A New On–Line Electrocardiographic Records Database and Computer Routines for Data Analysis

Carlos A. Ledezma[1], *Student Member, IEEE*, Erika Severeyn[1], Gilberto Perpiñán[1], Miguel Altuve[1], *Member, IEEE*, and Sara Wong[1]

***Abstract*— Gathering experimental data to test computer methods developed during a research is a hard work. Nowadays, some databases have been stored online that can be freely downloaded, however there is not a wide range of databases yet and not all pathologies are covered. Researchers with low resources are in need of more data they can consult for free. To cope with this we present an on-line portal containing a compilation of ECG databases recorded over the last two decades for research purposes. The first version of this portal contains four databases of ECG records: ischemic cardiopathy (72 patients, 3–lead ECG each), ischemic preconditioning (20 patients, 3–lead ECG each), diabetes (51 patients, 8–lead ECG each) and metabolic syndrome (25 subjects, 12–lead ECG each). In addition, one computer program and three routines are provided in order to correctly read the signals, and two digital filters along with two ECG waves detectors are provided for further processing. This portal will be constantly growing, other ECG databases and signal processing software will be uploaded. With this project, we give the scientific community a resource to avoid hours of data collection and to develop free software.**

I. INTRODUCTION

Gathering ECG records in order to test methods is a long and complicated work. Researchers in need of test subjects have a hard time looking for people that fill the requirements and people willing to be part of a control group. Additionally, if the disease being studied is a rare one, the finding of subjects is even more complicated. Finally, if one tries to conduct a research in a country were medical facilities are not well prepared, the data collection becomes nearly impossible.

All around the world people are conducting research using ECG records; there is no need for each researcher to do the data gathering. The development of on-line, free and high quality databases should be encouraged to help people carry on their research and conduct valuable scientific studies. The usefulness of on-line databases exceeds the scientific purpose, these tools may also be used to improve and adapt teaching methods [1].

Fortunately, a good number of databases exist on-line for researchers to have data to work with. Some of them, such as the ones provided by the Telemetric and Holter ECG Warehouse (http://thew-project.org) [2] and the Cardiac Safety Research Consortium (www.cardiac-safety.org) [3] are available on-line, but are membership based. Other databases, such as those found on PhysioNet (http://www.physionet.org) [4], are initiatives that help researchers find high quality on-line data to work with at no cost. PhysioNet has a very high acceptance among the scientific community; several tools are available on the website to process the data as well as annual competitions to encourage people to solve clinical problems; nevertheless, PhysioNet does not provide detailed information about the projects involved and does not have databases for some diseases such as diabetes and metabolic syndrome.

There are also older databases such as the American Heart Association Database for Evaluation of Ventricular Arrhythmia Detectors [5] and the Common Standards Electrocardiography Database [6] that have a good number of high quality recordings, but these are available on optical media types (CD-ROM), so they are not immediately available.

This work aims to contribute to such initiatives by placing at the disposal of the scientific community the databases and computer programs developed by our research group. For over two decades, we have designed and developed various ECG databases in order to study illnesses such as Chagas, ischemic cardiopathy, diabetes, metabolic syndrome, ... Hard work and collaboration between engineers and physicians has been necessary. Therefore, it is important to preserve this data and allow the scientific community to profit from this work.

In this paper, we present an on-line, free and high quality database containing, as a start-up, ECG of three common pathologies: ischemic cardiopathy, diabetes and metabolic syndrome. One hundred sixty subjects underwent ECG acquisition to produce 1044 ECG records (using different ECG-lead configurations) with different sampling rates and resolutions. Furthermore, computer routines for signal processing and data analysis are currently available on the website (three routines to read the signals, two digital filters and two ECG waves detectors).

The rest of this paper is organized as follows. In the next section, a detailed description of the portal is presented; databases and programs are carefully described. Later, the results are explained and discussed, and finally, conclusions and future works on this matter are outlined in the last section.

II. METHODOLOGY

A. Web-page description

The portal under the name of GBBANet was created using HyperText Markup Language (HTML) 4.0.1. In the web–page the user will find different spaces with links to navigate through the site and consult the different databases.

*This work was supported by the Deanship of Research and Development of Simón Bolívar University, under the project S1-IN-CAI-002-13.

[1]C.A. Ledezma, E. Severeyn, G. Perpiñán, M. Altuve, and S. Wong are with Grupo de Bioingeniería y Biofísica Aplicada, Universidad Simón Bolívar, Valle de Sartenejas, Baruta, estado Miranda, Venezuela. caledezma@gmail.com, severeynerika@usb.ve, maltuve@usb.ve, swong@usb.ve

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

In addition, each database has a description of the project where the data was acquired and the results of these studies. In the site, the user will also find references to papers, thesis and other publications that will lead to further information about the projects. This web–page is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. This means any user of the GBBANet is free to share and adapt the material here shown under the current following terms: i) First any user must give appropriate credit, providing the link to the license and indicating if changes were made; ii) also, this material may not be used to gain monetary profit; and finally iii) no additional legal terms may be applied that restrict others from doing what this license permits.

B. ECG databases

As a first implementation of the GBBANet, four databases were included and made public:

1) Ischemic cardiopathy database: The data for this database was collected in the rehabilitation unit of the University Hospital of Caracas [7]. ECG records from leads II, V5, and V6 were acquired for each patient.

The protocol used for the acquisition was: i) all patients performed two stress tests with one week between each test; ii) there was no medication variation between both tests and no cardiac events or symptom changes happened (patients with recent heart attacks were discarded); iii) each record pair was conducted at the same time with the same equipment (Quinton 3000 or Quinton 5000); iv) the stress tests were conducted following the Bruce protocol; and v) the tests were stopped in case of angina pectoris, ST level elevation over 3 mm, fatigue or hypertension.

For the data acquisition, the SISPAS prototype was used [7]. The acquisition was made through three channels of 12-bit resolution and 250 Hz sampling frequency.

2) Ischemic preconditioning database: Ischemic preconditioning is defined as a fast and adaptive response to a brief ischemic episode that minimizes the cellular death in a subsequent and longer episode of ischemia. The development of this database had the objective of studying to the detail the ischemic preconditioning phenomenon in order to better understand and document it [8].

The data was collected under the following conditions: i) each patient performed two stress test with 24 h separation; ii) the stress tests were conducted using a Quinton 5000 equipment; and iii) the data was acquired using the SISPAS prototype through three channels of 12-bit resolution and 250 Hz sampling frequency.

3) DICARDIA database: This database was conceived to study the cardiac autonomic neuropathy produced by diabetes [9]. The data sources for this project were routine procedures that took place at the University Hospital of Caracas. The collected data was introduced in a database called ecgML, which helps visualizing the demographic data and the ECG signal's characteristics with notes for each patient.

All the test subjects (excluding the control group) were patients presenting diabetes mellitus type II. Asymptomatic patients with no known pathology and repose ECG with no alterations formed the control group.

The data was collected under the following conditions: i) the subjects underwent an effort test following the Bruce protocol to which was added a three minute warm–up stage (in total there were 7 stages, 3 minute long each; for a non-trained subject the average test duration is 20 minutes); ii) the Ergocid–At–Plus effort test system was used to collect 8-lead ECG (500 samples per second) during the effort test and; iii) the demographic data was collected and saved for all patients.

4) Metabolic syndrome database: This ECG database is used to evaluate patients with metabolic syndrome, a condition associated with diabetes mellitus type II and cardiovascular disease [10]. For the diagnosis of metabolic syndrome, the National Cholesterol Education Program–Adult Treatment Panel (NCEP ATP III) criteria was used [11]. According to this criteria, the presence of three or more of the following characteristics is sufficient to diagnose metabolic syndrome: i) waist circumference greater than 102 cm for men and 88 cm for women, ii) blood pressure equal or higher than 130/85 mmHg, iii) HDL–cholesterol lower than 40 mg/dL for men and 50 mg/dL for women, iv) triglyceride count higher than 150 mg/dL and, v) fasting blood glucose equal to or higher than 110 mg/dL.

The database has two parts: the laboratory parameters and 12–lead ECG recording. The laboratory parameters consist of a blood test, anthropometric variables measurement, clinical variables measurement and oral glucose tolerance test (OGTT). The methodology applied in the collection of this data consisted of four parts: i) first an ECG record was performed, ii) then, a blood test was performed with the patient in basal state (while the patient was in fast), iii) later, the patient would take 75 gr of liquid glucose, and iv) finally, an ECG record followed by a blood test was taken every 30 minutes for 120 minutes. The ECG records had 15-minute duration in most of the cases. However, due to time limitations some samples were cut to 10 minutes.

The test subjects selected for this study were all male, aged between 20 and 44 years old, non–smokers, under no medical treatment, with no physical disabilities and no evident cardiovascular disease.

The data is provided in *.csv* files. For each patient various files are provided, all for the records taken in a basal state and after 30, 60, 90 and 120 minutes: i) complete ECG record (1000 samples per second), ii) QRS complex time of appearance, and iii) RR interval (cardiac cycle length) curve.

C. Useful Programs

Along with the databases, a number of routines are provided to correctly read and process the data.

The data from the ischemic cardiopathy and ischemic preconditioning databases may be read using the *read_reg* routine.

For the visualization of the DICARDIA data the *ecgML* program is provided. It uses a method that describes a document by inserting tags in it; ecgML was specifically designed for ECG data collection and analysis. The specifications are coded in XML language. The ecgML structure allows representing the data from the ECG following a hierarchical structure, taking into account the events, patient's data, diagnostic and other clinical data. The *readECG* routine is also provided in a *.m* file that allows to import the ECG signals to a *MATLAB* or *Octave* variable.

The *readMS* routine is provided in order to import the data from the metabolic syndrome database to a *MATLAB* or *Octave* variable; this routine can read all 12 leads of the recording or one specified by the user. The method also provides plots that show the different signals acquired from each subject.

Some general signal processing routines were coded in *MATLAB* and are provided. These routines can be used with any signal, even on other databases. Currently, a high-pass and a low-pass filters, a multi-lead QRS detector based on the Pan & Tompkins algorithm [12], and an R, P and T waves detector based on the wavelet transform method described in [13] can be found in the website.

III. RESULTS

The main result of this work was the creation of GBBANet, a web portal containing ECG signals and tools for signal processing and data analysis. The webpage is available in http://gbbanet.labc.usb.ve. Through this web page, any researcher may consult our growing database bank containing, at this moment, 1044 ECG records. In addition, one software (ecgML), three routines to read and visualize the signals (*ReadECG*, *read_reg*, and *readMS*), two digital filters (high-pass and low-pass) a QRS complex detector (multi-lead Pan and Tompkins) and an R, P and T waves detector (wavelet-based) are available on-line.

Fig. 1 shows the result of applying the *read_reg* routine along with the wavelet detector to a registry and Fig. 2 shows the result of applying the *readMS* routine along with the QRS complex detector (Pan and Tompkins) to a record. The *ReadECG* routine showed to successfully read all 8-leads from the files contained in the DICARDIA database.

The ischemic cardiopathy study resulted in two groups of male individuals of age 53 ± 9.3 years old: the first was a healthy group (23 subjects) and the other presented coronary disease (27 patients). These records are 30.05 ± 1.17 min long.

The ischemic preconditioning study resulted in two groups: a group of 14 individuals presenting ischemic cardiopathy of 59.00 ± 10.26 years old and a group of 6 asymptomatic individuals for the control group of 54.00 ± 9.66 years old. These records are 29.91 ± 1.69 min long.

The DICARDIA database was formed with 65 subjects divided in three groups: i) 51 diabetic subjects with cardiac complications of age 57.00 ± 10.00 years old and weight 73.00 ± 15.00 kg, ii) 3 diabetic subjects without cardiac complications of age 49.00 ± 12.00 years old and weight 79.00 ± 8.00 kg, and iii) 11 healthy subjects as a control group of age 50.00 ± 6.00 years old and weight 81.00 ± 20.00 kg. These records are 24.32 ± 7.25 min long.

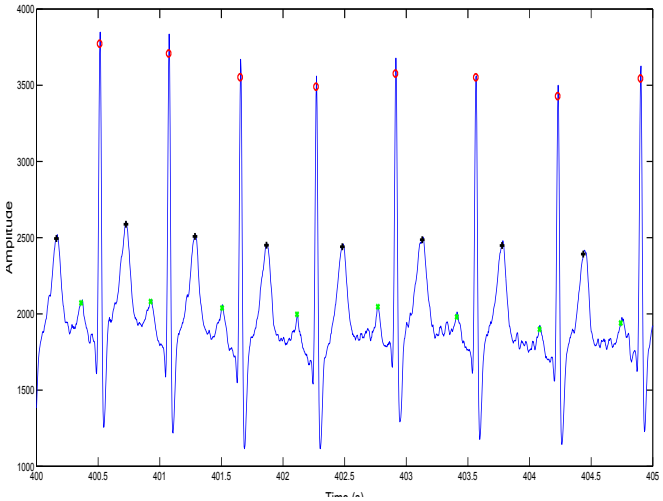

Fig. 1. Result of reading an ECG signal using *read_reg* and applying the wavelet detector. The positions of P (maximum), R, and T (maximum) waves, obtained using the wavelet-based detector, are indicated with characters ''x", ''o" and ''+", respectively.

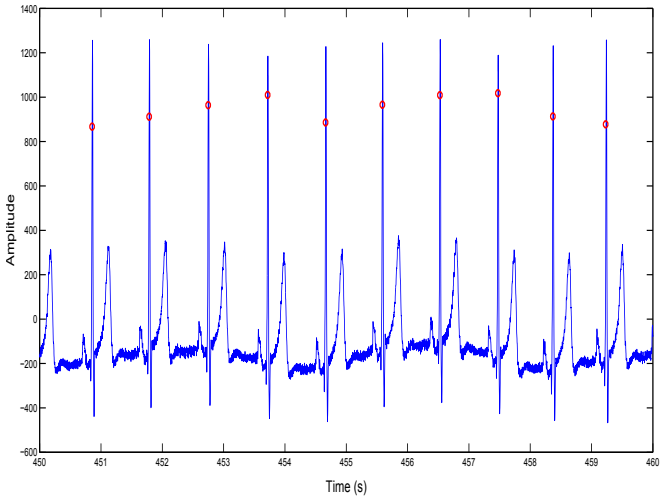

Fig. 2. Result of reading the aVF derivation using *readMS()* and applying the multi-lead QRS detector. Locations of R waves are indicated with character ''o".

The metabolic syndrome study produced 25 ECG recordings: 15 from patients presenting metabolic syndrome of age 31.40 ± 6.98 years old, height 1.74 ± 0.07 m, weight 104.66 ± 23.14 kg, waist circumference 113.63 ± 19.35 cm, systolic arterial pressure 135.13 ± 7.10 mmHg and diastolic arterial pressure 85.67 ± 10.59 mmHg; and 10 from healthy subjects for the control group of age 26.90 ± 4.18 years old, height 1.77 ± 0.09 m, weight 73.01 ± 13.56 kg, waist circumference 83.51 ± 10.75 cm, systolic arterial pressure 117.60 ± 9.23 mmHg and diastolic arterial pressure 72.40 ± 7.23 mmHg. These records are 13.84 ± 2.26 min long.

IV. DISCUSSION

GBBANet allows anyone to freely download different ECG signals used in the study of various diseases. Over 1000 ECG signals are being made public, this is an important step in the diversification of information. Through our database bank, detailed information on the conditions of acquisition and the project where the database was involved is provided. The ECG records uploaded to the website represent an important amount of data that may be used for scientific or academic purposes. This data is characterized by its high quality, it has been used in previous projects and has lead to important results already. An important feature of GBBANet is its originality, we are presenting ECG records that describe diseases from which there is very little public data on the internet.

Resources to read and process the signals are freely available as well. The routines developed during the construction of the GBBANet are useful to have a first approach to these databases. By using them, the user will be able to read, visualize and process the ECG signal.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we have dealt with the problem of lacking of testing ECG records for research purposes. As a start–up project, we have provided four ECG signals databases for the study of different widespread pathologies: ischemic cardiopathy, ischemic preconditioning, diabetes originated cardiac autonomic neuropathy, and the metabolic syndrome.

GBBANet is an important contribution to the world of online ECG databases. Many characteristics make GBBANet a high quality and innovative database. First, we are providing, for free, a bank of high quality signals that have been used in previous studies. Also, we provide detailed description of previous usages of the signals, which gives the researcher insight on possible applications. Additionally, we are providing ECG databases for the study of metabolic syndrome and diabetes which were not available, until now, in the main ECG databases.

This work represents an important step in the diversification of information. To our knowledge, this is the first initiative of this type in Latin America, it being an important step for our research community. We expect this database to help others accomplish their research and create free software.

As this is a public work targeting the benefit of the scientific community, it is protected by a Creative Commons license. This license ensures that the data offered in the website cannot be used for any commercial purposes and proper credit has to be given to the researchers involved in the data collection.

More databases are to be uploaded to the website. As a future work we expect to include other ECG databases developed in the GBBA [14], [15], medical histories and more programs.

Finally, we do not discard the possibility of uploading this data to Physionet. This would give our work a bigger reach and more credit would be given to the investigators. An additional advantage would come from the need of standardizing the data format, which would give us the possibility to use the Physiotoolkit to further process our signals.

ACKNOWLEDGMENT

Authors would like to thank everyone who has worked with GBBA and University Hospital of Caracas, gathering the information that is now being made public.

REFERENCES

[1] H. E. Wong, A. A. Kazzi, and M. I. Langdorf, "Utility of the CORD ECG Database in Evaluating ECG Interpretation by Emergency Medicine Residents," *The California Journal of Emergency Medicine*, vol. 3, no. 4, p. 59, 2002.

[2] J. Couderc, "The telemetric and holter ECG warehouse initiative (THEW): A data repository for the design, implementation and validation of ECG-related technologies," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2010, pp. 6252–6255.

[3] P. Kligfield and C. L. Green, "The cardiac safety research consortium ECG database," *Journal of electrocardiology*, vol. 45, no. 6, pp. 690–692, 2012.

[4] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[5] R. E. Hermes, D. B. Geselowitz, and G. C. Oliver, "Development, distribution, and use of the American Heart Association database for ventricular arrhythmia detector evaluation," *Computers in cardiology*, pp. 263–266, 1980.

[6] J. L. Willems, P. Arnaud, J. H. Van Bemmel, P. J. Bourdillon, R. Degani, B. Denis, I. Graham, F. M. Harms, P. W. Macfarlane, G. Mazzocca *et al.*, "A reference data base for multilead electrocardiographic computer measurement programs," *Journal of the American College of Cardiology*, vol. 10, no. 6, pp. 1313–1321, 1987.

[7] S. Wong, F. Mora, G. Passariello, J. Silva, and D. Almeida, "A new computerized approach for assessing myocardial ischaemia," *Revista Brasileira de Engenharia Biomédica*, 1996.

[8] F. Ng, S. Wong, D. Almeida, F. Mora, and G. Passariello, "Stress ECG analysis for ischemic preconditioning evidence," in *Computers in Cardiology, 1996*. IEEE, 1996, pp. 473–476.

[9] L. Quintero, S. Wong, R. Parra, J. Cruz, N. Antepara, D. Almeida, F. Ng, and G. Passariello, "Stress ECG and laboratory database for the assessment of diabetic cardiovascular autonomic neuropathy," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 2007, pp. 4339–4342.

[10] E. Severeyn, S. Wong, G. Passariello, J. L. Cevallos, and D. Almeida, "Methodology for the study of metabolic syndrome by heart rate variability and insulin sensitivity," *Revista Brasileira de Engenharia Biomédica*, vol. 28, no. 3, pp. 272–277, 2012.

[11] S. M. Grundy, H. B. Brewer, J. I. Cleeman, S. C. Smith, C. Lenfant *et al.*, "Definition of metabolic syndrome report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on scientific issues related to definition," *Circulation*, vol. 109, no. 3, pp. 433–438, 2004.

[12] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *Biomedical Engineering, IEEE Transactions on*, no. 3, pp. 230–236, 1985.

[13] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A wavelet-based ECG delineator: evaluation on standard databases," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 4, pp. 570–581, 2004.

[14] F. Mora, P. Gomis, and G. Passariello, "Señales electrocardiograficos de alta resolucion en chagas: el proyecto SEARCH," *Acta cient. venez*, vol. 50, no. 3, pp. 187–94, 1999.

[15] A. M. Chinea, C. Lollett, H. Herrera, G. Passariello, and S. Wong, "Protocol for cardiac assessment of recreational athletes," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 2132–2135.