# Temporal Abstraction for Long-term Changed Tests in the Hepatitis Domain

**Saori Kawasaki, Trong Dung Nguyen, and Tu Bao Ho**

Japan Advanced Institute of Science and Technology
Ishikawa, 923-1292 JAPAN
E-mail: {skawasa, nguyen, bao}@jaist.ac.jp

[Received April 30, 2003; accepted July 30, 2003]

**Temporal abstraction (TA) aims to transform temporal data into a symbolic interval-based representation of data. Most existing TA methods are applicable to data regularly collected in short periods. In this paper we proposed a TA method for temporal data irregularly collected in long periods. The core of our method is the notion of "change of states" and algorithms to detect them. The method has been applied to a database on hepatitis with results evaluated by medical experts.**

**Keywords:** temporal abstraction, long-term changed tests, hepatitis data.

## 1. Introduction

Temporal Abstractions (TA) is a special kind of data abstraction [6]. TA methods aim to transform time-stamp data into an interval-based representation of data by extracting their most relevant features. The TA process can be viewed in two steps of basic TA and complex TA. The former concerns with abstracting time-stamped data within episodes and it typically extract states (e.g., low, normal, high) and/or trends (e.g., increase, stable, decrease), while the latter concerns with temporal relationships between findings from a basic TA or from other complex TA. Typical TA works include the knowledge-based TA framework [9]; methods for context-sensitive and expectation-guided TA [5], [7]; methods for combining statistical and probability techniques with TA [1]. These methods, however, can apply only to short periods and/or irregular time-stamp data.

The hepatitis database collected during 1982-2001 at the Chiba university hospital was recently given to challenge the data mining research [2]. It contains results of 771 patients on 983 laboratory tests. The tests are divided into two groups: tests that can suddenly change in short-terms (several days or weeks) and tests that only smoothly change in long-terms (months and years). The hepatitis database is a large un-cleansed temporal relational database consisting of six tables of which the biggest has 1.6 million records. The typical characteristics of these data are their irregularity and long periods of gathering (e.g., twenty years). The doctors posed a number of problems on hepatitis that are expected to be investigated by KDD techniques [8].

This paper presents our TA method for mining long-term changed tests of the hepatitis data. The core of the method is the novel notion of "changes of state" to characterize the sequences of long-term changed tests as well as an algorithm to detect abstraction patterns of sequences. Part of the obtained results has been highly evaluated by medical doctors.

## 2. The Hepatitis Database and Preprocessing

### 2.1 The hepatitis database and problems

This relational database is composed of six tables: (T1) Basic information of patients (771 records); (T2) Results of biopsy (960 records); (T3) Information on interferon therapy (198 records); (T4.) Information about measurements in in-hospital tests (459 records); (T5) Results of out-hospital tests (30,243 records); (T6) Results of in-hospital tests (1,565,877 records)

The difficulties in investigating this database due to the fact that it contains many hospital tests each of which is a sequence of time related values for each patient. Also the durations, the sequences lengths, and time-stamp data differ from one another.

Among six problems P1-P6 challenged by the doctors to the KDD community [2], we focus on the three ones:

P1. Discover the differences in temporal patterns between hepatitis B and C.
P2. Evaluate whether the interferon therapy is effective or not.
P3. Discover the relationships between the stage of liver fibrosis and the onset of hepatocellular carcinoma.

Each of these problems requires a special sub-dataset derived from the original hepatitis database.

### 2.2 Preprocessing for hepatitis data

As the hepatitis database has been collected during a long period with progress in test equipments, it also contains inconsistent measurements, many missing values, and a large number of non unified notations. Such incompleteness has to be handled before any further analysis in order to avoid the misdirection. Our preprocessing includes data cleaning, integration, reduction, and transformation, as well a special preprocessing for extracting sub-datasets for problems under investigation. The followings are brief explanations for some of them.

*2.2.1 Feature selection and data reduction*

Each patient is described by 983 temporal sequences corresponding to the 983 hospital tests. As the complexity of learning generally increases with the number of tests under investigation, a small number of selected tests is expected. After selecting 41 tests from 983 tests by statistical frequency check and medical background knowledge, we firstly focus on the 15 most typical tests as suggested by medical experts. These tests can then be divided into two groups depending whether their values can change in a short term or long term.

(1) The short-terms changed tests: The tests in this group, GOT, GPT, TTT, and ZTT, in particular GOT and GPT, can rapidly change (within several days or weeks) their values to high or even very high values when liver cells were destroyed by inflammation.

(2) The long-term changed tests: The tests in this group can slowly change (within months or years). Liver has the reserve capacity so that some products of liver (T-CHO, CHE, ALB, and TP) do not have low values until its reserve capacity is exhaustive (the terminal state of chronic hepatitis, i.e., liver cirrhosis). Two kinds of tests in this group are: going down tests: T-CHO, CHE, ALB, TP, PLT, WBC, and HGB; and going up tests: D-BIL, I-BIL, T-BIL, and ICG-15.

*2.2.2 Extraction of data subsets*

The data extraction aims to create an appropriate dataset for solving each of problems P1-P3. These are supervised datasets where each patient's data correspond to temporal sequences of test values and the patient has a value of the class attribute such as "type B" or "type C" of hepatitis (problem P1); "F0", "F1", "F2", "F3" and "F4" of fibrosis stages (problem P2); "response", "partial response", "aggravation", and "no change" of the interferon therapy (problem P3).

The extraction of data subsets deeply depends on the domain context. It is important to recall that the result of temporal abstraction also strongly depends on how episodes on which data are abstracted were taken. In this research we usually investigated the episode of 5 years (with some other trials up to 10 years) according to the suggestion of medical doctors. As the goal of the problem P1 is to find the set of historical sequences of some tests to explain hepatitis B or C, the sequence of each test was forwardly taken from its beginning point to the point where the sequence reaches the episode length. In cases of problems P2 and P3, the sequences were backwardly taken from the biopsy date or the starting day of interferon therapy, respectively, to the point where the sequence reaches the episode length.

The complete solution to these problems requires different TA methods for processing short-term changed tests and long-term changed tests. Our work concerning a basic TA method for long-term changed tests will be presented in the next section. The TA algorithm method for short-term changed tests was introduced in [4] while some related notions are mentioned here.

## 3. Basic Temporal Abstraction Method for Long-Term Changed Tests

After having the preprocessed data subsets with test sequences taken in the given episodes, the task of basic TA is to find abstractions of those sequences. Our basic idea to abstract a sequence is to map it into a set of abstraction patterns viewed as combinations of TA primitives in subsequences. To this end, our method consists of two parts: (1) to determine typical abstraction patterns; and (2) to assign each test sequence into one of abstraction patterns.

### 3.1 Determination of Typical Abstraction Patterns

This problem can be stated as follows: Is it possible to determine a small number of typical abstraction patterns to be used to characterize most real sequences observed. In our opinion, the solution to such a problem can be well obtained by a combination of TA primitive computation with human inspection using visual tools.

*3.1.1 The TA primitives of long-term changed tests*

The TA primitives are commonly used in temporal abstraction [1], [5], [7], [9] and mainly concern with the states and trends of sequences.

A test value can lie in the normal region determined by medical doctors (with upper and lower bounds) or out of the normal region (high or low regions). Consequently, the state of any sequence or subsequence can be judged with the following TA state primitives: N (normal), L (low), VL (very low), XL (extreme low), H (high), VH (very high), and XH (extreme high). Note that the thresholds to determine these primitives are greatly different with respect to different tests. The thresholds to distinguish the state primitives of tests are given by medical doctors, for example, those to distinguish values N, H, VH, XH of "total protein" are 5.5, 6.5, 8.2, 9.2 where (5.5, 6.5) is the normal region. The long-term changed tests, whose values tend to gradually and smoothly change, usually have values of "N", "L" and "H" of the state primitives because the extremely far values hardly occur. The values of "VH", "XH", "VL" and "XL" basically can be observed in short-term changed tests.

In this work we consider the following TA trend primitives of long-term changed tests: S (stable), I (increasing), FI (fast increasing), D (decreasing), and FD (fast decreasing).

*3.1.2 Observation and determination of abstraction patterns*

In case of hepatitis data, the sequences taken in long episodes usually have complex changes. Our approach to determination of abstraction patterns is based on careful observations and analysis of various real sequences. To this end, we created a simple tool in MATLAB for visualizing the sequences of tests in a given episode. As preliminary work, we observed almost sequences of patient's test values in the hepatitis database to get an idea about the possible abstraction patterns.

In case of long-term changed tests, we have observed two main groups of sequences. One is a group of sequences that greatly fluctuate between the boundaries of the normal range. The other is a group of sequences that smoothly change between the normal, high or low regions. In case of short-term changed tests, the move is sometimes very rapid and the deviation from the base line of the sequence is very big. The baseline of a sequence often keeps a same level from the boundary in case it is out of the normal region but the distance from the normal boundary varies.

Our basic views in abstraction patterns is the proposed notions "change of state" to characterize long-term changed tests sequences, and the notion of "base state" and "peaks" to characterize the short-term changed tests sequences. For the short-term changed tests, a new state primitive, denoted by "P", indicates whether peaks appear in a given sequence.

*3.1.3 Relations between TA primitives*

By observations and visual analysis we detected and formulated four kinds of relations between primitives:

"**change state to**" (>): shows that the state of a sequence changes from one region to another region. This relation can occur in long-term changed tests.

"**and**" (&): connects a state primitive and the peak primitive, and occurs in short-term changed tests.

"**and then**" (–): connects a state primitive and a trend primitive in long-term changed tests. It expresses a trend in the sequence that does not change to another state ever after.

"**majority/minority**" (/): connects two adjoining states for describing sequences of long-term changed tests that fluctuate in the boundary of two regions. The state before the primitive "/" is the state where majority of points in the sequence belong to. For example, "X/Y" means that the majority of points are in state X and the minority of points is in state Y.

By analyzing various available sequences, we formulated four possible structures of abstraction patterns that can be used to represent most hepatitis sequences.

```
<pattern> ::= <state primitive>
<pattern> ::= <state primitive> <relation> | <state primitive>
                | <trend primitive>
<pattern> ::= <state primitive> <relation> <peak>
 <pattern> ::= <state primitive> <relation>
               <state primitive> <relation> | <state primitive>
                | <trend primitive>
```

**Fig.1**. Structure of abstraction patterns

Here are examples of abstraction patterns:

"ALB = N" (ALB is in the normal region),

"CHE = H–I" (CHE is in the high region and then increasing),

"GPT = XH&P" (GPT is extremely high and with peaks),

"I-BIL = N>L>N" (I-BIL is in the normal region, then changed to the low region, and finally changed to the normal region).

We developed and used the following procedure to identify typical abstraction patterns:

1. Consider the abstraction patterns structures as formulas and the <state primitive>, <trend primitive> and <relation> as their variables. Create all possible candidate abstraction patterns by replacing the <state primitive>, <trend primitive> and <relation> with their possible values.
2. Randomly take a large number of sequences from the datasets, visualize and manually match them with the candidate abstraction patterns to see each of them matches which candidate abstraction pattern.
3. Eliminate the candidate abstraction patterns that have no or a small number of matched sequences.

**Fig. 2**. Procedure for determining typical abstraction patterns

After applying the procedure and consulting medical doctors, we tentatively determined 22 typical patterns for long-term changed tests shown in **Fig**. **3**. (Also, 8 typical abstraction patterns were determined for short-term changed tests as shown in **Fig. 4**.).

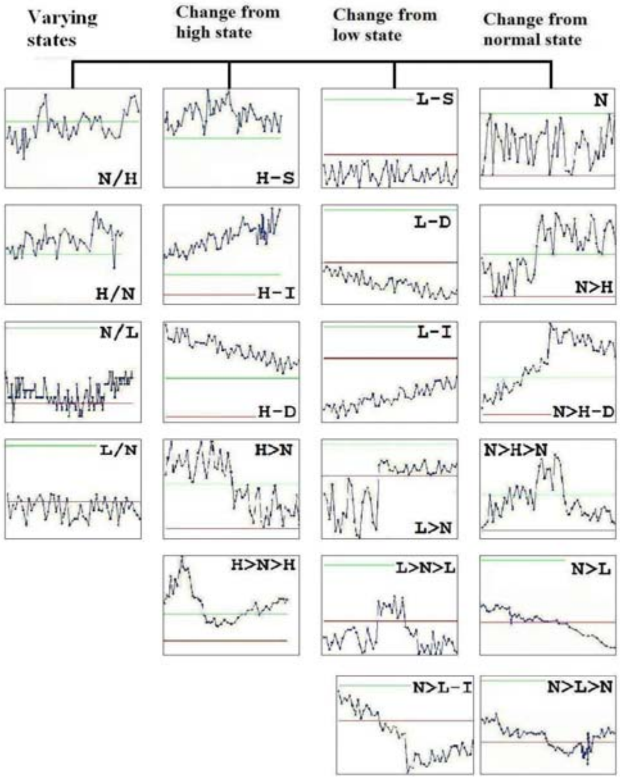


**Fig. 3.** Typical abstraction patterns for long-term changed tests

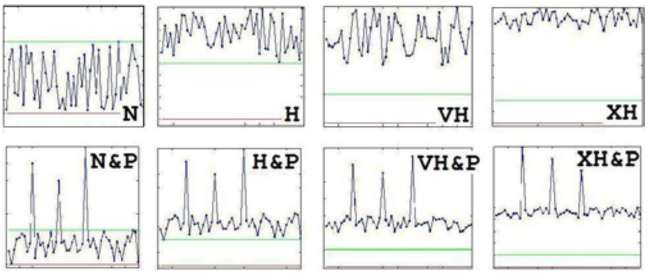


**Fig. 4.** Typical abstraction patterns for short-term changed test

Our key idea is to use the "change of state" to characterize sequences of the long-term changed tests. The "change of state" contains information of both state and trend, and can compactly characterize the sequence. We firstly distinguish two kinds of sequences: those fluctuate around the boundary of the normal region and those do not. The latter can be then observed if at their beginning, the first data points take which of the three states "N", "H", or "L". It will happen that either the sequence changes from one state to another state, smoothly or variably (at boundaries), or the sequence remains in its initial state without changing.

We determined four groups of abstraction patterns for long-term changed tests:

- "Varying states": Consisting of patterns that fluctuate around the boundary of the normal region (N/H, H/N, N/L, L/N):
- "Change from high state": Consisting of abstraction patterns H–S, H–I, H–D, H>N, H>N>H;
- "Change from low state": Consisting of abstraction patterns L–S, L–D, L–I, L>N, L>N>L;
- "Change from normal state": Consisting of abstraction patterns N, N>H, N>H–D, N>H>N, N>L, N>L>N, N>L–I.

### 3.2 Basic TA algorithm for long-term changed tests

The second step in our TA method is to assign each test sequence into one of abstraction patterns. The algorithm described in **Fig. 5,** which is intrinsically based on the properties of hepatitis long-term changed tests, allows to us to do this task.

*3.2.1 Notations and parameters used in the algorithm*

The basic idea of the algorithm is as follows: If the sequence is not fluctuated around the boundary of the normal region, specify its state at the beginning and at the end subsequences as well as the changes in the middle in order to match it with the abstraction patterns. To this end, a number of intermediate functions are defined on the sequence S.

*Functions*

- High(S): # points of S in the high region
- Low(S): # points of S in the low region
- Normal(S): # points of S in the normal region
- In(S) = Normal(S)/ (High(S) + Normal(S) + Low(S))
- Out(S) = 1 - In(S)
- Cross(S): # times S crosses the upper and lower boundaries of the normal region
- First$_\sigma$(S), Last$_\sigma$(S): State of the subsequence with length σ from the first point or the last point of S.
- State(S): State of S (one of the state primitives)
- Trend(S): Trend of S (one of trend primitives)

*Parameters*

The following parameters (thresholds) α, δ, ε, σ (integer), and β (real) are used in the algorithm whose values are dynamically and experimentally specified in relation with the length *n* of the sequence.

α: Threshold regarding the number of times the sequence crosses the boundaries of the normal region.

$$\alpha = \begin{cases} 0.3\times n & \text{if } n\in\{10,...,30\} \\ 0.2\times n & \text{if } n\in\{31,...,60\} \\ 0.11\times n & \text{if } n\in\{61,...,100\} \\ 0.1\times n & \text{if } n>100 \end{cases}$$

β: Threshold regarding the ratio of points in the sequence being in or out of the normal region.

$$\beta = \begin{cases} (n-2)/n & \text{if } n\in\{10,...,20\} \\ (n-3)/n & \text{if } n\in\{21,...,40\} \\ (n-5)/n & \text{if } n\in\{41,...,60\} \\ (n-6)/n & \text{if } n\in\{61,...,100\} \\ (n-8)/n & \text{if } n>100 \end{cases}$$

δ : Threshold regarding the number of sequence points that are in the high region. We employed δ = α.

ε: Threshold regarding the number of sequence points that are in the low region. We employed ε = α/2.

σ: Threshold regarding the length of the first or last subsequences. By default we take σ as 6 months.

*3.2.2 The algorithm*

This algorithm is context-sensitive and depends on the parameters. With the parameters taken as mentioned above, most testing sequences were correctly identified, and only a few sequences were undetermined.

**Input**: A sequence of patient's values of a test with length *n* denoted as $S_{00} = \{s_1, s_2, \ldots, s_n\}$ in a given episode.

**Output**: An abstracted pattern of the sequence derived from the sequence.

**Parameters**: α, δ, ε, σ (integer), β (real)

**Notations**: $S_{10}$ = [$s_1$, median], $S_{20}$ = [median, $s_n$], $S_{11}$ = [$s_1$, 1st quartile], $S_{12}$ = [1st quartile, median], $S_{21}$ = [median, 3rd quartile], $S_{22}$ = [3rd quartile, $s_n$]

*A. Identification of patterns with many crosses*

```
1. If $Cross(S_{00}) > \alpha \wedge In(S_{00}) > Out(S_{00}) \wedge High(S_{00}) > Low(S_{00})$
   then N/H
2. If $Cross(S_{00}) > \alpha \wedge In(S_{00}) > Out(S_{00}) \wedge High(S_{00}) < Low(S_{00})$
   then N/L
3. If $Cross(S_{00}) > \alpha \wedge In(S_{00}) < Out(S_{00}) \wedge High(S_{00}) > Low(S_{00})$
   then H/N
4. If $Cross(S_{00}) > \alpha \wedge In(S_{00}) < Out(S_{00}) \wedge High(S_{00}) < Low(S_{00})$
   then L/N
```

*B. Identification of patterns without many crosses*

```
5.  If $In(S_{00}) > \beta$ then N
6.  If $Out(S_{00}) > \beta \wedge State(S_{00}) = H \wedge Trend(S_{00}) = S$ then H–S
7.  If $Out(S_{00}) > \beta \wedge State(S_{00}) = H \wedge Trend(S_{00}) = I$ then H–I
8.  If $Out(S_{00}) > \beta \wedge State(S_{00}) = H \wedge Trend(S_{00}) = D$
      $\wedge Last(S_{22}) = H$ then H–D
9.  If $Out(S_{00}) > \beta \wedge State(S_{00}) = L \wedge Trend(S_{00}) = S$ then L–S
10. If $Out(S_{00}) > \beta \wedge State(S_{00}) = L \wedge Trend(S_{00}) = D$    then L–D
11. If $Out(S_{00}) > \beta \wedge State(S_{00}) = L \wedge Trend(S_{00}) = I$
      $\wedge Last(S_{22}) = L$ then L–I
```

*C. Identification of patterns with changes from the normal region*

12. **If** $\text{First}_\sigma$ $(S_{00}) = N \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{Last}_\sigma(S_{22}) = H$ $\wedge$ $\text{Trend}(S_{22}) = I \wedge \text{Low}(S_{00}) < \varepsilon$ **then** N>H
13. **If** $\text{First}_\sigma$ $(S_{00}) = N$ & $\text{Cross}(S_{00}) < \alpha$ & $\text{Last}_\sigma(S_{22}) = H$ & $\text{Trend}(S_{22}) = D \wedge \text{Low}(S_{00}) < \varepsilon$ **then** N>H–D
14. **If** $\text{First}_\sigma$ $(S_{00}) = N \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{High}(S_{00}) > \delta$ $\wedge$ $\text{Last}_\sigma(S_{22}) = N \wedge \text{Trend}(S_{22}) = D \wedge \text{Low}(S_{00}) < \varepsilon$ **then** N>H>N
15. **If** $\text{First}_\sigma$ $(S_{00}) = N \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{Last}_\sigma(S_{22}) = L$ $\wedge$ $\text{Trend}(S_{22}) = D \wedge \text{High}(S_{00}) < \varepsilon$ **then** N>L
16. **If** $\text{First}_\sigma$ $(S_{00}) = N \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{Last}_\sigma(S_{22}) = L$ $\wedge$ $\text{Trend}(S_{22}) = I \wedge \text{High}(S_{00}) < \varepsilon$ **then** N>L–I
17. **If** $\text{First}_\sigma$ $(S_{00}) = N \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{Low}(S_{00}) >$ $\delta \wedge \text{Last}_\sigma(S_{22}) = N \wedge \text{Trend}(S_{22}) = I \wedge \text{High}(S_{00}) < \varepsilon$ **then** N>L>N

*D. Identification of patterns with changes from the high region*

18. **If** $\text{First}_\sigma$ $(S_{00}) = H \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{Last}_\sigma(S_{22}) = N$ $\wedge$ $\text{Low}(S_{00}) < \varepsilon$ **then** H>N
19. **If** $\text{First}_\sigma$ $(S_{00}) = H \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{Normal}(S_{00}) > \delta$ $\wedge$ $\text{Last}_\sigma(S_{22}) = H \wedge \text{Trend}(S_{22}) = I \wedge \text{Low}(S_{00}) < \varepsilon$ **then** H>N>H

*E. Identification of patterns with changes from the low region*

20. **If** $\text{First}_\sigma$ $(S_{00}) = L \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{Last}_\sigma(S_{22}) = N$ $\wedge$ $\text{Low}(S_{00}) < \varepsilon$ **then** L>N
21. **If** $\text{First}_\sigma$ $(S_{00}) = L \wedge \text{Cross}(S_{00}) < \alpha \wedge \text{Normal}(S_{00}) > \delta \wedge$ $\text{Last}_\sigma(S_{22}) = L \wedge \text{Trend}(S_{22}) = D \wedge \text{High}(S_{00}) < \varepsilon$ **then** L>N>L
22. **If** NULL **Then** Undetermined.

**Fig. 5.** Basic TA algorithm for long-term changed tests

## 4. Complex Temporal Abstractions

### 4.1 The statistical significance of discovered knowledge

The complex temporal abstractions in our framework are achieved by applying data mining algorithms to the datasets obtained by basic TA algorithms. The data mining systems to be applied here are the rule induction program LUPC of system D2MS [4] and the association mining program in system Clementine. As usual, large numbers of association rules or prediction rules were found. A critical question is which of them are statistically significant and were not found due to chance?

**Table 1**. Ratio of statistical significant rules over rules found

| Sup | Conf | Significant/Total | Sup | Conf | Significant/Total |
|---|---|---|---|---|---|
| 1 | 75 | 341/4530 | 5 | 95 | 15/48 |
| 1 | 95 | 41/1158 | 7 | 75 | 124/256 |
| 2 | 75 | 239/1585 | 10 | 75 | 95/140 |
| 2 | 95 | 26/215 | 10 | 80 | 30/67 |
| 3 | 75 | 197/949 | 10 | 90 | 7/27 |
| 3 | 95 | 21/102 | 20 | 75 | 28/30 |
| 5 | 75 | 146/429 | 30 | 75 | 11/11 |
| 5 | 90 | 23/101 | 40 | 75 | 2/2 |

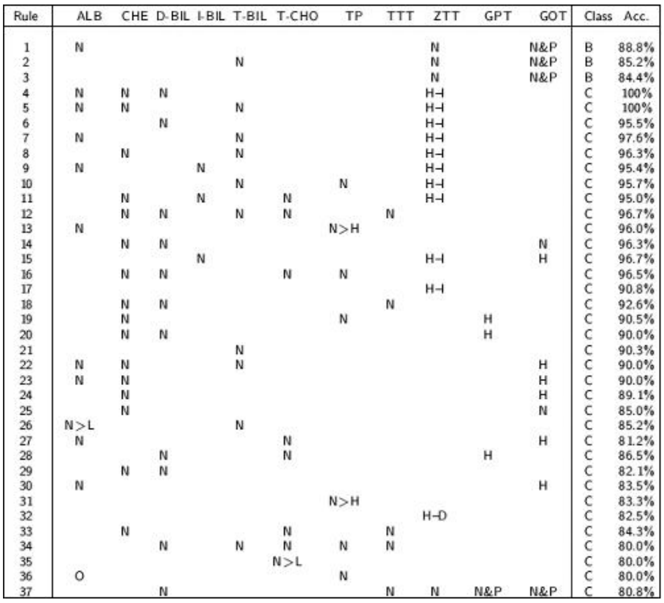
In this work we employed the method of evaluation of statistical significance in [2]. This method carries out three kinds of hypothesis testing: (1) the significance of the consequence of the rule; (2) the significance of antecedent of the rule; and (3) the significance of the accuracy/confidence of the rule. **Table 1** shows the ratio of rules verified to be statistically significant with 5% of significance level over the total number of rules found for the problem P1 on type B and C of hepatitis. The next subsections present rules found and checked their statistical significance.

### 4.2 Rules on hepatitis B and C

**Table 2** shows a set of rules concerning the difference between hepatitis B and C obtained by LUPC, from which we can draw some remarks as follows.

- The tests ALB, CHE, D-BIL, TP, and ZTT are frequent in rules.
- The appearance of tests GPT and GOT might not be necessary to distinguish types B and C of hepatitis, but GOT value of the rules describing the type B is “N&P”, while many are “H” for the type C.
- There are not many rules with large coverage for type B.
- Rule 32 is simple and interesting as it confirms that among four typical short-term changed tests, TTT and ZTT have sensitivity to inflammation but they do not have enough specificity to liver inflammation. The rule says that “if ZTT is high but decreasing we can predict the type C with accuracy 83%”.
- Rule 29 “IF CHE = N and D-BIL = N THEN Class = C” is also typical for type C as it covers a large population of the class (173/272 or 63.6%) with accuracy 82.08%.

**Table 2.** Rules for classifying hepatitis B and C



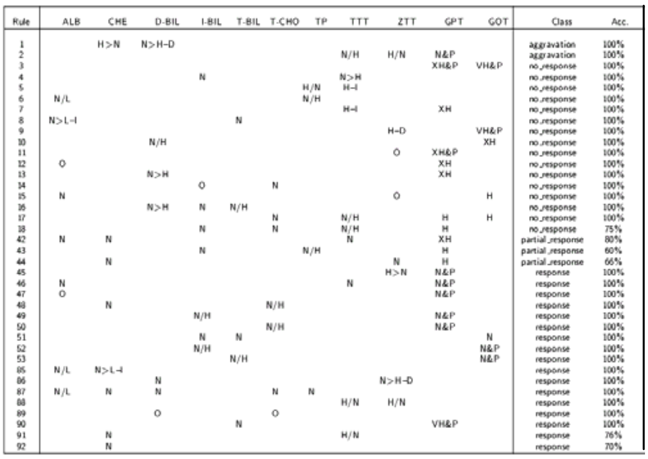
| Rule | ALB | CHE | D-BIL | I-BIL | T-BIL | T-CHO | TP | TTT | ZTT | GPT | GOT | Class | Acc. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | N | | | | | | | | N | | N&P | B | 88.8% |
| 2 | | | | | N | | | | N | | N&P | B | 85.2% |
| 3 | | | | | | | | | N | | N&P | B | 84.4% |
| 4 | N | N | N | | | | | | H-I | | | C | 100% |
| 5 | N | N | | | N | | | | H-I | | | C | 100% |
| 6 | | | N | | | | | | H-I | | | C | 95.5% |
| 7 | N | | | | N | | | | H-I | | | C | 97.6% |
| 8 | | N | | | N | | | | H-I | | | C | 96.3% |
| 9 | N | | | N | | | | | H-I | | | C | 95.4% |
| 10 | | | | | N | | N | | H-I | | | C | 95.7% |
| 11 | | N | | N | | N | | | H-I | | | C | 95.0% |
| 12 | | N | N | | N | N | | N | | | | C | 96.7% |
| 13 | N | | | | | | N>H | | | | | C | 96.0% |
| 14 | | N | N | | | | | | | | N | C | 96.3% |
| 15 | | | | N | | | | | H-I | | H | C | 96.7% |
| 16 | | N | N | | | N | N | | | | | C | 96.5% |
| 17 | | | | | | | | | H-I | | | C | 90.8% |
| 18 | | N | N | | | | | N | | | | C | 92.6% |
| 19 | | N | | | | | N | | | H | | C | 90.5% |
| 20 | | N | N | | | | | | | H | | C | 90.0% |
| 21 | | | | | N | | | | | | | C | 90.3% |
| 22 | N | N | | | N | | | | | | H | C | 90.0% |
| 23 | N | N | | | | | | | | | H | C | 90.0% |
| 24 | | N | | | | | | | | | H | C | 89.1% |
| 25 | | N | | | | | | | | | N | C | 85.0% |
| 26 | N>L | | | | N | | | | | | | C | 85.2% |
| 27 | N | | | | | N | | | | | H | C | 81.2% |
| 28 | | | N | | | N | | | | H | | C | 86.5% |
| 29 | | N | N | | | | | | | | | C | 82.1% |
| 30 | N | | | | | | | | | | H | C | 83.5% |
| 31 | | | | | | | N>H | | | | | C | 83.3% |
| 32 | | | | | | | | | H-D | | | C | 82.5% |
| 33 | | N | | | | N | | N | | | | C | 84.3% |
| 34 | | | N | | N | N | N | N | | | | C | 80.0% |
| 35 | | | | | | N>L | | | | | | C | 80.0% |
| 36 | O | | | | | | N | | | | | C | 80.0% |
| 37 | | | N | | | | | N | N | N&P | N&P | C | 80.8% |

### 4.3 Rule on effectiveness of interferon therapy

**Table 3** shows rules found for two classes of response and non-response cases to interferon. It can be observed that many rules describing the non-response class are with patterns on GPT and/or GOT having values “XH&P”, “VH&P”,

"XH", or "H", while many rules describing the response class are with patterns on GPT or GOT having values "N&P" or "H&P". The results allows us to hypothesize that the interferon treatment may have strong effectiveness on peaks (suddenly increasing in a short period) if the base state is normal or high. It can be hypothesized that when the base state is very high or extremely high, the interferon treatment is not clearly effective.

**Table 3** Rules on responses to IFN

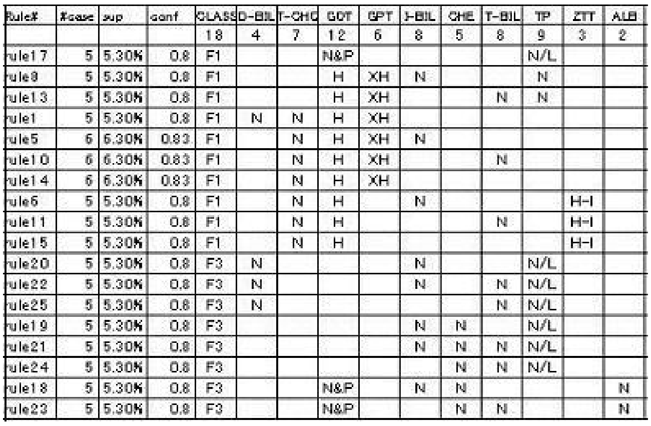


| Rule | ALB | CHE | D-BIL | I-BIL | T-BIL | T-CHO | TP | TTT | ZTT | GPT | GOT | Class | Acc. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | H>N | N>H–D | | | | | | | | | aggravation | 100% |
| 2 | | | | | | | | N/H | H/N | N&P | | aggravation | 100% |
| 3 | | | | | | | | | | XH&P | VH&P | no_response | 100% |
| 4 | | | | N | | | | N>H | | | | no_response | 100% |
| 5 | | | | | | | H/N | H–I | | | | no_response | 100% |
| 6 | N/L | | | | | | N/H | | | | | no_response | 100% |
| 7 | | | | | | | | H–I | | XH | | no_response | 100% |
| 8 | N>L–I | | | | N | | | | | | | no_response | 100% |
| 9 | | | | | | | | | H–D | | VH&P | no_response | 100% |
| 10 | | | N/H | | | | | | | | XH | no_response | 100% |
| 11 | | | | | | | | | O | XH&P | | no_response | 100% |
| 12 | O | | | | | | | | | XH | | no_response | 100% |
| 13 | | | N>H | | | | | | | XH | | no_response | 100% |
| 14 | | | | O | | N | | | | | | no_response | 100% |
| 15 | N | | | | | | | | O | | H | no_response | 100% |
| 16 | | | N>H | N | N/H | | | | | | | no_response | 100% |
| 17 | | | | | | N | | N/H | | H | H | no_response | 100% |
| 18 | | | | N | | N | | N/H | | H | | no_response | 75% |
| 42 | N | N | | | | | | N | | XH | | partial_response | 80% |
| 43 | | | | N | | | N/H | | | H | | partial_response | 60% |
| 44 | | N | | | | | | | N | H | | partial_response | 66% |
| 45 | | | | | | | | | H>N | N&P | | response | 100% |
| 46 | N | | | | | | | N | | N&P | | response | 100% |
| 47 | O | | | | | | | | | N&P | | response | 100% |
| 48 | | N | | | | N/H | | | | | | response | 100% |
| 49 | | | | N/H | | | | | | N&P | | response | 100% |
| 50 | | | | | | N/H | | | | N&P | | response | 100% |
| 51 | | | | N | N | | | | | | N | response | 100% |
| 52 | | | | N/H | | | | | | | N&P | response | 100% |
| 53 | | | | | N/H | | | | | | N&P | response | 100% |
| 85 | N/L | N>L–I | | | | | | | | | | response | 100% |
| 86 | | | N | | | | | | N>H–D | | | response | 100% |
| 87 | N/L | N | N | | | N | N | | | | | response | 100% |
| 88 | | | | | | | | H/N | H/N | | | response | 100% |
| 89 | | | O | | | O | | | | | | response | 100% |
| 90 | | | | | N | | | | | VH&P | | response | 100% |
| 91 | | N | | | | | | H/N | | | | response | 76% |
| 92 | | N | | | | | | | | | | response | 70% |

### 4.3 Rules on fibrosis stages

**Table 4** is a set of rules for problem P3 obtained by Apriori algorithm. It contains 10 rules discovered for fibrosis stages F1 and 8 rules for fibrosis stage F3. According to the table, the first rule describing fibrosis stage F1 can be read as "if GOT = N&P and TP = N/L then the class is F1". It is interesting that the rules describing fibrosis stage F1 and F3 are well separated:

**Table 4.** Discovered association rules and their coverage with min_sup = 5% and min_conf = 80%



| Rule# | #case | sup | conf | CLASS | D-BIL | T-CHO | GOT | GPT | I-BIL | CHE | T-BIL | TP | ZTT | ALB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 18 | 4 | 7 | 12 | 6 | 8 | 5 | 8 | 9 | 3 | 2 |
| rule17 | 5 | 5.30% | 0.8 | F1 | | | N&P | | | | | N/L | | |
| rule8 | 5 | 5.30% | 0.8 | F1 | | | H | XH | N | | | N | | |
| rule13 | 5 | 5.30% | 0.8 | F1 | | | H | XH | | | N | N | | |
| rule1 | 5 | 5.30% | 0.8 | F1 | N | N | H | XH | | | | | | |
| rule5 | 6 | 6.30% | 0.83 | F1 | | N | H | XH | N | | | | | |
| rule10 | 6 | 6.30% | 0.83 | F1 | | N | H | XH | | | N | | | |
| rule14 | 6 | 6.30% | 0.83 | F1 | | N | H | XH | | | | | | |
| rule6 | 5 | 5.30% | 0.8 | F1 | | N | H | | N | | | | H–I | |
| rule11 | 5 | 5.30% | 0.8 | F1 | | N | H | | | | N | | H–I | |
| rule15 | 5 | 5.30% | 0.8 | F1 | | N | H | | | | | | H–I | |
| rule20 | 5 | 5.30% | 0.8 | F3 | N | | | | N | | | N/L | | |
| rule22 | 5 | 5.30% | 0.8 | F3 | N | | | | N | | N | N/L | | |
| rule25 | 5 | 5.30% | 0.8 | F3 | N | | | | | | N | N/L | | |
| rule19 | 5 | 5.30% | 0.8 | F3 | | | | | N | N | | N/L | | |
| rule21 | 5 | 5.30% | 0.8 | F3 | | | | | N | N | N | N/L | | |
| rule24 | 5 | 5.30% | 0.8 | F3 | | | | | | N | N | N/L | | |
| rule18 | 5 | 5.30% | 0.8 | F3 | | | N&P | | N | N | | | | N |
| rule23 | 5 | 5.30% | 0.8 | F3 | | | N&P | | | N | N | | | N |

– The rules describing the fibrosis stage F1 except the first one are typically related to the combinations of "GOT = H and GPT = XH and (T-CHO = N or TP = N)", or "T–CHO = N and GOT = H and ZTT = H–I".

– The rules describing the fibrosis stage F3 can be distinguished from those of F1 by the combinations "TP = N/L and (D-BIL = N or CHE = N)", or "GOT = N&P and CHE = N".

## 5. Conclusions

We have presented a TA approach to mining long-term changed tests in temporal hepatitis data. Our TA approach differs from related temporal abstraction approaches in dealing with irregular time stamped data of multiple variables.

The findings by our temporal abstraction methods are positively evaluated by medical doctors in terms of novelty, acceptability and utility. They have evaluated many found patterns (rules) as new and interesting, and the sets of rules partially answered the problems under consideration (P1, P2, and P3).

**Acknowledgments:**

This research is supported by the project "Realization of Active Mining in the Era of Information Flood", Grant-in-aid for scientific research on priority areas (B). The authors would like to express their sincere thanks to doctors Katsuhiko Takabayashi and Hideto Yokoi of Chiba University Hospital for their guidance and collaboration.

**References**

[1] Bellazzi, R., Larizza, C., Magni, P., Monntani, S., and Stefanelli, M., "Intelligent Analysis of Clinic Time Series: An Application in the Diabetes Mellitus Domain", Artificial Intelligence in Medicine 20, pp. 37-57, 2000.

[2] Bruzzese, D. and Davino, C., "Statistical Pruning of Discovered Association Rules", Computational Statistics 16 (3), pp. 387 -398, 2001.

[3] Ho, T.B., Nguyen, T.D., Kawasaki, S., Le, S.Q., Nguyen, D.D., Yokoi, H., Takabayashi, K., "Mining Hepatitis Data with Temporal Abstraction", ACM International Conference on Knowledge Discovery and Data Mining KDD-03, Washington DC, 24-27 August (in press).

[4] Ho, T.B., Nguyen, T.D., Nguyen, D.D., and Kawasaki, S., "Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining", International Journal of Artificial Intelligence Tools, World Scientific, Vol. 10, No. 4, pp. 691-713, 2001.

[5] Horn, W., Miksch, S., Egghart, G., Popow, C., and Paky, F., "Effective Data Validation of High-Frequency Data: Time-Point-, Time-Interval-, and Trend-Based Methods", Computer in Biology and Medicine, Special Issue: Time-Oriented Systems in Medicine, 27(5), pp. 389-409, 1997.

[6] Lavrak, N., "Selected Techniques for Data Mining in Medicine", Artificial Intelligence in Medicine, 16, pp. 3-23, 1999.

[7] Miksch S., Horn W., Popow C., and Paky F., "Utilizing Temporal Data Abstraction for Data Validation and Therapy Planning for Artificially Ventilated Newborn Infants", Artificial Intelligence in Medicine, 8(6), pp. 543-576, 1996.

[8] Motoda, H., Active Mining: New directions of data mining (Ed.), IOS Press, 2002.

[9] Shahar, Y., "A Framework for Knowledge-based Temporal Abstraction", Artificial Intelligence, 90, pp. 79-133, 1997.

**Name:**
Saori Kawasaki

**Affiliation:**
School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST)

**Address:**
1-1 Asahidai, Tatsunokuchi, Ishikawa, 923-1292 Japan

**Brief Biographical History:**

1989: B.A., Kyushu University in 1989

2000: Master, Japan Advanced Institute of Science and Technology

2003 Ph.D., Japan Advanced Institute of Science and Technology

**Main Works:**

- Mining from Medical Data: Case-Studies in Meningitis and Stomach Cancer Domains, *KESS 2002, 6th International Conference on Knowledge-based Intelligent Information & Engineering Systems*, Crema, September 16-18, 547-551 (2002).
- Cluster-based Information Retrieval with a Tolerance Rough Set Model, *International Journal of Fuzzy Logic and Intelligent Systems*, KFIS, Vol. 2, No. 1, 26-32 (2002).

**Membership in Learned Societies:**

**Name:**
Trong Dung Nguyen

**Affiliation:**
Research Associate, School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST)

**Address:**
1-2 Asahidai, Tatsunokuchi, Ishikawa, 923-1292 Japan

**Brief Biographical History:**

2000: Ph.D., Japan Advanced Institute of Science and Technology

2000-2003, Associate, Japan Advanced Institute of Science and Technology

**Main Works:**

- Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining, *International Journal of Artificial Intelligence Tools*, World Scientific, Vol. 10, No. 4, 691-713 (2001)
- An Interactive-Graphic System for Decision Tree Induction, *Journal of Japanese Society for Artificial Intelligence,* Vol. 14, N. 1, 131-138 (1999).

**Membership in Learned Societies:**

**Name:**
Tu Bao Ho

**Affiliation:**
Professor, School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST)

**Address:**
1-3 Asahidai, Tatsunokuchi, Ishikawa, 923-1292 Japan

**Brief Biographical History:**

1987: Ph.D., University Paris 6

1991: Associate Professor, Institute of Information Technology, Vietnam

1998: Professor, Japan Advanced Institute of Science and Technology

**Main Works:**

- "A Knowledge Discovery System with Support for Model Selection and Visualization", *Applied Intelligence*, Vol. 19, No. 1-2, 125-141 (2003).
- "Chance Discovery and Learning Minority Classes", *Journal of New Generation Computing*, Vol. 21, No. 2, 147-160 (2003).
- "Nonhierarchical Document Clustering by a Tolerance Rough Set Model", *International Journal of Intelligent Systems*, Vol. 17, No. 2, 199-212 (2002).

**Membership in Learned Societies:**

- JSAI (Japan Society for Artificial Intelligence)
- IEEE (The Institute of Electrical and Electronics Engineers)