THE MIT-BIH ARRHYTHMIA DATABASE ON CD-ROM AND SOFTWARE FOR USE WITH IT

George B. Moody and Roger G. Mark

Massachusetts Institute of Technology, Cambridge, MA, USA

Summary

In August 1989 we produced a CD-ROM containing the original MIT-BIH Arrhythmia Database (developed between 1975 and 1979, and first released in 1980), as well as a large number of supplementary recordings assembled for various research projects between 1981 and 1989. In all, the CD-ROM contains approximately 600 megabytes of digitized ECG recordings, most with beat-by-beat annotations, having a total duration in excess of 200 hours. The CD-ROM format makes this substantial collection of ECGs accessible to researchers with PCs as well as those with larger computer systems.

We describe the contents of the CD-ROM and the issues involved in its production. We also describe software for use with the CD-ROM as well as for development of similar databases.

Background

In the long-term electrocardiogram, the diversity of waveform morphology and underlying cardiac rhythm, and the confounding influence of noise and artifact, conspire to make analysis of the signal difficult, and therefore interesting. Designs for automated arrhythmia detectors need to take this diversity of input into account. Users of arrhythmia detectors should be aware that the accuracy of detector outputs varies with characteristics of the input.

In the mid-1970s, it became apparent that increasingly ambitious designs for automated arrhythmia detectors could be evaluated and compared objectively only on the basis of their performance in standardized tests. During that period, our group at MIT and at Boston's Beth Israel Hospital developed the MIT-BIH Arrhythmia Database,[1] while K.L. Ripley, G.C. Oliver, and their colleagues at Washington University and many cooperating institutions developed the American Heart Association Database for the Evaluation of Ventricular Arrhythmia Detectors.[2]

It is fortunate in retrospect that the two groups which were developing long-term ECG databases did not share identical goals, because the products of their efforts largely complement rather than duplicate each other. Both databases are comprised of digitized excerpts from two-lead Holter ECG recordings, and each includes beat annotations (i.e., for each QRS complex, there is a computer-readable label indicating the beat type and its location within the recording). The recording formats are compatible.

The charter for the AHA Database emphasized ventricular ectopy, and to that end its developers sought to obtain recordings which fit one of eight sets of stringent selection criteria, each set defined in terms of the severity of ventricular ectopy. The result is a well-structured database with particularly strong representation of the most clinically important ventricular arrhythmias, some of which are rarely seen in Holter recordings. Overall, the signal quality is excellent, though a few recordings have moderate amounts of noise.

The MIT-BIH Arrhythmia Database

For the MIT-BIH Arrhythmia Database, we attempted to obtain a representative sample of the variety of recordings which are observed in clinical practice. This database consists of 48 annotated records, obtained from 47 subjects studied by the Arrhythmia Laboratory of Beth Israel Hospital in Boston between 1975 and 1979. About 60% of the records were obtained from inpatients. The database contains 23 records (the '100 series') chosen at random from a set of over 4000 24-hour Holter tapes, and 25 records (the '200 series') selected from the same set to include a variety of rare but clinically important phenomena which would not be well-represented by a small random sample. Several records in the 200 series were chosen specifically because features of the rhythm, QRS morphology, or signal quality may be expected to present significant difficulty to arrhythmia detectors. These records have gained considerable notoriety among users of the database. In the MIT-BIH Arrhythmia Database, there is considerable variation in signal quality, with significant portions of unreadable data in at least one lead.

Our observations, confirmed by those of many other users of both databases, suggest that arrhythmia detectors consistently perform better using the AHA Database rather than the MIT-BIH Arrhythmia Database, independent of which (if either) was used for training. Weaknesses in detector design are often exposed by the multiple challenges posed by the complexity of the rhythms, the beat morphologies, and the noise in the MIT-BIH Arrhythmia Database.

Although the MIT-BIH Arrhythmia Database contains many examples of ventricular ectopy, the AHA Database includes a more comprehensive selection of the most severe levels. On the other hand, the AHA Database includes very few examples of supraventricular ectopy or conduction abnor-

0276-6574/91/0000/0185$01.00 © 1991 IEEE

malities, none annotated. The MIT-BIH Arrhythmia Database includes many fully-annotated examples of these phenomena, which are important not only per se, but also as examples of signals which often provoke spurious detection of ventricular arrhythmias.

Each record in the MIT-BIH Arrhythmia Database is slightly over 30 minutes in length. Each signal file contains two signals sampled at 360 Hz. 'Header' files include information about the leads used, the patient's age, sex, and medications. The reference annotation files include beat, rhythm, and signal quality annotations. Each of the roughly 109,000 beats was manually annotated by at least two cardiologists working independently; their annotations were compared, consensus on disagreements was obtained, and the reference annotation files were prepared.

Four records (102, 104, 107, and 217) include paced beats. The original analog tapes do not represent the pacemaker artifacts with sufficient fidelity to permit them to be recognized by pulse amplitude (or slew rate) and duration alone, the method commonly used for real-time processing. The database records reproduce the analog recordings with sufficient fidelity to permit use of pacemaker artifact detectors designed for tape analysis, however.

Between 1980 and mid-1989, we distributed the MIT-BIH Arrhythmia Database to approximately 100 sites worldwide. Most of these received copies on digital 9-track tape. With the assistance of our colleagues at Washington University, in particular R. Hermes, we were able to devise a tape format compatible with that used for the AHA Database. A small number of sites requested and received copies on 4-channel IRIG-format FM analog tape (with annotations and time markers digitally encoded on the third and fourth channels). Since the initial release of the database in 1980, sixteen errors in beat annotations have been discovered and corrected. No such errors have been found since 1987.

We have continued to collect Holter recordings; as of mid-1990, the BIH Arrhythmia Laboratory's tape library had grown to approximately 17,000 recordings. Over the years, we received many requests for additional recordings to supplement those in the MIT-BIH Arrhythmia Database. Although we were often able to fulfill these requests, it was (and remains) difficult to do so, given the resources of a research laboratory. When CD-ROM first appeared in the mid-1980s, it appeared to be promising as a distribution medium for large databases such as ours. The capacity of a CD-ROM is roughly 680 megabytes; the MIT-BIH Arrhythmia Database occupies about 120 megabytes. The possibility of putting this substantial amount of data on a physically robust and compact medium was attractive.

Two significant obstacles to using CD-ROM were the lack of file-level format standards and the high development costs (initially estimated at US$50,000 or more). Within a few years, however, both of these obstacles disappeared. The adoption of the High Sierra format standard (subsequently the basis for the upwardly-compatible ISO 9660 standard) and the widespread availability of inexpensive CD-ROM drives (currently $500 to $1000 for PCs, usually somewhat more for UNIX workstations and other systems) made it safe to assume that others would be able to use a disk if we produced one. The production costs dropped rapidly, and for a database such as ours without the need for extensive indexing (most use of the records is sequential) it is now possible to produce 100 copies of a CD-ROM for about $2500. Almost all of this cost is for preparation of the disk master; additional copies of the disk are about $2 each.

Once we had decided to produce a CD-ROM, we took the opportunity to include many additional recordings on the disk. In addition to the original MIT-BIH Arrhythmia Database, the CD-ROM contains seven additional databases, mostly developed to support work previously reported in these pages. These are briefly described below.

Noise stress test database

This database consists of 15 thirty-minute records. Three of these (records 'bw', 'em', and 'ma') contain noise of the types typically observed in ECG recordings. They were obtained using a Holter recorder on an active subject, with leads placed so that the subject's ECG is not visible.[3] Two signals were recorded simultaneously. Record 'bw' contains primarily baseline wander, a low-frequency signal usually caused by motion of the subject or the leads. Record 'em' contains electrode motion artifact (usually the result of intermittent mechanical forces acting on the electrodes), with significant amounts of baseline wander and muscle noise as well. Record 'ma' contains primarily muscle noise (EMG), with a spectrum which overlaps that of the ECG, but which extends to higher frequencies. Electrode motion artifact is usually the most troublesome type of noise for arrhythmia detectors since it can closely mimic characteristics of the ECG.

The remaining records reproduce MIT-BIH Arrhythmia Database records 118 and 119 with 'em' noise added at various levels. Since the correct beat labels are known for these records, they may be used to test the noise tolerance of an arrhythmia detector. Records 118 and 119 were chosen for this purpose because they are not intrinsically noisy, and because most arrhythmia detectors can analyze them without significant errors. Thus any errors which occur in the analysis of the records to which noise has been added can be attributed to the noise, and not to any inherent difficulty in analyzing the ECG itself. The names of these records are of the form 'rrr_nn', where 'rrr' is the name of the original ECG record and 'nn' indicates the noise-to-signal ratio during the noisy periods (02 = noise with RMS amplitude 20% as large as the mean peak-to-peak amplitude of the QRS complexes, 12 = noise 120% as large as the QRS complexes, etc.).

ST Change Database

This database consists of 28 unannotated records ranging in length from 13 to 67 minutes, obtained from 28 subjects.[4] Records 300 to 323 were obtained during exercise stress tests, using an FM instrumentation tape recorder; these records exhibit transient ST depression in response to

exercise-induced ischemia. The header files for these records include information about the gain of the signals which can be used to calibrate ST measurements in terms of body surface potentials.

Records 324 to 327 were obtained from Holter tapes, and show ST elevation. Records 313 to 317 and 319 to 323 contain only one signal; the rest contain two signals. All signals were sampled at 360 Hz.

Malignant Ventricular Arrhythmia Database

This database consists of 22 thirty-five-minute records, obtained from Holter tapes of 16 subjects.[5,6] It is annotated only with respect to rhythm changes, which include 89 episodes of ventricular tachycardia, 60 episodes of ventricular flutter, and 42 episodes of ventricular fibrillation. The signal files contain two signals, each sampled at 250 Hz.

Atrial Fibrillation/Flutter Database

This database[7] may be useful for development and evaluation of atrial fibrillation/flutter detectors which rely on timing information only. It consists of 25 ten-hour records (obtained from Holter tapes of 25 subjects) containing about 300 episodes of atrial fibrillation and 40 episodes of atrial flutter. Because of space limitations, it is not feasible to include all 250 hours of the ECG signals on the disk. The CD-ROM contains two sets of annotation files for all 25 records, and a signal file for one record. The signal file contains two signals, sampled at 250 Hz. The reference annotation files contain only rhythm change annotations. The beat annotation files were produced by an automated QRS detector, and all beats are labelled normal; the R-R interval sequences may be recovered from these files and used as input to the atrial fibrillation/flutter detector to be tested. Note that the beat annotation files have not been audited, and contain a small number of errors.

ECG Compression Test Database

This database consists of 168 unannotated records, each 20.48 seconds in duration, obtained from Holter tapes from 38 subjects. The records exhibit a wide variety of arrhythmias, conduction disturbances, and noise. Many were selected because the characteristics of the signal or noise may be expected to pose a problem for an ECG compression method which is not exactly invertible. By comparing diagnoses made on the basis of compressed versions of these records with independent diagnoses made from the uncompressed versions, the ability of an ECG compression method to preserve clinically important waveform details can be assessed.[8] Each record contains two signals, sampled at 250 Hz.

Supraventricular Arrhythmia Database

This database contains the first 13 thirty-minute records of a database we are building to supplement the examples of supraventricular arrhythmias in the MIT-BIH Arrhythmia Database. The records were obtained from Holter tapes of 13 subjects. They have been annotated using a semi-automated method which gives highly accurate results, but the annotations have not been audited to the extent of those in the MIT-BIH Arrhythmia Database, and a small number of errors may be present. The reference annotation files include beat and signal quality annotations, but no rhythm annotations. Each record contains two signals, sampled at 128 Hz.

Long-Term Database

This database contains seven annotated long-term records ranging in length from 14 to 24 hours. These records are complete Holter tapes from seven subjects. As for the Supraventricular Arrhythmia Database, the records have been annotated using a semi-automated method, and a small number of errors may be present. The reference annotation files include beat and signal quality annotations, but no rhythm annotations. Six of the records contain two signals; the seventh contains three. All signals are sampled at 128 Hz.

Software for use with the CD-ROM

We have made generally available a substantial collection of software, developed to support our research, for use with our databases and with the AHA Database. In February, 1990, the annotation set was expanded to accommodate the needs of the European ST-T Database.[9] The software is sufficiently general, however, to be useful for dealing with any similar collection of digitized signals, which may or may not be annotated. It is written in a highly portable form compatible with ANSI C compilers or with K&R C compilers such as those on most UNIX systems. The software includes implementations of the beat-by-beat and run-by-run comparison algorithms specified for arrhythmia detector evaluation by the AAMI,[10] as well as programs for sample frequency conversion, filtering, printing annotated ECGs in 'chart recorder' or 'full disclosure' format, and waveform display.

We have recently developed an X Window System-based program called *WAVE* (Waveform Analyzer, Viewer, and Editor). *WAVE* is an extensible interactive graphical environment for manipulating multichannel digitized signals with optional annotations. It can run on a wide variety of UNIX systems; in addition, it can be accessed remotely using networked PCs or other systems for which X11 servers are available. Among its capabilities are fast display of waveforms and annotations at various calibrated scales, forward and backward searches for annotation patterns, graphical annotation editing, high-resolution printing of user-selected signal segments, and configurable control of external signal-processing and analysis programs. This last feature permits *WAVE* to be used as an interactive graphical 'shell' around an arbitrary application program which manipulates ECG database signal or annotation files.

This software is all built on a common library of database interface functions which provide clean and uniform access to signal and annotation files stored in a variety of formats, including those used on the CD-ROM and for the AHA Database. On the CD-ROM itself is a precompiled version of this library for MS-DOS users, who may link it with their own programs using any of a number of commercially available C compilers.

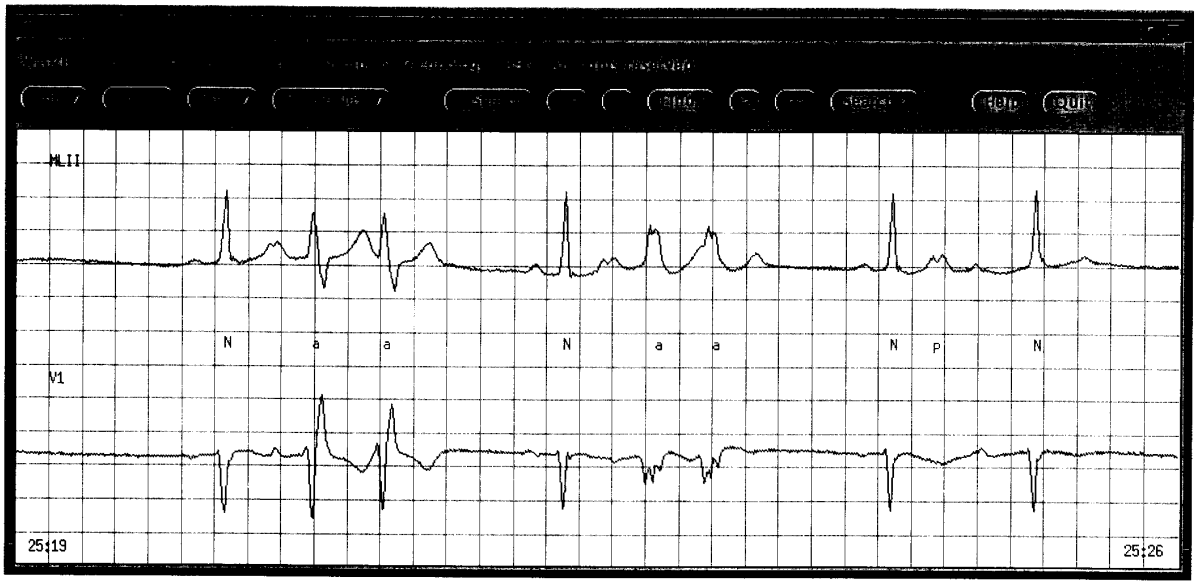

Figure 1. Screen dump of *WAVE*, illustrating a record from the MIT-BIH Arrhythmia Database.

The CD-ROM also includes MS-DOS executable versions of three programs for reading the data. Two of these convert user-selected portions of binary signal or annotation files into text form for review or processing by other programs; the third program interactively displays signals and annotations using any of the popular PC graphics adapters.

CD-ROM production

The production process is quite simple, and worth describing here to encourage others with large databases to follow our example. In our case, we prepared a set of twenty-five 1600 bpi nine-track tapes in ANSI X3.27 format, and a map of the directory hierarchy we wanted to have on the finished disk. One tape included a few binary MS-DOS executable programs, written as images in the same ANSI format.

We prepared artwork for the disk label and sent it with a sample tape to the disk mastering plant to be certain that we had the formatting details correct. Two weeks later, we completed the twenty-five tapes to be copied and sent them along; in another week, the finished CD-ROMs were delivered.

The software we have described above is a toolkit which can be used to develop and document a CD-ROM database of digitized signals. By making it generally available, we hope to stimulate the production of similar disks by others. Our experience demonstrates that CD-ROM technology can make it economically feasible to distribute databases which otherwise would be prohibitively expensive to duplicate on a small scale.

References

1. Mark, R.G., Schluter, P.S., Moody, G.B., Devlin, P.H., and Chernoff, D. An annotated ECG database for evaluating arrhythmia detectors. *Frontiers of Engineering in Health Care: Proceedings of the 4th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 205-210. New York: IEEE Press (1982).
2. Ripley, K.L., and Oliver, G.C. Development of an ECG database for arrhythmia detector evaluation. *Computers in Cardiology* **4**:203-209 (1977).
3. Moody, G.B., Muldrow, W.K., and Mark, R.G. A noise stress test for arrhythmia detectors. *Computers in Cardiology* **11**:381-384 (1984).
4. Albrecht, P. S-T segment characterization for long-term automated ECG analysis. MIT M.S. thesis (1983).
5. Greenwald, S.D., Albrecht, P., Moody, G.B., and Mark, R.G. Estimating confidence limits for arrhythmia detector performance. *Computers in Cardiology* **12**:383-386 (1985).
6. Greenwald, S.D. Development and evaluation of a ventricular fibrillation detector. MIT M.S. thesis (1986).
7. Moody, G.B., and Mark, R.G. A new method for detecting atrial fibrillation using R-R intervals. *Computers in Cardiology* **10**:227-230 (1983).
8. Moody, G.B., Mark, R.G., and Goldberger, A.L. Evaluation of the "TRIM" ECG data compressor. *Computers in Cardiology* **15**:167-170 (1988).
9. Taddei, A., Biagini, A., Distante, G., Marchesi, C., Mazzei, M.G., Pisani, P., Roggero, N., and Zeelenberg, C. An annotated database aimed at performance evaluation of algorithms for ST-T change analysis. *Computers in Cardiology* **16** (1989).
10. *Testing and Reporting Performance Results of Ventricular Arrhythmia Detection Algorithms*. Arlington, VA: Association for the Advancement of Medical Instrumentation; publication AAMI ECAR-1987.