Spiral Wave Clustering using Normalized Compression Distance

Celal Alagoz[1], Andrew R Cohen[1], Allon Guez[1], John Bullinga[2]

[1]Drexel University, Philadelphia, PA, USA
[2]Penn Cardiology, Penn Presbyterian Medical Center, Philadelphia, PA, USA

Abstract

Cardiac fibrillatory dynamics are identified with spiral waves in mathematical modeling of cardiac electrical propagation. Automatic identification of spiral wave dynamics is essential for patient specific cardiac modeling.

In our work we used normalized compression distance (NCD), an information theoretical distance measure, in order to cluster the simulated spiral waves as stable, meandering and break up. Different representation of the data was introduced to NCD in the form of raw time series, fast Fourier transform (FFT), feature summarization and symbolic quantization of the simulated electrograms. Clustering was done in an unsupervised way using spectral method. Clustering analysis was performed using different validation methods. Gap statistics was used to find optimal number of groups. Jaccard coefficient was used in order to evaluate accuracy of clustering.

We had a perfect evaluation results from the raw data representation and Fourier transformation with a jaccard index of 1, and a very good performance of feature summarization with a jaccard index of 0.98.

1. Introduction

Cardiac arrhythmias are one of the most common cause of morbidity in overall globe [1]. Fibrillation, a certain type cardiac arrhythmia, is identified with one or many rotating waves and vortices with higher frequency than observed in normal sinusoidal rhythm. Spiral waves play important role to represent fibrillatory mechanism observed both experimental and numerical studies [2].

As a matter of fact, defining detailed arrhythmia dynamics in cardiac tissue is extremely difficult in the mesoscopic scale. Hence, there is not much study addressing this issue. A classification study was previously done on real intracardiac bipolar electrograms data obtained from patients using Jeffries–Matusita distance and support vector machine (SVM) classifier (Nollo et al, 2008). Classification was based on type I, type II, and type III AF according to Wells' criteria on the certain intracardiac electrogram pattern and morphology (Wells et al.1978).

We used an information theoretical distance measure NCD [3] in different representations of simulated electrograms. In this sense, we used raw electrograms as a time series form. We then used both higher level representation which is fast fourier transformation and lower level representation which is feature summarized form of the electrograms. Further quantization was made by converting feature lists to symbols. Spectral clustering was then used to group the data. We used gap statistics to find optimal number of groups. Finally, assuming ground truth labels for the data, we evaluated the resulting clusters with Jaccard similarity coefficient. We obtained distinctive groupings for data representations raw electrograms, FFT, and feature summarization.

2. Methods

2.1. Simulating electrophysiology

To simulate cardiac electrical activity mono-domain reaction-diffusion equation is used and can be read as

$$C_m \frac{dV_m}{dt} = \nabla . \sigma \nabla V_m + I_{stim} - I_{ion} \qquad (1)$$

Where V_m is transmembrane potential, σ is conductivity tensor or scalar diffusion coefficient, I_{ion} is ionic current density determined by the cardiac model used. Unipolar electrograms are computed using a current source approximation [6]

$$\Phi_e(x,t) = \frac{1}{4\pi\sigma_e} \int \frac{I_m(y,t)}{|x-y|} dy \qquad (2)$$

Where Φ_e is extracellular potential, x is electrode location vector, y is current source location vector, σ_e is extracellular conductivity, and I_m is trans-membrane current per unit area of atrial tissue surface. I_m is also defined from mono-domain equation:

$$I_m = I_{ion} - I_{stim} + C_m \frac{dV_m}{dt} = \nabla . \sigma \nabla V_m \qquad (3)$$

We used minimal resistor model (MRM), a 3 variable version of Fenton-Karma model [7], for the ionic part, or in other words ODEs part, and finite difference method (FDM) for the PDEs part. The temporal step of 0.1 ms and spatial step of 0.5 mm was used. Simulation was

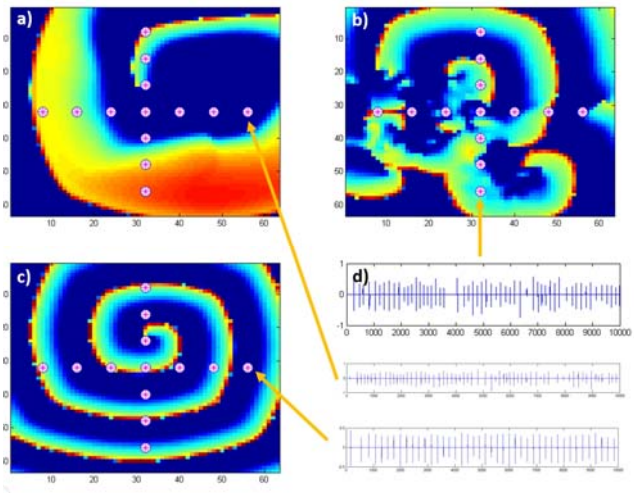

Figure 1. Simulated arrhythmia examples. a) Meandering spiral, b)Spiral breakup, c) Spiral breakup, and d) simulated electrograms. Electrogram channels were placed on the tissue (circles) in the way to mimic PentaRay catheter

performed on a 25.6 cm x 25.6 cm 2D grid. The diffusion rate or conductivity was assumed homogenous and was set to 0.00116 cm/S.

2.2. Normalized compression distance

The distance measure NCD reads as

$$NCD(x,y) = \frac{C(x,y) - \min\{C(x),C(y)\}}{\max\{C(x),C(y)\}} \quad (4)$$

where *C(.)* is a compression operator, *C(x)* is the file size (in bit length) of the compressed object *x* and *C(x,y)* is the file size of the concatenated objects *x* and *y*. We used bzip2 as the compressor which uses the Burrows-Wheeler algorithm.

2.3. Feature extraction and quantization

For comparison reasons, the data was presented to the distance function NCD in different formats. One of the formats is feature representation of the electrograms. For this reason, feature extraction hence data preprocessing was required. For features, cycle length (CL) and electrogram morphology was extracted.

CL was computed by estimating activation time, or spike time, using nonlinear energy operator [8]

$$E_j = x_j^2 - x_{j+1}x_{j-1} \quad (5)$$

The result then was filtered by a low-pass filter and the barycenter of the filtered signal was determined by a window and activation time was represented.

Electrogram morphology was computed same way as in [6] by detecting positive and negative deflections and then dividing difference of them with total amplitude.

We further summarized the data by using symbolic representation [9]. For our symbolic representation scheme, feature representation, CL and electrogram morphology, is replaced piecewise constant (PAA) representation as an intermediary step. We followed the same method as [10] when converting CL to the symbols by first getting CL histogram and from assumption they are normally distributed, assigning equiprobable regions with using inverse chi-squre distribution. We converted electrogram morphology directly by rounding them to the nearest tenth. We used 72 printable ascii characters for CL and 21 for electrogram morphology.

With feature and symbolic representation the time series of fixed length is reduced to arbitrary length.

2.4. Fourier transform and NCD

Higher representation of a time series data such as FFT and wavelet decomposition provides intrinsic frequency information. To get use of this, we introduced FFT of the simulated electrograms to NCD. We vertically buffered the electrogram from selected channels and performed FFT on them. In this case, L2 norm of FFT coefficients are used as compressing function *C(.)* and replaced bzip2.

2.5. Spectral clustering

Once having a distance matrix, it can be transformed, or mapped, to the spectral domain where eigenvector analysis can give further information about the data. First, data affinity matrix, or item-item similarity matrix, is calculated. Affinity matrix is calculated based on the distance values between the items and a free scale parameter and can be read as

$$A_{ij} = e^{-d(x_i,x_j)/2\sigma^2}$$

After getting affinity matrix, following step were followed which is adopted from [11]

Form spectral representation:

1. Form the affinity matrix *A*
2. Construct *D* by summing rows of *A* in diagonals of *D* where: $D_{ii} = \sum_j A_{ij}$
3. Form *N* by symmetric divisive normalization: $N = D^{-1/2}AD^{-1/2}$
4. Find the *k* largest eigenvectors of *N* which are $x_1, \ldots, x_k$ and form the matrix $X = [x_1, \ldots, x_k]$

For clustering:

5. Cluster into k-means
6. Assign the labels

2.6. Cluster analysis

To measure the cluster validity we used two different tools: jaccard index and gap statistics which are external and relative criteria respectively.

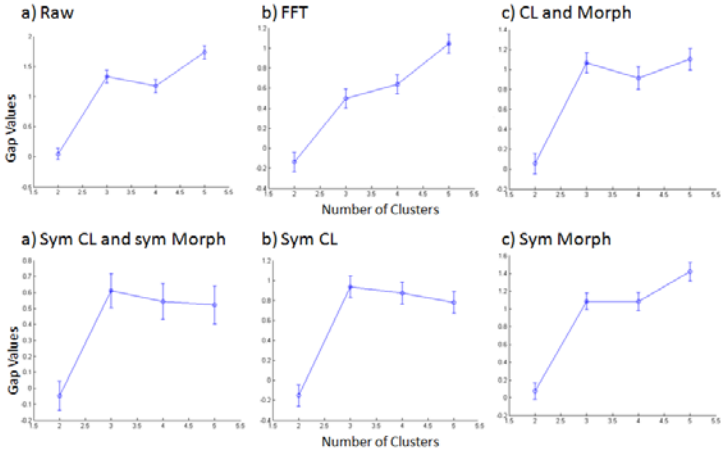

Figure 2. Gap values for different representations of the simulated electrogram data

Jaccard index as an external cluster validity criterion compares similarity and diversity between sample sets. It measures similarity by dividing the size of intersection divided by union of the samples

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (7)$$

In our case it measures the similarity between preimposed grouping structure and the groupings resulted from unsupervised clustering.

Gap statistic is presented to automatically assign the number of clusters by comparing them to random data generated by a uniform distribution [12]. The sum of distances within each cluster r

$$D_r = \sum_{i,j \in C_r} d_{i,j} \quad (8)$$

The summation of intracluster distances across all k clusters is then reads

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \quad (9)$$

The gap statistic is then the difference between intracluster distances of the data and the intracluster distances of B randomly generated uniformly distributed reference data

$$Gap(k) = \frac{1}{B} \sum_{b=1}^{B} \log(W_{kb}) - \log(W_k) \quad (10)$$

The simulation error for generation of B is defined as

$$s_k = \sigma_k \sqrt{1 + 1/B} \quad (11)$$

where σ_k standard deviation of set B. Then, k is chosen as the smallest value for

$$Gap(k) \geq Gap(k+1) - s_{k+1} \quad (12)$$

3. Results

In total 128 simulations were performed by changing the some of the parameters the MRM and each of them simulated a time duration of 10 seconds.

The simulated behaviors then labeled based on spiral wave dynamics such as: stable spiral, meandering spirals, and spiral wave breakup or also called as multiple wavelets. An example of the three behavior shown in Figure 1. In some parameter regimes the simulation was unstable and didn't give a meaningful result. Hence, they were taken out. Some spirals had a rather bigger tip trajectory and walked off the lattice. For practical reasons such as to have the electrogram vectors having the same size we eliminated them as well. Finally, the data set was constituted of 45 simulated spiral behavior 7 of which was stable spirals, 30 of which was meandering spirals, and remaining 8 of which was spiral wave breakups.

Table 1. Cluster analysis.

Data representation	Jaccard index (k=3)	Opt k	sigma
Raw	1	3	0.2
FFT	1	5	0.05
CL and Morph	0.98	3	0.23
Sym CL and Sym Morph	0.76	3	0.5
Sym CL	0.76	3	0.5
Sym Morph	0.93	3	0.25

After implementing NCD on the data, we used spectral clustering and gap statistic to calculate optimal number of clusters. Except FFT, all representations resulted in cluster number of 3 (Figure 2).

Then we imposed 3 spectral clusters on the data as shown in Figure 3. Finally, jaccard index to evaluate the cluster quality for each case. Results are summarized in Table 1. As we can see, raw electrograms and FFT gave perfect clustering. Feature summarization also gave a good result. Symbolic representation gave rather poor result. When we separately measured CL and electrogram morphology in symbolic representation, we saw that electrogram morphology is much more informative than CL. Using morphology alone almost give enough information to cluster data.

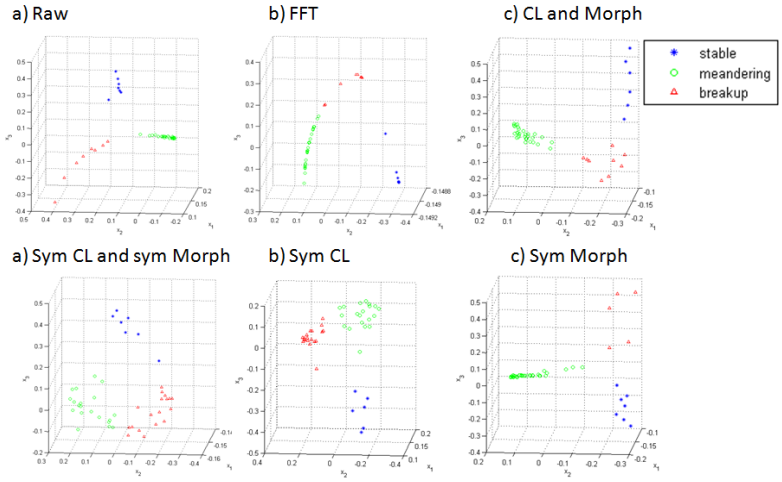

Figure 3. Spectral clustering results in 3D with imposed cluster number of 3

4. Discussion and future work

In this study we have tried to show that clustering of cardiac arrhythmia dynamics in terms of spiral waves is possible using electrogram data. The labels used for data clusters are intuitive and in agreement with phenomena of arrhythmia mechanisms such as multiple wavelets, leading circle and reentrant waves.

Different representation of the raw electrogram data which is in time series format was particularly important for compatibility of the data obtained from different means of measurements. For instance, for the same CL and electrogram morphology sequences an electrogram signal obtained from unipolar or monopolar recordings can be different.

Although simulated CLs were electrophysiologically realistic, we didn't take into account the conduction velocity and wavelength measurements. Given that the tissue size was bigger than the real case, the wavelength and conduction velocity values is not expected to be accurate. However, that doesn't mean the data used in this study is not compatible with the real data at all for the reason discussed in the previous paragraph.

References

[1] American Heart Association. Heart Disease and Stroke Statistics. 2013

[2] Davidenko JM, Pertsov AM, Salomonsz R, Baxter WT, Jalife J. Stationary and drifting spiral waves of excitation in isolated cardiac muscle. Nature (London) 1992; 355: 349–351

[3] Cilibrasi R, Vitanyi PMB. Clustering by compression. IEEE Trans. Information Theory. 2005; 51: 1523-1545.

[4] Nollo G, Marconcini M, Faes L, Bovolo F, Ravelli F, Bruzzone L. An automatic system for the analysis and classification of human atrial fibrillation patterns from intracardiac electrograms. IEEE Trans Biomed Eng 2008; 55:2275–85.

[5] Wells JL, Karp Jr RB, Kouchoukos NT, MacLean WA, James TN, and Waldo AL. Characterization of atrial fibrillation in man: Studies following open heart surgery. Pacing Clin Electrophysiol 1978; 1:426–438.

[6] Jacquemet V, Virag N, Ihrar Z, Dang L, Blanc O, Zozor S, Vesin JM, Kappenberger L, Henriquez C. Study of unipolar electrogram morphology in a computer model of atrial fibrillation. J Cardiovasc Electrophysiol 2003; 14:S172–S172.

[7] Fenton F, Karma A. Vortex dynamics in 3D continuous myocardium with fiber rotation: filament instability and fibrillation. Chaos 1998;8:20-47.

[8] Kaiser JF. On a simple algorithm to calculate the 'energy' of a signal. In Proc Int Conf Acoust, Speech, Signal Process. ICASSP 1990: 381–384.

[9] Lin J, Keogh E, Lonardi S, Chiu B. A symbolic representation of time series, with implications for streaming algorithms. Proc Eighth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery June 2003.

[10] Cohen A, Bjornsson C, Temple S, Banker G, Roysam B. Automatic summarization of changes in biological image sequences using algorithmic information theory. IEEE Trans Pattern Anal Machine Intell 2009; 31:1386–1403.

[11] Cohen AR, Gomes F, Roysam B, Cayouette M. Computational prediction of neural progenitor cell fates. Nature Methods 2010; 7:213 – 218.

[12] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap statistic. J Royal Statistical Soc 2001; 63: 411-423.

Address for correspondence.

Celal Alagoz
Electrical and Computer Engineering Department of Drexel University, Bossone Research Center Room 324A, 3140 Market Street, Philadelphia, PA 19104, USA
celal.alagoz@gmail.com