PhysioNet: Research Resource for Complex Physiologic Signals

George B. Moody

Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

Since 1999, PhysioNet (`http://physionet.org/`*) has offered free access via the web to large collections of recorded physiologic signals and related open-source software. This presentation offers an overview of PhysioNet and discusses its benefits to the global community of researchers, clinicians, educators, and students, with examples that include support of exploratory data analysis, development and validation of experimental methods, collaboration among geographically distant investigators, and stimulating rapid progress on specific questions via open research and engineering challenges.*

1. Introduction

This presentation introduces PhysioNet, a resource for researchers who study physiologic signals and time series.

The resource, intended to stimulate and support current research and new investigations in the study of complex biomedical and physiologic signals, has three closely interdependent components:

PhysioBank is a large and growing archive of well-characterized digital recordings of physiologic signals, time series, and related data for use by the biomedical research community.

PhysioToolkit is a large and growing library of open-source software for physiologic signal processing, analysis, and simulation.

PhysioNet is not only the name of the Resource, but also of its web site, physionet.org, which provides free access to PhysioBank, PhysioToolkit, and related research materials and facilities.

PhysioNet was established in 1999 by researchers at the Massachusetts Institute of Technology, Harvard Medical School, Boston University, and McGill University, with funding from the (US) National Center for Research Resources (NCRR) of the National Institutes of Health (NIH)[1, 2]. Since 2007, it has been funded by two other units of the US National Institutes of Health (NIH): the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Institute of General Medical Sciences (NIGMS). About 30,000 visitors use PhysioNet each month, and the main PhysioNet server at MIT supplies over one terabyte of data (about 4 million hits) each month. A global network of PhysioNet mirrors provides backup and additional network bandwidth to the worldwide community of PhysioNet users.

2. PhysioBank

PhysioBank currently includes databases of multiparameter cardiopulmonary, neural, and other biomedical signals from healthy subjects and patients with a variety of conditions with major public health implications, including sudden cardiac death, congestive heart failure, epilepsy, gait disorders, sleep apnea, and aging. These collections include data from a wide range of studies, as developed and contributed by members of the research community.

PhysioBank currently contains well over 7000 recordings of annotated, digitized physiologic signals and time series, organized in over 40 databases (collections of recordings). In this context, a database is simply a collection of recordings (records), available as a set of flat files. In contrast to typical relational databases, PhysioBank databases consist of relatively small numbers (tens to thousands) of records that may each be quite large (in some PhysioBank databases, records are as large as 500 Mb each, although typical sizes are a few Mb). Many of the databases currently in the PhysioBank Archives were developed at MIT and at Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center) and have previously been distributed in CD-ROM format[3]. All of these databases are available in their entirety from these archives.

Each database consists of a set of records (recordings), identified by the record name. In most cases, a record consists of at least three files, which are named using the record name followed by distinct suffixes (extensions) that indicate their contents. For example, the MIT-BIH Arrhythmia Database includes record 100; the three files 100.atr, 100.dat, and 100.hea together comprise record 100. Almost all records include a binary .dat (signal) file, containing digitized samples of one or more signals; these files can be very large. The .hea (header) file is a short text file that describes the signals (including the name or URL of the signal file, storage format, number and type of signals, sampling frequency, calibration data, digitizer characteristics, record duration and starting time). Most records include one or more binary annotation files (in the example, .atr denotes an annotation file). Annotation files contain sets of labels (annotations), each of which describes a feature of one or more signals at a specified time in the record; 100.atr, for example, contains an annotation for each QRS complex (heart beat) in the recording, indicating its location (time of occurrence) and type (normal, ventricular ectopic, etc.), as well as other annotations that indicate changes in the predominant cardiac rhythm and in the signal quality. In other databases, annotations mark other features of the signals.

Since PhysioBank currently occupies about 220 gigabytes and is growing, some PhysioNet mirrors provide only a subset of PhysioBank, known as the PhysioBank Core Collection. PhysioBank has been designed so that visitors to these mirrors are redirected to the master PhysioNet server when following a link to a PhysioBank record outside of the core collection. Currently the PhysioBank Core Collection includes all of PhysioBank except for the most recently added databases and a few extremely large databases.

3. PhysioToolkit

PhysioToolkit provides software for exploration and study of PhysioBank. Among its components are algorithms for detection of physiologically significant events using both classical techniques and novel methods based on statistical physics and nonlinear dynamics, interactive display and characterization of signals, creation of new databases, simulation of physiologic and other signals, quantitative evaluation and comparison of analysis methods, and analysis of nonequilibrium and nonstationary processes. A unifying theme of many of the research projects that contribute software to PhysioToolkit is the extraction of “hidden” information from biomedical signals, information that may have diagnostic or prognostic value in medicine, or explanatory or predictive power in basic research. All PhysioToolkit software is available in source form under the GNU General Public License (GPL).

4. PhysioNet web site

The PhysioNet web site is a public service of the PhysioNet Resource, established as its mechanism for free and open dissemination and exchange of recorded biomedical signals and open-source software for analyzing them. It provides free electronic access to PhysioBank data and PhysioToolkit software, and facilities for cooperative analysis of data and evaluation of proposed new algorithms.

In addition, the PhysioNet web site offers service and training via on-line tutorials to assist users at entry and more advanced levels. PhysioNet tutorials provide hands-on introductions to the data and software available from PhysioBank and PhysioToolkit. Currently, over 20 tutorials discuss topics such as heart rate variability, nonlinear dynamics, multifractal time series, multiscale entropy, human gait dynamics, physiologic signal processing and analysis, “how-to” guides for using and customizing PhysioToolkit software, and even setting up a mirror of PhysioNet.

In cooperation with the annual Computers in Cardiology (CinC) conference, PhysioNet hosts a series of challenges, inviting participants to tackle clinically interesting problems that are either unsolved or not well-solved.

5. PhysioNet/CinC Challenges

These challenges attract researchers and students worldwide, who have the opportunity to take a few weeks or months to investigate significant open problems using a common set of resources shared by all participants. Since the typically expensive and time-consuming tasks related to collecting these resources have already been done, challenge participants can focus their efforts on the research questions.

An ideal challenge problem is interesting, clinically important, and possible to study using available materials that have not been widely circulated previously. Moreover, there must be an objective way to evaluate the quality of a challenge entry. For an analysis problem, this usually means there must be a known set of correct analyses of the data, i.e., a ”gold standard” against which entries can be compared. Past topics have included detecting sleep apnea from the ECG, predicting paroxysmal atrial fibrillation and its spontaneous termination, RR interval time series modeling, distinguishing ischemic from non-ischemic ST changes, measuring the QT interval, estimating the location and extent of infarcts from body surface potential maps, and detecting and quantifying T-wave alternans.

In complementary ways, PhysioNet and CinC catalyze and support scientific communication and collaboration between basic and clinical scientists. The annual meetings of CinC are gatherings of researchers from many nations and disciplines, bridging the geographic and specialty chasms that separate understanding from practice, while PhysioNet provides on-line data and software resources that support collaborations of basic and clinical researchers throughout the year. The annual PhysioNet/CinC Challenges seek to provide stimulating yet friendly competitions, while at the same time offering both specialists and non-specialists alike opportunities to make rapid progress on significant open problems whose solutions may be of profound clinical value. The use of shared data provided via PhysioNet makes it possible for participants to work independently toward a common objective. At CinC, participants can make meaningful results-based comparisons of their methods; lively and well-informed discussions are the norm at scientific sessions dedicated to these challenges. Discovery of the complementary strengths of diverse approaches to a problem when coupled with deep understanding of that problem frequently sparks new collaborations and opportunities for further study.

A new challenge topic is announced each year. For each challenge, we assemble the raw materials needed to begin work, and we post them on PhysioNet. In many of these challenges, these raw materials consisted of a database of signals to be analyzed; the analyses were provided for half of the data (the ”learning set”) in each case, and the challenge was to analyze the other half (the ”test set”).

Each challenge begins when the announcement is posted on PhysioNet, and ends in September just prior to the CinC conference. An important milestone for participants is the deadline for submitting abstracts for Computers in Cardiology, which is usually 1 May each year. Those wishing to qualify as official entrants, with eligibility for awards, must submit an abstract describing their work as well as an entry for scoring by this deadline. A limited number of revised entries may be submitted between 1 May and the final challenge deadline in mid-September. Challenges are open to all. After the final challenge deadline, we post the names of the top scorers, their scores, the number of entries they submitted in order to achieve their scores, and (for the official entrants) the abstracts they submitted to Computers in Cardiology in order to qualify.

By presenting these challenges, we aim to stimulate work on important clinical problems and to foster rapid progress towards their solution. Collaborations among those who have developed complementary approaches to challenge problems are easily established. We consider it especially significant that many of those who have participated in these challenges would not otherwise have had access to the data needed to study these topics. By bringing with them the insights and methods they have acquired from their own areas of expertise, these researchers enrich our fields of interest.

PhysioNet aims to promote the creation of high-quality data collections and open-source software that can support research on, among other subjects, the challenge topics. These challenges have been highly successful in advancing this goal. Not only have participants contributed the products of their individual efforts, but also it has been possible to combine their analyses of challenge data to produce better characterizations of those data, and to combine their algorithms to produce more robust and capable algorithms (for examples, see [4, 5]).

6. Invitation

All data included in PhysioBank, and all software included in PhysioToolkit, are carefully reviewed. We invite clinicians and biomedical researchers worldwide to participate in the ongoing review process. By sharing common data sets, and software in source form, the biomedical community benefits from access to materials that have been rigorously scrutinized by many investigators. We further invite researchers to contribute data and software for review and possible inclusion in PhysioBank and PhysioToolkit.

Acknowledgements

The author is grateful to Prof. Noriaki Ikeda of Kitasato University, to Prof. Roger Mark of MIT, and to the JSCE for the opportunity to present this work to the JSCE. PhysioNet is funded by the National Institute of Biomedical Imaging and Bioengineering and by the National Institute of General Medical Sciences, under NIH cooperative agreement U01-EB-008577.

References

[1] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000 (June 13);101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[2] Moody GB, Mark RG, Goldberger AL. PhysioNet: a web-based resource for the study of physiologic signals. IEEE Eng in Med and Biol 2001 (May-June);20(3):70–75.

[3] Moody GB, Mark RG. The MIT-BIH Arrhythmia Database on CD-ROM and software for use with it. Computers in Cardiology 1990;17:185–188.

[4] Moody GB, Koch H, Steinhoff U. The PhysioNet/Computers in Cardiology Challenge 2006: QT interval measurement. Computers in Cardiology 2006;33:829–832. CinC on-line: http://cinc.mit.edu/archives/2006/pdf/0313.pdf.

[5] Penzel T, McNames J, de Chazal P, Raymond B, Murray A, Moody G. Systematic comparison of different algorithms for apnoea detection based on electrocardiographic recordings. Medical and Biological Engineering and Computing 2002; 40:402–407.

Address for correspondence:

George B. Moody
MIT E25-505A, Cambridge, MA 02139 USA
george@mit.edu