2007 IEEE International Conference on Bioinformatics and Biomedicine

Mining Clinical Data with a Temporal Dimension: a Case Study

Michele Berlingerio
IMT Lucca Insitute for
Advanced Studies, Italy

Francesco Bonchi Fosca Giannotti
Pisa KDD Laboratory
ISTI - CNR, Italy

Franco Turini
Computer Science Dept.,
University of Pisa, Italy

Abstract

Clinical databases store large amounts of information about patients and their medical conditions. Data mining techniques can extract relationships and patterns holding in this wealth of data, and thus be helpful in understanding the progression of diseases and the efficacy of the associated therapies.

A typical structure of medical data is a sequence of observations of clinical parameters taken at different time moments. In this kind of contexts, the temporal dimension of data is a fundamental variable that should be taken in account in the mining process and returned as part of the extracted knowledge. Therefore, the classical and well established framework of sequential pattern mining is not enough, because it only focuses on the sequentiality of events, without extracting the typical time elapsing between two particular events. Time-annotated sequences ($\mathcal{TAS}$), is a novel mining paradigm that solves this problem. Recently defined in our laboratory together with an efficient algorithm for extracting them, $\mathcal{TAS}$ are sequential patterns where each transition between two events is annotated with a typical transition time that is found frequent in the data.

In this paper we report a real-world medical case study, in which the $\mathcal{TAS}$ mining paradigm is applied to clinical data regarding a set of patients in the follow-up of a liver transplantation. The aim of the data analysis is that of assessing the effectiveness of the extracorporeal photopheresis (ECP) as a therapy to prevent rejection in solid organ transplantation. For each patient, a set of biochemical variables is recorded at different time moments after the transplantation. The $\mathcal{TAS}$ patterns extracted show the values of interleukins and other clinical parameters at specific dates, from which it is possible for the physician to assess the effectiveness of the ECP therapy.

We believe that this case study does not only show the interestingness of extracting $\mathcal{TAS}$ patterns in this particular context but, more ambitiously, it suggests a general methodology for clinical data mining, whenever the time dimension is an important variable of the problem in analysis.

1. Introduction

With the increasing proliferation of information systems in modern hospitals and health-care institutions, the volume of available medical and biological data is increasing accordingly. However, collecting the data is not enough: we need to develop appropriate data analysis tools to extract relevant information from this wealth of data. In medicine, overcoming the gap between data gathering and comprehension is particularly crucial since medical decision making needs to be supported by arguments based on the evidence of regularities, patterns and trends holding in the data. Medical data mining is an emerging and promising research field aiming to overcome this gap. The collaboration among physicians, biologists and computer scientist promises on the one hand to produce new important medical knowledge and, on the other hand, to devise new data analysis tools, whose relevance can only be proven by their concrete utility in the medical institutions, and their acceptability by the medical experts.

In this perspective, in Pisa (Italy) we have started an important data collection and analysis project, where a very large number of epidemiological, clinical, immunological and genetical variables collected before the transplantation of a solid organ, and during the follow-up assessment of the patients, are stored in a datawarehouse for future mining [2]. This on-going data collection involves all liver, kidney, pancreas and kidney-pancreas transplantations of the last five years of one of the largest (for number of transplantations) centers in Europe. In this paper we describe one of the case studies developed within this project. The presented case study is interesting both by a medical and a data analysis point of view.

The interestingness of the case study by the medical perspective lies in the uniqueness and relevance of the dataset in analysis. While by the data analysis perspective, the interestingness lies (i) in the structure of the data in analysis, (ii) in the importance of the temporal dimension, and (iii) in the repeatabil-

0-7695-3031-1/07 $25.00 © 2007 IEEE
DOI 10.1109/BIBM.2007.42

ity of this experience to other medical data analyses where the data exhibits the same structure. In fact, a typical structure of medical data is a sequence of observations of clinical parameters taken at different time moments. Here the time dimension is crucial: the focus is not only on the observed values, nor only in the sequence their compose, but it is also very important the typical time that elapses among two events. For instance, the time elapsed from the start of a certain therapy, to the appearance of a certain clinical phenomenon.

The main contribution of this paper is to provide a methodological approach to this kind of data, by means of Time-Annotated Sequences ($\mathcal{TAS}$) mining [5, 6]. $\mathcal{TAS}$ patterns extraction, is a novel mining paradigm defined in our laboratory together with an efficient algorithm for extracting them. $\mathcal{TAS}$ are sequential patterns where each transition between two events is annotated with a typical transition time that is found frequent in the data. In principle, this form of pattern is useful in several contexts: for instance, (i) in web log analysis, different categories of users (experienced vs. novice, interested vs. uninterested, robots vs. humans) might react in similar ways to some pages - i.e., they follow similar sequences of web access - but with different reaction times; (ii) in medicine, reaction times to patients' symptoms, drug assumptions and reactions to treatments are a key information. In all these cases, enforcing fixed time constraints on the mined sequences is not a solution. It is desirable that typical transition times, when they exist, emerge from the input data. $\mathcal{TAS}$ patterns have been also used as basic brick to build a truly spatio-temporal trajectory pattern mining framework [7]. The rest of the paper is organized as follows. In Section 2 we introduces the $\mathcal{TAS}$ paradigm and how this can be used for mining time-annotated data. Section 3 describes the real-world case study in which the $\mathcal{TAS}$ paradigm is applied to the photopheresis dataset. Section 4 summarizes the contributions and the results of this paper, and draws some future lines of research.

2. From Sequential Patterns to $\mathcal{TAS}$

In this section we recall Sequential Pattern Mining (introduced first in [1]), how it has evolved in the $\mathcal{TAS}$[5, 6] paradigm, and how this can be used as a general paradigm for time-annotated data.

2.1. Sequential Pattern Mining

Frequent Sequential Pattern mining (FSP) deals with the extraction of frequent sequences of events from datasets of transactions; those, in turn, are time-stamped sequences of events (or sets of events) observed in some business contexts: customer transactions, patient medical observations, web sessions, trajectories of objects moving among locations.

The *frequent sequential pattern* (FSP) problem is defined over a database of sequences $\mathcal{D}$, where each element of each sequence is a time-stamped set of objects, usually called items. Time-stamps determine the order of elements in the sequence. E.g., a database can contain the sequences of visits of customers to a supermarket, each visit being time-stamped and represented as the set of items bought together. Then, the FSP problem consists in finding all the sequences that are frequent in $\mathcal{D}$, i.e., appear as subsequence of a large percentage of sequences of $\mathcal{D}$. A sequence $\alpha = \alpha_1 \rightarrow ... \rightarrow \alpha_k$ is a subsequence of $\beta = \beta_1 \rightarrow ... \rightarrow \beta_m \quad (\alpha \preceq \beta)$ if there exist integers $1 \leq i_1 < ... < i_k \leq m$ such that $\forall_{1 \leq n \leq m}\ \alpha_n \subseteq \beta_{i_n}$. Then we can define the support $sup_{\mathcal{D}}(S)$ of a sequence S as the number of transactions $T \in \mathcal{D}$ such that $S \preceq T$, and say that S is frequent w.r.t. threshold σ is $sup_{\mathcal{D}}(S) \geq \sigma$.

Recently, several algorithms were proposed to efficiently mine sequential patterns, among which we mention PrefixSpan [13], that that employs an internal representation of the data made of database projections over sequence prefixes, and SPADE [15], a method employing efficient lattice search techniques and simple joins that needs to perform only three passes over the database. Alternative methods have been proposed, which add constraints of different types, such as *min-gap*, *max-gap*, *max-windows* constraints and regular expressions describing a subset of allowed sequences. We refer to [16] for a wider review of the state-of-art on sequential pattern mining.

2.2. The $\mathcal{TAS}$ mining paradigm

As one can notice, time in FSP is only considered for the sequentiality that it imposes on events, or used as a basis for user-specified constraints to the purpose of either preprocessing the input data into ordered sequences of (sets of) events, or as a pruning mechanism to shrink the pattern search space and make computation more efficient. In either cases, time is forgotten in the output of FSP. For this reason, the $\mathcal{TAS}$, a form of sequential patterns annotated with temporal information representing typical transition times between the events in a frequent sequence, was introduced in [6].

Definition 2.1 ($\mathcal{TAS}$) *Given a set of items $\mathcal{I}$, a* temporally-annotated sequence *of length $n > 0$, called n-$\mathcal{TAS}$ or simply $\mathcal{TAS}$, is a couple $T = (\overline{s}, \overline{\alpha})$, where $\overline{s} = \langle s_0, ..., s_n \rangle, \forall_{0 \leq i \leq n} s_i \in 2^{\mathcal{I}}$ is called the* se-

quence, *and* $\overline{\alpha} = \langle \alpha_1, ..., \alpha_n \rangle \in \mathbb{R}_+^n$ *is called the* (temporal) annotation. *$\mathcal{TAS}$ will also be represented as follows:*

$$T = (\overline{s}, \overline{\alpha}) = s_0 \xrightarrow{\alpha_1} s_1 \xrightarrow{\alpha_2} ... \xrightarrow{\alpha_n} s_n$$

Example 2.2 *In a weblog context, web pages (or pageviews) represent items and the transition times from a web page to the following one in a user session represent annotations. E.g.:*

$$(\langle \{'/'\}, \{'/papers.html'\}, \{'/kdd.html'\} \rangle, \langle 2, 90 \rangle) = \{'/'\} \xrightarrow{2} \{'/papers.html'\} \xrightarrow{90} \{'/kdd.html'\}$$

represents a sequence of pages that starts from the root, then after 2 seconds continues with page 'papers.html' and finally, after 90 seconds ends with page 'kdd.html'. Notice that in this case all itemsets of the sequence are singletons.

Similarly to traditional sequence pattern mining, it is possible to define a containment relation between annotated sequences:

Definition 2.3 (τ-containment ($\preceq_\tau$)) *Given a n-$\mathcal{TAS}$ $T_1 = (\overline{s_1}, \overline{\alpha_1})$ and a m-$\mathcal{TAS}$ $T_2 = (\overline{s_2}, \overline{\alpha_2})$ with $n \leq m$, and a time threshold τ, we say that T_1 is τ-contained in T_2, denoted as $T_1 \preceq_\tau T_2$, if and only if there exists a sequence of integers $0 \leq i_0 < ... < i_n \leq m$ such that:*

1. $\forall_{0 \leq k \leq n}.\ s_{1,k} \subseteq s_{2,i_k}$
2. $\forall_{1 \leq k \leq n}.\ |\ \alpha_{1,k} - \alpha_{*,k}\ | \leq \tau$

where $\forall_{1 \leq k \leq n}.\ \alpha_{*,k} = \sum_{i_{k-1} < j \leq i_k} \alpha_{2,j}$. As special cases, when condition 2 holds with the strict inequality we say that T_1 is strictly τ-contained *in T_2, denoted with $T_1 \prec_\tau T_2$, and when $T_1 \preceq_\tau T_2$ with $\tau = 0$ we say that T_1 is* exactly contained *in T_2. Finally, given a set of $\mathcal{TAS}$ D, we say that T_1 is τ-contained in D ($T_1 \preceq_\tau D$) if $T_1 \preceq_\tau T_2$ for some $T_2 \in D$.*

Essentially, a $\mathcal{TAS}$ T_1 is τ-contained into another one, T_2, if the former is a subsequence of the latter and its transition times do not differ too much from those of its corresponding itemsets in T_2. In particular, each itemset in T_1 can be mapped to an itemset in T_2. When two itemsets are consecutive in T_1 but their mappings are not consecutive in T_2, the transition time for the latter couple of itemsets is computed summing up the times of all the transitions between them, which is exactly the definition of annotations α_*. The following example describes a sample computation of τ-containment between two $\mathcal{TAS}$:

Example 2.4 *Consider two $\mathcal{TAS}$:*

$$T_1 = (\langle \{a\}, \{b\}, \{c\} \rangle, \langle 4, 9 \rangle)$$
$$T_2 = (\langle \{a\}, \{b, d\}, \{f\}, \{c\} \rangle, \langle 3, 7, 4 \rangle)$$

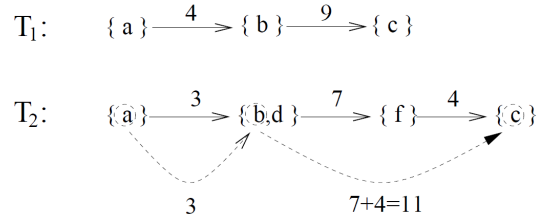

Figure 1. Example of τ-containment computation

also depicted in Figure 1, and let $\tau = 3$. Then, in order to check if $T_1 \preceq_\tau T_2$, we verify that:

- *$\overline{s_1} \subset \overline{s_2}$:in fact the first and the last itemsets of T_1 are equal, respectively, to the first and the last ones of T_2, while the second itemset of T_1 ($\{b\}$) is strictly contained in the second one of T_2 ($\{b,d\}$).*
- *The transition times between T_1 and its corresponding subsequence in T_2 are similar: the fist two itemsets of T_1 are mapped to contiguous itemsets in T_2, so we can directly take their transition time in T_2, which is equal to $\alpha_{*,1} = 3$ (from $\{a\} \xrightarrow{3} \{b,d\}$ in T_2). The second and third itemsets in T_1, instead, are mapped to non-consecutive itemsets in T_2, and so the transition time for their mappings must be computed by summing up all the transition times between them, i.e.: $\alpha_{*,2} = 7 + 4 = 11$ (from $\{b,d\} \xrightarrow{7} \{f\}$ and $\{f\} \xrightarrow{4} \{c\}$ in T_2). Then, we see that $|\ \alpha_{1,1} - \alpha_{*,1}\ | = |\ 4 - 3\ | < \tau$ and $|\ \alpha_{1,2} - \alpha_{*,2}\ | = |\ 9 - 11\ | < \tau$.*

Therefore, we have that $T_1 \preceq \tau T_2$. Moreover, since all inequalities hold strictly, we also have $T_1 \prec_\tau T_2$.

Now, frequent sequential patterns can be easily extended to the notion of frequent $\mathcal{TAS}$:

Definition 2.5 (Frequent $\mathcal{TAS}$) *Given a set $\mathcal{D}$ of $\mathcal{TAS}$, a time threshold τ and a minimum support threshold σ, we define the* τ-support *of a $\mathcal{TAS}$ T as $supp_{[\tau,\mathcal{D}]}(T) = |\ \{T^* \in \mathcal{D} \mid T \preceq_\tau T^*\}\ |$ and say that T is* frequent *in $\mathcal{D}$, given a minimum support threshold σ if $supp_{[\tau,\mathcal{D}]}(T) \geq \sigma$.*

It should be noted that a frequent sequence $\overline{s}$ may not correspond to any frequent $\mathcal{TAS}$ $T = (\overline{s}, \overline{\alpha})$: indeed, its occurrences in the database could have highly dispersed annotations, thus not allowing any single annotation $\overline{\alpha} \in \mathbf{R}_+^n$ to be close (i.e., similar) enough to a sufficient number of them. That essentially means $\overline{s}$ has no *typical* transition times.

Now, introducing time in sequential patterns gives rise to a novel issue: intuitively, for any frequent $\mathcal{TAS}$ $T = (\overline{s}, \overline{\alpha})$, we can usually find a vector $\overline{\epsilon}$ of small, strictly positive values such that $T' = (\overline{s}, \overline{\alpha} + \overline{\epsilon})$ is frequent as

well, since they are approximatively contained in the same $\mathcal{TAS}$ in the dataset, and then have very similar τ-support. Since any vector with smaller values than $\overline{\epsilon}$ (e.g., a fraction $\overline{\epsilon}/n$ of it) would yield the same effect, we have that, in general, the raw set of all frequent $\mathcal{TAS}$ is highly redundant (and also not finite, mathematically speaking), due to the existence of several very similar - and then practically equivalent - frequent annotations for the same sequence.

Example 2.6 *Given the following toy database of $\mathcal{TAS}$:*

$$a \xrightarrow{1} b \xrightarrow{2.1} c \quad a \xrightarrow{1.1} b \xrightarrow{1.9} c$$
$$a \xrightarrow{1.2} b \xrightarrow{2} c \quad a \xrightarrow{0.9} b \xrightarrow{1.9} c$$

if $\tau = 0.2$ and $s_{min} = 0.8$ we see that $T = a \xrightarrow{1} b \xrightarrow{2} c$. In General, we can see that any $a \xrightarrow{\alpha_1} b \xrightarrow{\alpha_2} c$ is frequent whenever $\alpha_1 \in [1, 1.1]$ and $\alpha_2 \in [1.9, 2.1]$.

This problem is solved by the $\mathcal{TAS}$ mining software developed in [6], by finding *"dense regions"* of annotations, and thus grouping together $\mathcal{TAS}$ patterns accordingly. The output of this process is a set of $\mathcal{TAS}$ patterns where the annotations are no longer points in $\mathbb{R}^n_+$, but instead are intervals: e.g.,

$$a \xrightarrow{[1,1.1]} b \xrightarrow{[1.9,2.1]} c$$

For more details see [5, 6].

3. Case Study: mining $\mathcal{TAS}$ from photopheresis data

In this section we describe the results obtained applying the $\mathcal{TAS}$ paradigm to medical data, coming from the application of the Photopheresis therapy to patients who had a liver transplant.

3.1. Extracorporeal photopheresis as a therapy against allograft rejection

Originally introduced for treatment of cutaneous T-cell lymphomas [4] and autoimmune diseases [10], extracorporeal photopheresis (ECP) has been reported to be effective to reverse acute heart, lung, and kidney allograft rejection episodes, as well as to treat acute and chronic graft-versus-host disease (GVHD) [8]. ECP is a novel immunomodulatory therapy performed through a temporary peripheral venous access. It has been speculated that ECP modulates alloreactive T-cell responses by multiple mechanisms: induction of apoptosis, inhibition of antigen-driven T-cell proliferation, and reduced expression of cell surface receptors. To date, the body of evidence supporting the use of ECP in the treatment of solid organ graft rejection stems from experiences with heart, lung, and renal transplant recipients.

In the setting of LT, acute graft rejection is nearly always amenable to reversal with steroid pulses. Severe graft rejection episodes or those not responding to steroid treatment may be reversed with lymphocytolytic antibody therapy. However, treatment of rejection is not devoid of complications, namely those related to high-dose immunosuppression and steroid pulses. Therefore, the use of ECP for allograft rejection in LT recipients might represent a valid alternative to overcome the side effects associated with current treatment modalities.

In [14], ECP showed no added morbidity or mortality, was well tolerated, and no procedure-related complications were observed. Its efficacy to reverse rejection was clinically remarkable. The use of ECP allowed reduction in immunosuppression in three out of five patients. Noteworthy, ECP was not associated with HCV or HBV flares in the present experience. Such data need long-term confirmation but may contribute to an expanded application of ECP for patients grafted for viral cirrhosis, thereby helping to reduce the use of steroids in this category of patients. In conclusion, the results in [14] suggest that ECP may represent a valuable alternative to treat graft rejection in selected recipients. Emerging issues awaiting clarification concern indications to ECP, timing and length of treatment, and its cost-effectiveness ratio.

In the case study described in the following, we analyze the dataset collected in [14]. Since prior to [14] only anecdotal reports describe the use of ECP in immunosuppressive protocols for LT recipients, and only one case of a LT recipient treated by ECP has appeared in the international literature [12], we can conclude that this is a unique dataset of this kind.

3.2. Dataset

The dataset in analysis contains information about 127 patients that had a liver transplant and, in order to prevent the allograft rejection, were subjected to the ECP therapy for a period of about 1 year.

Each of them has from few to about 40 observations along the therapy. For each observation 38 different continuous variables are recorded, including information on Red Blood Cells (RBC), Hemoglobin (HB), Hematocrit (HCT), Platelet (PLT), White Blood Cell (WBC), Neutrophils, Monocytes, Lymphocytes, Cluster Of Differentiation 3, 4, 8, 16, 19, 25, 40, Interferon Gamma (INF Gamma), Interleukins 2, 4, 5, 10, 12, Tumor Necrosis Factor (TNF Alfa), Intercellular Adhesion Molecule (ICAM-1), Human Leukocyte Antigen (HLA), Aspartate Aminotransferase (ASAT or GOT, Glutamic Oxalacetic Transferase), Alanine

Aminotransferase (ALAT or GPT, Glutamic Pyruvic Transaminase), Gamma Glutamyl Transpeptidase (GGT), Alkaline Phosphatase, Lactate DeHydrogenase (LDH), Creatine PhosphoKinase (CPK), Bilirubin, Serum bilirubin, Human C Virus (HCV), and Prothrombin Time (PT). Unfortunately, the data contains a lot of missing values, making a lot of observations, and patients, useless for the analysis objectives.

The dataset resulting from the removal of incomplete or useless data, contains complete information for 50 patients. The dataset in analysis also exhibits very peculiar characteristics, such as extremely diverse ranges for each variable for each patient, and non-standard distributions of the values.

For the above reasons, we decided to follow three different approaches, that needed three different kinds of preprocessing, and lead to three different types of results, that are shown below. Obviously, before being able to apply $\mathcal{TAS}$ mining to this dataset, each continuous variable must be discretized somehow, to produce the vocabulary of items. In other terms, in our analysis an item is always a bin of the discretization of a variable. For instance, an item could be "Interleukin 12 $\in [25, 50]$".

For each analysis we received a feedback from the physicians, who suggested how to proceed to the subsequent stages, moving the focus from some interesting patterns to others, including other variables and so on. Tables 1, 2 and 3 show some of the results obtained for the approaches 1, 2 and 3, respectively.

In each table we have 7 columns: in the first one we have the ID of the $\mathcal{TAS}$ found; in the second column we find the values of the involved variables, encoded and discretized for being processed by the $\mathcal{TAS}$ miner; columns 3 and 4 show the support of the $\mathcal{TAS}$ found, both in percentage (0.0 to 1.0) and in absolute value (number of patients presenting the pattern); in column 5 and 6 we have the lower and upper coordinates, respectively, of the pattern (i.e., the minimal and the maximal elapsed time along the sequence); in the last column we have the density of the pattern (i.e., the minimal number of occurrences of the pattern).

Each $\mathcal{TAS}$ found can have several time annotations, corresponding to the same sequence found with different frequent time-stamps. For example, the $\mathcal{TAS}$ n.1 in Table 1 should be read (according to our encoding and discretization of the values) this way: an increment by $10^3 - 10^5$ times of the interleukin 4 was followed by a decrement by at least 85% of the same variable, after 13 to 14 days (according to the first frequent temporal annotation) or after 1 to 2 days (according to the second frequent temporal annotation). $\mathcal{TAS}$ n.6 in Table 1 contains a sequence of length 3 and should be read: a quite stable (around 105%) value of the interleukin 5 followed by (almost) the same value after 91 to 93 or 0 to 2 days, followed again by (almost) the same value after 0 to 1 day or 35 to 36 days.

3.3. First Analysis

Data preparation. n the first analysis, for each variable and for each observation, we focus on the variation w.r.t. the same variable at the previous observation. This has been done with the hope of finding association patterns of "trends": for instance a small decrement of variable V_1 appearing together with a strong increment in variable V_2 is frequently followed after a week by a strong increment of variable V_3 with V_1 staying almost constant.

For this first analysis we only took in consideration the values of the interleukins (2, 4, 5, 10, 12), on explicit request of the physicians.

Results. Table 1 shows some of the results obtained with the approach 1. As one can see, the focus in the first approach was to find some interesting patterns involving the interleukins: in this way we had a first feedback for the adequateness of the $\mathcal{TAS}$ paradigm. As an example, in $\mathcal{TAS}$ n.1, we found an interesting sequence of increase and decrease of the same variable in the 70% of the patients, which is quite a big support. The pattern was supported with two different frequent temporal annotations: the first one finding the elapsed time around 14 days, and the second one very early in the therapy (1 to 2 days after the beginning).

Unfortunately, these patterns were not enough meaningful for the physicians, because the patterns does not say anything about the initial values of the variables, nor about the exact day when the increase or decrease happened. For these two reasons, we decided to change the focus as done in the second analysis.

3.4. Second Analysis

Data preparation. In the second analysis we focused, for each variable, on its variation w.r.t. the clinical "normal" value, without any reference to the variation w.r.t. the previous observation. In order to keep track not only of the elapsed time among the observations, we added as item within all the observations (itemsets) also the discretized information about the day when the observation was taken. The discretization steps for the date follow the ECP therapy milestones: i.e., 7, 14, 30, 90, 180, 270 and 365 days. In this analysis, we took in consideration also other vari-

ID	Pattern	Support (%)	Support (abs)	Low Cords	Up Cords	Density
1	(100000IL4) (15IL4)	0.70	33	13	14	10
				1	2	10
2	(15IL4) (100000IL4)	0.70	33	7	8	10
3	(15IL4) (105IL5)	0.61	29	8	9	10
4	(105IL5) (105IL5)	0.63	30	21	22	10
				19	20	10
				6	8	10
				14	15	10
5	(100000IL4) (15IL4 45IL10)	0.17	8	15	16	5
6	(105IL5) (105IL5) (105IL5)	0.42	20	91 0	93 1	5
				0 35	2 36	5
7	(105IL5 60IL10) (100000IL4)	0.31	15	59	60	5

Table 1. Some of the results of the first analysis

ables: GOT, GPT, GGT, Alkaline Phosphatase, HCV and Bilirubin.

Results. Table 2 reports some of the patterns extracted during the second analysis. Adding more variables and changing the items meaning, we obtained more detailed and meaningful results. For example, the $\mathcal{TAS}$ n.1 in Table 2 shows a very low value of the interleukin 12, followed by an even lower value, after 7 to 10 days.

However, the main limit of this analysis is the lack of focus on interesting, surprising patterns, due to the strong prevalence of "normal" values that hide all the other patterns. Based on this consideration we moved to another analysis.

3.5. Third Analysis

Data preparation. In the third analysis, with the aim of making interesting patterns arise, we changed the discretization by creating for each variable 7 (almost) equally populated bins. Moreover, after applying a couple of runs of the $\mathcal{TAS}$ miner, we also decided to remove the classes with the values in the most normal ranges, in order to make the particular values come out more easily.

Results. Table 3 shows some of the results obtained during the third analysis. With this approach, we divided the values for each variable in several classes, putting (almost) the same amount of occurrences per class. We also removed most of the normality range values to get more interesting patters. As also done in the second analysis, we also put the date information itself as a variable (DD). In this way we obtained very meaningful patterns, showing interesting values of the variables and the date when the values appeared in the medical observations. For example, $\mathcal{TAS}$ n.2 and 3 in Table 3 showed very strange values of the interleukins. During the pattern evaluation, the physicians guessed that this strange behaviour could represent a early warning for allograft rejection. Checking the patients that support the two patterns, we found out that physicians' guess was actually correct.

This third analysis raised a lot of interest in the physicians and the biologists since it produced some frequent $\mathcal{TAS}$ which were interesting for them. In the future analyses we will take in consideration more variables, and hopefully the physicians will provide us with a larger number of patients. We are also studying how to cross-over these patterns with other model of extracted knowledge regarding information that do not change with time, e.g., some particular genetic patterns of the patients. One possible objective is to develop a prediction model able to rise early warnings for patients for whom the ECP therapy is not working adequately, and thus could result in allograft rejection.

4. Conclusions and Future Work

In this paper we have described a pattern discovery analysis that we conducted on clinical data of patients in the follow-up of a liver transplantation. The aim of the data analysis was that of assessing the effectiveness of the extracorporeal photopheresis as a therapy to prevent rejection in solid organ transplantation. To this aim the discovery of associative frequent patterns, describing trends of different biochemical variables along the time dimension is a key step.

The analysis conducted has led to the discovery of some patterns already known by the physicians, and other surprising ones. The feedback we received by the physicians and biologists was very positive: they consider the extracted patterns as additional evidences

ID	Pattern	Support (%)	Support (abs)	Low Cords	Up Cords	Density
1	(50IL12) (25IL12)	0.71	35	7	10	15
2	(50IL12) (1000IL10)	0.61	30	12	13	15
				8	10	10
3	(50IL12) (50IL12)	0.77	38	9	12	15
				13	14	10
4	(50IL12) (1000IL4)	0.69	34	9	10	15
5	(1000BIL) (1000IL10)	0.42	21	12	13	10
6	(1000BIL) (50IL12)	0.44	22	11	12	10
7	(225AP) (50IL12)	0.57	28	8	9	10
8	(300AP) (25IL12)	0.51	25	8	9	10
9	(25IL12) (1000AP)	0.26	13	6	8	10
10	(50IL12) (225AP)	0.53	26	24	25	10
				19	21	10
				1	2	10
11	(25IL12) (1000AP)	0.26	13	6	8	10
12	(50IL12) (225AP)	0.53	26	24	25	10
				19	21	10
				1	2	10
13	(1000IL10) (1000AP)	0.32	16	8	9	10
				6	7	10

Table 2. Some of the results of the second analysis

ID	Pattern	Support (%)	Support (abs)	Low Cords	Up Cords	Density
1	(7DD 113-337GPT) (29-45IL10)	0.38	19	6	7	8
2	(45-672,5IL10) (11,8-19,7IL10 90DD)	0.18	9	36	37	8
3	(0-30IL12) (94-131IL12 90DD)	0.18	9	32	33	8
4	(25,7-54,6IL4) (1-4,2IL4 90DD)	0.22	11	34	35	8
5	(7DD 4,5-41,8BIL) (0,1-1IL4)	0.38	19	6	7	8
6	(4,5-41,8BIL) (14DD 0,1-1IL4)	0.18	9	6	7	8
7	(25,7-54,6IL4) (1,1-1,96BIL 90DD)	0.32	16	36	37	8
				28	29	8
				23	24	8
8	(0,1-1IL4) (1,1-1,96BIL 90DD)	0.26	13	20	22	8
9	(7DD 4,5-41,8BIL) (14DD 0,1-1IL4)	0.18	9	6	7	8

Table 3. Some results of the third analysis

(other than the clinical evidence) of the effectiveness of the ECP therapy.

However, the main contribution of this analysis is methodological:

- it proved adequateness of frequent $\mathcal{TAS}$ pattern discovery in a new domain i.e., medical data mining;
- it described a general methodology that can be replicated in other analyses in the medical domain, whenever the data exhibits a temporal dimension, and the temporal development of the clinical or biochemical variables is part of the knowledge of interest.

In our forthcoming analyses we are going to apply various pattern discovery techniques to different datasets describing different aspects related to solid organ transplantation. The common ambitious goal is to gain deeper insights in all the phenomena related to solid organ transplantation, with the aim of improving the donor-recipient matching policy used nowadays.

Among these analyses, one that is felt particularly promising by the physicians, is aimed at extracting both qualitative and quantitative knowledge about the phenomenon of the *chimerism* [3] in the blood of the recipient in the immediate follow-up of the transplantation. Also in this case, the data is made of a set of biochemical variables recorded at different time mo-

ments, and the typical time elapsing between two relevant events is a very important knowledge. Thus, there are all the conditions to apply again the $\mathcal{TAS}$ mining paradigm. The patterns extracted then can be used as basic bricks to build more complex model, for instance by enriching them with genetical information.

Other analyses we plan to perform by means of pattern discovery techniques, are:

- analyze the polymorphism of the major histocompatibility complex genes in the donor-recipient pairs;
- analyze pairs the polymorphism of the minor histocompatibility complex genes in the donor-recipient;
- assess the influence of the genetic polymorphism of the IL-10 gene [11];
- analyze the genetic polymorphism of genes associated with the chronical reject (ICAM-1, VEGF) [9].

Acknowledgments

We are indebted to Mirco Nanni and Fabio Pinelli for the $\mathcal{TAS}$ mining software. We also wish to thank the biologists and physicians of U.O. Immunoematologia, Azienda Ospedaliero-Universitaria Pisana, for providing us the data and the mining problem described in this paper, in particular Irene Bianco, Michele Curcio, Alessandro Mazzoni and Fabrizio Scatena.

This collaboration between physicians, biologists and computer scientist is carried out within the project *"La Miniera della Salute"*, supported by *"Fondazione Cassa di Risparmio di Pisa"*.

References

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. S. P. Chen, editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

[2] M. Berlingerio, F. Bonchi, S. Chelazzi, M. Curcio, F. Giannotti, and F. Scatena. Mining HLA patterns explaining liver diseases. In *19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006)*, pages 702–707, 2006.

[3] S. Bretagne, M. Vidaud, M. Kuentz, C. Cordonnier, T. Henni, G. Vinci, M. Goossens, and J. Vernant. Mixed blood chimerism in t cell-depleted bone marrow transplant recipients: evaluation using dna polymorphisms. *Circulation Research*, 70:1692–1695, 1987.

[4] R. Edelson, C. Berger, F. Gasparro, B. Jegasothy, P. Heald, B. Wintroub, E. Vonderheid, R. Knobler, K. Wolff, and G. Plewig. Treatment of cutaneous t-cell lymphoma by extracorporeal photochemotherapy. In *N. Engl. J. Med.*, volume 316, pages 297–303, 1987.

[5] F. Giannotti, M. Nanni, and D. Pedreschi. Efficient mining of temporally annotated sequences. In *Proceedings of the Sixth SIAM International Conference on Data Mining*, 2006.

[6] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Mining sequences with temporal annotations. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC)*, pages 593–597, 2006.

[7] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Trajectory patter mining. In *The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.

[8] H. Khuu, R. Desmond, S. Huang, and M. Marques. Characteristics of photopheresis treatments for the management of rejection in heart and lung transplant recipients. In *J. Clin. Apher.*, volume 17(1), pages 27–32, 2002.

[9] I. Kim, S.-O. Moon, S. K. Park, S. W. Chae, and G. Y. Koh. Angiopoietin-1 reduces vegf-stimulated leukocyte adhesion to endothelial cells by reducing icam-1, vcam-1, and e-selectin expression. *Circulation Research*, 89:477–481, 2001.

[10] R. Knobler, W. Graninger, A. Lindmaier, and F. Trautinger. Photopheresis for the treatment of lupus erythematosus. In *Ann. NY Acad. Sci.*, volume 636, pages 340–56, 1991.

[11] T. Kobayashi, I. Yokoyama, S. Hayashi, M. Negita, Y. Namii, T. Nagasaka, H. Ogawa, T. Haba, Y. Tominaga, and K. U. H. Takagi. Genetic polymorphism in the il-10 promoter region in renal transplantation. *Transplant Proc.*, 31:755–756, 1999.

[12] M. Lehrer, E. Ruchelli, K. Olthoff, L. French, and A. Rook. Successful reversal of recalcitrant hepatic allograft rejection by photopheresis. In *Liver Transpl.*, volume 6(5), pages 644–7, 2000.

[13] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *Proceedings of the 17th International Conference on Data Engineering*, pages 215–224, 2001.

[14] L. Urbani, A. Mazzoni, G. Catalano, P. D. Simone, R. Vanacore, C. Pardi, M. Bortoli, G. Biancofiore, D. Campani, V. Perrone, F. Mosca, F. Scatena, and F. Filipponi. The use of extracorporeal photopheresis for allograft rejection in liver transplant recipients. In *Transplant Proc.*, volume 36(10), pages 3068–70, 2004.

[15] M. J. Zaki. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60, 2001.

[16] Q. Zhao and S. Bhowmick. Sequential pattern mining: a survey. In *Technical Report. Center for Advanced Information Systems, School of Computer Engineering, Nanyang Technological University, Singapore*, 2003.