On Computation of Calcium Cycling Anomalies in Cardiomyocytes Data

Martti Juhola, Henry Joutsijoki, Kirsi Varpa, Jyri Saarikoski, Jyrki Rasku, Kati Iltanen, Jorma Laurikkala, Heikki Hyyrö, Jorge Ávalos-Salguero, Harri Siirtola
School of Information Sciences
University of Tampere
Tampere, Finland

Kirsi Penttinen, Katriina Aalto-Setälä
BioMediTech
University of Tampere
Tampere, Finland

Abstract**—***Induced pluripotent stem cell (iPSC) lines derived from skin fibroblasts of patients suffering from cardiac disorders were differentiated to cardiomyocytes and used to generate a data set of Ca^{2+} transients of 136 recordings. The objective was to separate normal signals for later medical research from abnormal signals. We constructed a signal analysis procedure to detect peaks representing calcium cycling in signals and another procedure to classify them into either normal or abnormal peaks. Using machine learning methods we classified signals into normal or abnormal signals on the basis of peak findings in them. We compared classification results obtained to those made visually by an expert biotechnologist who assessed the signals independent of the computer method. Classification accuracies of around 85% indicated high congruence between two modes denoting the high capability and usefulness of computer based processing for the present data.*

Keywords**—***Calcium cycling, cardiomyocytes, signal analysis, classification*

I. Introduction

Calcium cycling (Ca^{2+}) is vital for cardiac functionality. Variability of this biochemical cycling occurs in cardiac disorders and dysfunction. Consequently, cardiac functionality and disorders could be studied more thoroughly with the investigations of Ca^{2+} data analysis.

Spontaneously beating cardiomyocytes were differentiated from induced pluripotent stem cells originated from patients suffering from cardiac disorders. Cardiac signaling anomalies appear in the shape and frequency of time series data. It is imperative to investigate such forms to gain more information about cardiac functionality and disorders and to study the influence of medication on these cells. To our knowledge, such investigations have only been made subjectively and visually.

In the present research, we developed computational methods for the evaluation and prediction of Ca^{2+} signaling data in order to create efficient tools for medical and biotechnological researchers. Our approach is on the basis of signal analysis and data mining methods. The former was applied to detect peaks or cycles in a signal data set the results of which were run with those of latter to classify peaks or entire signals into two classes, either normal or abnormal.

II. Cell Line Data and its Preprocessing

A. Generation of cell data

Induced pluripotent stem cell (iPSC) lines from skin fibroblasts of patients suffering different cardiac disorders were established with retroviruses encoding for *OCT4*, *SOX2*, *KLF4* and *MYC* [1]. Differentiation into cardiomyocytes was carried out by co-culturing iPSCs with murine visceral endoderm-like cells as described earlier [2]. The beating areas of the cell colonies were mechanically and enzymatically dissociated for further analysis [2].

Ca^{2+} imaging was conducted in spontaneously beating, Fura-2 AM (Invitrogen, Molecular Probes) loaded dissociated cardiomyocytes perfused with extracellular solution as described earlier [3]. Ca^{2+} measurements were done on an inverted IX70 microscope (Olympus Corporation, Hamburg, Germany) and cardiomyocytes were visualized with a UApo/340 x20 air objective (Olympus). Images were recorded with an ANDOR iXon 885 CCD camera (Andor Technology, Belfast, Northern Ireland) synchronized with a Polychrome V light source by a real time DSP control unit and TILLvisION or Live Acquisition software (TILL Photonics, Munich, Germany). Fura-2 was excited at 340 nm and 380 nm light and the emission was recorded at 505 nm. For Ca^{2+} analysis, regions of interests were selected for spontaneously beating cells and signals were acquired as the ratio of the emissions at 340/380nm wavelengths.

B. Preprocessing of data

The present data set included short signals of no more than a few hundred samples since toxic UV and Fura-2 exposure on cell lines did not permit longer measurements. Signals consisted of different sampling frequencies: 8.3, 10.4, 11.4 or 22.3 Hz. Their durations also varied from around 11 to 24 s. Subject to the highest sampling frequency, signals were lowpass filtered with a standard median filter of window length 5 to suppress random impulse-type noise [4].

The research was sponsored by the Council of Tampere region under the auspices of European Regional Development Fund.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

At first, a rough amplitude estimate of large peaks in a signal was calculated by means of the amplitude distribution of samples. Before this task a linear trend was removed from a signal since frequently there was a linearly decreasing trend in the present data. In addition, the minimum of a signal was computed and subtracted from all values to obtain the zero minimum. Note that these operations were only used for the peak detection. After the detection, all further computation was made for the original (filtered) data. Note also that the samples had no unit since they were ratios of two measurement values as mentioned in Section II.A.

Next, all samples (amplitude values of the data) were counted to form a histogram mapping to represent their distribution. From their maximum index, i.e., the last histogram bar down to the location of 80% in the histogram bars was determined from which the mean of all amplitude values up to maximum index was calculated to estimate a rough lower bound of (large) peak maxima in a signal. This estimate A gave us preliminary information from the average quantity of large peaks in a signal such as in Fig. 1. It was used mainly to regulate some threshold values as described below.

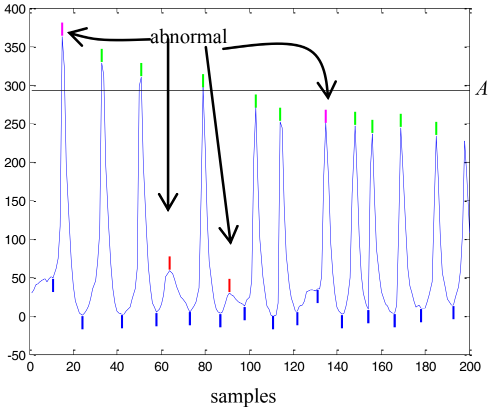

Fig. 1. A signal of 19.2 s after removing linear trend with four abnormal peaks. A is a rough estimate for large peak amplitudes in the signal. Because of the abnormal peaks (with the purple and red marks) the signal was assessed to be abnormal.

Next, the first derivative signal was differentiated (approximated) from a preprocessed signal by computing linear regression through successive samples with the window length of w equal to 3 or 7 depending on the above sampling frequencies. Slope values of linear regression were used to approximate the first derivative of the original signal.

III. Detection of Signal Peaks

The peak detection consists of the determination of locations of the beginning, maximum (top) and end of each peak. Note that all peaks are positive, i.e., their tops are always local maxima. On the other hand, beginnings and ends are local minima. See Fig. 1.

A. *Detection of peak candidates*

The distribution of signal samples after the preceding preprocessing was approximately from interval [60,300]. Each signal was searched for linearly from the signal beginning to its end as follows. To detect a peak beginning by means of the first derivative values, a segment was searched for in which there were derivative values less than threshold t_1 equal to 30 and after these values there were more than one value greater than or equal to t_1. To find the maximum of the current peak, there had to occur at least one derivative value again less than t_1. When right sides of peaks were often less steep than left sides, a smaller threshold of 0.6 t_1 was applied to search for peak ends. These threshold values were found experimentally on the basis of our datasets.

After detecting the segment of a peak beginning, the exact location of the beginning was determined to be the least signal sample of the signal segment just before the derivative value changing above t_1. For the peak end, the exact location was conversely just after a first derivative value dropped below the threshold. To exactly find the maximum at the top of the peak, the maximum sample was searched for from the signal segment found above for the first derivative maximum. Note that the exact extremum locations were searched for from the signal samples, but the segments to give their approximate locations were first determined with the derivative values.

B. *Discarding erroneous peak candidates*

Not all waveforms are acceptable and interesting peaks related to the cardiomyocyte data source with which they are associated in a signal. All waveforms are initially considered as peak candidates, and peaks are discarded based on the following criteria. A peak amplitude (Fig. 2) was calculated from the peak maximum to either beginning or end choosing the higher side (greater value). Sometimes these sides differed considerably from each other.

- If the first or last peak candidate of a signal was only partial, in other words, the first included no beginning (minimum), but merely the maximum and end, or the last peak candidate contained merely the beginning and maximum, such partial peak candidates were discarded.
- If the amplitude of a peak candidate was exceptionally small, such a candidate was deemed to be of noise or other irrelevant waveform and was excluded. Threshold t_2 of 10% (modified from [3]) related to the amplitude estimate A was experimentally found to be appropriate for most signals.
- Sometimes there were exceptional peak candidates being within either the left or right side of a larger peak. In such a situation, this additional, smaller peak consisted of much higher left side than its right side followed by the maximum of the larger peak containing the small peak in its own left side. The smaller peak and larger one had the same beginning. In the other situation, when the smaller peak and the larger had the same end, the right side of a larger peak contained the smaller peak which had a tiny left side but a larger right

side. If the smaller side of such an “inner” peak candidate was less than threshold t_3 equal to 20% from the higher side, this small inner peak candidate was discarded (left as such inside the larger peak).

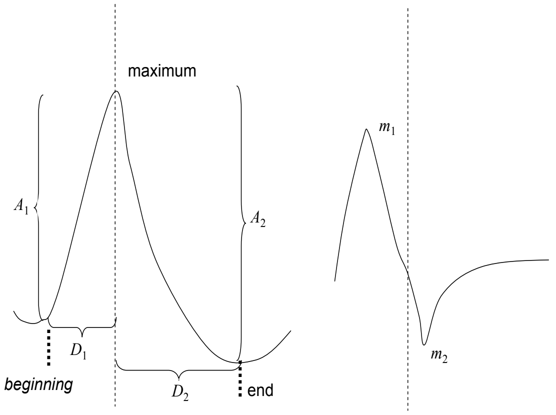

Fig. 2. A fictional peak on the left and its first derivative curve on the right show six variables computed: amplitudes A_1 and A_2, durations D_1 and D_2 and maximum derivative values m_1 and m_2 of the left and right side.

C. Computation of peak variables

In the beginning, six peak variables were computed from those called accepted peaks that remained after the rejection of erroneous peak candidates. The variables are shown in Fig. 2. They are the amplitudes of the left and right sides of a peak, the durations of both sides, and their maxima of the first derivative values. Since there existed more or less curvature in peaks, we used these maxima to evaluate the curvature property. Later, we added the seventh variable that was the interval (in time) of a peak calculated from the maximum of the preceding peak to the maximum of the current peak or from the beginning of the signal in the situation of the first peak when there had not been a rejected partial (large) peak before it. This was reasonable because it shows whether the peaks of a signal appear regularly or irregularly subject to time. The purpose of the use of peak variables was to classify individual peaks and entire signals.

IV. Classification of Accepted Peaks

Ultimately, we classified the signal peaks that were composed of two subsets including somewhat different types of signals as to their sampling frequencies, number of samples and number of peaks. All signals had been assessed visually by an expert familiar with the data and who had conducted the measurements. She had given the label of either normal or abnormal for every entire signal. In addition, the below algorithm was programmed to automatically screen all accepted peaks and to classify individual peaks to be either normal or abnormal. All the data were considered in two ways: processing with the below peak classification and as visually labeled signals. The expert did not take part in the development of the algorithm and had performed her visual labeling earlier. Thus, these two actions were independent of each other.

The peak classification programmed in Matlab was run for all accepted peaks of all signals given seven peak variables as follows.

- If the larger side of a peak was greater than t_3 equal to 70% (modified from [3]) from that of the preceding peak provided that the preceding one was either normal or greater than t_4 equal to 50% of estimate A, the current peak was classified as normal. If the condition was not satisfied, the current peak was classified to be abnormal. If the current peak was the first in a signal or there was neither normal nor large enough predecessor peak, it was compared to estimate A applying t_4. The second and third abnormal peaks in Fig. 1 were found according to this rule.
- The asymmetry of peak sides was checked. If one of the peak sides was clearly less than the other, it was classified to be abnormal. The threshold was t_5 equal to 88% (modified from [3]). The first and fourth abnormal peaks in Fig. 1 were identified with the present rule.

Not all of the peak variables were used in the peak classification above. However, they were used later in the classification with data mining methods. A signal classified both visually and automatically as normal is seen in Fig. 3. There was large variability among signals. For instance, a part of them contained a few peaks only. The signal in Fig. 1 included an average number of accepted peaks, 13.

V. Results

The signals were processed according to Sections III and IV to classify the peaks. We tested an earlier subset of 93 signals sampled at 8.3, 10.4 or 11.4 Hz and a later subset of 43 signals sampled at 22.3 Hz, and then jointly all the 136 signals. The approach was chosen since the signal properties varied between two subsets and, thus, we were also interested in seeing their separate results. The subset of 43 signals contained 234 peaks and 93 signals included 1690 peaks. The number of peaks in all signals varied from 1 to 43 (mean 14.1, median 12 and mode 5 peaks). Before classification all seven variables of the data were standardized to have the zero mean and unit variance.

We classified the peaks of signals using leave-one-out validation that is appropriate to small data sets. In leave-one-out, a single signal formed the test set and all others were used as a learning set to build a computational model, i.e., the peaks of a single signal were used in testing one at a time whereas the peaks of the other signals were used in training of the classification method. This way, all signals were run. Classification accuracy means here the percentage of correct decisions into the classes of either abnormal or normal peaks or signals. True positive and negative rates were also computed.

We ran K-nearest neighbor method with odd K equal to {1,3,…,21}, linear and quadratic discriminant analysis, naïve Bayes rule and classification trees [5,6]. Table I contains the best accuracies of the methods in the peak classification. The best K varied within subsets (with K=1,3,5,9,11).

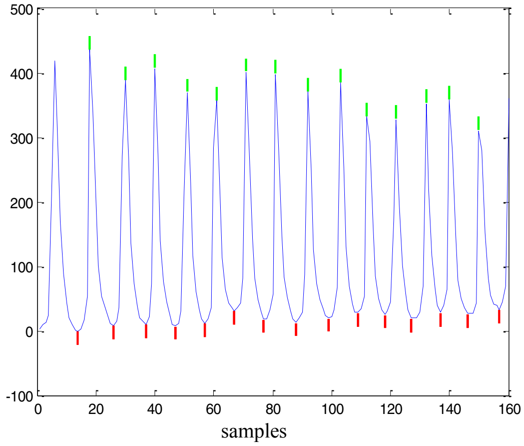

Fig. 3. The signal of 19.2 s was classified to be normal including no abnormal peak.

Taking into account both true positive and negative rates and accuracy, discriminant analysis produced the best results. Nevertheless, the smaller subset of the 43 signals was difficult for all classification, obviously for the sake of the imbalanced class distribution. Signals labelled as abnormal also included normal peaks and not only abnormal. This could achieve false negative corresponding to normal peaks in abnormal signals.

Second, we classified the signals after their peak classifications (normal or abnormal) with each method. Again, the peaks of one signal were used as test cases and those of the others as the training set. If even a single peak of a signal was classified to be abnormal, the whole signal was interpreted to be such. These signal classification results were compared to the visually determined signal classes (seen correct). If a classification test gave the same result for a signal as was determined visually, the classification was correct, otherwise incorrect. The classification results are shown in Table II. In the signal level classification, the best K was 3 or 5. The best accuracies in peak and signal classifications with different sets were yielded by discriminant analysis. Again, the smaller subset of the 43 signals was difficult. The probable reason was its highly imbalanced class distribution and small size. In the entire data set, the class sizes were virtually equal.

VI. Discussion and Conclusion

We obtained useful results subject to the new way of considering calcium cycling anomalies in the present cardiomyocyte data and to the development of their automatic assessment. Our classification results were good even if our small data set was rather heterogeneous in regard with signal properties. In addition, automatic classification is difficult for these data because of short signals frequently including only small numbers of peaks.

The results reported consisted of preliminary tests only. We will develop the method introduced. Computer based methods will play an essential role in the future once the medical research of this field will emerge as applications in practice.

TABLE I. Classification results of individual peaks (classes of normal as positive and abnormal peaks as negative) with K-nearest neighbor searching (K-NN), (linear or) quadratic discriminant analysis (DA), naïve Bayes (NB) rule and classification trees

	True positive *TP* and negative *TN* rates and accuracy *A* [%]								
	1924 peaks in 136 signals			***1690 peaks in 93 signals***			***234 peaks in 43 signals***		
peaks	1496	428	1924	1292	398	1690	204	30	234
Method	***TP***	***TN***	***A***	***TP***	***TN***	***A***	***TP***	***TN***	***A***
K-NN	97.5	40.7	84.9	97.6	41.2	84.3	99.5	6.7	87.6
DA	94.8	72.9	89.9	95.0	73.9	90.1	80.9	33.3	74.8*
NB	92.6	38.6	80.6	90.7	39.4	78.6	99.5	0.0	86.8
trees	87.0	60.3	81.1	86.3	60.3	80.2	94.1	13.3	83.8

* Quadratic discriminant analysis not used since a positive-definitive matrix was not obtained.

TABLE II. Classification results (classes of normal as positive and abnormal signals as negative) with K-nearest neighbor searching (K-NN), (linear or) quadratic discriminant analysis (DA), naïve Bayes (NB) Rule and classification trees compared with visual rating of signals

	True positive *TP* and negative *TN* rates and accuracy *A* [%]								
signals	67	69	136	29	64	93	38	5	43
Method	***TP***	***TN***	***A***	***TP***	***TN***	***A***	***TP***	***TN***	***A***
K-NN	88.1	76.8	82.4	79.3	78.1	78.5	97.4	40.0	90.7
DA	80.6	91.3	86.0	79.3	90.6	87.1	84.2	60.0	81.4*
NB	88.1	59.4	73.5	69.0	71.9	71.0	100.0	20.0	90.7
trees	71.6	91.3	81.6	41.4	89.1	74.2	84.2	60.0	81.4

* Quadratic discriminant analysis not used since a positive-definitive matrix was not obtained.

References

[1] K. Takahashi, K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka et al., ”Induction of pluripotent stem cells from adult human fibroblasts by defined factors”, *Cell*, vol. 131, pp. 861-872, 2007.

[2] C. Mummery, D. Ward-van Oostwaard, P. Doevendans, R. Spijker, S. van den Brink et al., “Differentiation of human embryonic stem cells to cardiomyocytes: role of coculture with visceral endoderm-like cells”, *Circulation*, vol. 107, pp. 2733-2740, 2003.

[3] K. Kujala, J. Paavola, A. Lahti, K. Larsson, M. Pekkanen-Mattila, M. Viitasalo, A.M. Lahtinen, L. Toivonen, K. Kontula, H. Swan, M. Laine, O. Silvennoinen, K. Aalto-Setälä, “Cell model of catecholaminergic polymorphic ventricular tachycardia reveals early and delayed afterdepolarizations”, *PLoS ONE*, vol. 7, no 9, 2012.

[4] M. Juhola, J. Katajainen, T. Raita, ”Comparison of algorithms for standard median filtering,” *IEEE Transactions on Signal Processing*, vol. 39, pp. 204-208, 1991.

[5] A. Webb, *Statistical Pattern Recognition*, 2nd ed. Chichester, England: John Wiley & Sons, 2002.

[6] I.H. Witten, E. Frank, *Data Mining*, San Francisco, CA, USA: Morgan Kaufmann Publishers, 2000.