 26-28, 2012. pp 895-906.

[5] C.G. Nevill-Manning and I.H. Witten. (1997). Identifying Hierarchical Structure in Sequences: A linear-time algorithm. Journal of Artificial Intelligence Research, 7, 67-82.

[6] Keogh, E,. Chakrabarti, K,. Pazzani, M. & Mehrotra "Dimensionality reduction for fast similarity search in large time series databases", Journal of Knowledge and Information Systems. (2000).

[7] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover. (2009). Exact Discovery of Time Series Motifs. In proceedings of the 2009 SIAM International Conference on Data Mining. April 30- May 2. Sparks, NV.

[8] Agrawal, R., Lin, K. I., Sawhney, H. S. & Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In proceedings of the 21st Int'l Conference on Very Large Databases. Zurich, Switzerland, Sept. pp 490-50.

[9] Chan, K. & Fu, A. W. (1999). Efficient time series matching by wavelets. In proceedings of the 15th IEEE Int'l Conference on Data Engineering. Sydney, Australia, Mar 23-26. pp 126-133.

[10] Dragomir Yankov, Eamonn Keogh, Jose Medina, Bill Chiu, Victor Zordan, Detecting time series motifs under uniform scaling, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, August 12-15, 2007, San Jose, California, USA.

[11] Arvind Balasubramanian, Balakrishnan Prabhakaran, Technical Report UTDCS-02-13, Department of Computer Science, The University of Texas at Dallas, February 2013. http://www.utdallas.edu/~arvind/publications/techreport_UTDCS-02-13.pdf

[12] E. Keogh. The UCR Time Series Data Mining Archive.

[13] Eibe Frank . Sequitur Java port (http://sequitur.info/java/)

[14] T. Armstrong, E. Drewniak, Unsupervised Discovery of Motifs Under Amplitude Scaling and Shifting in Time Series Databases, Proceedings of the 7th International Conference on Machine Learning and Data Mining in Pattern Recognition, 2011.

[15] Lin, J., E. Keogh, and S. Lonard, "Visualizing and discovering non-trivial patterns in large time series databases," Information Visualization, 4(2):61-82, July, 2005.

[16] Hochheiser H and Shneiderman B. Interactive Exploration of Time Series Data. the 4th Int'l Conference on Discovery Science 2001 (Washington D.C.), Springer-Verlag; 441-446.