PhysioNet: An NIH Research Resource for Physiologic Datasets and Open-Source Software

IC Henry
Beth Israel Deaconess Medical Center
ihenry@mit.edu
AL Goldberger
Beth Israel Deaconess Medical Center
agoldber@caregroup.harvard.edu

GB Moody
Massachusetts Institute of Technology
george@mit.edu
RG Mark
Massachusetts Institute of Technology
rgmark@mit.edu

Abstract

PhysioNet (http://www.physionet.org/) is an Internet resource supplying well-characterized physiologic datasets and related open-source software. PhysioNet maintains a growing collection of multi-parameter 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. The data is supported by a collection of open-source software for biomedical signal processing and data management. PhysioNet is also an online forum of discussions and articles pertaining to the posted datasets and software. The utility of PhysioNet was demonstrated by the Sleep Apnea Challenge, which resulted in several new algorithms for the previously unsolved problem of detecting obstructive sleep apnea from the electrocardiogram.

1. Introduction

PhysioNet, with support from the NIH's National Center for Research Resources (NCRR), offers an online forum for free exchange of data and software, with facilities for cooperative analysis of data and analytic methods. PhysioNet uses the open-source software movement as a model for Internet-based collaboration in the biomedical science community [1]. The challenge is to adapt Internet technology to achieve large-scale, distributed, cooperation in three areas:

- Creation and distribution of high-quality datasets, and repositories for data used in peer-reviewed publications.
- Organization of these datasets into searchable databases, to facilitate data-mining applications.
- Development of high quality open-source software for the scientific community. Research in data analysis and modeling techniques is aided by the ability to share source-code easily.

2. The Open-Source Model for Dataset Creation and Publication

Under the open-source model, programmers use the Internet to collaborate on software supporting open standards, making the source code for their applications publicly available. The result has been the surprisingly rapid creation of high quality software, both free and

0-7695-1004-3/01 $10.00 © 2001 IEEE

commercial [2]. The scientific community stands to benefit from the extension of open-source techniques to dataset creation. The challenge is to merge the power and energy of open-source software development techniques with the quality and discriminating analysis of the scientific peer-review process. PhysioNet aims to act as a collaborative center for the biomedical community. The web site and its products have value to the biomedical research community because active members of the community develop them.

2.1 Benefits of open-source software and published data

The availability of well-characterized physiologic data sets and open source software is a primary objective of PhysioNet. More than simple collections of signals, the data sets on PhysioNet have been carefully annotated and reviewed for quality. Sharing well-described data sets and open-source software is desirable for several reasons:

- Careful review by many users permits rapid discovery of errors, and eventually high confidence in the accuracy and completeness of a data set or piece of software.
- The availability of open-source software and well-chosen and well-characterized data lower the barriers to research, permitting investigators to conduct many types of exploratory studies at low cost.
- Using a data set or piece of software that is already well known reduces the burden on the researcher to demonstrate that his or her conclusions are based on good data and methods. Within size constraints of typical journal articles, this means that more detailed, hence more useful, expositions of innovative methods and new results become possible.
- Publishing the data and software that supports 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. Published algorithms can be more readily assessed and compared when software implementations are made available. Such comparisons motivate developers to improve their devices' performance and thus stimulate advances in the state of the art.

2.2 Comparison to existing resources

The biomedical communities have been early adopters of Internet technology, spawning numerous online resources, the two of the most widely used of which are GenBank and PubMed. PhysioNet draws inspiration from both these resources and the open-source model to provide a more complete solution to scientific publishing.

Genbank [3,4], the NIH genetic sequence database, has proven the efficacy and power of open source models in dataset creation. However, Genbank is limited to one branch of biological science and a specific type of data, DNA sequences; consequently Genbank was able to enforce data formats and software requirements on the community. PhysioNet seeks to extend the Genbank paradigm by hosting a wide variety of physiologic data, but without attempting to impose standards on the research community. A single data format or collection of software is unlikely to meet the needs of the majority of researchers. PhysioNet requires only that data be published in a well-described data format, and that the source code

for accompanying software be provided. This permissive standard requires additional features not found in GenBank, but resemble online journals.

The online publication database PubMed [5] is a widely used resource for medical science research. PubMed allows investigators to quickly find citations for a given topic. Many articles indexed in PubMed reference data published on GenBank, but many more are based on data or software that is unavailable. Clearly there is a gap between GenBank's model for publishing data and PubMed's online citation index. PhysioNet proposes to fill the gap by using an open source model to provide a place to post data, software and the information necessary to use them. Therefore it is not necessary to enforce strict data formatting standards. While, this could result in an undesirable proliferation of formats, with proper management, experience shows that open standards will evolve naturally. PhysioNet facilitates the process by providing a forum for community discussion and a set of open standards on which to build.

3. The Structure of PhysioNet

All software used to support PhysioNet is freely available and open-source. PhysioNet runs on the Apache web server under Linux. The ma ster server is located at MIT, with several mirror sites around the world. Using only open-source software to build the resource ensures equal and unfettered access for all researchers. The PhysioNet web site has three components:

3.1. PhysioNet

PhysioNet is an online community providing discussion groups, tutorials, and articles from peer-reviewed journals. PhysioNet is virtual laboratory for collaborative work leading to the creation of new algorithms and databases. Researchers use PhysioNet as a forum to discuss and develop datasets and software. There is a growing collection of peer-reviewed articles with accompanying data [8..15] or software [16..18]. In addition, the tutorial section provides hands-on introductions to the data and software on the resource.

3.2. PhysioBank

PhysioBank is a large and growing archive of well-characterized digital recordings of physiologic signals and related data for use by the biomedical research community. PhysioBank currently includes datasets of multi-parameter 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 databases will grow in size and scope, and will eventually include signals from selected in vitro and in vivo experiments, as developed and contributed by members of the research community.

Building a large repository of data is only the first step, to provide an efficient way of locating data that may be relevant to a researcher's interests, a searchable index of PhysioBank is being developed. The index is a relational database built with open-source PostgreSQL technology. The index has been designed to be easily extensible, so that new data types can be readily incorporated. A web-based interface allows users who are not relational database experts not only to search current collections but also to index new datasets.

3.3. PhysioToolkit

PhysioToolkit is a large and growing library of software for physiologic signal processing and analysis, and detection of physiologically significant events. A unifying theme 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). The core of PhysioToolkit is the WFDB ANSI C-language library, which provides all functions necessary for creating applications for reading and analyzing PhysioBank data. The software has been designed to run on many platforms and is available in binary form for Unix systems and MS-Windows. PhysioToolkit also includes a growing collection of software contributed to PhysioNet.

4. The Sleep Apnea Challenge

One of the primary goals of PhysioNet is to stimulate advances in knowledge by encouraging the examination of data and algorithms by researchers other than their creators. New insights can be obtained by independent studies of a common data set or algorithm using a variety of approaches, and rapid progress is possible when such studies complement and support each other. The benefits are well illustrated by an experiment hosted by PhysioNet between February and September 2000, in which researchers were challenged to design and evaluate methods for detecting obstructive sleep apnea from the electrocardiogram (ECG) [19]. Sleep apnea (intermittent cessation of breathing) is a serious condition affecting millions of people; diagnoses of sleep apnea, and evaluation of therapy for it, typically require one or more nights in a sleep laboratory, an intrusive and expensive process. Since long-term ambulatory ECG recording of outpatients is a far more widely available and much less expensive technique for physiologic monitoring, a reliable method for detecting apnea based on the ECG alone may offer significant benefits. Although such methods have been known since the mid-1980s, clinical acceptance of these approaches depends on careful quantitative comparisons of their accuracy with that obtained from conventional analysis of polysomnographic recordings made in a sleep laboratory.

To spur interest in the topic, PhysioNet sponsored an apnea detection competition (http://www.physionet.org/cinc-challenge-2000.shtml). A database consisting of ECGs extracted from conventional polysomnograms of subjects with and without obstructive sleep apnea was made available via PhysioNet [20]. A learning set, consisting of half of the ECG recordings together with samples of the respiration signals for several recordings, was accompanied by detailed reference annotations of the apneas based on expert analysis of the full polysomnograms. Entrants produced their own apnea annotations for the remaining recordings, and their entries were scored against the (unpublished) reference annotations for this group. Participants implemented a variety of techniques to compete in two events: screening, in which the goal was to identify which recordings came from subjects with clinically significant obstructive sleep apnea, and quantification, in which the goal was to identify which one-minute periods in each case coincided with episodes of apnea. Fifteen teams from nine countries participated, and thirteen of these presented their methods and results at Computers in Cardiology 2000 (see the proceedings of this conference, and http://www.physionet.org/cinc-top-scores.shtml for details of the results).

Notably, most participants were able to obtain scores of 90% or better in the screening event, and four obtained perfect scores in this event, demonstrating that a variety of approaches can be employed for reliable apnea screening based on the ECG. Remarkably, top scores in the quantification event exceeded 92%, comparable to independent estimates of human inter-observer concurrence of around 90%, suggesting that even for the much more difficult task of identifying individual apnea episodes, ECG-based methods can perform as reliably as conventional analysis of full polysomnograms. Although the challenge has ended, several teams are now collaborating on a new algorithm that combines the strengths of their independently developed methods to achieve even better results. The success of the apnea detection challenge has encouraged us to propose another: that of predicting paroxysmal atrial fibrillation (for background on this problem, and information for entrants, see http://www.physionet.org/cinc-challenge-2001.shtml).

Serious work on these challenges, or on any number of similarly interesting and clinically significant problems, requires resources that are typically unavailable to non-specialist researchers. The process of gathering the data needed to work on such problems, and to characterize these data, can occupy years of effort. Resources such as GenBank or PhysioNet permit researchers to make important contributions in subjects where access to data has been a barrier to entry in the past, and to do so in a matter of months rather than years. By leveraging the opportunities presented by PhysioNet, we seek to focus attention on challenging and clinically important problems, and to foster rapid progress towards their solution.

5. Conclusion

The open source movement, born from traditions of sharing and peer review in the academic research environment, has taught us much about how research can benefit from distributed open development. Ubiquitous Internet access makes it possible for researchers to share their data and algorithms as never before; as a community, we are still learning the profound cultural implications of ``open science.'' We invite the research community to make use of PhysioNet not only as a source, but also as a repository, of data and algorithms. The contributions of many researchers around the world are making PhysioNet (http://www.physionet.org) a unique and valuable laboratory without walls.

Acknowledgement

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

[1] Moody GB, Mark RG, Goldberger AL. PhysioNet: A Research Resource for Studies of Complex Physiologic and Biomedical Signals. Computers In Cardiology 2000;27:179-182

[2] Perens P. The Open Source Definition. In *Open Sources*, Dibona C, ed. Oreilly Press, Sebastop, 1999.

[3] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. Genbank. Nucleic Acids Res. 2000;28(1):15-18.

[4] McEntyre J, Lipman DJ. GenBank: a model community resource. Nature Web Debates, [http://www.nature.com/nature/debates/e-access/Articles/lipman.html] 6 May 2001.

[5] Roberts RJ. PubMed Central: The GenBank of the published literature. Proc. Natl. Acad. Sci. 2001;98(2):381-382.

[8] Al-Aweel IC, Krishnamurthy KB, Hausdorff JM, Mietus JE, Ives JR, Blum AS, Schomer DL, and Goldberger AL. Post-ictal heart rate oscillations in partial epilepsy. Neurology 1999;53(7):1590-1592.

[9] Hausdorff JM, Zemany L, Peng C-K, and Goldberger AL. Maturation of gait dynamics: stride-to-stride variability and its temporal organization in children. Journal of Applied Physiology 1999;86:1040-1047.

[10] Laguna P, Mark RG, Goldberger AL, and Moody GB. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. Computers in Cardiology 1997;24:673-676.

[11] Jané R, Blasi A, Garcia J, and Laguna P. Evaluation of an automatic threshold based detector of waveform limits in Holter ECG with the QT database. Computers in Cardiology 1997;24:295 -298.

[12] Moody GB and Mark RG. A database to support development and evaluation of intelligent intensive care monitoring. Computers in Cardiology 1996;23:657-660.

[13] Peng C-K, Mietus JE, Liu Y, Khalsa G, Douglas PS, Benson H, and Goldberger AL. Exaggerated heart rate oscillations during two meditation techniques. International Journal of Cardiology 1999;70:101-107.

[14] Jager F, Moody GB, Taddei A, Antolic G, Zabukovec M, Skrjanc M, Emdin M, and Mark RG. Development of a long-term database for assessing the performance of transient ischemia detectors. Computers in Cardiology 1996;23:481-484.

[15] Jager F, Taddei A, Emdin M, Antolic G, Dorn R, Moody GB, Glavic B, Smrdel A, Varanini M, Zabukovec M, Bordigiao S, Marchesi C, and Mark RG. The Long-Term ST Database: a research resource for algorithm development and physiologic studies of transient myocardial ischemia. Computers in Cardiology 2000;27:841-848.

[16] Moody GB, Mark RG, Zoccola A, and Mantero S. Derivation of respiratory signals from multi-lead ECGs. Computers in Cardiology 1985;12:113-116.

[17] Moody GB, Mark RG, Bump MA, Weinstein JS, Berman AD, Mietus JE, and Goldberger AL. Validation of the ECG-derived respiration (EDR) technique. Computers in Cardiology 1986;13:507-510.

[18] Rosenstein MT, Collins JJ, and De Luca, CJ. A practical method for calculating largest Lyapunov exponents from small data sets. Physica D 1993;65:117-134.

[19] Moody GB, Mark RG, Goldberger AL, Peter JH. Stimulating rapid research advances via focused competition: The Computers and Cardiology Challenge 2000. Computers In Cardiology 2000;27:207-209.

[20] Penzel T, Moody GB, Mark RG, Goldberger AL, Peter JH. The Apnea-ECG Database. Computers In Cardiology 2000;27:255-258.