Mining Clinical, Immunological, and Genetic Data of Solid Organ Tansplantation

Michele Berlingerio[1,2], Francesco Bonchi[2], Michele Curcio[3], Fosca Giannotti[2], and Franco Turini[4]

[1] IMT School for Advanced Studies
Via San Micheletto, 3 - Lucca, Italy
[2] Pisa KDD Laboratory ISTI - CNR, Area della Ricerca di Pisa
Via Giuseppe Moruzzi, 1 - Pisa, Italy
[3] Unità di Immunoematologia 2, Azienda Ospedaliero Universitaria Pisana
Via Paradisa, 2 - Pisa, Italy
[4] Pisa KDD Laboratory - Computer Science Dept., University of Pisa
Largo Pontecorvo, 3 - Pisa, Italy

Summary. Clinical databases store large amounts of information about patients and their medical conditions. Data mining techniques can extract relationships and patterns implicit in this wealth of data, and thus be helpful in understanding the progression of diseases and the efficacy of the associated therapies. 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 genetic 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. This on-going data collection involves all liver, kidney, pancreas and kidney-pancreas transplantations of the last five years of one of the largest (as to number of transplantations) centers in Europe. The project 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. In this chapter we report in details two different data mining activities developed within this project. The first analysis involves mining genetic data of patients affected by terminal hepatic cirrhosis with viral origin (HCV and HBV) and patients with terminal hepatic cirrhosis with non-viral origin (autoimmune): the goal is to assess the influence of the HLA antigens on the course of the disease. In particular, we have evaluated if some genetic configurations of the class I and class II HLA are significantly associated with the triggering causes of the hepatic cirrhosis. The second analysis involves clinical data of 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 both analyses we describe in details, the medical context and goal, the nature and structure of the data. We also discuss which kind of data mining technique is the most suitable for our purposes, and we describe the details of the knowledge discovery process followed and extracted knowledge.

A.S. Sidhu et al. (Eds.): Biomedical Data and Applications, SCI 224, pp. 211–236.
springerlink.com

1 Introduction

With recent proliferation of information systems in modern hospitals and health care institutions, there is an increasing volume of medical-related information being collected. Appropriate tools are needed to extract relevant and potentially fruitful knowledge from this wealth of medical data.

Traditionally, statistical data analysis was the final phase of experimental design that, typically, included a careful selection of patients, their features and the definition of the hypothesis to test. With the introduction of data warehouses, such a selective approach to data collection is altered and data may be gathered with no specific purpose in mind. Yet, medical data stored in warehouses may provide a useful resource for potential discovery of new knowledge. The activity of analyzing, for the purpose of knowledge discovery, data that has been collected with no clear pre-defined objective in mind, is usually named Data Mining.

Data Mining is an emerging technology aimed at unlocking the knowledge lying dormant in huge databases, thus closing the gap between data generation and data comprehension. In medicine, overcoming this gap is particularly crucial since medical decision making needs to be supported by arguments based on basic medical knowledge as well as knowledge, regularities and trends extracted from data. There are two main aspects that define the significance of and the need for intelligent data analysis in medicine:

- The first aspect concerns the support of specific knowledge-based problem solving activities (diagnosis, prognosis, monitoring, treatment planning, etc.) through the intelligent analysis of individual patients' raw data. Data are mostly numeric and often quite noisy and incomplete. The aim is to glean, in a dynamic fashion, useful abstractions of the patient's (past, current, and hypothesized future) situation which can be matched against the relevant (diagnostic, prognostic, monitoring, etc.) knowledge for the purposes of the particular problem solving activity.
- The second important aspect concerns the discovery of new medical knowledge that can be extracted through data mining of representative collections of example cases, described by symbolic or numeric descriptors. The available datasets are often incomplete (missing data) and noisy (erroneous). Of particular value to medicine is the requested accuracy and interpretability of the results of data mining. The interpretability may be achieved by representing the results of data mining graphically or by symbolically expressed rules or relationships. To increase the chances of getting useful and interpretable results, data mining can benefit from medical experts who may specify additional (background) knowledge, interact with the mining process, and evaluate its results. Only the accurate patterns and relationships that are expressed at the right level of abstraction in the vocabulary used by medical experts may be of use for a practitioner who will decide whether to adopt and use the extracted knowledge in daily decision making.

In this perspective, we have started an important data collection and mining project, where a very large number of epidemiological, clinical, immunological and genetic 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. In this chapter we report in details two different data mining activities developed within this project. For both analyses we describe in details, the medical context and goal, the nature and structure of the data. We also discuss which kind of data mining technique is the most suitable for our purposes, and we describe the details of the knowledge discovery process followed and extracted knowledge. We believe that this experience can be repeated in other medical data analyses where the data exhibits similar characteristics and structure.

2 The *Health Mine* Project

While the immunosuppressive drugs developed in the past two decades have improved the short-term survival of organ allografts, the effects of these regimens on long-term outcome has not yet been determined. A significant shortcoming of current anti-rejection therapies is that recipients require life-long treatment on an immunosuppressive regimen and are left at greater risk of serious side effects. Thus, it is of significant importance to properly assess the best donor-recipient pair in the transplant context. Human Leukocyte Antigens (HLA, described in deeper details later in Section 4), also known as histocompatibility antigens, are molecules found on all nucleated cells in the body. Histocompatibility antigens help the immune system to recognize whether or not a cell is foreign to the body. Hence the success of an organ transplantation is strongly connected to the HLA systems of the donor-recipient pair. Beyond this important role, the HLA system seems to influence also the clinical course of, for example, hepatic cirrhosis, both on viral and autoimmune basis. However, not only different antigens have different importance w.r.t. hepatic cirrhosis, but, to make things more complicated, other not yet well characterized factors could play an important role. It is thus important to assess the relationships that can hold between HLA patterns, together with other medical and non-medical factors, and the success of a solid organ transplantation.

The "Unità di Immunoematologia 2" of the "Azienda Ospedaliero Universitaria Pisana", a high-specialized Analysis Laboratory, is in charge of studying the patients to be treated with solid organ transplantation. The strong synergy between the laboratory and the transplant group has lead in the last years to excellent results of national relevance. This also thanks to several research projects aiming at understanding post-transplant reject issues. As said before, although the application of last generation immunosuppressive therapies, a minimal percentage of the patients fall into reject. Unfortunately, understanding the the causes and the processes leading to an allograft reject is a difficult task due to the large variety of variables involved,

and to the fact that tipically the physicians miss the analytical instruments needed to explore all the possible combinations of such variables.

In this context we have started a collaboration between the "Unità di Immunoematologia 2" and the Knowledge Discovery and Delivery Laboratory of the ISTI - CNR (National Council of Research), aimed at analyzing by means of data mining techniques the information collected about patients who had a solid organ transplant, with the final goal of extracting knowledge useful for understanding the genetical mechanisms involved in the transplant success. This collaboration as been named the *Health Mine* project, and it as been structured in three subprojects, reflecting different speed in the data collection phase, and different challenges in the data analysis, thus shaping short term and long-term objectives:

1. HLA and liver diseases,
2. Photopheresys assessment,
3. HLA and kidney transplantation.

HLA and liver diseases. Here the medical problem is to find possible associations between HLA alleles combinations and serious liver diseases that lead the patients to organ transplantation. The main objective of this analysis is to find *patterns* involving DNA, gender, age and other variables, which are more frequent in the patients with liver diseases than in the healthy population. In this way we can make hypotheses on possible associations between particular genetic and demographics characteristics and particular kinds of cirrhosis. In this analysis we used Frequent Pattern Discovery techniques, as further explained in Section 4.

Photopheresys assessment. In the last few years, a new type of therapy, the extracorporeal photopheresys (ECP) has started emerging as a helpful therapy when the reject is detected. This treatment can reduce the amount of lymphocytes that sustain the reject process. The objective of this data mining activity is to assess the efficacy of the photopheresys in patients with liver transplant, by analyzing at the same time several kinds of genetical and biochemical variables, in order to individuate which of them can positively influence the success of the photopheretic therapy.

HLA and kidney transplantation. The *chimerism* and the *microchimerism* are phenomena that can occur in the recipient after a transplant. In poor words, they consist in the acquisition, by the recipient's body, of some cells of the donor, due to the organ transplantation. In particular, in this analysis we focus on kidney transplantation, and the objectives are:

- to create a database containing all the clinical variables collected during the follow-up of the recipients,
- to set-up a technique for assessing both from the qualitative and the quantitative point of view the systemic microchimerism by using the HLA polymorphism and the Real-Time technology,

- to assess the relationship among epidemiological, clinical, genetical and immunological variables w.r.t. the study of the polymorphism of the minor histocompatibility antigens.

In the next section we discuss the challenges related to the data collection phase which is still in progress. While in section 4 and 5 we present the two analysis for which the data collection and mining has been concluded in terms of applied methodologies, results and patterns evaluation.

3 Data Collection

The data collection phase in the medical domain is inherently much more challenging that in other domains. It is the nature of the data itself that makes it on the one hand very difficult to be collected (and in reasonably size), and on the other hand very precious for the kind of knowledge that we can acquire by analyzing it. In particular, the reasons behind the hardness of medical data collection are various:

- *Privacy and legal constraints*: due to the highly sensitive nature of the data, the outsourcing of the data mining analysis to researchers external to the medical institution is usually quite complex or in some cases almost impossible. On the one hand the data can hardly be moved *outside* the medical institution, on the other hand external researcher can be hardly put at work *inside* the medical institution.
- *Ownership of the data*: it is very difficult to access the medical data owned by different laboratories, unless they do not cooperate in the same research project, and sometime even if they do.
- *Money-cost of data acquisition*: most of the data come from results of expensive chemical-clinical analyses that are not always possible to perform.
- *Time-cost of data acquisition*: there are no machines that take the blood from the refrigerator, perform automatically the needed analyses, report the results in a standard form, sort, merge and clean the data, and so on. All these operations have to be done manually by people that are not full-time dedicated to this process. Sometimes already existing information are not stored digitally, but physically written on paper, thus even more difficult to access, understand and acquire.

For the reasons above, the most involving data collection, i.e., the one for the third subproject (*HLA and kidney transplantation*) is still going on, while for the other two subprojects is finished.

For the first analysis the database contained 2 datasets: the first one containing data from 534 patients with autoimmune or viral cirrhosis, collected since 1996. For each of the patients information about their age, gender, geographic origin and the alleles corresponding to the six HLA antigens corresponding to loci A, B and DRB1, plus some information about the disease are collected. In the second dataset, used as control population, we had data

of 4718 healthy individuals (coming from the same geographic area), collected since 2001, for which we had the same information (except, obviously, for the disease-related information).

The second database, collected since 2001, contained information about 127 patients that received a liver transplant and, in order to prevent the allograft rejection, were subjected to the ECP therapy for a period of about 1 year. For each patient we had several observations of the same set of variables, taken at different times.

The datasets in analysis are unique in terms of kind of information contained, global amount of patients and variables taken into account. Another unique feature is the various kind of information contained in the datasets. In fact, the first one was aimed at recording all the patients that were starting the entire process towards the organ transplantation. For these patients, the information needed was the HLA characterization, the blood group, the kind of diseases and some personal information like age, gender, and so on. For these patients, only one record containing one single value for each variable was needed. In terms of type of information, this corresponds to the *market basket analysis* case: several baskets (patients) with several items (variable), a single instance for each item.

On the other hand, the second database contained a completely different kind of information. Basically, the medical context in which the data was collected was different: here the physicians kept track of the response of the patients to the ECP treatment. Thus, in addition to the personal and genetical information, that do not change during the time, we had several observations for the same set of variables for each patients. In this case the temporal dimension played an important role, because the focus is not on a particular pattern at a specific time, but rather on the "trend" of some variables, how frequent a kind of response to the treatment can be.

As one can see, these different data need to be treated in different ways, and the next two sections explain the analyses performed on such data, together with the Data Mining techniques used for this purpose.

4 Mining HLA Patterns Associated with Liver Diseases

As stated in Section 2 not only different antigens have different importance w.r.t. hepatic cirrhosis, but, to make things more complicated, other not yet well characterized factors could play an important role. It is thus important to assess the existence of associations between haplotypic settings (possibly mixed with clinical and demographic information) and hepatic cirrhosis. Algorithms developed for frequent patterns discovery can help us to achieve this goal. Thanks to their ability in handling a very large number of variables and the associated exponential search space, such techniques can discover interesting haplotypic patterns beyond the capabilities of traditional analysis methods.

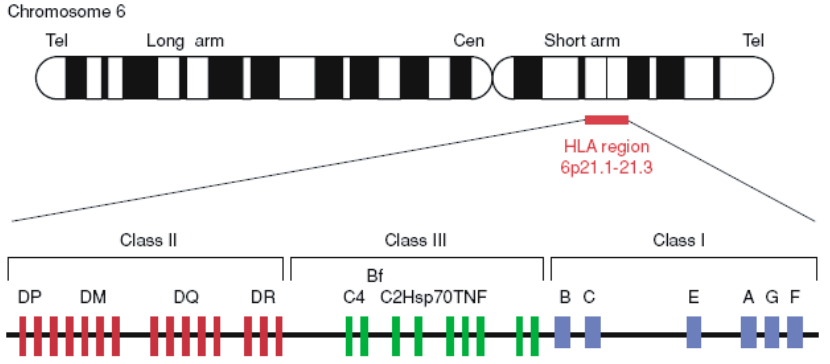

Fig. 1. Gene map of the human leukocyte antigen (HLA) region. The HLA region spans 410^6 nucleotides on chromosome 6p21.1 to p21.3, with class II, class III and class I genes located from the centromeric (Cen) to the telomeric (Tel) end.

This section presents the analisys done in this direction, both from the medical and from the data mining techniques points of view.

4.1 Background and Motivations

The *Major Histocompatibility Complex* (MHC) is a genetic region (i.e., a set of genes), that is essential for a correct autoimmune response. Such genetic region, located on the short arm of the sixth chromosome, includes the HLA system. The main function of HLA genes consists of discriminating cells that are your own (self) from those that are foreign (non-self). The genetical organization of the HLA system is rather simple (see Figure 1). There are three classes of HLA antigens: six major antigens, two for each locus HLA-A, HLA-B and HLA-C, which are the most important among class I antigens; other four groups of antigens, named HLA-DR, HLA-DQ, HLA-DM and HLA-DP, which form the class II antigens; and finally the HLA class III region, which contains many gene encoding proteins that are unrelated to cell-mediated immunity but that nevertheless modulate or regulate immune responses in some way.

Despite the rather simple organization, the nomenclature defining HLA antigens is rather complicated, because the definition system must take in account the high level of *allelic polymorphism*, i.e., the high number of different possible values (*alleles*) for a *locus* (the position or the site of a gene), present in human population. Due to the extreme polymorphism of the HLA loci, only a few individuals out of the family are identical with respect to this system. For these reasons, the HLA was identified as the genetic locus whose products are directly responsible of the rapid rejection in solid organ and marrow transplantation. Beyond this important role, the HLA system seems to influence also the clinical course of hepatic cirrhosis. In fact, although several infectious factors could influence the clinical course of the

hepatic cirrhosis (e.g., HCV, HBV), the genetical setting of the host patient seems to play a central role. The first genes that have been studied, due to their important function in controlling the immune response, were the class I and class II HLA gene. The association between the class II HLA molecules (HLA-DRB1, HLA-DQB1) and, mainly, the chronic infection by HCV and HBV has firstly been reported in [27, 26]. However, the HLA has also an important role in determining the susceptibility to autoimmune based diseases. Most of the diseases associated with the HLA system are, in fact, pathologies related to an autoimmune response against self antigens: in practice, in some genetically predisposed individuals could start some uncontrollable immune phenomena, direct against self antigens that determine a physiologic damage to the target organ.

While some associations between liver diseases due to viral infections and HLA antigens are known in literature, it is still not clear which role is played by which antigens in the development of autoimmune hepatic cirrhosis. It is thus important to assess the existence of associations between haplotypic settings (possibly mixed with clinical and demographical information) and hepatic cirrhosis. Algorithms developed for frequent patterns discovery can help us achieve this goal. Thanks to their ability in handling a very large number of variables and the associated exponential search space, such techniques can discover interesting haplotypic patterns beyond traditional analysis methods capabilities.

In this section we report a frequent pattern discovery analysis on genetical data of patients affected by terminal hepatic cirrhosis with viral origin (HCV and HBV) and patients with terminal hepatic cirrhosis with non-viral origin (autoimmune), conducted with the aim of assessing the influence of the HLA antigens on the course of the disease. In particular, we have evaluated if some genetical configurations of the class I and class II HLA are significantly associated with the triggering causes of the hepatic cirrhosis. The analysis has been performed on two datasets: the first one contains data of 534 patients with terminal cirrhosis that led to liver transplantation, while the second one used as a control population, contains the data of 4718 healthy individuals coming from the same geographic area (this dataset has been previously used in [22]).

The frequent pattern analysis has led to the discovery of some associations already known in the medical literature, and of some not previously known interesting ones, which are object of further biological investigation. The main contribution of this analysis is however methodological: it has proven the adequateness of the frequent pattern discovery methods on this kind of data.

4.2 Frequent Pattern Discovery

The collation of large electronic databases of scientific and commercial information has led to a dramatic growth of interest in methods for discovering

structures in such databases. These methods often go under the general name of *data mining*. However, recently two different kinds of structures sought in data mining have been identified: *models* and *patterns*. The first of these, models, are high level, global, descriptive summaries of data sets. Patterns, on the other hand, are local descriptive structures. Patterns may be regarded as *local models*, and may involve just a few points or variables; that is, they are descriptions of some small part of the data, instead of overall descriptions. Accordingly, *Pattern Discovery* has a distinguished role within data mining technology. In particular, since *frequency* provides support to any extracted knowledge, it is the most used and maybe the most useful measure of interest for the extracted patterns. Therefore during the last decade a lot of researchers have focussed their studies on the computational problem of *Frequent Pattern Discovery*, i.e., mining patterns which satisfy a user-defined minimum threshold of frequency [2, 13].

The simplest form of a frequent pattern is the frequent *itemset*.

Definition 1 (Frequent Itemset Mining). *Let $\mathcal{I} = \{x_1, ..., x_n\}$ be a set of distinct items, an itemset X is a non-empty subset of $\mathcal{I}$. A transaction database $\mathcal{D}$ is a bag of itemsets $t \in 2^{\mathcal{I}}$, usually called transactions. The support of an itemset X in a database $\mathcal{D}$, denoted $sup_{\mathcal{D}}(X)$, is the number of transactions which are superset of X. Given a user-defined minimum support, denoted σ, an itemset X is called frequent in $\mathcal{D}$ if $sup_{\mathcal{D}}(X) \geq \sigma$. The Frequent Itemset Mining Problem requires to compute all the itemsets which are frequent in a transaction database:*

$$\mathcal{F}(\mathcal{D}, \sigma) = \{\langle X, sup_{\mathcal{D}}(X)\rangle | X \in 2^{\mathcal{I}} \wedge sup_{\mathcal{D}}(X) \geq \sigma\}$$

The identification of sets of items, products, symptoms, characteristics, and so forth, that often occur together in the given database, can be seen as one of the most basic tasks in data mining. Although frequent itemsets are *per se* meaningful, the original motivation to extract them was given by the well known *association rules analysis* [1], where the analyst is interested in finding rules describing customers behavior in buying products. Their direct applicability to business problems together with their inherent understandability, even for non data mining experts, made association rules a popular mining method, and frequent itemsets mining one of the hottest research themes in data mining. However frequent itemsets are meaningful not only in the context of association rules mining: they can be used as basic elements in many other kind of analysis, and in particular, they can be used to build global models, ranging from classification [20, 19] to clustering [25, 32].

In the biological domain where our analysis applies, we are not interested in all possible frequent itemsets. Intuitively, we search for patterns describing HLA configurations and their association with the diseases, thus we are mainly interested in itemsets covering as many as possible of the variables regarding HLA loci and the other variables. In practice, if we got a pattern (a set of characteristics) $\{a, b, c\}$ which has the same frequency of a larger

pattern, e.g., $\{a, b, c, d, e\}$, then we are only interested in the larger one. This practical consideration justifies the need for *closed frequent itemsets*.

Closed itemsets were first introduced in [23] and received a great deal of attention especially from an algorithmic point of view [34, 31]. They are a concise and lossless representation of all frequent itemsets, i.e., they contain the same information without redundancy. Intuitively, a closed itemset groups together all its subsets that have the same support; or in other words, it groups together itemsets which identify the same group of transactions.

Definition 2 (Closure Operator). *Given the function* $f(T) = \{i \in \mathcal{I} \mid \forall t \in T, i \in t\}$*, which returns all the items included in the set of transactions* T*, and the function* $g(X) = \{t \in \mathcal{D} \mid \forall i \in X, i \in t\}$ *which returns the set of transactions supporting a given itemset* X*, the composite function* $c = f \circ g$ *is the closure operator.*

Definition 3 (Closed Itemset). *An itemset* I *is closed if and only if* $c(I) = I$*. Alternatively, a closed itemset can be defined as an itemset whose supersets have a strictly smaller support. Given a database* $\mathcal{D}$ *and a minimum support threshold* σ*, the set of frequent closed itemsets is denoted:* $\mathcal{C}l(\mathcal{D}, \sigma) = \{\langle X, sup_{\mathcal{D}}(X)\rangle \in \mathcal{C}_{freq} \mid \nexists Y \supset X s.t. \langle Y, sup_{\mathcal{D}}(X)\rangle \in \mathcal{C}_{freq}\}$.

In particular we adopted the DCI_Closed software [21] for mining frequent closed itemsets. In the next section we describe in details the analysis we performed.

4.3 Mining HLA

As stated before, the objective of our analysis is to find possible associations between combinations of frequent HLA alleles and serious liver diseases that led the patient to liver transplantation.

We have worked on two databases: the first one containing data from 534 patients with autoimmune or viral cirrhosis. For each of them we had the data about their age, gender, geographic origin and the alleles corresponding to the six HLA antigens corresponding to loci A, B and DRB1, plus some characterization and information about the disease. In the second database, used as control population, we had data on 4718 healthy individuals (coming from the same geographic area) for which we had the same information (except, obviously, for the disease-related information).

Before the mining phase, we have performed some *pre-processing* on the patients database, in order to make it suitable for the frequent closed itemset discovery algorithm. During this first phase we have removed noisy data and patients for which too much information was missing. At the end of this phase the database of patients was reduced to 410 individuals. Since the algorithms work on integer numbers, we have then mapped all the available information to integers: an integer number (an item) as been assigned to each single possible value of each variable. In order to make it more significant, we have

discretized the age information in buckets: the first bucket ranging from 1 to 30 years, and the following ones having a size of 5 years.

An important problem addressed in this phase has been that of homozygous alleles, i.e., the situation in which the two alleles for a given locus are identical. Since we are dealing with information structured in transactions, i.e., plain sets and not bags, no duplicate items are admitted. Therefore, in order not to lose information about homozygous alleles, we have created a special item for each locus indicating the presence of homozygous alleles.

Following the physicians' indications we have then divided the pathologies in two main groups, with a total of three item codes:

- Hepatic cirrhosis due to viral infection:
  - one item code for HCV+;
  - one item code for HBV+;
- Hepatic cirrhosis with non viral origin (mainly autoimmune): one unique item code for cryptogenetic cirrhosis, primary biliary cirrhosis, necrosis VBI and VBE, alcoholic cirrhosis, HCC, etc.

Example 1. Table 1 shows a sample of the input file accepted by the algorithm. The file is composed by several transactions (sets of integers), that can have different length: a patient can have different types of cirrhosis at the same time (this is the case of the patient in the third transaction, which has two disease codes, 1002 and 1004), or it could have some genetic variables missing. The first transaction in Table 1 regards a female patient, up to 30 years old, with autoimmune cirrhosis (code 1004) and the following HLA characterization: $A_1 = 24$, $A_2 = 25$, $B_1 = 8$, $B_2 = 18$, $DRB1_1 = DRB1_2 = 1$. Here the DRB1 is a homozygous locus, i.e. it has the same value in each of the two chromosomes; this is represented by the presence of the item code 2499.

We ran the software on the input database and produced the frequent closed itemsets. After the mining phase we performed a *post-processing* phase, in which the extracted patterns were automatically selected w.r.t. their interestingness. Here by interesting pattern we mean a pattern with an exceptionally high frequency in the patients database w.r.t. the frequency of the same pattern, without disease information, in the control database. In fact, this situation would correspond to a possible association between a specific pattern and a specific class of diseases.

Table 2 reports some of the results obtained with this method. In the second column we have the pattern which if taken together with the disease code in the third column, constitutes a frequent closed itemset in the patients

Table 1. A sample of the input database

2	30	2024	2025	2208	2218	2401	2499	1004
1	30	2001	2002	2203	2297	2403	2499	1004
2	35	2024	2031	2214	2251	2414	2404	1004 1002

Table 2. Some interesting patterns discovered

ID	Pattern	Disease	patients DB	control DB	ratio
1	Sex=F; $A = 1; B = 35$	autoim.	1.463%	0.169%	8.656
2	Age in [56,60]; $A = 2; B = 11$	autoim.	1.219%	0%	$+\infty$
3	$A_1 = 1; A_2 = 2; B = 18; DRB1 = 11$	autoim.	1.463%	0.169%	8.656
4	Sex=M; $A = 1; B = 18; DRB1 = 11$	autoim.	1.463%	0.169%	8.656
5	Age in [51,55]; $A = 1; B = 35; DRB1 = 11$	any	1.219%	0.233%	5.231
6	Sex=M; $A = 2; DRB1_1 = DRB1_2 = 11$	autoim.	1.463%	0.254%	5.759
7	Sex=M; $A = 2; B = 51$	HCV+, autoim.	2.439%	0.701%	3.476
8	Sex=M; $A = 2; B = 51; DRB1 = 11$	autoim.	1.463%	0.445%	3.287
9	Age in [56,60]; $A = 1; B = 18; DRB1 = 11$	HBV+	1.219%	0%	$+\infty$
10	Sex=F; Age in [41,45]; $A = 2; B = 18; DRB1 = 11$	any	1.219%	0.042%	29.023
11	Age in [56,60]; $A = 2; DRB1 = 7$	HCV+, autoim.	1.951%	0%	$+\infty$

database, whose relative frequency is given in the fourth column. In the fifth column we have the relative frequency of the pattern (without the disease information) in the control database. In the last column is reported the ratio between the relative frequency in the patients and the relative frequency in the control database. The higher this ratio, the more interesting the pattern.

The patterns exhibiting a surprisingly high ratio, automatically selected in our post-processing phase, have been then evaluated by the domain experts.

4.4 Patterns Evaluation

In the following we report some bio-medical interpretation of a few patterns which have been considered particularly interesting by the domain experts.

Patterns 3 and 4: these two patterns, sharing the common haplotype A=1, B=18, DRB1=11, exhibit a very high ratio w.r.t. the control population. This could be explained by assuming that, as reported for some autoimmune diseases (Grave's disease, Hashimoto's disease), particular alleles (HLA-DRB1 = 11 in our case) have a higher capability in presenting the antigens to T cells.

Patterns 7 and 8: these two patterns share the subpattern A=2, B=51 with autoimmune cirrhosis. In the biological literature the association among A=2, B=51 and the hepatic cirrhosis with autoimmune origin was already known (see [29] and [30]). The discovery of these patterns confirms the adequateness of the proposed methodology.

Pattern 11: describes a set of patients 56 to 60 years old, with A = 2, DRB1 = 7, with hepatic cirrhosis from viral HCV+ infection on hepatic cirrhosis with autoimmune origin. Such a pattern has a relative frequency of 1,951% in the patients database, while it never appears in the much larger control database. This situation caught our attention and we further analyzed it. Surprisingly, we discovered that while no individual older than 55 is present in the healthy population, we found a 2,607% of healthy individuals with the same haplotype (A = 2, DRB1 = 7), but younger than 55. This could point out that patients with such haplotype are not capable to eliminate the HCV+ virus but they are predisposed to the

development of a chronic cirrhosis, leading to transplantation or to death. This would explain why over a certain age threshold, no individual is present in the healthy population.

The obtained results showed that the HLA antigens connected to high level hepatic damage are different in accordance to the cause of the disease. The pattern discovery techniques we used proved their capability in bringing to light particular associations in the haplotypic setting that could remain hidden to the traditional data analysis methods used in biomedicine.

Another interesting aspect acknowledged by the domain experts is that the HLA pattern discovered are, indeed, *haplotypes*, while the input data were containing *genotypes* (genotype is the entire allelic combination of an individual, while the haplotype is the allelic sequence inherited as a block from one of the parents). This fact, on one hand strengthens the biological significance of the patterns obtained, on the other hand, it suggests that frequent pattern discovery techniques can be applied to the *haplotype inference problem*, i.e., the problem of reconstructing haplotypes of individuals, given their genotypes. This problem has been widely studied in bioinformatics, and to date is still a hot algorithmic problem [5, 11, 10, 12].

5 Mining Temporal Patterns Assessing the Effectiveness of a Therapy

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 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 [9].

In this section 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. The temporal information contained in the $\mathcal{TAS}$ patterns extracted is a fruitful knowledge that helps the physicians to evaluate the outcome of the ECP therapy even during the therapy itself.

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.

The interestingness of the case study by the medical perspective lies in the uniqueness and relevance of the dataset under analysis. While by the data analysis perspective, the interestingness lies (*i*) in the structure of the data under analysis, (*ii*) in the importance of the temporal dimension, and (*iii*) in the repeatability of the experience to other medical data analyses where the data exhibit 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 they compose, but the typical time that elapses among two events is also very important. 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 section is to provide a methodological approach to this kind of data, by means of Time-Annotated Sequences ($\mathcal{TAS}$) mining [7, 8].

The rest of the section is organized as follows. In section 5.1 we describe the biological problem studied in this section. In Section 5.2 we introduce the $\mathcal{TAS}$ paradigm and how this can be used for mining time-annotated data. Section 5.3 describes the real-world case study in which the $\mathcal{TAS}$ paradigm is applied to the photopheresis dataset.

5.1 Extracorporeal Photopheresis as a Therapy against Allograft Rejection

In this subsection we describe the biological background to the problem presented in this section.

Originally introduced for treatment of cutaneous T-cell lymphomas [6] and autoimmune diseases [16], 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) [14]. ECP is 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 liver transplantation (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 by means of lymphocytolytic antibody therapy. However, the 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 [28], 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 they 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 [28] 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 [28]. Since prior to [28] 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 [18], we can conclude that this is a unique dataset of this kind.

5.2 From Sequential Patterns to $\mathcal{TAS}$

In this section we summarize the main aspects of the Sequential Pattern Mining (introduced first in [3]), how it has evolved in the $\mathcal{TAS}$[7, 8] paradigm, and how this can be used as a general paradigm for time-annotated data.

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 $sup(S)$ 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 $sup(S) \ \geq \sigma$.

Recently, several algorithms were proposed to efficiently mine sequential patterns, among which we mention PrefixSpan [24], that that employs an internal representation of the data made of database projections over sequence prefixes, and SPADE [33], 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 [35] for a wider review of the state-of-art on sequential pattern mining.

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 [8].

Definition 4 ($\mathcal{TAS}$). *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* sequence*, 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. 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 5 (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. $\forall_{1 \leq k \leq n}.\ \mid \alpha_{1,k} - \alpha_{*,k} \mid \leq \tau$

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 $\mathcal{TAS}$:

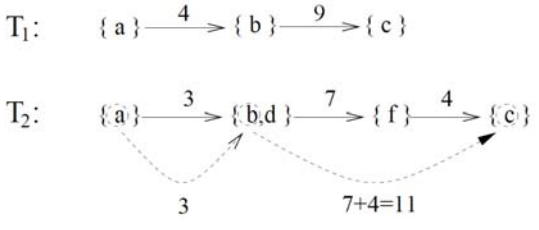

Fig. 2. Example of τ-containment computation

Example 3. Consider two $\mathcal{TAS}$:

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

also depicted in Figure 2, 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 $\mid \alpha_{1,1} - \alpha_{*,1} \mid = \mid 4 - 3 \mid < \tau$ and $\mid \alpha_{1,2} - \alpha_{*,2} \mid = \mid 9 - 11 \mid < \tau$.

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 6 (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* $supp_{[\tau,\mathcal{D}]}(T) = \mid \{T^* \in \mathcal{D} \mid T \preceq_\tau T^*\} \mid$ *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 4. 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 [8], 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 [7, 8].

5.3 Case Study: Mining $\mathcal{TAS}$ from Photopheresis Data

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

Dataset. The dataset under 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 many missing values, thus making many 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 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 in order to produce the vocabulary of items. In other terms, in our analysis an item is always valued by 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, by moving the focus from some interesting patterns to others, by including other variables and so on. Tables 3, 4 and 5 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 as an 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 3 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 3 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.

First Analysis

Data preparation. In the first analysis, for each variable and for each observation, we focus on the variation w.r.t. the value of the same variable in 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 considered the values of the interleukins (2, 4, 5, 10, 12), on explicit request of the physicians.

Results. Table 3 shows some of the results obtained by means of 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 70% of the patients, which is quite a large 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 do 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.

Table 3. Some of the results of the first analysis

ID	Pattern	Support (%)	Support (abs)	Interval		Density
1	(100000IL4) (15IL4)	70	33	13	14	10
				1	2	10
2	(15IL4) (100000IL4)	70	33	7	8	10
3	(15IL4) (105IL5)	61	29	8	9	10
4	(105IL5) (105IL5)	63	30	21	22	10
				19	20	10
				6	8	10
				14	15	10
5	(100000IL4) (15IL4 45IL10)	17	8	15	16	5
6	(105IL5) (105IL5) (105IL5)	42	20	91 0	93 1	5
				0 35	2 36	5
7	(105IL5 60IL10) (100000IL4)	31	15	59	60	5

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 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 considered also other variables: GOT, GPT, GGT, Alkaline Phosphatase, HCV and Bilirubin.

Results. Table 4 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 4 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.

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 unusual values come out more easily.

Results. Table 5 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.

Table 4. Some of the results of the second analysis

ID	Pattern	Support (%)	Support (abs)	Interval		Density
1	(50IL12) (25IL12)	71	35	7	10	15
2	(50IL12) (1000IL10)	61	30	12	13	15
				8	10	10
3	(50IL12) (50IL12)	77	38	9	12	15
				13	14	10
4	(50IL12) (1000IL4)	69	34	9	10	15
5	(1000BIL) (1000IL10)	42	21	12	13	10
6	(1000BIL) (50IL12)	44	22	11	12	10
7	(225AP) (50IL12)	57	28	8	9	10
8	(300AP) (25IL12)	51	25	8	9	10
9	(25IL12) (1000AP)	26	13	6	8	10
10	(50IL12) (225AP)	53	26	24	25	10
				19	21	10
				1	2	10
11	(25IL12) (1000AP)	26	13	6	8	10
12	(50IL12) (225AP)	53	26	24	25	10
				19	21	10
				1	2	10
13	(1000IL10) (1000AP)	32	16	8	9	10
				6	7	10

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 5 showed very unusual values of the interleukins. During the pattern evaluation, the physicians guessed that this unusual behaviour could represent an early warning for allograft rejection. Checking the patients that support the two patterns, we found out that physicians' guess was actually correct.

Table 5. Some results of the third analysis

ID	Pattern	Supp. (%)	Supp. (abs)	Interval		Density
1	(7DD 113-337GPT) (29-45IL10)	38	19	6	7	8
2	(45-672,5IL10) (11,8-19,7IL10 90DD)	18	9	36	37	8
3	(0-30IL12) (94-131IL12 90DD)	18	9	32	33	8
4	(25,7-54,6IL4) (1-4,2IL4 90DD)	22	11	34	35	8
5	(7DD 4,5-41,8BIL) (0,1-1IL4)	38	19	6	7	8
6	(4,5-41,8BIL) (14DD 0,1-1IL4)	18	9	6	7	8
7	(25,7-54,6IL4) (1,1-1,96BIL 90DD)	32	16	36	37	8
				28	29	8
				23	24	8
8	(0,1-1IL4) (1,1-1,96BIL 90DD)	26	13	20	22	8
9	(7DD 4,5-41,8BIL) (14DD 0,1-1IL4)	18	9	6	7	8

This third analysis raised much interest in the physicians and the biologists. 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 models of extracted knowledge regarding information that does 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.

6 Conclusions and Future Work

In this chapter we have described two pattern discovery analyses that we conducted on clinical data of patients in the follow-up of a liver transplantation. The two pattern discovery analyses used different techniques for different objectives.

The first pattern discovery analysis was conducted on genetical data of patients with serious liver diseases, with the aim of discovering associations holding between some genetic configuration and the arising of some serious form hepatic cirrhosis. In particular we have focussed on the HLA genetic region, since it is supposed to play an important role in both viral and autoimmune hepatic cirrhosis. For this kind of analysis, associative frequent patterns where the adopted technique.

The aim of the second data analysis was to assess 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.

Both analyses lead to the discovery of some patterns already known by the physicians, and other novel ones: the feedback we received by the physicians and biologists was very positive. However, the main contribution of these analyses is methodological: they proved adequateness of pattern discovery techniques in medical data mining, and they described a general methodology that can be replicated in other analyses in the medical domain, whenever the data exhibits similar characteristics. 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* [4] in the blood of the recipient in the immediate follow-up of the transplantation. Also in this case, the data is formed by a set of biochemical variables recorded at different time moments, and the typical time elapsing between two relevant

events is a very important piece of 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 models, 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 [17];
- analyze the genetic polymorphism of genes associated with the chronical reject (ICAM-1, VEGF) [15].

Acknowledgments

We 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 section. In particular we are indebted with Irene Bianco, Michele Curcio, Alessandro Mazzoni and Fabrizio Scatena. We must also thank Mirco Nanni and Fabio Pinelli for the $\mathcal{TAS}$ mining software.

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. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Proceedings ACM SIGMOD (1993)
2. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th VLDB (1994)
3. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Yu, P.S., Chen, A.S.P. (eds.) Eleventh International Conference on Data Engineering, Taipei, Taiwan, pp. 3–14. IEEE Computer Society Press, Los Alamitos (1995)
4. Bretagne, S., Vidaud, M., Kuentz, M., Cordonnier, C., Henni, T., Vinci, G., Goossens, M., Vernant, J.: Mixed blood chimerism in t cell-depleted bone marrow transplant recipients: evaluation using dna polymorphisms. Circulation Research 70, 1692–1695 (1987)
5. Clark, A.: Inference of haplotypes from pcr-amplified samples of diploid populations. In: Molecular Biology and Evolution, pp. 111–122 (1990)
6. Edelson, R., Berger, C., Gasparro, F., Jegasothy, B., Heald, P., Wintroub, B., Vonderheid, E., Knobler, R., Wolff, K., Plewig, G.: Treatment of cutaneous t-cell lymphoma by extracorporeal photochemotherapy. N. Engl. J. Med. 316, 297–303 (1987)
7. Giannotti, F., Nanni, M., Pedreschi, D.: Efficient mining of temporally annotated sequences. In: Proceedings of the Sixth SIAM International Conference on Data Mining (2006)

8. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F.: Mining sequences with temporal annotations. In: Proceedings of the 2006 ACM Symposium on Applied Computing (SAC), pp. 593–597 (2006)
9. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F.: Trajectory patter mining. In: The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2007)
10. Gusfield, D.: A practical algorithm for deducing haplotypes in diploid populations. In: Press, A. (ed.) Proceedings of the Eigth International Conference on Intelligent Systems in Molecular Biology, pp. 915–928 (2000)
11. Gusfield, D.: A practical algorithm for optimal inference of haplotypes from diploid populations. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pp. 183–189. AAAI Press, Menlo Park (2000)
12. Gusfield, D.: Inference of haplotypes from samples of diploid populations: complexity and algorithms. J. of Computational Biology 8(3) (2001)
13. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proceedings of ACM SIGMOD (2000)
14. Khuu, H., Desmond, R., Huang, S., Marques, M.: Characteristics of photopheresis treatments for the management of rejection in heart and lung transplant recipients. J. Clin. Apher. 17(1), 27–32 (2002)
15. Kim, I., Moon, S.-O., Park, S.K., Chae, S.W., Koh, G.Y.: 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)
16. Knobler, R., Graninger, W., Lindmaier, A., Trautinger, F.: Photopheresis for the treatment of lupus erythematosus. Ann. NY Acad. Sci. 636, 340–356 (1991)
17. Kobayashi, T., Yokoyama, I., Hayashi, S., Negita, M., Namii, Y., Nagasaka, T., Ogawa, H., Haba, T., Tominaga, Y., Takagi, K.U.H.: Genetic polymorphism in the il-10 promoter region in renal transplantation. Transplant Proc. 31, 755–756 (1999)
18. Lehrer, M., Ruchelli, E., Olthoff, K., French, L., Rook, A.: Successful reversal of recalcitrant hepatic allograft rejection by photopheresis. Liver Transpl. 6(5), 644–647 (2000)
19. Li, W., Han, J., Pei, J.: CMAR: Accurate and efficient classification based on multiple class-association rules. In: Proceedings of the 2001 IEEE International Conference on Data Mining, ICDM 2001 (2001)
20. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: 4th Int. Conf. Knowledge Discovery and Data Mining (KDD 1998), New York, pp. 80–86 (1998)
21. Lucchese, C., Orlando, S., Perego, R.: Dci closed: A fast and memory efficient algorithm to mine frequent closed itemsets. In: FIMI (2004)
22. Marroni, F., Curcio, M., Fornaciari, S., Lapi, S., Mariotti, M., Scatena, F., Presciuttini, S.: Microgeographic variation of hla-a, -b, and -dr haplotype frequencies in tuscany, italy: implications for recruitment of bone marrow donors. Tissue Antigens 64, 478–485 (2004)
23. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: Beeri, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 398–416. Springer, Heidelberg (1998)
24. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: Prefixspan: Mining sequential patterns by prefix-projected growth. In: Proceedings of the 17th International Conference on Data Engineering, pp. 215–224 (2001)

25. Pei, J., Zhang, X., Cho, M., Wang, H., Yu, P.: Maple: A fast algorithm for maximal pattern-based clustering. In: Proceedings of the Third IEEE International Conference on Data Mining, ICDM 2003 (2003)
26. Thursz, R., Kwiatowski, D., Allsopp, C., Greenwood, B., Thomas, H., Hill, A.: Association between a mhc class ii allele and clerance of hepatitis b virus in gambia. New England Journal of Medicine 332, 1065–1069 (1995)
27. Thursz, R., Yallop, R., Goldins, R., Trepo, C., Thomas, H.: Influence of mhc class ii genotype on outcome of infection with hepatitis c virus. Lancet. 354, 2119–2124 (1999)
28. Urbani, L., Mazzoni, A., Catalano, G., Simone, P.D., Vanacore, R., Pardi, C., Bortoli, M., Biancofiore, G., Campani, D., Perrone, V., Mosca, F., Scatena, F., Filipponi, F.: The use of extracorporeal photopheresis for allograft rejection in liver transplant recipients. Transplant Proc. 36(10), 3068–3070 (2004)
29. Verity, D., Marr, J., Ohno, S.: Behçet's disease, the silk road and hla-b51: historical and geographical perspectives. Tissue Antigens 54, 213–220 (1999)
30. Verity, D., Wallace, G., Delamaine, L.: Mica allele profiles and hla class i associations in behçet's disease. Immunogenetics 49, 613–617 (1999)
31. Wang, J., Han, J., Pei, J.: Closet+: searching for the best strategies for mining frequent closed itemsets. In: KDD 2003: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 236–245. ACM Press, New York (2003)
32. Yiu, M.L., Mamoulis, N.: Frequent-pattern based iterative projected clustering. In: Proceedings of the Third IEEE International Conference on Data Mining, ICDM 2003 (2003)
33. Zaki, M.J.: SPADE: An efficient algorithm for mining frequent sequences. Machine Learning 42(1/2), 31–60 (2001)
34. Zaki, M.J., Hsiao, C.-J.: Charm: An efficient algorithm for closed itemset mining. In: SDM (2002)
35. Zhao, Q., Bhowmick, S.: Sequential pattern mining: a survey. Technical Report. Center for Advanced Information Systems, School of Computer Engineering, Nanyang Technological University, Singapore (2003)