Motif Discovery in Physiological Datasets: A Methodology for Inferring Predictive Elements

Published in final edited form as:
ACM Trans Knowl Discov Data. 2010 January ; 4(1): 2. doi:10.1145/1644873.1644875.

Motif Discovery in Physiological Datasets: A Methodology for Inferring Predictive Elements

Zeeshan Syed,
University of Michigan

Collin Stultz,
Massachusetts Institute of Technology

Manolis Kellis,
Massachusetts Institute of Technology

Piotr Indyk, and
Massachusetts Institute of Technology

John Guttag
Massachusetts Institute of Technology

Zeeshan Syed: zhs@umich.edu

Abstract

In this article, we propose a methodology for identifying predictive physiological patterns in the absence of prior knowledge. We use the principle of conservation to identify activity that consistently precedes an outcome in patients, and describe a two-stage process that allows us to efficiently search for such patterns in large datasets. This involves first transforming continuous physiological signals from patients into symbolic sequences, and then searching for patterns in these reduced representations that are strongly associated with an outcome.

Our strategy of identifying conserved activity that is unlikely to have occurred purely by chance in symbolic data is analogous to the discovery of regulatory motifs in genomic datasets. We build upon existing work in this area, generalizing the notion of a regulatory motif and enhancing current techniques to operate robustly on non-genomic data. We also address two significant considerations associated with motif discovery in general: computational efficiency and robustness in the presence of degeneracy and noise. To deal with these issues, we introduce the concept of active regions and new subset-based techniques such as a two-layer Gibbs sampling algorithm. These extensions allow for a framework for information inference, where precursors are identified as approximately conserved activity of arbitrary complexity preceding multiple occurrences of an event.

We evaluated our solution on a population of patients who experienced sudden cardiac death and attempted to discover electrocardiographic activity that may be associated with the endpoint of

Z. Syed was previously affiliated with the Massachusetts Institute of Technology.

death. To assess the predictive patterns discovered, we compared likelihood scores for motifs in the sudden death population against control populations of normal individuals and those with non-fatal supraventricular arrhythmias. Our results suggest that predictive motif discovery may be able to identify clinically relevant information even in the absence of significant prior knowledge.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.5.4 [**Pattern Recognition**]: Applications; J.3 [**Life and Medical Sciences**]

General Terms

Algorithms; Design; Experimentation; Performance

Additional Key Words and Phrases

Gibbs sampling; knowledge discovery; data mining; motifs; physiological signals; inference

1. INTRODUCTION

The subject of finding predictive elements has been extensively studied in a wide variety of contexts including geodesic, medical, and financial data. In this article, we present a motif discovery methodology for discovering precursors. While we focus mainly on physiological datasets, we present general techniques that may be broadly applicable to a wider group of signals.

We model prediction as the problem of identifying activity that consistently precedes an event of interest. In the absence of any prior knowledge, this activity can be discovered by observing multiple occurrences of the event and detecting statistically significant commonalities in the data preceding it, by searching for conserved elements unlikely to occur purely by chance prior to the event of interest (Figure 1). To handle noise, we further adopt a relaxed view of conservation, whereby precursors may approximately match or be altogether absent on some observations of the event. A further practical consideration is that the search be computationally efficient to handle large amounts of data resulting from multiple observations.

This model of prediction is similar to the search for regulatory motifs in the setting of computational biology. Motif discovery techniques operate on genomic datasets and search for DNA sequences that are conserved across genomes. We generalize this model and describe how the search for precursors to acute clinical events can be carried out in an analogous manner, by first converting continuous physiological signals into an alphabetical representation, and then mining this representation for conserved activity. A variety of randomized greedy algorithms can be used to efficiently carry out the search for such patterns. We use techniques such as TCM and Gibbs sampling as the foundation of our work, and enhance them to operate on data with highly divergent background distributions of symbols, frequent noise and patterns of increased degeneracy relative to genomic data.

The rest of this article describes the proposed unsupervised inference methodology. While the techniques we suggest can be used on a variety of signals and are sufficiently general-purpose, we motivate them in the more concrete setting of searching for predictive activity

in physiological signals. We detail the challenges associated with such an approach and describe its benefits and limitations.

Section 2 details the concept and challenges of representing continuous physiological signals as symbolic strings. Section 3 presents a similar discussion of the problem of detecting predictive motifs in string data. Section 4 describes existing computational biology algorithms for motif detection, while Section 5 proposes data transformations and algorithms (including a two-level Gibbs sampling technique) that have been augmented to search for motifs in a computationally-efficient manner in the presence of noise and degeneracy. An application of our work to sudden cardiac death data is discussed in Section 6. Related work is presented in Section 7. Finally, a summary and conclusions appear in Section 8.

2. SYMBOLIZATION

2.1 Symbolic Representation of Physiological Data

The notion of representing physiological signals as symbolic sequences follows from the quasi-periodic nature of many important signals. For example, data from the heart and lungs often comprises units such as heart beats or breaths, which are repetitive. It is often more natural to analyze physiological signals in terms of these units than at the level of raw samples.

We use the property of quasi-periodicity in physiological signals to determine appropriate boundaries for segmentation, and then replace each unit with a single symbol. In doing so, we exploit the underlying repetitive structure and redundancy to obtain a layer of data reduction. The raw physiological data is reexpressed to retain salient differences between units of quasi-periodic activity while abstracting away the common structure. For example, as shown in Figure 2, raw ECG data can be partitioned at the level of heart beats into different equivalence classes, each of which is assigned a unique alphabetic label for identification. This reduces the data rate from around 4000 bits/second (for a beat lasting one second in a signal sampled at 360 Hz with 11 bit quantization) to n bits/second (where n depends upon the number of bits needed to differentiate between symbols, two for this case).

The data reduction introduced by symbolization reduces the search space for the detection of interesting activity and provides a significant computational advantage over working in the original space of the raw signal. A further advantage of using symbolization is that it implicitly abstracts away some of the time-normalization issues that complicate the use of cross-correlation and other techniques that operate on raw time samples.

2.2 Creating Symbolic Representations

To transform continuous waveforms into a string representation that can be mined for patterns more efficiently, we propose segmenting the original signal into intervals and then assigning an alphabetic label to each token. This effectively transforms the original data into a sequence of symbols, and maps the problem into the domain of string algorithms.

The task of assigning labels can be carried out in a number of different ways. One approach is to use clinical information to partition segmented tokens into equivalence classes. This approach provides a set of symbols that have a fixed meaning in a medical context and can be shared across a population. For example, the ECG signal in Figure 3 can be decomposed into RR-intervals as shown (each RR-interval corresponds to the period between two successive contractions of the ventricles of the heart: the period between successive sharp spikes in the raw ECG tracings). Each RR-interval can then be labeled using existing annotations for electrophysiological activity. RR-intervals associated with normal heart

beats are labeled *N*, while those associated with abnormal contractions originating from ventricular regions are labeled *V*.

The approach of using clinical annotations is restricted to detecting predictive activity that expresses itself in terms of known clinical classes. It does not allow for the isolation of changes at the level of variations within particular classes. This is important because the granularity of later analysis is constrained by the granularity of labeling.

From a knowledge discovery goal, it is appealing to derive the alphabet for symbolization directly from the data itself. Techniques such as those in Syed et al. [2007] can be employed to achieve this goal. While the approach of generating a patient-specific symbolic representation is powerful in its ability to capture significant changes across a patient, it poses the problem that the clusters are derived separately for each patient. This restricts comparisons across a population. A possible means for addressing this issue is to use a semisupervised approach where the symbols derived for each patient are related by a human expert. This allows for the symbols to be dynamically derived based on characteristics inherent in the data itself, and for these symbols to be related and compared across a population.

At present, registering patient-specific symbols in a fully automated manner across a population represents an area of continuing work. The discussion that follows therefore focuses on the use of clinical annotations (or of semisupervised symbols related manually across patients) despite the possible benefits of patient-specific symbols.

3. PHYSIOLOGICAL MOTIFS

3.1 Physiological Motifs in Symbolic Data

In the setting of computational biology, regulatory motifs correspond to short DNA sequences that regulate gene expression. This notion of a genetic switch that controls activity further downstream is well-suited to our model for prediction. We generalize this idea and choose to model regulatory motifs as sequential triggers that precede abrupt clinical events and are conserved across a population of patients owing to an association with the event.

A recent strategy for regulatory motif discovery that has gained popularity is to make use of comparative genomics [Kellis et al. 2003]. This allows for the discovery of regulatory elements by exploiting their evolutionary conservation across related species. Under this approach, regulatory motif discovery can be viewed computationally as finding sequences that are recurrent is a group of strings, upstream of specified endpoints.

The problem of regulatory motif discovery can be stated more formally in either a combinatorial or probabilistic framework [Jones and Pevzner 2004]. While the two frameworks both attempt to identify similar preceding subsequences, they may lead to slightly different results and require distinct algorithmic techniques.

- *Combinatorial*. Given a set of sequences $\{s_1,\ldots, s_N\}$, find a subsequence $m_1,\ldots, m_W$ that occurs in all s_i with k or fewer differences.
- *Probabilistic*. Given a set of sequences $\{s_1,\ldots,s_N\}$, find a set of starting positions $\{p_1,\ldots,p_N\}$ in the sequences that lead to the best (as defined in the following) $A \times W$ profile matrix M (where A is the number of different symbols in the data and W is the length of the motif).

For the probabilistic case, the profile matrix is derived from the subsequences of length W immediately following the starting positions $p_1,\ldots, p_N$ in each of $s_1,\ldots, s_N$. These

subsequences are lined up and the probability of each of the A unique symbols at every one of the W motif positions is estimated. $M(x, y)$ then gives the probability that the motif has character x at position y. The resulting profile matrix can be scored using different criteria with the implicit goal of seeking a nontrivial profile that is strongly conserved at each position and best explains the data. The scoring function most often used is the log-odds likelihood:

$$score = \sum_{i=1}^{N} \sum_{j=1}^{W} \log \left[\frac{M(s_i(p_i + j - 1), j)}{B(s_i(p_i + j - 1))} \right], \quad (1)$$

where B gives the background distribution of each unique symbol in the data. Effectively, this calculates the log-likelihood of a motif while compensating for trivial occurrences that would be seen in the data merely due to the frequent occurrence of certain symbols.

3.2 Challenges Associated with Motif Detection in Symbolic Signals

The problem of motif discovery gives rise to a number of issues in the physiological setting. This section discusses the major challenges faced when modeling acute clinical events as physiological motifs.

3.2.1 Symbol Distribution Skews—A complication arising in the context of physiological signals is that of the sparsity of abnormal activity. Periods with interesting events are typically separated by long, variable-sized runs of normal behavior—the distribution of the symbols is significantly skewed in favor of normal labels. This increases the number of trivial motifs in the data and consequently the running time of the motif discovery algorithms. In addition, for algorithms such as TCM and Gibbs sampling, discussed in Section 4, a secondary effect resulting from the presence of long stretches of normal behavior is that the starting locations chosen randomly may often correspond to uninteresting regions of the signal, further increasing time to convergence.

3.2.2 Motif Degeneracy—The issue of degeneracy is frequently encountered in DNA sequences and assumes a critical role for physiological motifs as well. Predictive patterns may be approximately conserved across some patients in a population, while in others, they may be missing altogether. This results from a variety of factors, including differences in the age, gender, clinical history, medications, and lifestyle of patients, as well as noise obscuring predictive patterns in some recordings.

The goal of detecting imperfectly conserved activity represents a significant challenge to the task of discovering precursors. Since patterns can vary, the process of determining whether a pattern appears in a patient, is required to explore a large search space, spanning all possible variations. Similarly, the fact that some patients may have the predictive activity obscured due to noise requires recognizing these cases and preventing motif discovery algorithms from forcibly incorporating this data in the search process.

4. COMPUTATIONAL BIOLOGY ALGORITHMS FOR MOTIF DISCOVERY

In this section, we review three popular algorithms for finding regulatory motifs using comparative genomics; the Two Component Mixture (TCM) algorithm using expectation-maximization, Gibbs sampling, and Consensus. TCM and Gibbs sampling attempt to solve the probabilistic formulation of motif discovery, while Consensus focuses on the combinatorial problem.

4.1 Two Component Mixture (TCM)

TCM is an enhancement to the basic EM algorithm [Bailey and Elkan 1995], which essentially reduces the search into two smaller, decoupled problems. The first (the M-step) involves constructing the profile for a motif given a set of fuzzy starting positions $p_1,\ldots, p_N$ in the input sequences (the M-step). The second (the E-step) then uses this matrix profile representation to score all possible starting positions in every sequence and then update the initial $p_1,\ldots,p_N$.

The overall TCM algorithm operates in the following manner.

TCM Algorithm

TCM($\{s_1, \ldots, s_N\}$, W):

1 Set random starting positions $p_1, \ldots, p_N$

2 Do

- **I** M-step to update profile matrix
- **II** E-step to update starting positions

Until the change in the score of M is less than some threshold ε.

The M-step of TCM estimates the profile matrix using the probability Z_{ij} that the motif starts in sequence i at position j. As a first step, the values $n_{c,k}$, which indicate how often the character c occurs at position k in the motif, are estimated.

$$n_{c,k}=\begin{cases}\sum_{i}\sum_{j\mid s_{i,j}=c} Z_{ij} & k>0\\ n_c-\sum_{j=1}^{W} n_{c,j} & k=0\end{cases} \tag{2}$$

$k = 0$ represents the case where character c occurs in the sequence outside the motif, while n_c gives the total number of times c occurs in the data. Using these values, we can obtain a profile matrix M as follows:

$$M_{c,k}=\frac{n_{c,k}+d_{c,k}}{\sum_{a}(n_{a,k}+d_{a,k})}, \tag{3}$$

where $d_{c,k}$ denotes the pseudocount for character c and helps ensure that the probability of c at position k is not zero while estimating frequencies from finite data [Bailey and Elkan 1995].

In addition to computing the profile matrix during the M-step, TCM also calculates a prior probability that a motif might start arbitrarily at any position in the data. This is denoted by λ and is obtained by taking the average of Z_{ij} across all sequences and positions.

TCM primarily differs from other EM approaches to motif discovery in its E-step. For every sequence s_i in the dataset, TCM assigns a likelihood L_{ij} to the W-mer starting at each position j:

$$L_{ij}(1)=\Pr(s_{ij}|Z_{ij}=1,M,b)=\prod_{k=j}^{j+W-1} M_{k-j+1,c_k} \tag{4}$$

and

$$L_{ij}(0)=\Pr(s_{ij}|Z_{ij}=0,M,b)=\prod_{k=j}^{j+W-1} b_{c_k}, \tag{5}$$

where b gives the background probability for each character in the dataset. For iteration t of TCM, the values of Z_{ij} can then be estimated using:

$$Z_{ij}^{(t)}=\frac{L_{ij}^{(t)}(1)\lambda^{(t)}}{L_{ij}^{(t)}(0)\,[1-\lambda^{(t)}]+L_{ij}^{(t)}(1)\lambda^{(t)}}. \tag{6}$$

4.2 Gibbs Sampling

Gibbs sampling [Gert et al. 2002] can be viewed as a stochastic analogue of EM for finding regulatory motifs and is less susceptible to local minima than EM. It is also much faster and uses less memory in practice. This is because unlike EM, the Gibbs sampling approach keeps track only of the starting locations, $p_1,\ldots,p_N$, of the motif in each sequence and does not maintain a distribution over all possible starting positions for the motif (the Z_{ij} in TCM representing fuzzy starting positions, are replaced by hard $p_1,\ldots,p_N$).

The Gibbs sampling algorithm for motif discovery can then be written as follows.

Gibbs Sampling Algorithm

GIBBS($\{s_1, \ldots, s_N\}$, W):

1. Set random initial values for p
2. Do
   - i. Select sequence s_i at random
   - ii. Estimate M from set $\{s_1, \ldots, s_N\} - s_i$
   - iii. Use M to score all starts in s_i
   - iv. Pick start p_i with probability proportional to its score

   Until the change in the score of M is less than some threshold ε.

Gibbs sampling is less dependent on the initial parameters than TCM and therefore more versatile. However, it is dependent on all sequences having the motif. This is an inefficiency we address in our work.

4.3 Consensus

Consensus [Stormo and Hartzell 1989] is a greedy motif clustering algorithm that picks out two sequences at random, finds the most conserved pairs among them and then iterates over

all the remaining sequences adding the *W*-mers that best match the results of the previous iteration at every stage.

The Consensus algorithm is as follows.

Consensus Algorithm

CONSENSUS($\{s_1, \ldots, s_N\}$, *W*):

1 Pick sequences s_i and s_j at random

2 Find most similar *W*-mers in s_i and s_j

3 For each unprocessed sequence s_k

i Expand solution set with *W*-mers from s_k that match best with previous ones.

5. DATA TRANSFORMATIONS AND SUBSET-BASED TECHNIQUES

5.1 Active Regions

The issue of skewed symbol distributions can be addressed by removing long stretches of activity that are known to be uninteresting. By definition, a predictive motif is associated with an acute clinical event and must be associated with abnormal activity. As a result, trivial motifs comprising normal activity can be trimmed away to reduce the running time associated with the motif-discovery algorithms. For example, given the sequence:

V JV J J N N N N N N N N N N V N V N B B r,

a possible reduction of this data would be:

V J V J J N+V N+V N+B B r.

This technique is associated with a significant loss of information. Specifically, the search for motifs proceeds in the transformed space, and the N+ regular expression may occur in motifs without a consistent meaning (it may be arbitrarily long in some patients). The more general issue here is that conservation of a pattern in the transformed space does not imply conservation in the original signals.

To avoid this issue, we identify regions of abnormal activity—active regions—by splicing out trivial periods in the signal. Given a motif length *W*, this involves iterating over the data and removing all normal symbols that would occur only in trivial motifs. This approach preserves the temporal structure of abnormal stretches of the signal, ensuring that the motifs correspond to patterns that are conserved in all of the original signals. For example, using this approach for a motif of length 3, the original example pattern would map to:

V J V J J N N V N V N B B r.

5.2 Gibbs^2 and Seeded Consensus

The Gibbs sampling algorithm in Section 4 assumes that a motif is present in all sequences. To deal with the issue of degeneracy, where noise may obscure the predictive pattern completely for some patients, we propose a new algorithm that provides a layer of robustness while dealing with a population where activity may be altogether absent in some of the observed examples. This is achieved by adding a second layer of Gibbs sampling to the original algorithm, leading to the Gibbs^2 algorithm presented here.

The Gibbs^2 algorithm operates at any time on a working subset $V = \{\upsilon_1,\ldots, \upsilon_C\}$ of the original sequences $\{s_1,\ldots, s_N\}$. Sequences are dynamically swapped into and out of this set with the goal of replacing poor matches with potentially better options. The underlying goal is to arrive at a cluster of sequences that share a strongly conserved motif.

The initial subset of sequences is chosen at random, and at each iteration, a single sequence, υ_i, in the working set is scored at every position p_i, using the profile generated from $V-\upsilon_i$, i.e.:

$$score\,(\upsilon_i\,(p_i)) = \sum_{j=1}^{W} \log\left[\frac{M(s_i(p_i+j-1),\,j)}{B(s_i(p_i+j-1))}\right]. \tag{7}$$

With some probability, υ_i is swapped out and replaced by one of the sequences outside the working set. The probability of being swapped out varies inversely with the maximum score seen for the sequence at any position, the score at the position that corresponds most strongly to the profile matrix:

$$\log[\Pr(swap)] \propto -\max_{pi}[score(\upsilon_i(p_i))]. \tag{8}$$

The proportionality factor depends on the length of the motifs being searched for.

The intuition behind the Gibbs^2 algorithm is that if a sequence scores high for a motif, it matches quite well with other sequences used to derive the profile and is retained with a higher probability. Conversely, if a sequence does not score highly, it matches poorly with the remaining sequences in the working set used to derive the profile.

Ideally, the sequence swapped out should be replaced by one that scores highest on the profile matrix being used. This approach is computationally intensive since all outstanding sequences need to be scored before the optimal one can be chosen. To avoid this, once a sequence is swapped out, it is replaced by any of the sequences outside the working set at random. This avoids the need to score all previously excluded sequences to find the one with the best match. Furthermore, after each swap, further swapping is temporarily disabled to allow the new sequence to be absorbed and contribute to the profile matrix.

The Gibbs^2 algorithm can be written as follows (with C denoting the size of the working set and K representing the number of iterations, swapping is disabled after a sequence is replaced from one outside the working set):

Gibbs^2 Algorithm

GIBBS^2($\{s_1, \ldots, s_N\}$, W, C, K):

1. Choose C sequences at random from $\{s_1, \ldots, s_N\}$
2. Set random initial values for p
3. Do
   - i. Select sequence υ_i at random
   - ii. Estimate M from set $V - \upsilon_i$
   - iii. Use M to score all starts in υ_i
   - iv. Swap out υ_i with Pr($swap$) and replace it with a random sequence outside the working set

Gibbs^2 Algorithm

v If swap occurs,

a Disable swapping for K iterations

vi Pick start p_i with probability proportional to its score

Until the change in the score of M is less than some threshold ε.

The Gibbs^2 approach can be used to iteratively partition the data into a set containing a strongly conserved motif and an outstanding set that can be broken into further subsets sharing a common pattern. This allows for the discovery of multiple predictive motifs occurring in subsets of the population.

We propose choosing the working set size by studying how the log-odds likelihood of motifs changes for different selections of C. The average contribution to the log-odds likelihood by each sequence in the working set can be measured using (1) as:

$$\frac{1}{C}\sum_{i=1}^{C}\sum_{j=1}^{W}\log\left[\frac{M(s_i(p_i+j-1),j)}{B(s_i(p_i+j-1))}\right]. \quad (9)$$

As sequences are added to the working set, the average contribution measured in (9) decreases significantly if the addition of a further sequence prevents the working set from sharing a common motif—if the additional sequence does not allow a strong motif to be identified. The size of the working set for the Gibbs^2 algorithm can therefore be determined by searching for a knee in the curve relating the average contribution to the log-odds likelihood by each sequence with C. This process may be approximated by a binary search to reduce computation.

The use of Gibbs^2 also allows for the performance of the Consensus algorithm from Section 4 to be improved. Specifically, Consensus can be seeded using a strongly conserved pattern obtained by Gibbs^2. This reduces the likelihood that Consensus will be affected by a poor choice of the initial two strings.

6. EVALUATION

6.1 Testing Methodology

We applied our techniques to the Physionet Sudden Cardiac Death Holter Database (SDDB) [Goldberger et al. 2000]. This database contains several hours of ECG data recorded using Holter monitors from 23 patients who experienced sudden cardiac death. The recordings were obtained in the 1980s in Boston area hospitals and were compiled as part of a later study of ventricular arrhythmias. Owing to the retrospective nature of this collection, there are important limitations. Patient information including drug regimens and dosages is limited, and sometimes completely unavailable. Furthermore, sudden cardiac death may result from a variety of underlying causes and it is likely that among the 23 patients there are multiple groups sharing different regulatory factors. Despite these shortcomings, the SDDB ECG signals represent an interesting dataset since they represent a population sharing a common acute event. In addition, the recordings are sufficiently long (up to 24 hours prior to death in some cases) that it is likely the predictive factors occurred during the recording period. Finally, the signals in SDDB are generally well-annotated, with cardiologists

providing labels at the level of each beat, and this yields a source of clinically relevant symbols that can be used to search for motifs.

For the 23 SDDB patients TCM, Gibbs sampling, Gibbs[2], and Consensus were used to discover potentially predictive motifs of lengths 4, 10, and 16. Since TCM, Consensus and the variants of the Gibbs sampling algorithms, are stochastic in nature, a hundred runs were executed with the strongest motifs being automatically returned as the solution. The scoring function used was the log-likelihood score described in Section 3.

In each case, the endpoint used to signify the acute event associated with death was the occurrence of ventricular fibrillation (VF). This was annotated for all patients and only regions preceding VF were searched for conserved motifs.

For visualization purposes, we used WebLogo [Crooks et al. 2004] to display the motifs returned by our algorithms. This uses the profile matrix to represent motifs as sequence logos, which are graphical representations consisting of stacks of symbols. For each position in the motif, the overall height of the stack indicates how strongly the motif is conserved at that position, while the height of symbols within the stack indicates the relative frequency of each symbol at that position. For example, for the length 10 motif in Figure 4, the sequence logo shows that the motif is strongly conserved at positions 8 and 10, where the predictive sequence was found to contain normal beats across patients. The motif is also conserved at positions 1, 3, and 5, where ventricular activity was seen for most patients, with some occurrences of normal beats (position 1) and supraventricular beats (positions 3 and 5) as well.

For position j in the motif, the height of symbol i at that location is given by:

$$M(i, j)[2 - H(j)], \tag{10}$$

where:

$$H(j) = -\sum_{k} M(k, j)\log_2(M(k, j)). \tag{11}$$

For Consensus, where a profile matrix is not explicitly constructed, the best-matching subsequences were used to derive a profile that could be represented using WebLogo. This allowed for results to be consistently visualized, irrespective of the algorithm used to discover motifs.

More information on WebLogo can be found at their Web site.[1]

6.2 Data Reduction

The transformations discussed in Section 5 can be evaluated in terms of the data compression realized using these approaches. This allows for an appreciation of the extent to which the original data contains long runs of normal activity that can be compacted. The original sequences across the 23 patients contained 1,216,435 symbols in total, each corresponding to a single beat annotated by a skilled cardiologist. Using the notion of active regions and stripping away uninteresting normal motifs reduced the size of the data to 257,479 characters—a reduction of 78.83%.

[1] http://weblogo.berkely.edu

6.3 TCM, Gibbs Sampling, and Consensus

Figures 4–6 present the results returned by TCM, Gibbs sampling, and Consensus, as sequence logos. Commonly occurring labels are N = normal, V = premature ventricular contraction, and S = supraventricular premature or ectopic beats.

The motifs discovered by all three algorithms were similar and comprised runs of premature ventricular contractions. For each choice of motif length, TCM returned more strongly conserved motifs than both Gibbs sampling and Consensus. This can be explained by the fact that TCM scores all starting positions in every sequence during each iteration, and is stochastic only in the choice of an initial profile matrix. It employs significantly more computation than either Gibbs sampling or Consensus and is able to find more strongly conserved patterns as a result. On the other hand, the Gibbs sampling algorithm depends on both a random set of initial starting positions and probabilistic choices during each iteration to select a string, s_i, and a new starting position within that string. Consensus is similar to TCM in that it is stochastic only in its initial choice of sequences to use as seed, but unlike TCM, where a poor initial choice can be corrected during subsequent iterations, in the case of Consensus, the effects of a poor initial choice propagate all the way through.

Although TCM produced the best results in this case, the process of scoring every starting position in each sequence was considerably more time consuming and took an order of magnitude more time than either Gibbs sampling or Consensus.

6.4 Gibbs^2 and Seeded Consensus

Figure 7 shows the motifs discovered by the Gibbs^2 algorithm with an initial working set of size 12, containing sequences chosen at random. The size of the initial working set was determined from the average contribution of each sequence to the log-odds likelihood of the best scoring motif, as described in Section 5.2. Figure 8 illustrates how the average contribution of the log-odds likelihood changed with increased values of C.

In this case, the predictive motif once again found comprised runs of premature ventricular contractions, but was more strongly conserved than the best results produced earlier by TCM, Gibbs sampling, and Consensus. Specifically, comparing Figures 4–7, the stack of symbols in Figure 7 shows the premature ventricular activity figuring more prominently at positions within the motifs.

This effect may be attributed to the ability of Gibbs^2 to select a group of patients who had matching motifs comprising premature ventricular activity, unlike TCM, Gibbs sampling, and Consensus, which were constrained to find a less conserved intermediate that was a best fit for data from all the different patients in the population. For this reason, Gibbs^2 provided an improvement not only over the original Gibbs sampling algorithm but also the more computationally intensive TCM. The Gibbs^2 algorithm has the same basic structure as the original Gibbs sampling technique, but is able to outperform TCM by addressing the issue of subsets of the population exhibiting different regulatory activity. Section 6.5 explores this aspect of subset-based algorithms in more detail, and motivates the idea of searching for different motives in subpopulations iteratively.

Figure 9 presents the result of using Seeded Consensus to detect motifs of length 4 relative to the original Consensus algorithm. In this case, the Gibbs^2 algorithm with a working set of size 5 was used to first find an initial seed for the Consensus algorithm. As the data shows, Seeded Consensus produced more strongly conserved results than the original Consensus algorithm. This effect followed from reducing the chance that a poor initial choice of sequences would propagate and adversely affect the search for motifs.

The motif found using Seeded Consensus in Figure 9 is not as strongly conserved as the one discovered by Gibbs^2 in Figure 7. This can be explained by the fact that Seeded Consensus uses Gibbs^2 to discover an initial seed but otherwise still operates on all the sequences in the data. The issue of motifs occurring only in a subset of patients does not therefore get addressed, although Seeded Consensus is still able to produce results that are comparable with TCM without the need for intensive computation.

The results of these experiments suggest that subset based techniques using Gibbs^2 either to search for motifs directly, or for the purpose of providing seeds that can be fed into the Consensus algorithm, may allow for more strongly conserved motifs to be discovered than through use of TCM, Gibbs sampling, and the original Consensus algorithm. Moreover, the improvement provided by the Gibbs^2 algorithm proposed in our work is not associated with a significant computational overhead. In addition, the ability to partition the data into groups with homogenous motifs allows for the discovery of more than one predictive pattern, each of which may be associated with the outcome in a different group of patients. We explore this idea in more detail in the next section.

6.5 Two-Stage Gibbs^2

For the motif of length 4, the sequences remaining outside the working set at the termination the of Gibbs^2 algorithm were searched for a second motif common to this group. Figure 10 shows the results of this approach.

In this case, a second motif comprising runs of supraventricular premature or ectopic beats was found among this subgroup of the population. Notably, these patients did not show a motif similar to the ones found earlier, comprising premature ventricular beats, during any of the multiple executions of the motif discovery algorithm. This suggests that the subset of patients left outside the working set by Gibbs^2 did not exhibit regulatory activity similar to the ones for whom a premature ventricular motif was discovered. Including these patients in the search for a predictive motif, as would be the case for non-subset-based techniques, would therefore lead to a less informative motif and would obscure the fact that different groups of patients show varied predictive patterns associated with an endpoint.

6.6 Motif-Event Delay

Using the motif of length 10 shown in Figure 7, for each sequence, the time delay between the starting location of the motif: p_i, and the clinical endpoint (the occurrence of VF in the patients) was calculated for the Gibbs^2 algorithm. For one of the 23 patients in the dataset, the motif occurred less than a minute prior to the event itself. In all other cases, the motif discovered preceded the actual event by at least 20 minutes or more. The median motif-event delay was 60 minutes, while the 25% and 75% quartile times were 42 and 179 minutes respectively. The maximum time separation of the motif and the event was 604 minutes.

These results suggest that the motif occurred sufficiently in advance of the endpoint not to be considered merely an extension of the final event itself. Furthermore, the fact that the motif may occur at a wide range of times prior to the endpoint reinforces the need to carry out the search for predictive patterns in an automated manner, which is able to relate information across a range of positions within each sequence.

6.7 Comparison with Controls

For each patient in the SDDB population, the log-likelihood score was calculated for each starting position in the ECG label sequence. The overall score for the patient was the maximum log-likehood score found. Intuitively, this strategy assigns each patient the risk score associated with the occurrence of the discovered motif of length 10 shown in Figure 7

at any point during the recording: if activity similar to the motif associated with sudden death is seen anywhere, the patient is assumed to be at higher risk for the event.

Figure 11 shows the probability density function that can be estimated from the scores for the SDDB population. A similar strategy was adopted to score patients in two control datasets: the Physionet Normal Sinus Rhythm Database (NSRDB) and the Physionet Supraventricular Arrhythmia Database (SVDB). The decision to use SVDB data in addition to normal individuals was owing to the fact that the SVDB signals contained the same labels as the SDDB data with a higher background frequency of abnormal symbols. This ensured that a difference in scores across populations did not result from an absence of labels, but more so because activity was organized in different forms. Specifically, 1.45% of the beats in the SDDB data were premature ventricular contractions. By comparison, 5.39% of the beats in the SVDB signals and 0.002% of the NSRDB beats fell into the same category. This suggests that if the motifs seen in the SDDB population were random occurrences, then they would be expected to be seen more frequently in the SVDB dataset. With this in mind, the fact that SVDB patients had a higher percentage of premature ventricular activity but still scored lower on the discovered motifs provides further indication that the motif corresponded to activity that was not purely a random occurrence in the sudden death population.

Using a maximum likelihood separator, we were able to use our motif to correctly identify 70% of the patients who suffered sudden cardiac death during 24 hours of recording while classifying none of the normal individuals, and only 8% of the patients from the supraventricular dataset as being at risk. The small number of patients in the dataset, however, does not allow for us to make statistically significant clinical statements about these findings.

7. RELATED WORK

In this section, we review existing knowledge-discovery work to detect potentially predictive activity. A discussion of aspects of our work extending computational biology techniques appears earlier in Sections 3 and 4.

An extensive literature exists in the areas of data mining and machine learning on the subject of prediction. A common approach is to infer prediction rules from data of the form:

$$\text{IF } cond_1 \text{ AND } \ldots\ cond_i\ \ldots \text{ AND } cond_m \text{ THEN } pred.$$

These rules correspond to a set of conditions associated with a specific outcome. The challenge in this case is to select conditions that are able to distinguish between whether an event occurs or not, but do not overfit available training data. A number of different techniques exist for this purpose, ranging from decision trees [Helmbold and Schapire 1997] to more recent work using evolutionary algorithms [Freitas 2001].

We supplement this work by finding precursors that exist at a lower level of the data. As an alternative to rules based on the outcomes of a series of diagnostic tests or a sophisticated feature set, we attempt to find interesting patterns by analyzing the specific sequences a system moves through. The motivation for such an approach is provided by the ever-increasing amounts of data collected in various fields, for example, medicine, geodesic studies, space, and planetary sciences. In many of these cases, well-formulated predictive attributes do not exist. Unsupervised techniques can, however, be used to decompose signals into stationary or periodic tokens. These can then be assigned labels to reexpress the original data as a sequence of symbols. Our work allows for the discovery of a specific class of

regulatory activity (occurring as subsequences) in this representation without assuming higher-level features for classification.

The idea of transforming time-series data into symbols has been proposed in different forms earlier (an excellent review on this subject can be found in Daw et al. [2003] with some additional techniques for symbolization described in Giles and Lawrence [2001]; Lin et al. [2003]; and Syed et al. [2007]). We build upon this work and use symbolization in the context of large physiological datasets to make the search for predictive patterns more efficient.

While symbolization is an integral component of our approach, we note that there has also been much promising work in recent years on the discovery of motifs directly in time-series signals. These approaches search for patterns either in the raw signal without any transformation [Chiu et al. 2003], or by using piecewise aggregate approximation (PAA) [Patel et al. 2002; Lin et al. 2002] to achieve a simple yet efficient symbolization. The focus in both these cases is on the discovery of patterns that occur frequently in the data. These methods do not explicitly address the goal of prediction and are not designed to find patterns that may occur infrequently but have a consistent regulatory effect. The discovery of such patterns, e.g., those associated with an acute endpoint, is a potentially more computationally intensive problem, whereas symbolization plays an important role in making the search process more scalable.

Our pattern discovery algorithms add to an extensive body of existing work to analyze symbolic sequences. A fairly rich literature can be found on techniques to discover local patterns [Mannila et al. 1999; Jin et al. 2002], frequent sequential associations [Harms et al. 2002], and generalized episodes [Mannila et al. 1997]. These methods search for sequences of discrete events or symbols that occur frequently in the data with a partially specified order. This information can help make future predictions about the behavior of symbolic sequences. For example, the observation that symbols A and B frequently occur in the data and are always followed by symbol C allows for the symbol C to be predicted whenever A and B are next encountered. Such techniques address a similar goal to our work and offer the advantage of finding more general predictive patterns than our methods. However, these methods are fundamentally focused on the discovery of frequent patterns in single symbolic sequences and on learning structure between symbols that occur close to each other. This is in contrast to the techniques proposed in our work, which searches for patterns that may not be frequent but are associated with an endpoint in multiple symbolic sequences, and where the symbols constituting the predictive pattern and the acute event can be far apart. In this way, the motifs discovered by our algorithms can be viewed as global patterns, spanning both a subsequence of symbols and an annotated event that may be variably delayed.

Our analysis of sequential signals is also similar to the use of Markov models to study systems [Durbin et al. 1998]. Our work differs from a purely Markovian approach in that we do not attempt to develop a model explaining the data and focus instead on explicitly identifying predictive elements. Furthermore, in many cases, including the sudden death study conducted as part of this project, the regulatory activity may occur well in advance of the event. Developing a Markov model containing sufficient memory and complexity to handle these cases would prove to be challenging in such situations.

A different form of prediction in learning theory is to approach the task in an online manner and consistently refine a hypothesis based on errors and temporal differences [Cesa-Bianchi and Lugosi 2006]. This approach is similar to the inference of prediction rules in that decisions are made on attributes or features and not individual sequences. Our techniques

further differ in that they attempt to exploit the availability of batch data and do not address the issue of online learning.

In addition to suggesting methods to discover motifs, we propose subset-based techniques that can isolate subsets of the data that share common predictive motifs. For example, in Section 6.5, we discuss how two-stage Gibbs^2 can find subpopulations sharing different predictive sequences. This is important since the same event may be associated with different causes. We consider the selection of sequences that share regulatory activity as being internal to the problem of motif discovery. Specifically, partitioning the data in advance without information on the specific predictive pattern is difficult—subsets of the sequences sharing a motif can only be isolated once the motif is known. For this reason, we address the issue of degeneracy and heterogeneous predictive patterns as part of motif discovery and tailor our algorithms to automatically recognize and handle these cases.

A key component of our approach to address motif degeneracy is to randomly swap symbolic sequences into and out of a working subset. The notion of randomized swaps has also been proposed earlier, in the context of data mining to assess the significance of mining results on high-dimensional data [Gionis et al. 2006]. The focus in that work is on randomly swapping 0–1 features within a matrix to create a new matrix with preserved row and column margins that can be used for testing. The process of random swapping does not attempt to isolate a subset of features with desired properties but is used to randomly perturb the matrix instead. Finally, our work is similar to unary classification techniques [Scholkopf et al. 2001] in that the algorithms proposed do not require the presence of both positive and negative examples. Instead, they are geared towards selecting subsequences of labels that can be found across a population in a form unlikely to occur purely by chance. The goal is to better understand similarities that can be analyzed for a predictive relationship with the acute event being considered.

8. SUMMARY AND CONCLUSIONS

In this article, we propose a framework for discovering potentially predictive activity preceding acute events. We generalize the notion of regulatory motifs from computational biology and adapt existing algorithms to operate robustly and efficiently on a broad set of data. We develop and evaluate this work in the context of physiological signals, detailing the challenges associated with fitting a motif-detection model to signals other than DNA. We also describe the performance of subset-based techniques to discover activity associated with sudden cardiac death, comparing discovered patterns against control populations comprising normal individuals and those with supraventricular arrhythmias.

Our work represents a fully-automated approach for discovering a specific class of possible precursors: patterns that are sequential in that a given ordering of different classes is associated with an end result. We impose no restrictions on the patterns to be discovered, and our tools are able to identify sequences of arbitrary complexity that occur in a possibly degenerate form across a population sharing an event.

A central requirement for the techniques described in this article is that the data being mined is symbolic. In the context of physiological signals, this requires transforming continuous waveforms into alphabetical sequences. Creating a set of labels that can be applied to the data can be achieved in a number of different ways. In the work described here, we use clinical labels that have a fixed meaning and can be applied across patients. It is possible that potentially predictive activity may occur at a more subtle level, where differences within clinical classes are important. For this reason, an important future direction of this

research is to extend approaches to annotate signals in a patient-specific, data-derived manner to achieve symbolization.

Finally, it is important to stress that although our initial results on detecting a predictive pattern associated with sudden cardiac death appear promising, the small number of patients in the dataset and limited patient histories means that further investigation on a larger set of ECG signals is necessary.

Acknowledgments

We would like to thank Ebad Ahmed for helping implement the software routines used in this work. We also thank the editors and the anonymous reviewers for their feedback in helping improve the presentation of this work.

This work was supported in part by the Center for the Integration of Medicine and Innovative Technology (CIMIT) and the Harvard-MIT Division of Health Sciences and Technology (HST).

REFERENCES

Bailey, T.; Eklan, C. The value of prior knowledge in discovery motifs with MEME. Proceedings of the International Conference on Intelligence Systems in Molecular Biology; 1995. p. 21-29.

Chiu, B.; Keogh, E.; Lonardi, S. Probabilistic discovery of time series motifs. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2003.

Cesa-Bianchi, N.; Lugosi, G. Prediction, Learning and Games. Cambridge, UK: Cambridge University Press; 2006.

Crooks G, Hon G, Chandonia J, Brenner S. WebLogo: a sequence long generator. Genome Res 2004;14:1188–1190. [PubMed: 15173120]

Daw C, Finney C, Tracy E. A review of symbolic analysis of experimental data. Rev. Sci. Instr 2003;74:915–930.

Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. Biological Sequence Analysis. Cambridge, UK: Cambridge University Press; 1998.

Freitas, A. Data Mining and Knowledge Discovery with Evolutionary Algorithms. Berlin, Germany: Springer-Verlag; 2001.

Gert T, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J. Comput. Biol 2002;9:447–464. [PubMed: 12015892]

Giles C, Lawrence S. Noisy time series prediction using a recurrent neural network and grammatical inference. Mach. Learn 2001;44:161–183.

Gionis, A.; Mannila, H.; Mielikainen, T.; Tsaparas, P. Assessing data mining results via swap randomization. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2006.

Goldberger A, Amaral L, Glass L, Hausdorff J, Invanov P, Mark R, Mietus J, Moody G, Peng C, Stanley H. Components of a new research resource for complex physiologic signals. Circulation 2000;101:215–220.

Harms, S.; Deogun, J.; Tadesse, T. Discovering sequential association rules with constraints and time lags in multiple sequences. Proceedings of the 13th International Symposium on Foundation of Intelligent Systems; 2002.

Helmbold D, Schapire R. Predicting nearly as well as the best pruning of a decision tree. Mach. Learn 1997;27:51–68.

Jin, X.; Wang, L.; Lu, Y.; Shi, C. Indexing and mining of the local patterns in sequence database. Proceedings of the 3rd International Conference on Intelligent Data Engineering and Automated Learning; 2002.

Jones, N.; Pevzner, P. An Introduction to Bioinformatics Algorithms. Cambridge, MA: The MIT Press; 2004.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander E. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 2003;423:241–254. [PubMed: 12748633]

Lin, J.; Keogh, E.; Lonardi, S.; Patel, P. Finding motifs in time series. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2002.

Lin, J.; Keogh, E.; Lonardi, S.; Chiu, B. A symbolic representation of time series with implications for streaming algorithms. Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery; 2003.

Mannila H, Toivonen H, Verkamo A. Discovery of frequent episodes in event sequences. Data Mining and Knowl. Discov 1997;1:259–289.

Mannila, H.; Pavlov, D.; Smyth, P. Prediction with local patterns using cross-entropy. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 1999.

Patel, P.; Keogh, E.; Lin, J.; Lonardi, S. Mining motifs in massive time series databases. Proceedings of the International Conference on Data Mining; 2002.

Scholkopf B, Platt J, Shawe-Taylor J, Smola A, Williamson R. Estimating the support of a high-dimensional distribution. Neural Comput 2001;13:1443–1471. [PubMed: 11440593]

Stormo G, Hartzell G. Identifying protein-binding sites from unaligned DNA fragments. Proc. Nat. Acad. Sciences 1989;86:1183–1187.

Syed Z, Stultz C, Guttag J. Clustering and symbolic analysis of cardiovascular signals: Discovery and visualization of medically relevant patterns in long-term data using limited prior knowledge. EURASIP J. Advances Sig. Proc. 2007. 2007 Article ID 67938.

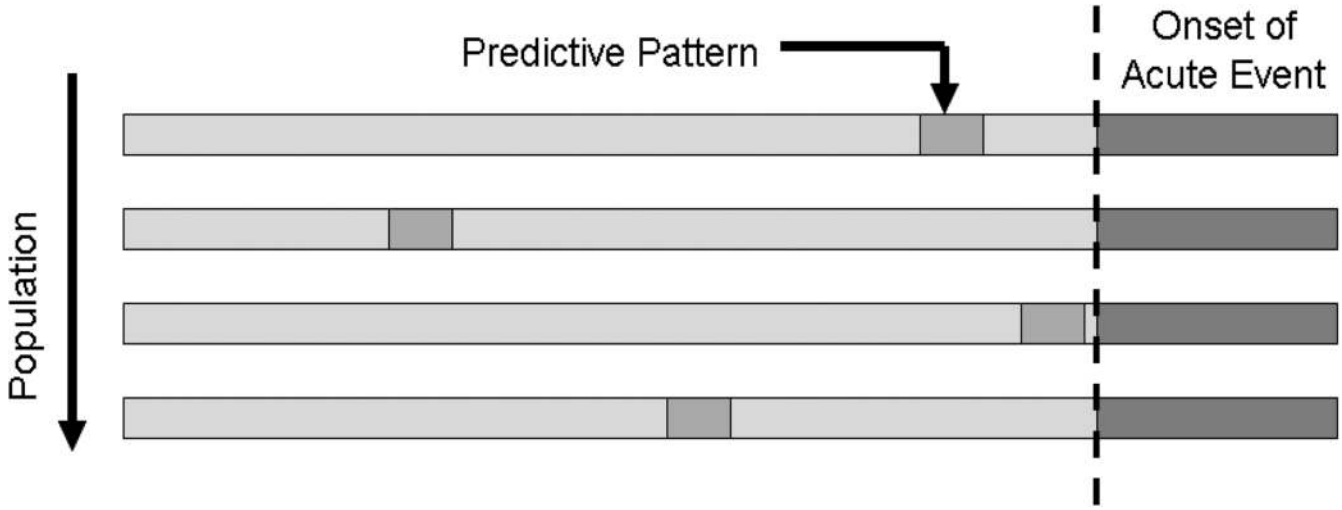

Fig. 1.
Prediction through conservation in the context of a population of patients affected by a common acute clinical event.

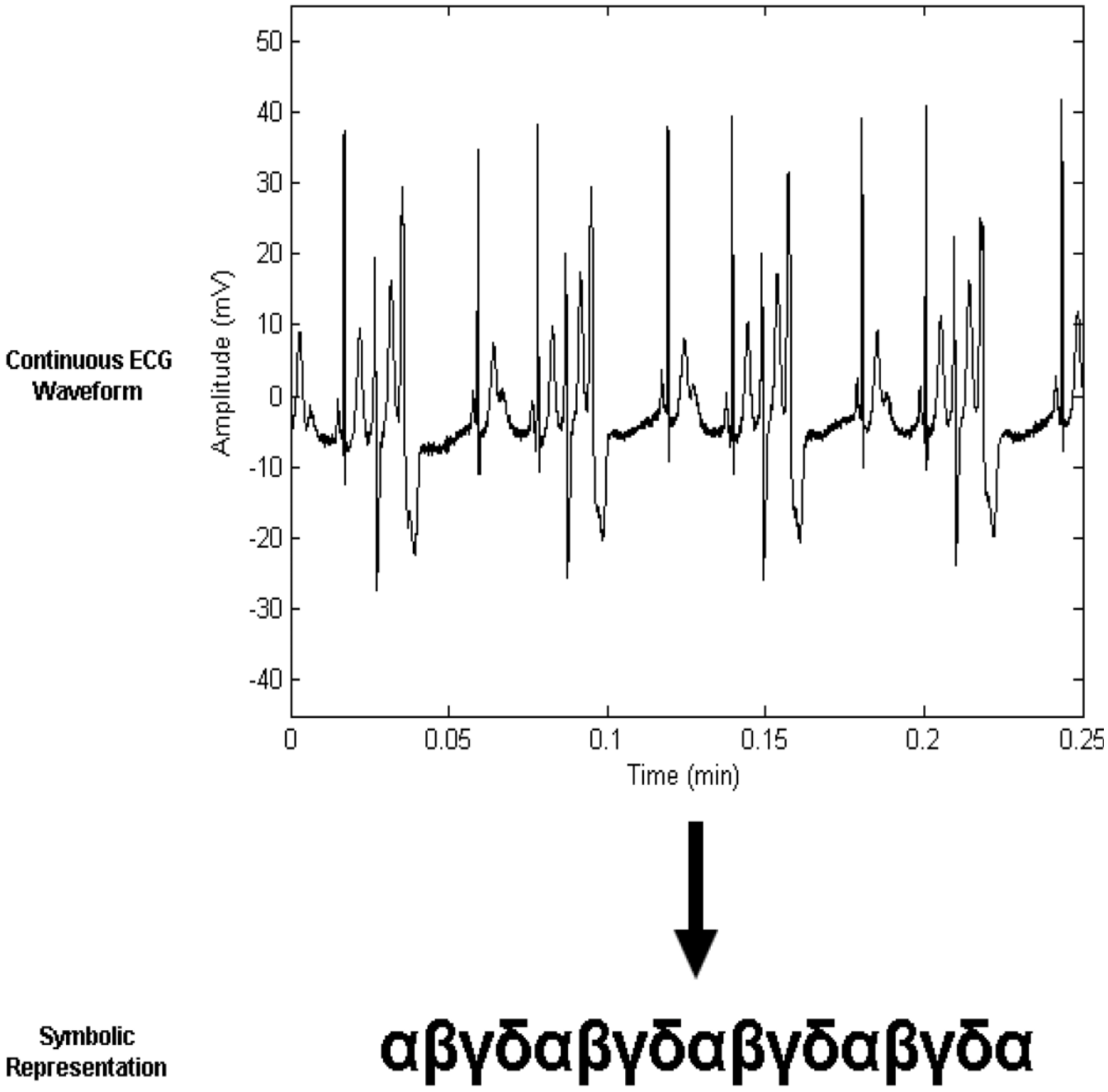

Fig. 2.
Example transformation of continuous ECG waveform to a string of symbols. Each of the symbols shown corresponds to a different class of electrophysiological activity.

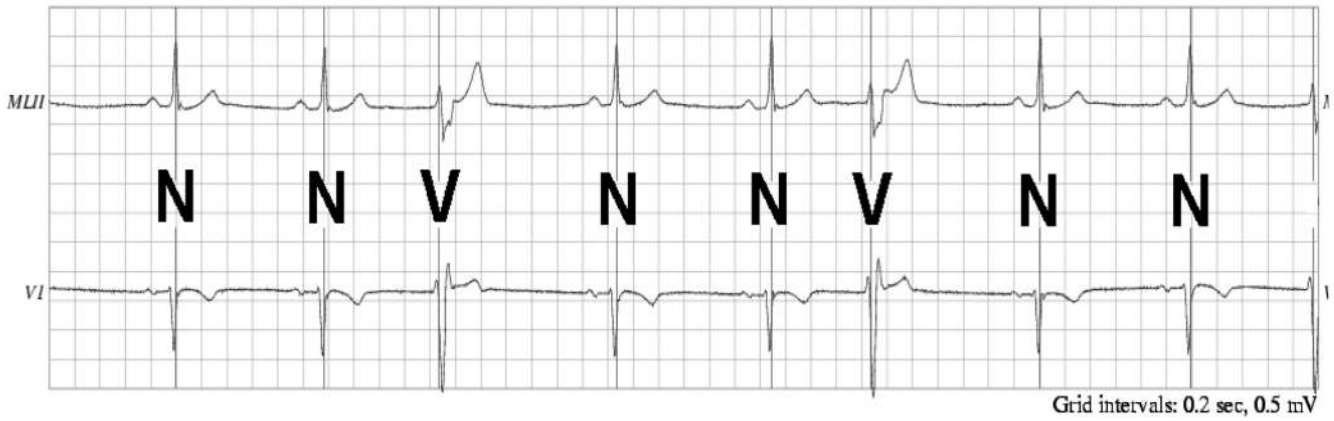

Fig. 3.
Example symbolization of continuous ECG waveforms using clinical annotations (N = normal, V = premature ventricular contraction).

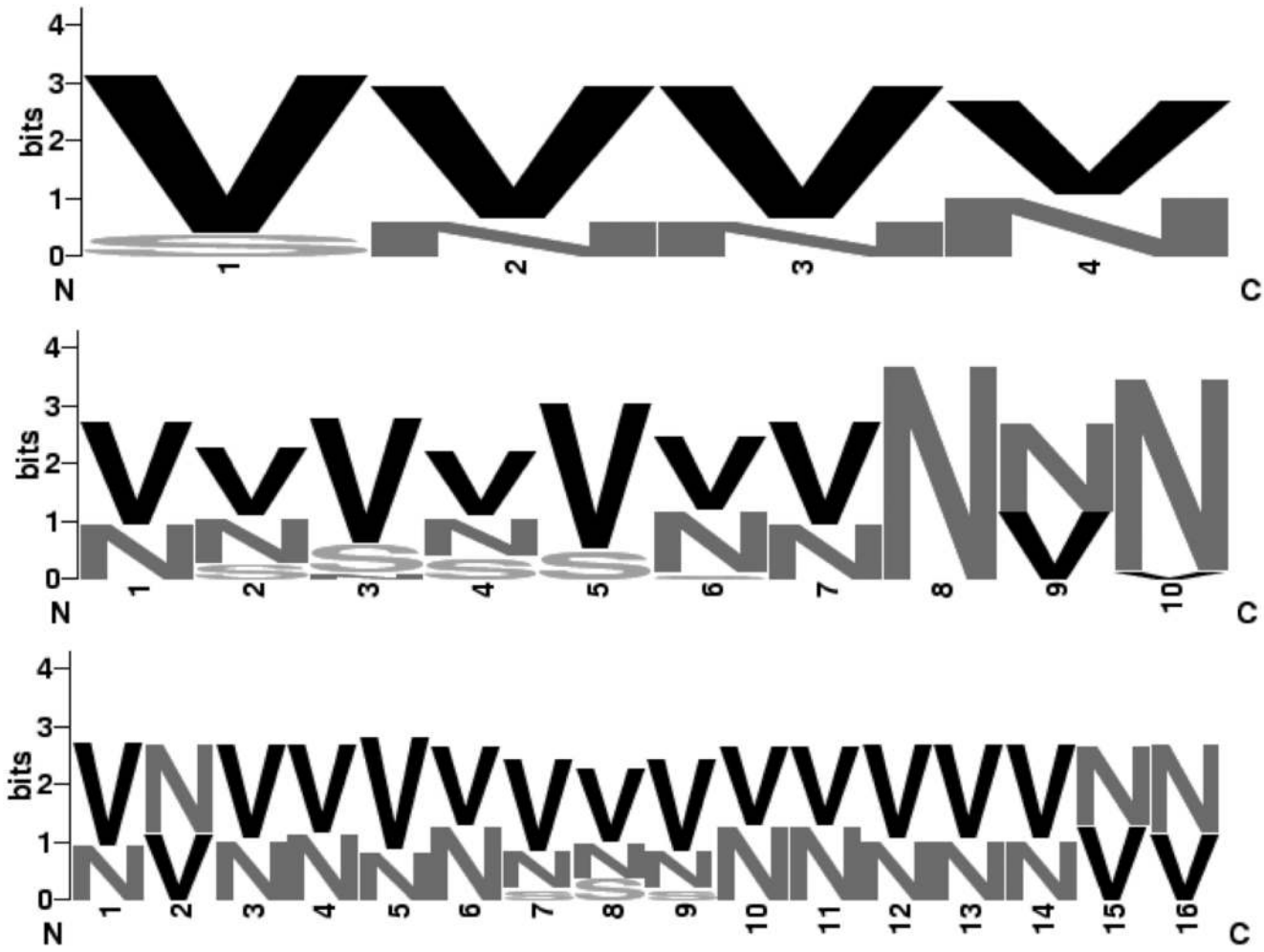

Fig. 4.
Motifs of length 4, 10, and 16 found using TCM.

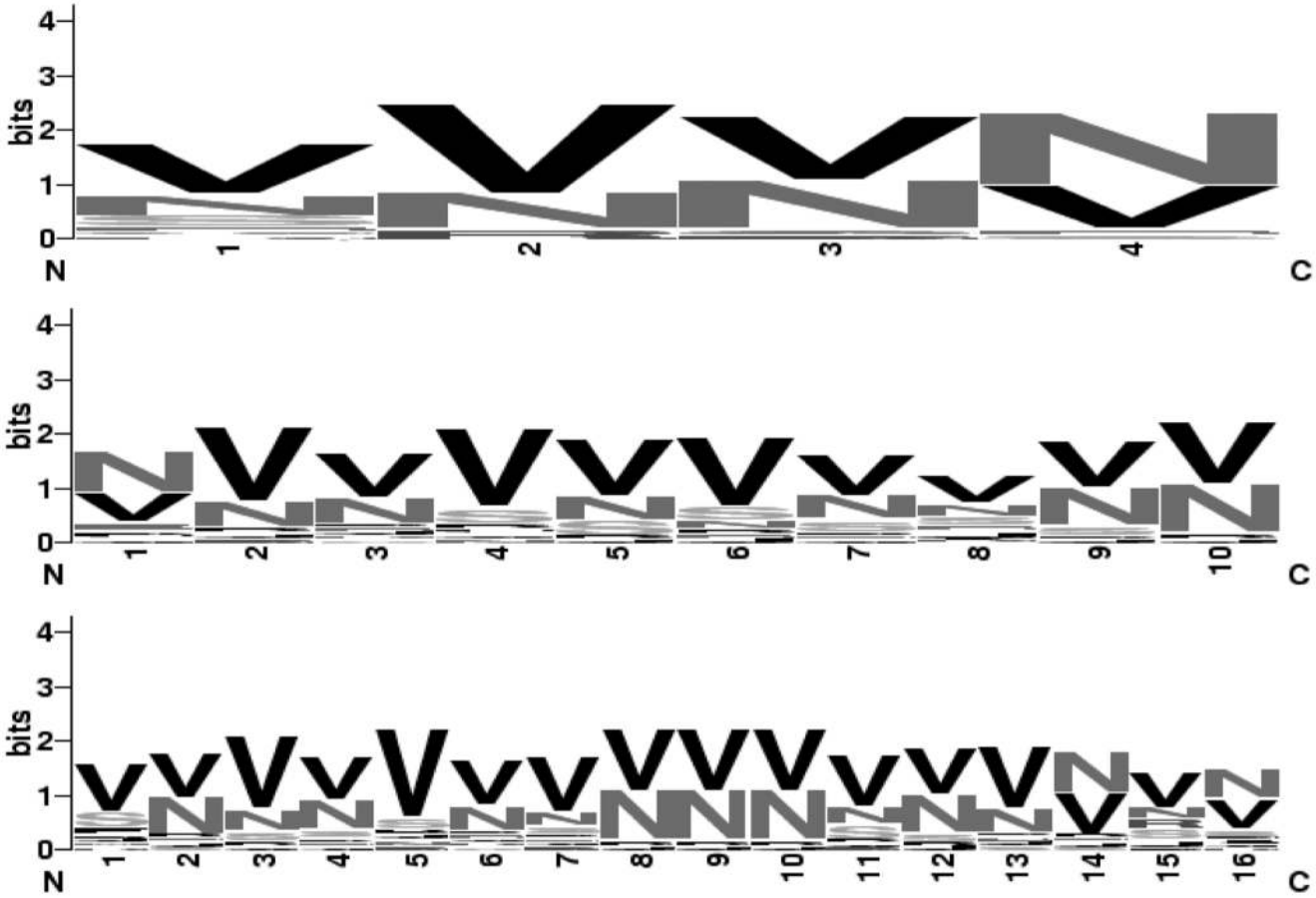

Fig. 5.
Motifs of length 4, 10, and 16 found using Gibbs sampling.

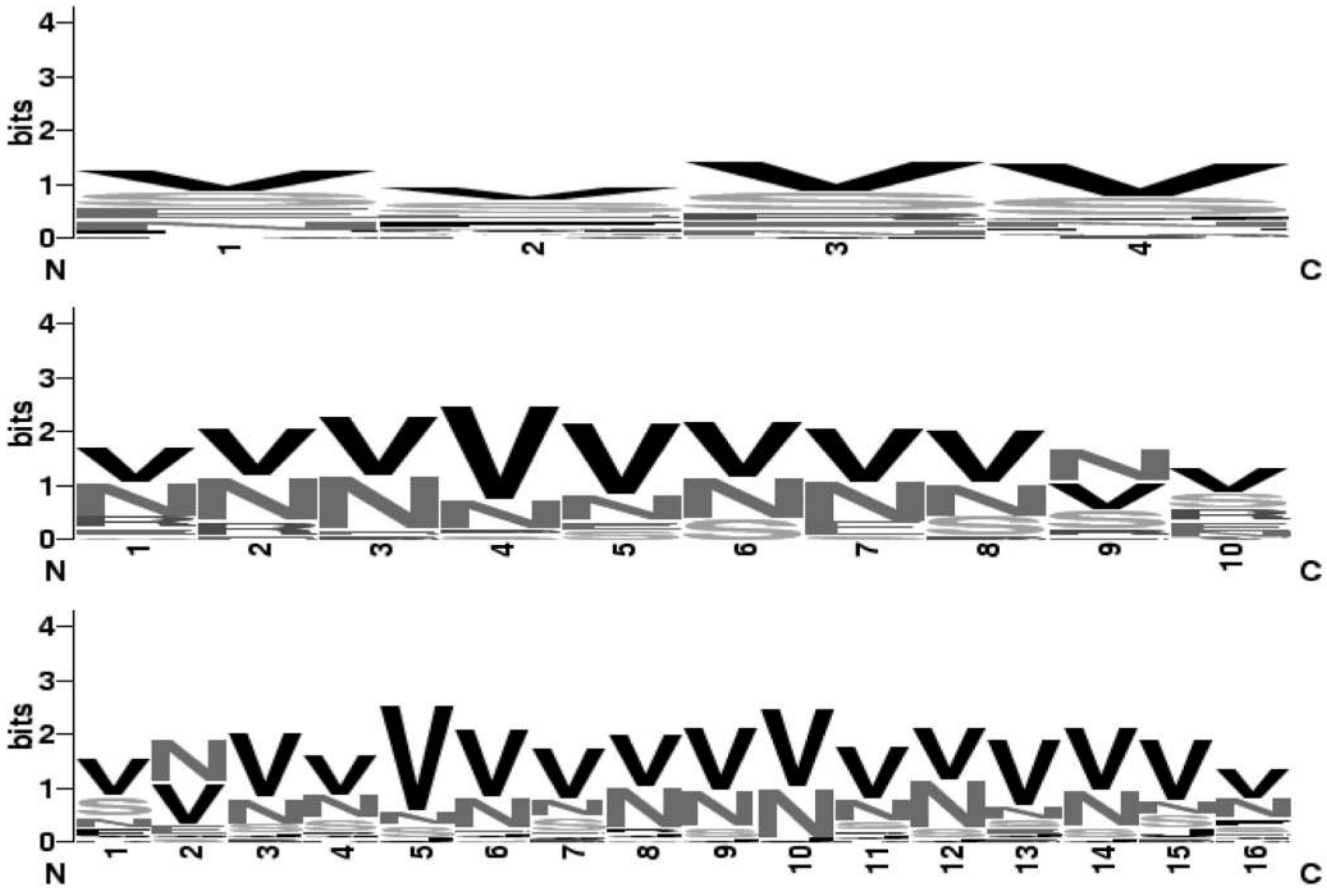

Fig. 6.
Motifs of length 4, 10, and 16 found using Consensus.

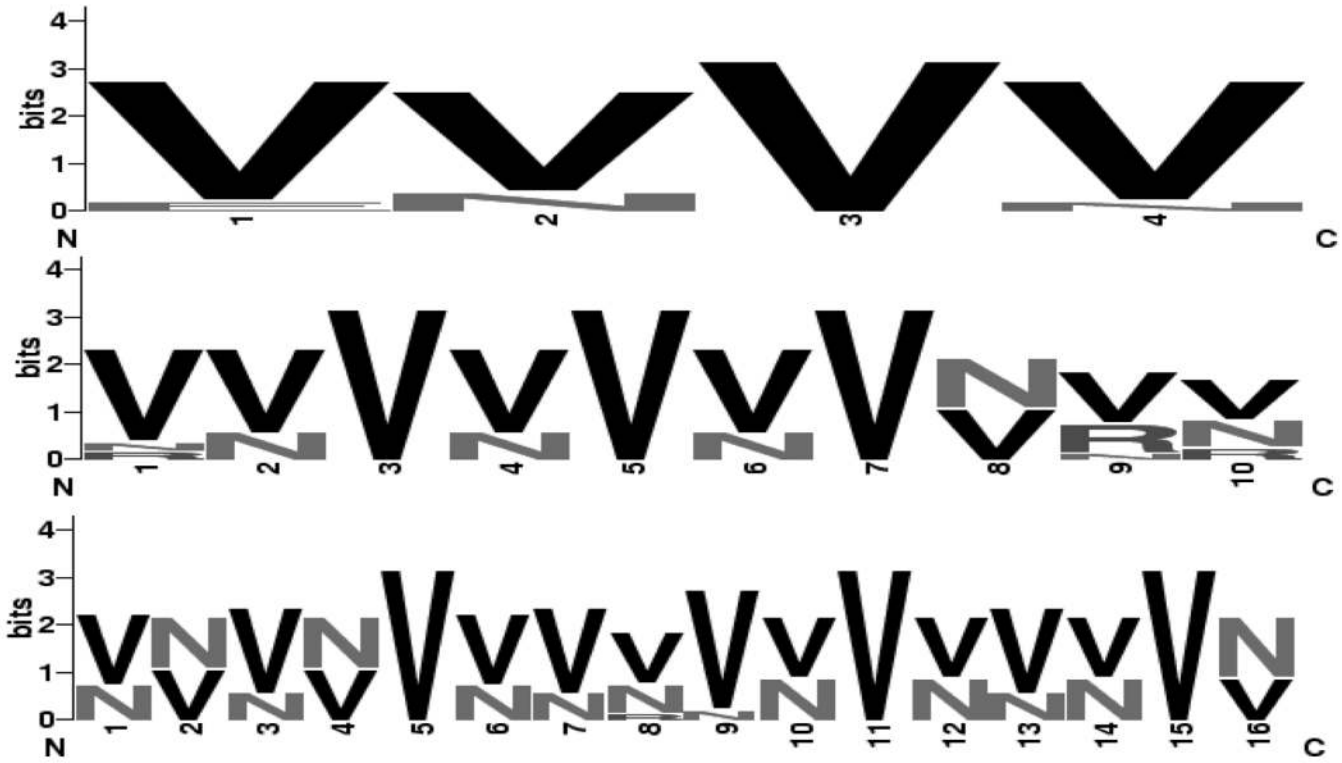

Fig. 7.
Motifs of length 4, 10, and 16 found using Gibbs[2].

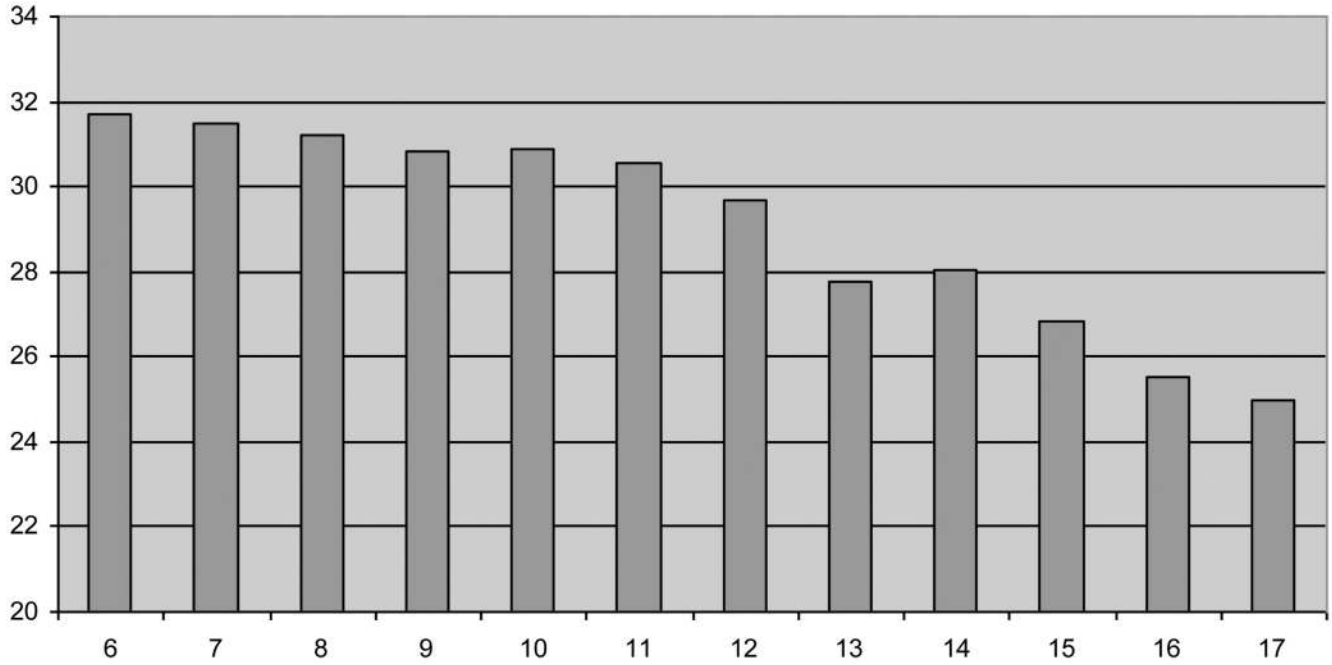

Fig. 8.
Relation of the average contribution of each sequence to the log-odds likelihood for the best scoring motif with increasing values of *C*.

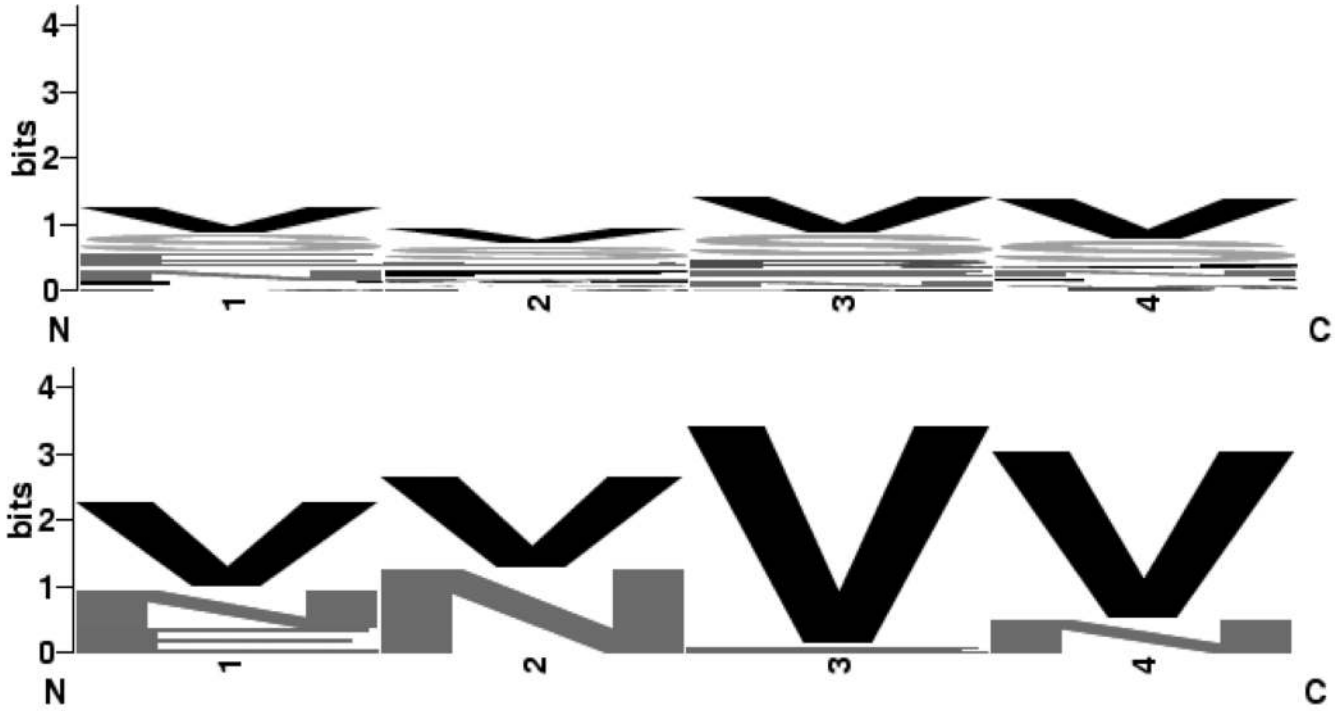

Fig. 9.
Motifs of length 4 found using Consensus (top) and Seeded Consensus (bottom).

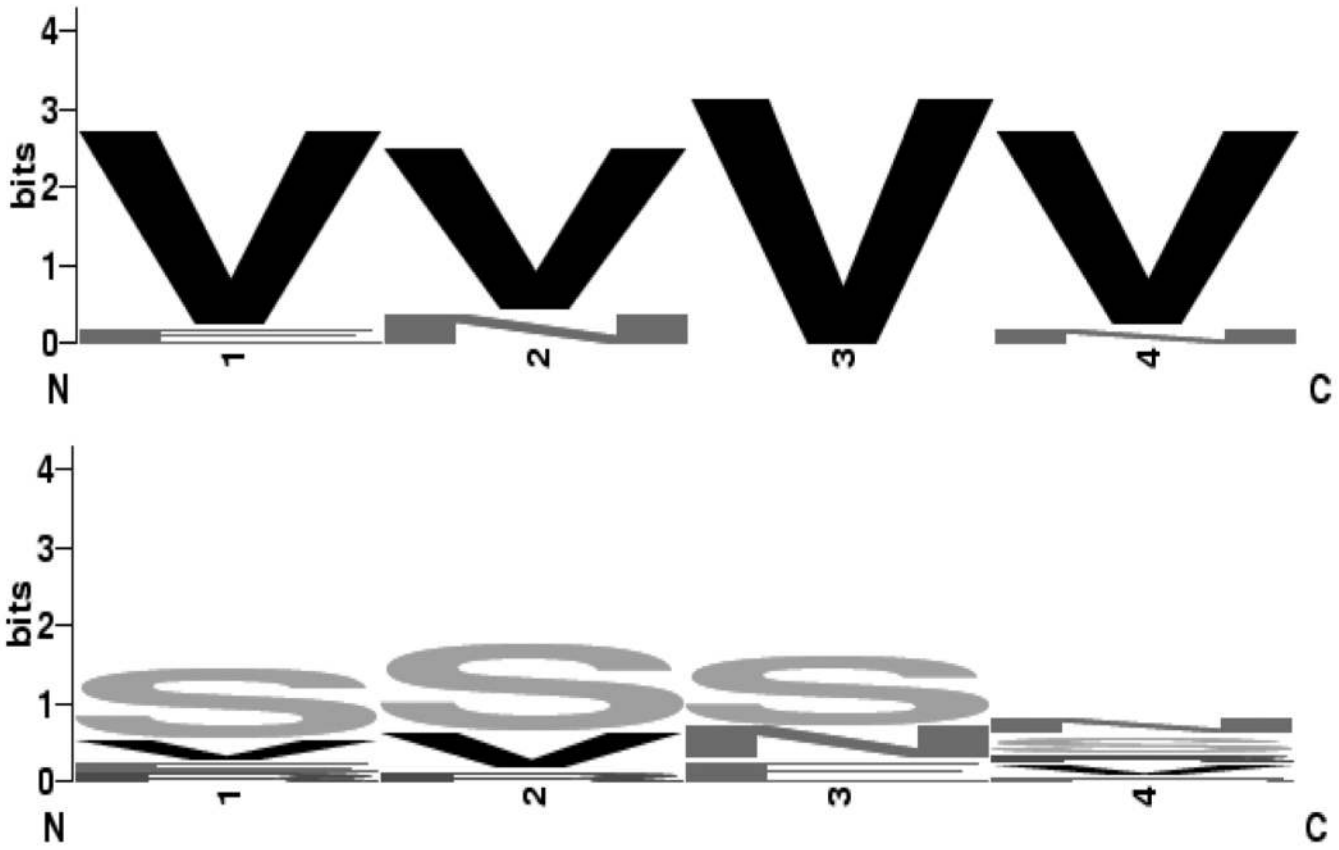

Fig. 10.

Two-stage Gibbs^2 motifs of length 4. The top motif comprises a working set of size 12, while the second motif corresponds to those 11 sequences (from a total population of 23) that were not included in the original working set.

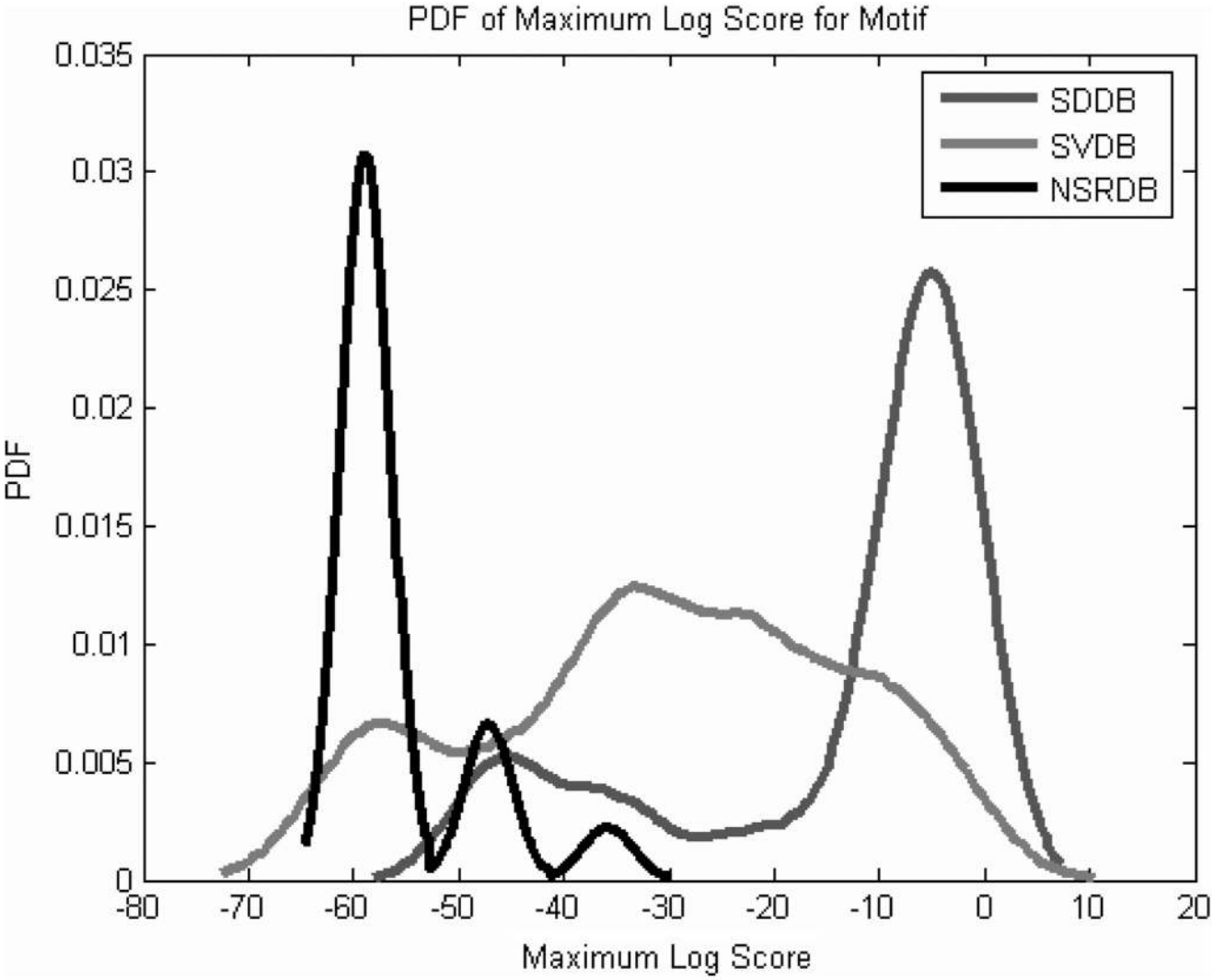

Fig. 11.

Motif-matching scores for patients in the Sudden Death Database (SDDB), Supraventricular Arrhythmia Database (SVDB), and Normal Sinus Rhythm Database (NSRDB). The graph shows the probability distributions estimated using kernel density estimation.