PhysioNet: A Research Resource for Studies of Complex Physiologic and Biomedical Signals

GB Moody, RG Mark, AL Goldberger

Harvard-M.I.T. Division of Health Sciences and Technology, Cambridge, MA, USA
Beth Israel Deaconess Medical Center, Boston, MA, USA

Abstract

PhysioNet (http://www.physionet.org/) is a web-based resource supplying well-characterized physiologic signals and related open-source software to the biomedical research community. Inaugurated in September 1999 under the auspices of the NIH's National Center for Research Resources (NCRR), PhysioNet provides an on-line forum for free dissemination and exchange of research data and software, with facilities for cooperative analysis of data and evaluation of new analytic methods. As of September 2000, PhysioBank, the data archive made available via PhysioNet, contained roughly 35 gigabytes of recorded signals and annotations. PhysioNet is a public service of the Research Resource for Complex Physiologic Signals, a cooperative project initiated by researchers at Boston's Beth Israel Deaconess Medical Center/Harvard Medical School, Boston University, McGill University, and MIT.

1. Introduction

The recently-established Research Resource for Complex Physiologic Signals is intended to stimulate current research and new investigations in the study of complex biomedical signals. PhysioNet is the major public outreach component of this NIH NCRR-sponsored resource. PhysioNet is a web-based forum for dissemination and exchange of recorded biomedical signals and open-source software for analyzing them, by providing facilities to support wide-ranging collaborations among researchers, including development of new analytic methods and reference databases. PhysioNet provides on-line tutorials to orient new users and to introduce new algorithms and data sets to experienced users.

2. Origins and objectives of PhysioNet

Our interest in automated arrhythmia detection motivated us to create the MIT-BIH Arrhythmia Database between 1975 and 1980, and as we have continued to pursue related topics, we have collected many more databases of physiologic signals. More than simple collections of signals, many of these databases have been meticulously annotated (for example, the MIT-BIH Arrhythmia Database contains about 110,000 beat, rhythm, and signal quality annotations, created by at least two expert annotators working independently with review by a third to resolve discrepancies). Recognizing their value and importance to other researchers, we have made these databases available to the research community since 1980 [1,2]. Sharing data in this way is desirable for several important reasons:

- Careful review by many users, each motivated to correct any errors that may remain, permits rapid discovery of any such errors, and eventually high confidence in the accuracy and completeness of the annotations.
- The availability of well-chosen and well-characterized data lowers the barrier to research, permitting investigators to conduct many types of exploratory studies at very low cost. Limited funds can be invested in pursuing new ideas rather than in redundant data collection efforts.
- Using a data set that is already well-known reduces the burden on the researcher to demonstrate that his or her conclusions are based on good data.
- Publishing the data that support a paper invites the most rigorous peer review, and conclusions that are supported by such a review acquire an extra measure of credibility.
- When several publications refer to the same data set, the relative merits of differing analytic methods can be assessed. Within the narrower field of medical device evaluation, reference databases provide the essential tools for obtaining reproducible performance measurements that can be compared. Such comparisons motivate developers to improve their devices' performance and thus stimulate advances in the state of the art.

Since successful use of these databases requires specialized software, we also distributed portable software toolkits that contained the components needed for researchers to read and manipulate the signals and annotations. The principal benefit of sharing software in this way is becoming widely appreciated: particularly when dealing with complex algorithms, the open-source model (i.e., peer review) is the only way to acquire confidence that an implementation of such an algorithm is correct (or usable).

0276-6547/00 $10.00 © 2000 IEEE

Computers in Cardiology 2000;27:179-182.

PhysioNet was established with NIH support to provide free access via the web to these databases (PhysioBank) and software (PhysioToolkit). These materials require rigorous review so that researchers can use of them with confidence. We assign each database, and each software component, to one of three classes to indicate to what extent further review may be necessary. As stated on the PhysioNet web site:

Class 1 databases and software are fully supported. Class 1 databases have been carefully scrutinized and have been thoroughly annotated. Class 1 software has been extensively and rigorously tested. We will correct and document any remaining errors, and encourage users to bring these to our attention.

Class 2 databases and software are archival copies of materials that support published research, contributed by authors or journals. We will maintain copies of the original data and software together with corrections submitted by the authors. We encourage users to report errors directly to the authors.

Class 3 databases include collections of data that may have been less thoroughly studied than those in class 1, but that may be of interest to the research community. These databases include works in progress, to which users are invited to contribute. In some cases, these databases may be archived on their creators' web sites. Class 3 software includes code that may need further testing or development; again, users are invited to dig in and help their creators transform these works in progress into robust and useful tools for research.

3. PhysioBank

As of September, 2000, about 35 gigabytes of freely available digitized physiologic signals, most with annotations, have been organized into 21 major collections (databases) within PhysioBank. Our collections have been supplemented by generous contributions from colleagues (see table 1), and we welcome further contributions.

4. PhysioToolkit

PhysioNet's open-source software archive, PhysioToolkit, is based on the WFDB (WaveForm DataBase) software package written in C and portable between Linux, Unix, and MS-Windows. The package includes applications for sampling frequency conversion, digital filtering, signal averaging, power spectral density estimation, QRS detection, graphical annotation editing (currently under Linux and Solaris only), printing high-quality "chart recordings", and much more. PhysioToolkit also includes peer-reviewed implementations of new techniques for time series analysis, simulation, modeling, and digital signal processing, such as software for deriving a respiration signal from the ECG [22,23], and for detrended fluctuation analysis [24,25].

Most of this software is built on a common foundation (the WFDB library), which supplies methods for reading and writing signals and annotations in many formats, and can be linked to user-written applications in C, C++, or Fortran. The library can read input directly from remote web servers as well as local files.

5. PhysioNet infrastructure

All of the software used to support PhysioNet is freely available and open-source, just as are the PhysioToolkit components. PhysioNet is indexed and searchable using a customized version of ht://Dig (http://www.htdig.org/). Application services are provided using PhysioToolkit software and additional custom CGI scripts.

PhysioNet servers run the Apache web server under Linux, on 200 MHz or faster x86-compatible PCs, typically connected to the Internet with T1 (10 Mbps) or faster connections. During PhysioNet's first year, the PhysioNet servers at MIT provided over 160 Gb of data to visitors from over 15,000 locations in 90 countries.

PhysioNet's master web servers are located at MIT in Cambridge, Massachusetts. Mirrors are located in Italy, Israel, Spain, Slovenia, Taiwan, China, and elsewhere in the US (see http://www.physionet.org/mirrors/ for a complete list). To avoid overloading the master server, mirroring is performed using rsync (http://rsync.samba.org/). The mirror sites are operated by volunteers; they provide improved access especially in areas where connections to the US are often slow. Additional mirror sites are welcome. The total cost of a mirror can be under US$1000, since the hardware consists of standard commodity components and the software is free.

6. Future directions

We invite the research community to contribute to PhysioNet, by participating in the ongoing review process, and by sharing common data sets and software in source form for peer review and possible inclusion in PhysioBank and PhysioToolkit. PhysioNet has received a number of contributions of physiologic signals, time series data, and software from published studies.

As demonstrated by the Computers in Cardiology Challenge 2000, the the unique aspects of collaboration among widely-distributed researchers using the same data or software can enable rapid progress on well-defined problems.

Acknowledgement

This work was supported by a grant from the National Center for Research Resources of the National Institutes of Health (P41 RR13622).

Database	*Class*	*Duration*	*Contents*
MIT-BIH Polysomnographic DB [3]	3	80	Sleep studies (EEG, ECG, BP, respiration, etc.)
MIMIC DB * [4]	3	4658	Intensive care monitor recordings (72 of 121 records complete; remainder will be completed during 2001)
Apnea-ECG DB [5]	3	560	ECG (some with respiration, SaO_2); contributed by T Penzel, initially for the CinC Challenge 2000 [6]
MIT-BIH Arrhythmia DB * [1,2]	1	14	ECGs with beat, rhythm, and signal quality annotations (available selections include 25 complete records, first 10 minutes of all others, complete annotation files) hours)
MIT-BIH Noise Stress Test DB [7]	1	7	ECGs from the MIT-BIH Arrhythmia DB with calibrated amounts of noise, for assessing noise tolerance of arrhythmia detectors
European ST-T DB * [8]	1	96	ECGs with beat, rhythm, signal quality, and ST and T change annotations; 48 complete records, and reference annotation files for all 90 records, contributed by the European Society of Cardiology and the creators of the database
QT DB [9]	2	26	ECGs with manual and algorithm-derived waveform boundary (P, QRS, T, U) annotations; contributed by P Laguna
Post-ictal heart rate oscillations in partial epilepsy [10]	2	16	ECGs prior to and during seizures
Exaggerated heart rate oscillations during two meditation techniques [11]	2	220	RR interval series only (no ECGs); includes additional data from spontaneously and metronomically breathing controls, and from highly trained athletes
BIDMC Congestive Heart Failure DB [12]	2	300	ECGs from patients with severe (NYHA class 3-4) heart failure
Fantasia DB [13]	2	20	RR interval series only; rigorously-screened healthy young and elderly volunteers in normal sinus rhythm
MIT-BIH Malignant Ventricular Arrhythmia DB [14]	3	11	ECGs containing sustained ventricular tachycardia, flutter, and fibrillation
Creighton University Ventricular Tachyarrhythmia DB	3	5	Digitally recorded ECGs showing onset of ventricular fibrillation, collected by FM Nolle
MIT-BIH Supraventricular Arrhythmia DB [15]	3	39	Supplement to the MIT-BIH Arrhythmia DB
MIT-BIH Atrial Fibrillation DB (250 hours) [16]	3	250	RR intervals only (one ten-hour record also includes ECGs)
MIT-BIH ST Change DB [17]	3	25	ECGs during exercise stress tests
MIT-BIH Normal Sinus Rhythm DB	3	380	20-24 hour recordings from subjects (ages 20 to 50) without diagnosed heart disease
MIT-BIH ECG Compression Test DB [18]	3	1	168 short recordings for testing compression algorithms
MIT-BIH Long-Term DB	3	150	Seven 14- to 24-hour ECGs
Gait Maturation DB [19]	2	5	Stride intervals in healthy children ages 3-14
Gait in Aging and Disease DB [20, 21]	2	3	Stride intervals in healthy young and elderly subjects, and in subjects with Parkinson's disease.

Table 1. Summary of the contents of PhysioBank as of September 2000 (DB: Database). The database classes are described in the text; durations are totals for all available records in each database, in hours. Asterisks indicate incomplete databases; all others are complete.

References

[1] Mark RG, Schluter PS, Moody GB, Devlin P, Chernoff D. An annotated ECG database for evaluating arrhythmia detectors. In Frontiers of Engineering in Health Care 1982; Proc. 4th Ann. Conf. IEEE EMBS. Long Beach, California: IEEE Computer Society Press, 1982; 205–210.

[2] Moody GB, Mark RG. The MIT-BIH arrhythmia database on CD-ROM and software for use with it. In Computers in Cardiology 1990. Los Alamitos, California: IEEE Computer Society Press, 1991; 185–188.

[3] Ichimaru Y, Moody GB. Development of the polysomnographic database on CD-ROM. Psychiatry and Clinical Neurosciences 1999 (April);53(2):175–177.

[4] Moody GB, Mark RG. A database to support development and evaluation of intelligent intensive care monitoring. In Computers in Cardiology 1996. Piscataway, New Jersey: IEEE Press, 1996; 657–660. [http://www.physionet.org/physiobank/database/mimicdb/mimic-cic96/.].

[5] Penzel T, Moody GB, Mark RG, Goldberger AL, Peter JH. The apnea-ecg database. In Computers in Cardiology 2000. Piscataway, New Jersey: IEEE Press; 2000. [this volume].

[6] Moody GB, Mark RG, Goldberger AL. Stimulating rapid research advances via focused competition: The Computers in Cardiology Challenge 2000. In Computers in Cardiology 2000; 2000. [this volume].

[7] Moody GB, Muldrow WK, Mark RG. A noise stress test for arrhythmia detectors. In Computers in Cardiology 1984. Long Beach, California: IEEE Computer Society Press, 1985; 381–384.

[8] Taddei A, Benassi A, Biagini A, Bongiorni MG, Contini C, Distante G, Landucci L, Mazzei MG, Pisani P, Roggero N, Varanini M, Marchesi C. ST-T change analysis in ECG ambulatory monitoring: a European standard for performance evaluation. In Computers in Cardiology 1987. Long Beach, California: IEEE Computer Society Press, 1988; 63–68.

[9] Laguna P, Mark RG, Goldberger AL, Moody GB. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In Computers in Cardiology 1997. Piscataway, New Jersey: IEEE Press, 1997; 673–676. [http://www.physionet.org/physiobank/database/qtdb/doc/].

[10] Al-Aweel I, Krishnamurthy KB, Hausdorff JM, Mietus JE, Ives JR, Blum AS, Schomer DL, Goldberger AL. Post-ictal heart rate oscillations in partial epilepsy. Neurology 1999 (October 22);53(7):1590–1592. [http://www.physionet.org/physiobank/database/szdb/szdb.shtml].

[11] Peng CK, Mietus JE, Liu Y, Khalsa G, Douglas PS, Benson H, Goldberger AL. Exaggerated heart rate oscillations during two meditation techniques. International Journal of Cardiology 1999;70:101–107. [http://www.physionet.org/physiobank/database/meditation/].

[12] Baim DS, Colucci WS, Monrad ES, Smith HS, Wright RF, Lanoue A, Gauthier DF, Ransil BJ, Grossman W, Braunwald E. Survival of patients with severe congestive heart failure treated with oral milrinone. J American College of Cardiology 1986 (March);7(3):661–670.

[13] Iyengar N, Peng CK, Morin R, Goldberger AL, Lipsitz LA. Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. American Journal of Physiology 1996;271:1078–1084.

[14] Greenwald S, Albrecht P, Moody GB, Mark RG. Estimating confidence limits for arrhythmia detector performance. In Computers in Cardiology 1985. Washington, DC: IEEE Computer Society Press, 1986; 383–386.

[15] Greenwald SD. Improved Detection and Classification of Arrhythmias in Noise-Corrupted Electrocardiograms Using Contextual Information. Ph.D. thesis, Harvard-MIT Division of Health Sciences and Technology, 1990.

[16] Moody GB, Mark RG. A new method for detecting atrial fibrillation using RR intervals. In Computers in Cardiology 1983. Long Beach, California: IEEE Computer Society Press, 1984; 227–230.

[17] Albrecht P. S-T Segment Characterization for Long-Term Automated ECG Analysis. Master's thesis, MIT Dept. of Electrical Engineering and Computer Science, 1983.

[18] Moody GB, Mark RG, Goldberger AL. Evaluation of the "TRIM" ECG data compressor. In Computers in Cardiology 1988. Los Alamitos, California: IEEE Computer Society Press, 1989; 167–170.

[19] Hausdorff JM, Zemany L, Peng CK, Goldberger A. Maturation of gait dynamics: Stride-to-stride variability and its temporal organization in children. Journal of Applied Physiology 1999;86:1040–1047. [http://www.physionet.org/physiobank/database/gait-maturation-db/jap/].

[20] Hausdorff J, Purdon PL, Peng CK, Ladin Z, Wei JY, Goldberger AL. Fractal dynamics of human gait: Stability of long-range correlations in stride interval fluctuations. Journal of Applied Physiology 1996;80:1448–1457.

[21] Hausdorff JM, Cudkowicz ME, Firtion R, Wei JY, Goldberger AL. Gait variability and basal ganglia disorders: Stride-to-stride variations of gait cycle timing in Parkinson's and Huntington's disease. Movement Disorders 1998; 13:428–437.

[22] Moody GB, Mark RG, Zoccola A, Mantero S. Derivation of respiratory signals from multi-lead ECGs. In Computers in Cardiology 1985. Washington, DC: IEEE Computer Society Press, 1986; 113–116. [http://www.physionet.org/physiotools/edr/cic85/].

[23] Moody GB, Mark RG, Bump MA, Weinstein JS, Berman AD, Mietus JE, Goldberger AL. Clinical validation of the ECG-derived respiration (EDR) technique. In Computers in Cardiology 1986. Washington, DC: IEEE Computer Society Press, 1987; 507–510. [http://www.physionet.org/physiotools/edr/cic86/].

[24] Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. On the mosaic organization of DNA sequences. Physical Review E 1994;49:1685–1689.

[25] Peng CK, Havlin S, Stanley HE, Goldberger AL. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 1995;5:82–87.

Address for correspondence:

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