Generating Explanations with the Help of Temporal Constraints

Margo Guertin
Computer Science Dept., Boston University, Boston, MA. 02215
guertin@cs.bu.edu

Abstract

This paper describes HOLMES, a new system for generating explanations for cyclic events using temporal constraints. The domain is electrical conduction in the heart, with the temporal constraints coming from the trace of an electrocardiogram (ECG), and an explanation being a laddergram — a diagram of the paths of the beat impulses through the heart which could have produced that ECG. I have combined a method for propagating temporal constraints with an object-oriented model of heart conduction, and it has resulted in a remarkably quick and efficient search of the vast space of possible explanations. This type of system, which combines temporal reasoning and abductive inference, may well point the way to much fruitful research in the future.

1 Introduction

My original motivation for developing the HOLMES system was to build a program which could use ECG information to automatically trace the paths of beats through the heart. This is something which is currently done by hand when a doctor suspects heart conduction problems. As I worked on the system, I realized it was also an interesting and promising hybrid of temporal reasoning and abductive inference. It is this aspect I concentrate on here.

Abduction, the inference of the best explanation for the observed evidence, is a small but steadily growing field of AI research. A clear and thorough discussion of it appears in John and Susan Josephson's recent book on abductive inference [3], in which they describe past and current systems for doing automated abduction. All these systems must deal with the problem of an enormous search space of explanations which might account for a body of evidence. Since there are far too many possible hypotheses to look at them all, much effort has been expended on ways of directing the search towards the hypotheses most likely to produce the "best" explanations.

Josephson points out that a system can avoid these problems of search complexity if it has enough of the right kind of information and an effective strategy for using it.[2] I believe that the temporal intervals which represent the start, stop and duration times of events in the HOLMES system are an example of this right kind of information. Almost all of the vast search space of hypotheses are ruled out quickly by a form of quantitative temporal constraint propagation, based on Isaac Kohane's Temporal Utility Package (TUP)[4], a domain independent utility which makes use of interval arithmetic to propagate quantitative temporal constraints. TUP, in turn, is based upon upon James F. Allen's interval-based logic [1].

In my domain, a hypothesis is a chronologically ordered list of events, each of which may account for one of the "bumps" on the ECG. For example, the event of the left ventricle of the heart being activated by an impulse from the adjacent left bundle branch at time x could be used to explain an ECG spike (or QRS wave) at time x. An explanation is a causal network of events which can account for all the ECG evidence. Each attempted explanation is generated incrementally and in order, incorporating one event at a time, beginning at the time of the first piece of evidence and ending with the last piece of evidence, until an inconsistency is detected or an explanation has been successfully completed. Any inconsistency which arises is either because an event necessary to the particular explanation being attempted is inconsistent with the evidence or with some event(s) previously incorporated into it. Because events are incorporated in chronological order, failure can only arise from a conflict with the evidence or previous events already part of the explanation being formed. This means that if an ordered hypothesis of events A, B, C, D, E, F, ... fails on the attempt to incorporate D, then *all* hypotheses whose first four events are A, B, C, and D can be ruled out. In a search space where each hypothesis consists of many events, a failure that occurs early in the hypothesis list allows a shallow cutoff in the search tree, eliminating a significant number of hypotheses from consideration. One is particularly likely to encounter this kind of efficient pruning in a cyclic domain where the pieces of evidence are often repetitive. In such a domain, the

chances are high that any inconsistency which may occur near the end of a list of hypothesized events, is a repeat of the same inconsistency occuring much earlier in the list, where the pruning is more drastic.

Thus, in the domain of heart conduction, it is possible to take full advantage of the cyclic temporal information available, propagate temporal constraints to rule out almost all of the possible explanations early, and consider only those remaining. This ability to generate all the explanations would be important in any scenario where more evidence might be gathered later which could rule out what was formerly rated the "best" explanation. It would also be important to any system meant to discover new explanations, not normally considered.

In the following sections, I discuss the domain of heart conduction in more detail, give an overview of its control with a brief example of how it uses temporal constraint propagation, and finally give my preliminary statistics on search efficiency from generating explanations for five different (normal) ECGs.

2 The Heart Conduction Domain

The information on heart physiology and electrocardiography below comes primarily from three sources [5, 6, 7]. The heart can by thought of as a fist-sized electric pump for moving blood through the body (see Fig. 1) . When all is going well, the sino-atrial

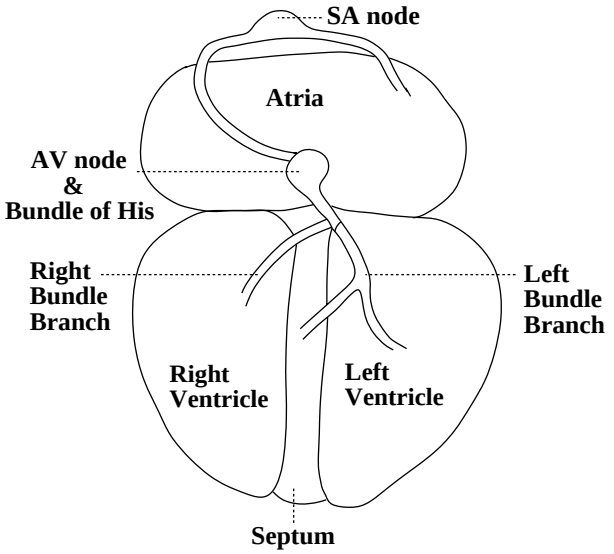

Figure 1: Diagram of the main parts of the heart

node (SA node) at the top generates a steady electrical pulse which is conducted down through the other parts of the heart. The SA node's current first activates the upper chambers of the heart (the atria), which contract and squeeze the blood into the lower chambers (the left and right ventricles). Very soon after the SA node activates the atria, it activates the atrioventricular node (AV node) and the "Bundle of His", and the current travels through them to the left and right bundle branches. From the bundle branches it travels into the ventricles and the septum between them, normally activating them all simultaneously. As the current travels through the ventricles, they contract powerfully, forcing the blood just received from the atria on out into the body.

When any part of the heart is activated, it begins to *depolarize*, which has the effect of spreading the current throughout, and soon to adjacent parts. Once the process of depolarization is complete, the longer process of *repolarization* begins. This can be thought of as a recovery process, during which the heart part cannot be activated by any impulse. Once the repolarization process is complete, the part is in its initial resting state, ready to receive and carry an impulse comes along. Each heart part can be thought of as a little finite state machine (FSM), and the whole heart as a bunch of parallel, interacting FSMs.

In the HOLMES heart model, an *event* is defined to be the activation of one heart part, either by an adjacent part, or occasionally, by itself. (It is possible for any heart part to activate itself by producing its own electrical impulse spontaneously, although normally this only happens in the SA node.) The representation of an event in the model includes temporal interval parameters for the start of its depolarization, duration of depolarization, start of repolarization, duration of repolarization, and start of its rest state.

The *electrocardiogram* (see Fig. 2) is a trace of the heart's electrical activity as measured from different points on the surface of the body. It provides incomplete evidence of the impulse paths through the heart, because some heart parts produce a current strong enough to be measured from the outside, while others do not. The atria depolarizing give rise to the P wave, the ventricles depolarizing produce the QRS complex, and the T wave is the result of the ventricles repolarizing. All other events are invisible to the ECG and must be inferred by HOLMES.

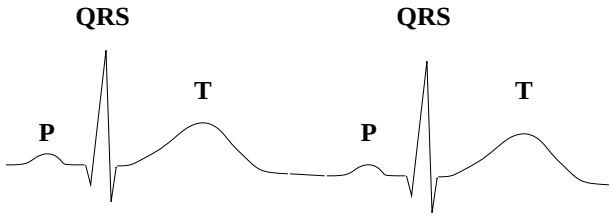

Figure 2: Two heartbeats in a typical ECG

A *hypothesis* is a list of assumed events, each of which could account for one bump on the ECG. In Fig. 2, for instance, each P wave could be due to the atria being activated by the SA node, or the atria activating themselves spontaneously. Each QRS could be the left ventricle being activated by the left bun-

dle branch, the left ventricle being activated by the septum, or the left ventricle activating itself spontaneously. So, even for a 2-beat ECG, we would have $2 \times 3 \times 2 \times 3$, or 36 4-event hypotheses.

In this domain, an *explanation* is a complete causal network of events which could have resulted in the ECG. Following is an example of a possible explanation for a P wave, followed by a QRS complex, followed by a T wave. The SA node generated a beat which activated the atria and the AV node and Bundle of His. Then the Bundle of His activated the left and right bundle branches, and soon after, the left bundle branch activated the left ventricle and septum, while the right bundle branch activated the right ventricle. Any explanatory network must be internally consistent, and include all the assumed events of the hypothesis, as well as all invisible events (within the ECG's time span) which either lead to or resulted from the hypothesis events. All events in the explanation must, of course, be consistent with the ECG evidence.

3 The HOLMES System

3.1 Control

The HOLMES system, shown in Fig. 3, contains a general model of the heart which is individualized with information from the ECG. For example, normal heart rates range anywhere from 55 beats per minute to 120 beats per minute, but if the ECG shows a normal heart rate, in this case varying from, say, 80 to 95, the interval range for the heart rate parameter is restricted accordingly. As many parameters as possible are tightened using the ECG information. The hypothesis former, a set of rules for forming all possible hypotheses from an ECG, swings into action next, producing a very large number of hypothesis lists which go into the search space. (In the case of an apparently normal n-beat ECG, 6^n lists of length $2n$ are produced.) From this point on, all hypotheses in the search space (which have not been pruned away) are fed to the explanation, one at a time, until the search space is exhausted.

The explanation developer attempts to form all explanations possible for each hypothesis it receives. (There can be several explanations for the same hypothesis because the hypothesis is a list of only *visible* events, and there can be a number of unique causal networks of events which include the same set of visible events.) For each attempted explanation, the explanation developer starts a chronologically ordered list of events to be incorporated, which initially contains only the events of the hypothesis, but is expanded, with each event successfully incorporated, to include all events which would result from it. Each event incorporated must be consistent with the ECG and with the events already incorporated in the explanation. When it is added, any constraints it imposes are propagated back through the model. If failure occurs, the explanation is abandoned and its failure point noted. All successful explanations resulting from the hypothesis are put into the pool of completed explanations. If all attempts to explain the current hypothesis fail, the *latest* hypothesis event responsible for failure is noted and sent on to the search space pruner, which prunes the search space accordingly.

3.2 Simplified Example of Temporal Constraint Propagation

Following is a simplified example of constraint propagation through a very small causal network of four events — the activations of four adjacent heart parts, say, A, B, C, and D. It is exactly this kind of thing which happens, over and over, as HOLMES tries to develop an explanation, and each event being incorporated in the explanation is checked for internal consistency with those already there.

Given:

1. The activations of heart parts A, B, and C do not show up on the ECG; the activation of part D *does*.
2. It has been inferred from the model and the assumptions of the current hypothesis that the activation of part A occurred some time between 0 and 20 msec. We represent this by the interval (0, 20).
3. Part A is adjacent to part B.
4. It takes (20, 40) for a current to travel from A to B.
5. Part B is adjacent to parts C and D (as well as A).
6. It takes (5, 15) for a current to travel from B to C.
7. It takes (10, 20) for a current to travel from B to D.

So initially:

1. A is activated at (0, 20).
2. B is activated at (20, 60).
3. C is activated at (25, 75).
4. D is activated at (30, 80).

But then we see from the ECG that part D was activated at 45 (or (45, 45) in interval notation). This restricted start time for the activation of D constrains the start of B, which in turn constrains the starts of A and C. Using interval arithmetic [4], we end up with the following constrained information:

1. A is activated at (0, 15).
2. B is activated at (25, 35).
3. C is activated at (30, 50).
4. D is activated at (45, 45).

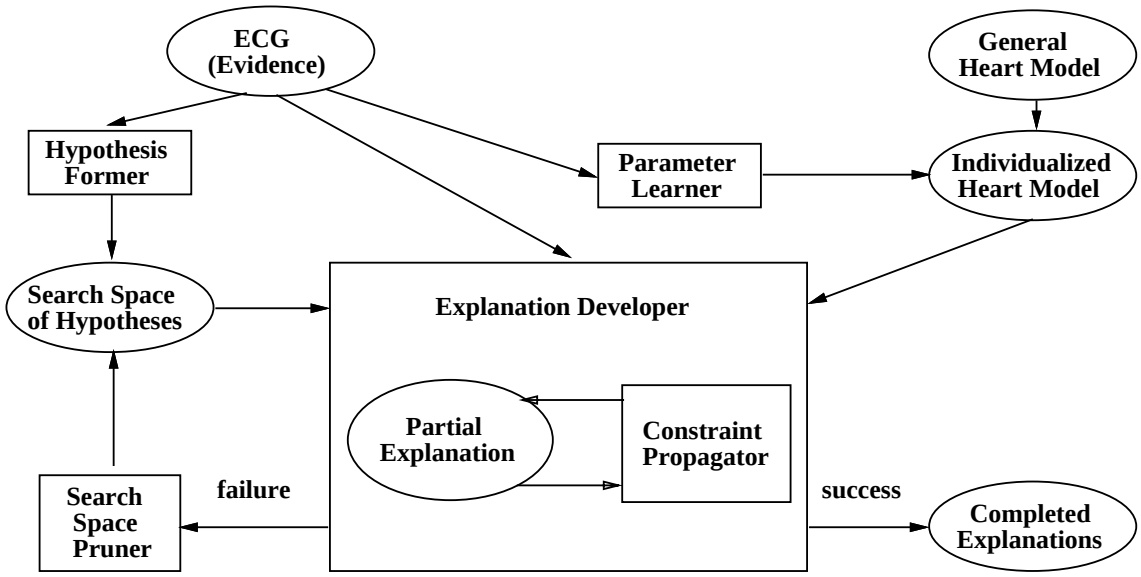

Figure 3: Architecture of the HOLMES System

4 Results and Conclusions

At this point in time, the HOLMES system is fully operational for normal (i.e. healthy) ECGs, and I have run it successfully on five such ECGs so far. I am now in the process of fine tuning the general heart model to allow it to handle abnormal ECGs as well, since one of my main motivations in developing this system was to make a physicians' tool for diagnosing heart conduction problems.

I have gotten good results with the five ECGs I have tried. For each, HOLMES has come up with the simplest explanation (that of a normal sinus rhythm with no aberrations). For two of them, it has also found some more — for ECG #1, three other explanations, and for ECG #5, two others. While these alternative explanations are much less likely, it is quite interesting to be able to see what aberrant events *could* could be occuring in a heart's conduction system, when a normal-looking ECG trace is being produced.

The efficiency of performing a complete search is very high. I measure efficiency with the ratio of the number of hypotheses which need to be examined in a complete search to the search space size (about 6^n for an n-beat ECG). In my first five runs, the efficiency ratio is $34 : 6^{10}$, in the least efficient search, and $135 : 6^{17}$, in the most efficient.

I think that these results are due, in large part, to the pruning enabled by temporal constraint propagation, and are further improved by the cyclic nature of the domain. I hope they will encourage more researchers to incorporate temporal information and constraint propagation techniques into their models, and to start work on some of the many unexplored cyclic domains of medical and environmental science.

Acknowledgments

This paper has benefitted greatly from the insightful comments and questions of John Josephson and Michael Weintraub on an earlier draft. I am also very grateful to Bipin Indurkya and Scott O'Hara for their time and thoughtful criticism throughout its development.

References

[1] James F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[2] John R. Josephson. Personal communication.

[3] John R. Josephson and Susan G. Josephson. *Abductive Inference: Computation, Philosophy, Technology.* Cambridge University Press, 1994.

[4] Isaac S. Kohane. *Temporal Reasoning in Medical Expert Systems.* PhD thesis, Boston University, 1987.

[5] Henry J. L. Marriott. *ECG/PDQ.* Williams & Wilkins, 1987.

[6] Lionel H. Opie. *The Heart: Physiology and Metabolism*, chapter 5, pages 102–126. Raven Press, New York, second edition, 1991.

[7] Allen M. Scher. The electrocardiogram. In Harry D. Patton, Albert F. Fuchs, Bertil Hille, Allen M. Scher, and Robert Steiner, editors, *Textbook of Physiology: Circulation, Respiration, Body Fluids, Metabolism, and Endocrinology*, chapter 38, pages 796–819. W.B. Saunders Company, 21st edition, 1989.